Finding Answers to Definition Questions on the Web

Dissertation
zur Erlangung des akademischen Grades eines
Doktors der Philosophie
der Philosophischen Fakultäten
der Universität des Saarlandes

vorgelegt von
Alejandro G. Figueroa A.

aus Chile

Saarbrücken, Juli 2010

Prof. Dr. Erich Steiner - *Dekan*

Promotionskommission:

PD Dr. Günter Neumann - *Doktorvater & Erstberichterstatter*
Univ.-Prof. Dr. Dietrich Klakow - *Zweitberichterstatter*
Univ.-Prof Dr. M. Crocker - *Vorsitz*
Dr. Yi Zhang

Tag der letzten Pruefungsleistung: 9. Juli 2010

to Adonai

Acknowledgements

First of all, I am very thankful to all the reviewers of my thesis; in particular, to my advisor Guenter Neumann and Prof. Dr. Klakow. I, additionally, owe a great debt of gratitude to all the reviewers who anonymously and unknowingly pointed out interesting comments on my work, and ergo helped me in its refinement.

Also, many thanks to all the people who had interesting talks about definition question answering with me during these three years including Cesar de Pablo and Eduard Hovy. An special mention to Mihai Surdeanu, who supervised the work presented in chapter seven. In this respect, I am also profoundly indebted to Ricardo Baeza-Yates, who provided me the opportunity of visiting Yahoo! Research for three months as an intern, and starting this work.

I must sincerely thank to the people who made my doctoral study possible in economic terms: Guenter Neumann, Michael Moossen, Thierry Declerk and Stephan Busemann. In special manner, the last three assisted me in the most difficult part of this project, that is, when the forces of evil tried to prematurely stop it.

Last but not least, I enjoyed all the good talks that I had with Sergio Roa, Hendrik Zender and Dennis Stachowicz. They make the stay of a foreigner in the DFKI LT-Lab nicer. In the same way, I must highlight all the aid given by Hendrik in dealing with the German system. Finally, I want to express a very deep gratitude to Laurel Loch for her proofreading.

Contents

Zusammenfassung

Frage-Antwort-Systeme sind im Wesentlichen dafür konzipiert, von Benutzern in natürlicher Sprache gestellte Anfragen automatisiert zu beantworten. Der erste Schritt im Beantwortungsprozess ist die Analyse der Anfrage, deren Ziel es ist, die Anfrage entsprechend einer Menge von vordefinierten Typen zu klassifizieren. Traditionell umfassen diese: Faktoid, Definition und Liste. Danach wählten die Systeme dieser frühen Phase die Antwortmethode entsprechend der zuvor erkannten Klasse. Kurz gesagt konzentriert sich diese Arbeit ausschließlich auf Strategien zur Lösung von Fragen nach Definitionen (z.B. „*Wer ist Ben Bernanke?*"). Diese Art von Anfrage ist in den letzten Jahren besonders interessant geworden, weil sie in beachtlicher Zahl bei Suchmaschinen eingeht.

Die meisten Fortschritte in Bezug auf die Beantwortung von Fragen nach Definitionen wurden unter dem Dach der Text REtrieval Conference (TREC) gemacht. Das ist, genauer gesagt, ein Framework zum Testen von Systemen, die mit einer Auswahl von Zeitungsartikeln arbeiten. Daher, zielt Kapitel eins auf eine Beschreibung dieses Rahmenwerks ab, zusammen mit einer Darstellung weiterer einführender Aspekte der Beantwortung von Definitionsanfragen. Diesen schließen u.a. ein: (a) wie Definitionsanfragen von Personen gestellt werden; (b) die unterschiedlichen Begriffe von Definition und folglich auch Antworten; und (c) die unterschiedlichen Metriken, die zur Bewertung von Systemen genutzt werden.

Seit Anbeginn von TREC haben Systeme vielfältige Ansätze, Antworten zu entdecken, auf die Probe gestellt und dabei eine Reihe von zentralen Aspekten dieses Problems beleuchtet. Aus diesem Grund behandelt Kapitel vier eine Auswahl einiger bekannter TREC Systeme. Diese Auswahl zielt nicht auf Vollständigkeit ab, sondern darauf, die wesentlichen Merkmale dieser Systeme hervorzuheben. Zum größten Teil nutzen die Systeme Wissensbasen (wie z.B. Wikipedia), um Beschreibungen des zu definierenden Konzeptes (auch als *Definiendum* bezeichnet) zu erhalten. Diese Beschreibungen werden danach auf eine Reihe von möglichen Antworten projiziert, um auf diese Art die richtige Antwort zu ermitteln. Anders ausgedrückt nehmen diese Wissensbasen die Funktion von annotierten Ressourcen ein, wobei die meisten Systeme versuchen, die Antwortkandidaten in einer Sammlung von Zeitungsartikeln zu finden, die diesen Beschreibungen ähnlicher sind.

Den Grundpfeiler dieser Arbeit bildet die Annahme, dass es plausibel ist, ohne annotierte Ressourcen konkurrenzfähige, und hoffentlich bessere, Systeme zu entwickeln. Obwohl dieses deskriptive Wissen hilfreich ist, basieren sie nach Überzeugung des Autors auf zwei falschen Annahmen:

1. Es ist zweifelhaft, ob die Bedeutungen oder Kontexte, auf die sich das *Definiendum* bezieht, dieselben sind wie die der Instanzen in der Reihe der Antwortkandidaten. Darüber hinaus erstreckt sich diese Beobachtung auch auf die Tatsache, dass nicht alle Beschreibungen innerhalb der Gruppe der mutmaßlichen Antworten notwendigerweise von Wissensbasen abgedeckt werden, auch wenn sie sich auf dieselben Bedeutungen und Kontexte beziehen.

2. Eine effiziente Projektionsstrategie zu finden bedeutet nicht notwendigerweise auch ein gutes Verfahren zur Feststellung von deskriptivem Wissen, denn es verschiebt die Zielsetzung der Aufgabe hin zu einem „mehr wie diese Menge" statt zu analysieren, ob jeder Kandidat den Charakteristika einer Beschreibung entspricht oder nicht. Anders ausgedrückt ist die Abdeckung, die durch Wissensbasen für ein spezifisches

Definiendum gegeben ist, nicht umfassend genug, um alle Charakteristika, die für seine Beschreibungen kennzeichnend sind, zu erlernen, so dass die Systeme in der Lage sind, alle Antworten innerhalb der Kandidatenmenge zu identifizieren. Eine konventionelle Projektionsstrategie kann aus einem anderen Blickwinkel als Prozedur zum Finden lexikalischer Analogien betrachtet werden.

Insgesamt untersucht diese Arbeit Modelle, die Strategien dieser Art in Verbindung mit annotierten Ressourcen und Projektion außer Acht lassen. Tatsächlich ist es die Überzeugung des Autors, dass eine robuste Technik dieser Art mit traditionellen Methoden der Projektion integriert werden und so eine Leistungssteigerung ermöglichen kann.

Die größeren Beiträge dieser Arbeit werden in den Kapiteln fünf, sechs und sieben präsentiert. Es gibt mehrere Wege diese Struktur zu verstehen. Kapitel fünf, beispielsweise, präsentiert einen allgemeinen Rahmen für die Beantwortung von Fragen nach Definitionen in mehreren Sprachen. Das primäre Ziel dieser Studie ist es, ein leichtgewichtiges System zur Beantwortung von Fragen nach Definitionen zu entwickeln, das mit Web-Snippets und zwei Sprachen arbeitet: Englisch und Spanisch. Die Grundidee ist, von Web-Snippets als Quelle deskriptiver Information in mehreren Sprachen zu profitieren, wobei der hohe Grad an Sprachunabhängigkeit dadurch erreicht wird, dass so wenig linguistisches Wissen wie möglich berücksichtigt wird. Genauer gesagt berücksichtigt dieses System statistische Methoden und eine Liste von Stop-Wörtern sowie eine Reihe von sprach-spezifischen Definitionsmustern.

Im Einzelnen teilt sich Kapitel fünf in zwei spezifischere Studien auf. Die erste Studie zielt im Grunde darauf ab, aus Redundanz für die Ermittlung von Antworten Kapital zu schlagen (z.B. Worthäufigkeiten über verschiedene Antwortkandidaten hinweg). Obwohl eine solche Eigenschaft unter TREC Systemen weit verbreitet ist, legt diese Studie den Schwerpunkt auf die Auswirkungen auf verschiedene Sprachen und auf ihre Vorteile bei der Anwendung auf Web-Snippets statt Zeitungsartikeln. Eine weitere Motivation dahinter, Web-Snippets ins Auge zu fassen, ist die Hoffnung, Systeme zu studieren, die mit heterogenen Corpora arbeiten, ohne es nötig zu machen, vollständige Dokumente herunterzuladen. Im Internet, beispielsweise, steigt die Zahl verschiedener Bedeutungen für das *Definiendum* deutlich an, was es notwendig macht, eine Technik zur Unterscheidung von Bedeutungen in Betracht zu ziehen. Zu diesem Zweck nutzt das System, das in diesem Kapitel vorgestellt wird, einen unüberwachten Ansatz, der auf der Latent Semantic Analysis basiert. Auch wenn das Ergebnis dieser Studie zeigt, dass die Unterscheidung von Bedeutungen allein anhand von Web-Snippets schwer zu erreichen ist, so lässt es doch auch erkennen, dass sie eine fruchtbare Quelle deskriptiven Wissens darstellen und dass ihre Extraktion spannende Herausforderungen bereithält.

Der zweite Teil erweitert diese erste Studie durch die Nutzung mehrsprachiger Wissensbasen (d.h. Wikipedia), um die möglichen Antworten in eine Rangfolge einzureihen. Allgemein ausgedrückt profitiert sie von Wortassoziationsnormen, die von Sätzen gelernt werden, die über Wikipedia hinweg zu Definitionsmustern passen. Um an der Prämisse festzuhalten, keine Artikel mit Bezug auf eine spezifisches *Definiendum* zu nutzen, werden diese Sätze anonymisiert, indem der Begriff mit einem Platzhalter ersetzt wird, und die Wortnormen werden von allen Sätzen der Trainingsmenge gelernt, statt nur von dem Wikipedia-Artikel, der sich auf das spezielle *Definiendum* bezieht. Die Ergebnisse dieser Studie zeigen, dass diese Nutzung dieser Ressourcen ebenfalls vorteilhaft sein kann; speziell zeigen sie auf, dass Wortassoziationsnormen eine kosteneffiziente Lösung darstellen. Allerdings nehmen die Corpusgrößen für andere Sprachen als Englisch deutlich ab, was auf deren Unzulänglichkeit für die Konstruktion von Modellen für andere Sprachen hinweist.

Kapitel sechs, weiter hinten, wird spezieller und handelt ausschließlich von der Einord-

nung von Antwortkandidaten in englischer Sprache in eine Rangfolge. Der Grund dafür, hier Spanisch außer Acht zu lassen, ist die geringe beobachtete Dichte, sowohl in Bezug auf redundante Information im Internet als auch in Bezug auf Trainingsmaterial, das von Wikipedia erworben wurde. Diese geringe Dichte ist deutlich stärker ausgeprägt als im Fall der englischen Sprache und erschwert das Erlernen mächtiger statischer Modelle. Dieses Kapitel präsentiert einen neuartigen Weg, Definitionen zu modellieren, die in n-gram Sprachmodellenverankert sind, die aus der lexikalisierten Darstellung des Abhängigkeitsbaumes des in Kapitel fünf erworbenen Trainingsmaterials gelernt wurden. Diese Modelle sind kontextuell in dem Sinne, dass sie in Bezug auf die Semantik des Satzes konstruiert werden. Im Allgemeinen können diese Semantiken als unterschiedliche Typen von *Definienda* betrachtet werden (z.B. Fußballer, Sprache, Künstler, Krankheit und Baum). Diese Studie untersucht zusätzlich die Auswirkungen einiger Eigenschaften (nämlich benannter Entitäten und Part-of-speech-Tags) auf diese Kontextmodelle. Insgesamt sind die Ergebnisse, die mit diesem Ansatz erhalten wurden, ermutigend, insbesondere in Bezug auf eine Steigerung der Genauigkeit des Musterabgleichs. Indes wurde höchstwahrscheinlich experimentell beobachtet, dass ein Trainingscorpus, das nur Positivbeispiele (Beschreibungen) enthält, nicht ausreicht, um perfekte Genauigkeit zu erreichen, da diese Modelle die Charakteristika nicht ableiten können, die für nicht-deskriptiven Inhalt kennzeichnend sind. Für die weitere Arbeit ermöglichen es Kontextmodelle zu untersuchen, wie unterschiedliche Kontexte in Übereinstimmung mit deren semantischen Ähnlichkeiten verschmolzen (geglättet) werden können, um die Leistung zu verstärken.

Kapitel sieben wird anschließend sogar noch spezieller und sucht nach der Menge von Eigenschaften, die dabei helfen kann, Beschreibungen von anderen Textarten zu unterscheiden. Dabei sollte beachtet werden, dass diese Studie alle Arten von Beschreibungen berücksichtigt, einschließlich derer, die Definitionsmustern nicht genügen. Dadurch werden Maximum-Entropy-Modelle konstruiert, die auf einen automatisch akquirierten Corpus von großem Umfang aufsetzen, der Beschreibungen von Wikipedia und Nicht-Beschreibungen aus dem Internet umfasst. Grob gesagt werden unterschiedliche Modelle konstruiert, um die Auswirkungen verschiedenerlei Merkmale zu untersuchen: Oberfläche, benannte Entitäten, Part-of-speech-Tags, Chunks und, noch interessanter, von den lexikalisierten Abhängigkeitsgraphen abgeleitete Attribute. Im Allgemeinen bestätigen die Ergebnisse die Effizienz von Merkmalen, die Abhängigkeitsgraphen entnommen sind, insbesondere Wurzelknoten und n-gram-Pfaden. Experimente, die mit verschiedenen Testmengen diverser Charakteristika durchgeführt wurden, legen nahe, dass auch angenommen werden kann, dass Attribute gefunden werden, die sich auf andere Corpora übertragen lassen.

Es gibt zwei weitere Kapitel: zwei und drei. Ersteres untersucht unterschiedliche Strategien, das Netz nach deskriptivem Wissen zu durchforsten. Im Wesentlichen analysiert dieses Kapitel einige Strategien, die darauf abzielen, die Trefferquote (den Recall) deskriptiver Sätze über Web-Snippets hinweg zu verstärken, insbesondere solcher Sätze, die weit verbreiteten Definitionsmustern genügen. Diese Studie ist eine Nebenstudie, die jedoch für den Kern dieser Arbeit dienlich ist, da es für auf das Internet gerichtete Systeme notwendig ist, effektive Suchstrategien zu entwickeln. Im Gegensatz dazu hat Kapitel drei zwei Ziele: (a) einige Komponenten vorzustellen, die in den Strategien benutzt werden, die in den letzten drei Kapiteln dargestellt werden, was dabei hilft, sich auf die Schlüsselaspekte der Rankingstrategien zu konzentrieren und so die relevanten Aspekte der Ansätze in diesen drei Kapiteln klar darzulegen; (b) einige Charakteristika auszuarbeiten, die die Trennung wirklicher Antwortkandidaten von irreführenden schwierig machen; insbesondere unter Sätzen, die Definitionsmustern entsprechen. Kapitel drei ist hilfreich, um einen Teil der linguistischen Phänomene zu verstehen, von denen die späteren Kapitel handeln.

Als abschließende Bemerkungen dieser Arbeit, da es ja eine Unzahl an Methodologien

gibt, beginnen die Kapitel sechs und sieben mit den jeweiligen Strategien verwandte Arbeiten zu analysieren. Der Hauptbeitrag des jeweiligen Kapitels beginnt in den Abschnitten 6.5 bzw. 7.6. Diese beiden Abschnitte beginnen mit einer Diskussion und einem Vergleich zwischen den vorgeschlagenen Methoden und den verwandten Arbeiten wie sie in den entsprechenden vorhergehenden Abschnitten vorgestellt wurden. Dieser Aufbau zielt auf eine Vereinfachung der Kontextualisierung der vorgeschlagenen Ansätze, da es unterschiedliche Frage-Antwort-Systeme mit vielfältigen Charakteristika gibt.

Abstract

Fundamentally, question answering systems are designed for automatically responding to queries posed by users in natural language. The first step in the answering process is query analysis, and its goal is to classify the query in congruence with a set of pre-specified types. Traditionally, these classes include: factoid, definition, and list. Systems thereafter chose the answering method in concert with the class recognised in this early phase. In short, this thesis focuses exclusively on strategies to tackle definition questions (e.g., "*Who is Ben Bernanke?*"). This sort of question has become especially interesting in recent years, due to its significant number of submissions to search engines.

Most advances in definition question answering have been made under the umbrella of the Text REtrieval Conference (TREC). This is, more precisely, a framework for testing systems operating on a collection of news articles. Thus, the objective of chapter one is to describe this framework along with presenting additional introductory aspects of definition question answering including: (a) how definition questions are prompted by individuals; (b) the different conceptions of definition, and thus of answers; and (c) the various metrics exploited for assessing systems.

Since the inception of TREC, systems have put to the test manifold approaches to discover answers, throwing some light onto several key aspects of this problem. On this account, chapter four goes over a selection of some notable TREC systems. This selection is not aimed at completeness, but rather at highlighting the leading features of these systems. For the most part, systems benefit from knowledge bases (e.g., Wikipedia) for obtaining descriptions about the concept being defined (a.k.a. *definiendum*). These descriptions are thereafter projected onto the array of candidate answers as a means of discerning the correct answer. In other words, these knowledge bases play the role of annotated resources, and most systems attempt to find the answer candidates across the collection of news articles that are more similar to these descriptions.

The cornerstone of this thesis is the assumption that it is plausible to devise competitive, and hopefully better, systems without the necessity of annotated resources. Although this descriptive knowledge is helpful, it is the belief of the author that they are built on two wrong premises:

1. It is arguable that senses or contexts related to the *definiendum* across knowledge bases are the same senses or contexts for the instances across the array of answer candidates. This observation also extends to the fact that not all descriptions within the group of putative answers are necessarily covered by knowledge bases, even though they might refer to the same contexts or senses.

2. Finding an efficient projection strategy does not necessarily entail a good procedure for discerning descriptive knowledge, because it shifts the goal of the task to a "*more*

like this set" instead of analysing whether or not each candidate bears the characteristics of a description. In other words, the coverage given by knowledge bases for a specific *definiendum* is not wide enough to learn all the characteristics that typify its descriptions, so that systems are capable of identifying all answers within the set of candidates. From another angle, a conventional projection methodology can be seen as a finder of lexical analogies.

All in all, this thesis investigates into models that disregard this kind of annotated resource and projection strategy. In effect, it is the belief of the author that a robust technique of this sort can be integrated with traditional projection methodologies, and in this way bringing about an enhancement in performance.

The major contributions of this thesis are presented in chapters five, six and seven. There are several ways of understanding this structure. For example, chapter five presents a general framework for answering definition questions in several languages. The primary goal of this study is to design a lightweight definition question answering system operating on web-snippets and two languages: English and Spanish. The idea is to utilise web-snippets as a source of descriptive information in several languages, and the high degree of language independency is achieved by making allowances for as little linguistic knowledge as possible. To put it more precisely, this system accounts for statistical methods and a list of stop-words, as well as a set of language-dependent definition patterns.

In detail, chapter five branches into two more specific studies. The first study is essentially aimed at capitalising on redundancy for detecting answers (e.g., word frequency counts across answer candidates). Although this type of feature has been widely used by TREC systems, this study focuses on its impact on different languages, and its benefits when applied to web-snippets instead of a collection of news documents. An additional motivation behind targeting web-snippets is the hope of studying systems working on more heterogenous corpora, without incurring the need of downloading full-documents. For instance, on the Internet, the number of distinct senses for the *definiendum* considerably increases, ergo making it necessary to consider a sense discrimination technique. For this purpose, the system presented in this chapter takes advantage of an unsupervised approach premised on Latent Semantic Analysis. Although the outcome of this study shows that sense discrimination is hard to achieve when operating solely on web snippets, it also reveals that they are a fruitful source of descriptive knowledge, and that their extraction poses exciting challenges.

The second branch extends this first study by exploiting multilingual knowledge bases (i.e., Wikipedia) for ranking putative answers. Generally speaking, it makes use of *word association norms* deduced from sentences that align definitions patterns across Wikipedia. In order to adhere to the premise of not profiting from articles related to a specific *definiendum*, these sentences are anonymised by replacing the concept with a placeholder, and the word norms are learnt from all training sentences, instead of only from the Wikipedia page about the particular *definiendum*. The results of this study signify that this use of these resources can also be beneficial; in particular, they reveal that *word association norms* are a cost-efficient solution. However, the size of the corpus markedly decreases for languages different from English, thus indicating their insufficiency to design models for other languages.

Later, chapter six gets more specific and deals only with the ranking of answer candidates in English. The reason for abandoning the idea of Spanish is the sparseness observed across both the redundancy from the Internet and the training material mined from Wikipedia. This sparseness is considerably greater than in the case of English, and it makes learning powerful statistical models more difficult. This chapter presents a novel way of modeling definitions grounded on n-gram language models inferred from the lexicalised dependency tree representation of the training material acquired in the study of chapter five. These mod-

els are contextual in the sense that they are built in relation to the semantic of the sentence. By and large, these semantics can be perceived as the distinct types of *definienda* (e.g., footballer, language, artist, disease, and tree). This study, in addition, investigates the effect of some features on these *context models* (i.e., named entities, and part-of-speech tags). Overall, the results obtained by this approach are encouraging, in particular in terms of increasing the accuracy of the pattern matching. However, in all likelihood, it was experimentally observed that a training corpus comprising only positive examples (descriptions) is not enough to achieve perfect accuracy, because these models cannot deduce the characteristics that typify non-descriptive content. More essential, as future work, *context models* give the chance to cogitate on the idea of amalgamating (smoothing) various contexts in agreement with their semantic similarities in order to ameliorate the performance.

Subsequently, chapter seven gets even more specific and it searches for the set of properties that can aid in discriminating descriptions from other kinds of texts. Note that this study regards all kinds of descriptions, including those mismatching definition patters. In so doing, Maximum Entropy models are constructed on top of an automatically acquired large-scale training corpus, which encompasses descriptions from Wikipedia and non-descriptions from the Internet. Roughly speaking, different models are constructed as a means of studying the impact of assorted properties: surface, named entities, part-of-speech tags, chunks, and more interestingly, attributes derived from the lexicalised dependency graphs. In general, results corroborate the efficiency of features taken from dependency graphs, especially the root node and n-gram paths. Experiments conducted on testing sets of varied characteristics suggest that it is also plausible to find attributes that can port to other corpora.

There are two extra chapters: two and three. The former examines different strategies to trawl the Web for descriptive knowledge. In essence, this chapter touches on several strategies geared towards boosting the recall of descriptive sentences across web snippets, especially sentences that align widespread definition patterns. This is a side, but instrumental study to the core of this thesis, as it is necessary for systems targeted at the Internet to develop effective crawling techniques. On the contrary, chapter three has two goals: (a) presenting some components used by the strategies outlined in the last three chapters, this way helping to focus on key aspects of the ranking methodologies, and hence to clearly present the relevant aspects of approaches laid out in these three chapters; and (b) fleshing out some characteristics that make separating the genuine from the misleading answer candidates difficult; particularly, across sentences matching definition patterns. Chapter three is helpful for understanding part of the linguistic phenomena that the posterior chapters deal with.

On a final note about the organisation of this thesis, since there is a myriad of techniques, chapter six and seven start dissecting the related work closer to each strategy. The main contribution of each chapter begins at section 6.5 and 7.6, respectively. These two sections start with a discussion and comparison between the proposed methods and the related work presented in their corresponding preceding sections. This organisation is directed at facilitating the contextualisation of the proposed approaches as there are different question answering systems with manifold characteristics.

Chapter 1

What is a definition?

> "an explanation of the meaning of a word or phrase, especially in a dictionary; the act of stating the meaning of words and phrases." (Oxford Dictionary)
>
> "The meaning of a word is what it is explained by the explanation of its meaning." I.e.: If you want to understand the use of the word *meaning*, look for what are called *explanations of meanings*. (Philosophical Investigations I §560) [Wittgenstein, 1953].

1.1 Introduction

The continuous growth and diversification of online text information on the Web represents a continuing challenge, affecting both the designers and the users of information processing systems. The development of systems that assist users in finding relevant pieces of information across large text collections is a key task, because they transform a static set of stored text files into accessible and searchable knowledge.

Situated at the frontier of Natural Language Processing (NLP) and modern Information Retrieval (IR), open-domain Question Answering (QA) is an appealing choice for the retrieval of full-length documents. Users of QA systems specify their information needs in the form of natural-language questions, ergo eliminating any artificial constraints sometimes imposed by a special input syntax (e.g., boolean operators). The QA system satisfies the need of the user by returning brief answer strings extracted from the text collection. To be more precise, QA systems capitalise on the fact that answers are often concentrated in small fragments of text documents. It is up to the QA system to analyse the content of full-length documents and identifying these small, pertinent text fragments.

One prominent type of query prompted by users is definition questions (e.g., "*What is ..?*" and "*Who is ..?*"). The motivation behind studying definition queries versus other kinds is due to their increasing number actually submitted on the Internet. This sort of queries enquires the system about a topic or concept (a.k.a. *definiendum*), that is definition QA systems must search for explanations of meanings of the *definiendum* across the target collection. As for answers, definition QA systems do not solely account for these direct descriptions, but they also usually embrace an array of important and/or factual document snippets (a.k.a. nuggets) about the *definiendum*. However, in order to provide enough context and ensure readability, definition QA systems produce a set of sentences carrying these nuggets.

The standard strategy for answering these questions using the Web or a textual corpus involves a combination of IR and Named Entity Recogniser (NER) techniques [Voorhees, 2003].

Until recently, definition queries remained a largely unexplored area for QA. Standard factoid QA technology, designed to extract single answers, cannot be straightforwardly applied to this task.

The aim of this chapter is presenting the fundamentals of definition QA systems. It is organised as follows: the next section deals at greater length with the trade-off between coverage and dependability of distinct sources of answers, section 1.3 presents the archetypal components of a definition QA system, section 1.4 compares two distinct international assessment frameworks, section 1.5 introduces issues related to query analysis, section 1.6 describes some characteristics of definitions in detail, section 1.7 fleshes out diverse evaluation metrics, and section 1.8 offers a conclusion to this chapter.

1.2 Definitions in Natural Language Texts across the Web

Strictly speaking, any document can be a potential provider of answers to definitions questions. There are, however, types of documents that are richer in definitions than others. For example, a fertile source of descriptions is online dictionaries and encyclopedias, such as Wikipedia and answers.com. Needless to say, this sort of resource is typically exploited by definition QA systems, despite their unevenly reliability and their frequent lack of considerable coverage.

KNOWLEDGE BASES

To be more specific, on-line commercial dictionaries like the Oxford Dictionary offer almost unequivocal definitions in terms of reliability, whereas on-line encyclopedias (e.g., Wikipedia) supply more disputable and/or unreliable pieces of information. On the other hand, it is equally important to underscore that the coverage of these authoritative resources varies markedly from one *definiendum* to the other. More precisely, this variation can consist in the absence of entries in these knowledge bases, or in terms of the amount and class of information they convey. The reader can inspect a list of the most widely used Knowledge Bases (KB) in table 2.4 in section 2.1 on page 25.

HOMEPAGES

Another rich source of descriptions is homepages. This kind of resource is very important as it, once in a while, yields biographical information about the owner of the page, who is the potential *definiendum*. Above all, they are a useful resource when tackling persons or organisations as *definienda*. Without a shadow of doubt, these resources also suffer from the same problems: reliability and coverage. Specifically, the owner can alter or omit the publication of interesting facts that are inconvenient for their carrier or business, which can be of special interest to the user of a definition QA system. A good example of biographical information found in homepages is shown in figure 1.1.

NEWSPAPERS

By extension, newspapers are also a fruitful source of short definitions as it is a common practice to provide a brief description when a person or organisation is mentioned -usually for first time- in the text. The next is a delineative surrogate of a news article which renders a succinct and precise description of "*Chuck Berry*":

Johnny McCain No Goode for Chuck Berry - Politics News Summaries | Newser
American guitarist and singer Chuck Berry is seen in this 1980 file photo.
Berry has said he wants to see Barack Obama win the White House.
www.newser.com/story/29615/johnny-mccain-no-goode-for-chuck-berry.html

Occasionally, news articles can play the role of knowledge bases, because a piece of news can sometimes focus its attention on a special topic (*definiendum*) giving interesting information about several of its chief aspects. The advantage of newspapers articles over knowledge bases is two-fold: (a) they can supply supplementary titbits and up-to-date information as well as the latest facts; and (b) from the viewpoint of definition QA systems, they expand the

Chuck Berry's music has transcended generations. He earns respect to this day because he is truly an entertainer. Berry, also known as "The Father of Rock & Roll", gained success by watching the audience's reaction and playing accordingly, putting his listeners' amusement above all else.
For this reason, tunes like "Johnny B. Goode," "Maybellene" and "Memphis" have become anthems to an integrated American youth and popular culture. Berry is a musical icon who established rock and roll as a musical form and brought the worlds of black and white together in song.
Born in St. Louis on October 18, 1926 Berry had many influences on his life that shaped his musical style. He emulated the smooth vocal clarity of his idol, Nat King Cole, while playing blues songs from bands like Muddy Waters. For his first stage performance, Berry chose to sing a Jay McShann song called "Confessin' the Blues." It was at his high school's student musical performance, when the blues was well-liked but not considered appropriate for such an event. He got a thunderous applause for his daring choice, and from then on, Berry had to be onstage.
Berry took up the guitar after that, inspired by his partner in the school production. He found that if he learned rhythm changes and blues chords, he could play most of the popular songs on the radio at the time. His friend, Ira Harris, showed him techniques on the guitar that would become the foundation of Berry's original sound. Then in 1952, he began playing guitar and singing in a club band whose song list ranged from blues to ballads to calypso to country. Berry was becoming an accomplished showman, incorporating gestures and facial expressions to go with the lyrics.
Berry continued his success with such hits as "Brown-Eyed Man," "Too Much Monkey Business," "Memphis," "Roll Over, Beethoven!" and "Johnny B. Goode." "Johnny B. Goode" is Berry's masterpiece, as it brought together all the elements of Berry's unique musical sound. It cemented his place in rock history and led to fame in the 1950s. His popularity garnered him television and movie appearances and he toured frequently.

Figure 1.1: Some pieces of biographical information about *"Chuck Berry"* excerpted from www.chuckberry.com/about/bio.htm (As of October 2009).

coverage as they address some topics that are more unlikely to be found across knowledge bases. Typical cases are pieces of news reporting on investigations into issues, such as new virus breakthroughs, medicines or technology advances. The factor, however, that makes newspaper articles less attractive is their inherent bias, unbalance, and propagandist nature, which make them, to a smaller extent, trustworthy.

DOCUMENTS

As a means of broadening the coverage and boosting the reliability of answers, definition QA systems take advantage of the hierarchy returned by the IR engine for selecting the most propitious documents. In general, the required exhaustiveness of the response or the coverage yielded by the already mentioned resources necessitated taking into account full documents as sources of definitions. However, processing full-documents typically demands a marked increase in the processing time, ergo definition QA systems prefer analysing web snippets to process full-documents. More precisely, web snippets are the brief surrogates returned by commercial search engines describing the local contexts of the documents that best match the submitted search string. The following two web snippets sketch how descriptive information can also be found across these document surrogates:

WEB-SNIPPETS

The Official Site of Chuck Berry
Chuck Berry is one of rock & roll's great lyricists and developed some of its earliest trademark guitar licks; represented by CMG Worldwide.
www.chuckberry.com/index.php

Chuck Berry - Discover music, videos, concerts, & pictures at ...
Chuck Berry is an influential figure and one of the pioneers of rock and roll ... as I like Elvis I must admit that Chuck Berry is the real King of Rock n' Roll.
www.last.fm/music/Chuck+Berry

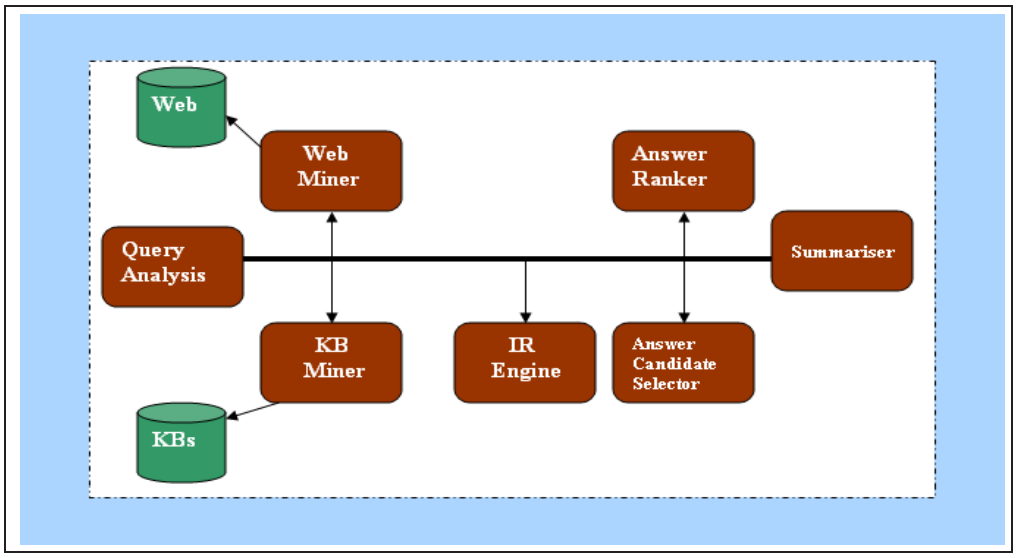

Figure 1.2: General architecture of a definition QA system.

With web-snippets, it is notable that: (a) these hits vary from one search engine to the other as they privilege different features while ranking documents in concert with a given query, and (b) they can also include excerpts from newspapers and homepages as depicted by the first illustrative surrogate. But, on the flip side, web-snippets are noisier and less reliable than authoritative resources, as most of the hits returned by search engines are aimed at boosting the similarity of their respective documents to the submitted query. The crucial issue here is that this similarity is not necessarily in the best interests of definition QA systems. The bottom line is, search engines are biased towards returning in the top position of their ranking hits that are more likely to raise their advertising revenues.

BLOGS AND FORUMS

Incidentally, forums and blogs are the diametrical opposite of knowledge bases, because they touch on wide-ranging topics and allow their users to make comments of various natures such as opinions, advertisements, suggestions, descriptions, and general text snippets. This sort of resource demands more efforts to separate the wheat from the chaff. But they are, nonetheless, a rich source of explanations of some new and/or more specific *definienda*, which are barely found in online dictionaries and/or encyclopedias.

To sum up, there is a trade-off between coverage and reliability, in which knowledge bases are the most authoritative, thus trustworthy, but they provide limited coverage. On the contrary, ordinary web pages/snippets yield broad coverage, but they are less reliable.

1.3 Archetypal Modules of a Definition QA System

The architecture of a definition QA system can sharply vary from one system to the other. There are, nonetheless, some modules that transpire most architectures. The following is a list of these components (figure 1.2):

- **Query Analysis** is the step in charge of discriminating the kind of question prompted by the user, and in the event of a definition query, it is responsible for identifying the *definiendum*(*s*).

- **KB Miner** extracts from specific resources (e.g., dictionaries and encyclopedias) articles related to the *definiendum*.

- **Web Miner** discovers information about the *definiendum* across the Web.
- **IR engine** is the interface between the definition QA system and the target collection of documents. It returns the top ranked documents in congruence with a specified similarity measure.
- **Answer Candidate Selector** is the module that pre-processes the most promising documents (e.g. sentence splitting, and co-reference resolution). This then singles out the most propitious contexts as putative answers.
- **Answer Ranker** scores answer candidates and selects the definitive array of answers.
- **Summariser** abridges the set of definitive answers by reducing redundancy and eliminating potentially irrelevant pieces of information.

1.4 Definition Questions in TREC and CLEF

Fundamentally, there are international assessments established for the purpose of evaluating and comparing distinct strategies devised to automatically answer natural language questions. With specific regard to definition QA systems, evaluations can sharply differ as systems have multiple facets, and accordingly, different assessments can highlight varied aspects of systems (e.g., corpus, language, user requirements and interactions). Basically, the most widely known evaluation forums are the Text REtrieval Conference and the Cross Language Evaluation Forum:

Text REtrieval Conference (TREC) is a workshop hosted by the National Institute of Standards and Technology (NIST). This takes place yearly and supplies the necessary infrastructure to assess IR systems operating on a large collection of documents. More specifically, Text REtrieval Conference (TREC) has built a variety of massive test collections, including the AQUAINT corpus, which is the collection utilised for question answering tasks.

In this forum, QA systems compete for correctly answering the highest possible number of questions that are members of a given test set determined by the organisers. More often than not, this pre-determined test set encompasses a variety of queries, like factoid and definitions as well as list. In the special case of answering definition questions, QA systems are encouraged to find document snippets (nuggets) across the AQUAINT corpus that render relevant facts and/or information about the *definiendum* (see, for instance, table 1.1). Definition QA systems must thereafter remove all redundant content by producing a sort of summary of these selected nuggets. In other words, systems have to distinguish duplicates together with snippets that express the same information, and those conveying descriptions already subsumed in other text fragments.

The appropriateness of a nugget as a part of the response is verified in congruence with the opinion of a group of assessors. Hence, nuggets are perceived as a hierarchical gold standard, in which they are deemed to be “vital" when their inclusion in the output is indispensable, whereas interesting, but dispensable, nuggets are labelled as “okay". Evidently, uninteresting text snippets are excluded from this sort of ground truth. Definition QA systems are therefore rated in agreement with their outputs and this pre-defined gold standard. To be more precise, systems are rewarded for returning “vital" nuggets, while not penalised for “okay" nuggets, and punished for the incorporation of any extra text fragment omitted from the pre-determined ground truth.

Cross Language Evaluation Forum (CLEF) provides evaluation tracks to test a variety of aspects in relation to cross-language information retrieval systems. A noteworthy feature of this assessment framework is the fact that QA systems target at heterogeneous collection of documents, which comprises news articles and Wikipedia. Another fundamental element of this evaluation is that it consists of various monolingual (non-English) and cross-language sub-tasks. That is to say, questions are given in one language, but the target corpus can be an array of news or Wikipedia documents in a different language. Essentially, these language pairs are picked from more than nine distinct European languages.

Like TREC, the QA track takes into consideration several kinds of questions, and it also includes definition queries. However, unlike TREC, an answer to a definition question is seen as a short string that briefly encapsulates the essence of the *definiendum* (e.g., "*President of Spain*"). The performance of definition QA systems is basically measured in tandem with the number of correctly answered queries versus the size of the test set of definition questions.

1.5 Types of Definitions Questions

To a great extent, definition questions have become especially interesting in recent years, because of its number of submissions to search engines, namely about 25% of queries in real search engine logs are requests for definitions [Rose and Levinson, 2004]. To begin with, the first phase in the answering process is distinguishing the type of query (information needed) that the user is demanding (e.g., a factoid, definition or list question). Despite dealing with a wide variety of natural language inputs, there are some distinctive structures utilised by users for indicating definition queries. The following is a list of a few of the most common across query logs,

COMMON STRUCTURES

- *What does <definiendum> mean?*
- *Who is/are <definiendum>?*
- *What is/are <definiendum>?*
- *What is the meaning of <definiendum>?*
- *Define <definiendum>*

Contrary to what these five cases might suggest, the subsequent categorisation of definition queries suffices for showing that this query analysis stage cannot be seen as an easy task. To state it more clearly, the following list dissects some of the main issues relating to query analysis and definition questions:

QUERY ANALYSIS FEATURES

1. **Multiple *definienda*.** In some cases, users want to find definitions of several *definienda*, thus they utilise the flexibility provided by natural language for putting them together into a sole, but more compact, query. Two delineative structures of this class include: "*What does <definiendum_1> and <definiendum_2> mean?*" and "*What is the definition of/for <definiendum_1> and <definiendum_2>?*".

   The degree of relation between the distinct *definienda* is uncertain, hence definition QA systems must normally tackle each of them independently. The user, nevertheless, can sometimes enter two synonyms or orthographical variations of the same underlying *definiendum*, because he/she is unsure of its correct spelling or what is the more appropriate entry to get the best possible outcomes from the system.

Equally, the connectors "*and*" and "*or*" bring about ambiguity to the query analysis. Take for instance the *definienda*: "*Akihito and Michiko*", "*Tom and Jerry*", and "*Trance and Dance*". In these depictive cases, it is dubious whether the user wants to find information about "*Tom*" and "*Jerry*", and/or about the cartoon "*Tom and Jerry*". Assuredly, this ambiguity occurs, whenever several potential *definienda* are concatenated forming a new valid *definiendum*. Definitively, in many cases, the most likely option is interpreting the input as a sole *definiendum*: "*The Fool and His Money*", "*Hatfield and the North*", "*For Love or Money*", and "*Deal or No Deal*". Frequently, the user is aware of this ambiguity, and therefore provides the *definiendum* with quotation marks.

CONNECTOR AMBIGUITY

2. **Misspellings/Ungrammaticality.** Occasionally, ungrammatical queries formulated by users can make the question analysis step harder, causing misleading inferences. Two examples of ungrammaticality found across query logs are: "*What does <definiendum> means?*" and "*What do it mean <definiendum>?*". Note that not all users that pose questions have mastered the English language.

3. **Multiple clauses.** Another observable phenomenon across query logs is two-clause questions. This type of input is comprised of a statement and a query, or of two questions. Users sometimes treat QA systems as dialog systems, hence they input colloquial constructions. As often than not, these constructions are geared towards making clearer what they are looking for to the machine. Besides, these constructions can consist of extra phrases directed at boosting the chances of discovering documents satisfying the information need of the user. The following two inputs are representative samples: "*What does <definiendum> means or what is it?*" and "*I need info on <definiendum>. Can you find me some?*". The former additionally illustrates that users can make grammatical mistakes when typing these colloquial-styled questions.

4. **Indirect requests.** In some cases, users implicitly request for definitions by asking for the location of internet resources, and/or their links thereof, that contain the solicited information:

   - *Where can I find the/a (good) definition for/of <definiendum>?*
   - *Where can/could/do I find (out) about <definiendum>?*
   - *Where can I find a biography of <definiendum>?*
   - *Where do I find facts about/on <definiendum>?*
   - *Where can I find a/an web site/article for/about/on <definiendum>?*
   - *Where can I find a definitive biography of <definiendum> online?*
   - *Where can I find an explanation of <definiendum_1> and <definiendum_2>?*

   The last illustrative query also underlines that this class of definition question can be intermixed with previous kinds (e.g., the first type). It is worth highlighting, nonetheless, that the response to this sort of query can be a ranking of links to knowledge bases containing all the required definitions. However, it is crystal clear that many *definienda* will not be found across knowledge bases, causing the definition QA system to produce answers from web documents as a fallback.

5. **Explicitly specified coverage.** At times, the user also enters salient hints about the expected length of the response that the definition QA system should return. This extra information is normally, but not necessarily, verbalised in the form of explicit keywords inserted into the query (e.g., "*some*" and "*anything*"). This is a potentially decisive factor

as the system could only output the most reliable answers to the user, while he/she is soliciting a concise description. This can also involve picking only entries to the most authoritative knowledge bases, and therefore, avoiding a more extensive search across a larger array of web documents. In like manner, cues such as "*new about*" signify that the definition QA system should explore only recently updated documents. Anyhow, this is arguable, because determining which are the novel pieces of information might entail the analysis of "old" resources or answer logs. Accordingly, some queries in this category are listed below:

- *Everything I can find out about <definiendum>*
- *Find meaning and picture for <definiendum>*
- *Can you find anything on <definiendum>?*
- *What do I need to know about <definiendum>?*
- *Can you find me some information about/on <definiendum>?*
- *Can you please tell me everything about <definiendum>?*
- *Where can I find out (more/something) about <definiendum>?*
- *Can you find basic information on <definiendum>?*
- *What is new about <definiendum>?*

In these illustrative definition questions, keywords such as "*everything*" signal that the intention behind the user is discovering as much information as possible about the target *definiendum*. By the same token, cues like "*some*" and "*anything*" are more indicative of the desire for a more succinct and precise definition, whereas other keywords or expressions, including "*basic*" and "*What do I need to know*", are more inclined to aim at essential attributes of the *definienda*. There also several cues that imply explicit special requirements such as pictures.

6. **Verbose.** Some queries are typified by their length. In particular, by the addition of unnecessary words, which make them more long-winded than informative. Some samples found across the query logs include:

   - *Can you find any information on/about <definiendum> (for me)?*
   - *What info/information (do) you have on/about/in <definiendum>?*
   - *Can you find (me) (some) information on/about <definiendum>?*
   - *I need to find some info/information about/on <definiendum>*
   - *Can you please tell me the definition of <definiendum>?*

   In some occasions, this verbiage can generate noise during the query analysis phase, and/or while the definition QA system is fetching promising documents.

7. **Short queries.** In contrast with the previous category, some definition questions are very concise:

   - *<definiendum> means*
   - *Describe <definiendum>*
   - *Tell me/find out about <definiendum>*
   - *The meaning/definition of <definiendum>*

- *<definiendum>*

The first example is aimed at discovering explicit descriptive content with an exact wording across the target collection (i.e., "*<definiendum> means meaning ...*"), while the last one is of ordinary usage. In this last case, the user enters only the *definiendum*. It can be conjectured that search engines are likely to return entries to knowledge bases at the top of their search results, due to the common utilisation of this last structure; particularly, every time they detect that the search string could be a plausible *definiendum*. This recognition can be allowed by checking the existence of the *definiendum* in the list of knowledge base entries. It is worth mentioning here that about 25% of queries sent to search engines are a request for definitions.

8. **Colloquial queries.** Some users structure their definition queries very colloquially: "*What the hell is <definiendum>?*" and "*What was going on in <definiendum>?*".

9. **Contextual queries.** Many times, the user is aware of the potential ambiguity of the *definiendum*, ergo he/she enriches the query with information about the target domain with the hope of obtaining more accurate results. Some delineative questions of this category are as follows:

   - *What does the abbreviation <definiendum> stand among <domain>?*
   - *What does <domain> abbreviation <definiendum> stand for?*
   - *From a/an <domain> standpoint, what is <definiendum>?*
   - *What does <definiendum> mean in relation to <domain>?*
   - *What does concept <definiendum> mean in <domain>?*
   - *What does <domain> define as/say about <definiendum>?*

   Another way of expressing context is through the second type. Take for instance the *definiendum*: "*The Bible and homosexuality*". This kind of query subordinates the facts about homosexuality to the context of or in consonance with "*the Bible*". This depictive *definiendum* also offers more insight into how diverse sorts of features can be fused to form more complex definition questions.

   Concatenation as Contextual Queries

10. **Abbreviations.** This sort of query is at the border between factoid and definition questions, because the answer can be a simple string, wherewith the acronym is resolved or some sentences that describe the essence of the *definiendum*. As a concrete example; the phrase "*The NSA stands for the principle of individual rights*". At any rate, in the context of definition QA systems, it is a widespread fact that acronym resolution is interpreted as part of a desirable explanation.

    - *What does the abbreviation of <definiendum> stand for?*
    - *What does the abbreviation <definiendum> mean in <domain>?*
    - *What does <definiendum> stand for on <domain>?*

1.6 What is a Definition?

One of the most intriguing aspects of definitions is the fact that people use terms that they do not necessarily know how to explain their meanings. This inherent separation between elucidating a meaning and usage is evidence suggesting that words are mere instruments of language. As tools of language, one of their crucial function is aiding humans in forming the idea of a concept of reality in their minds [Wittgenstein, 1953]. According to [Swartz, 1997], the reason why people care about definitions is the continuing necessity for redefining concepts. Substantially, this redefinition happens in three distinct ways:

1. The expansion of vocabulary, this means the transmission of the (new) meaning of a (new) term from one person to the other.
2. The reduction or elimination of ambiguity. On numerous occasions, different possible meanings for a given term exist, thus it is necessary to explicitly condition the usage of a word to a new or to one of their meanings.
3. The diminution of vagueness.

With this in mind, it can be said that dictionaries do not define or create new terms, they rather report on the usages that people give to them. In so doing, dictionaries explicate the standard and most common usages, sometimes by giving hints. At any rate, dictionaries are not aimed at having the last word on the meaning of terms. In reality, there is no such thing called real meanings independent of the language and the persons [Wittgenstein, 1953].

CONTEXT AND MEANING

In other words, all terms have their meaning subject to how they are utilised in a specific context, namely persons and language. Then, a word may be given various (a family of) meanings in consonance with its usage. An archetypal term is "*God*", this word has diverging connotations as they are several disparate religions (e.g., *Buddhism*, *Christianity*, and *Islam*). Thus, each religion has its own contextual meaning of "*God*". On further examination of the various denominations (e.g., *Christianism*), several diverging stipulations can be found which are directed at disambiguating or specifying their different conceptions (e.g., *Catholics*, *Protestants*, and *Lutherans*). Furthermore, it is still possible to zoom in, and elucidate the meaning of "*God*" in consonance with the various Catholic denominations, and what is more; different meanings can still be distinguished across various Catholic home churches within each specific denomination. From another angle, the usage of the words "*God*" and "*ace*" can be synonymously found in the context of sports. As a natural consequence of these distinct conceptions, people have difficulty in understanding each other. Incidentally, it is notable that the context definition queries, mentioned in section 1.5, in conjunction with the disambiguation pages supplied by KBs, such as Wikipedia, provide collaborative evidence supporting this marriage between a definition and a context. As a result of this alliance, the meaning of words may change in accordance with changes in the context. Roughly speaking, definitions emerge from the culture and society in which they are used [Wittgenstein, 1953].

The attachment of definitions to a context consequently stresses the great relevance of definition QA systems as it is desirable to account for a strategy capable of automatically discovering potential meanings of a term (*definiendum*) in all its plausible contexts. There are two aspects worth emphasising here: (1) the number of feasible contexts is not stipulated by the entries in KBs, but rather by the target collection of documents; and (2) this makes methodologies that exploit a limited set of KBs or glossaries less attractive as they are more likely to specify the most prominent and prevalent meanings in the most predominant contexts. Since users might know the denotations in KBs, it cannot be ruled out that they might be especially targeted at discovering usages stipulated in narrow and precise contexts that they might ignore (e.g. in a particular piece of legislation).

As stated in the following quote, apart from using words, humans also create terms and names in a certain context (e.g., "*Climate-Gate*"). As a matter of fact, it is a natural activity to name objects, or to attach labels to things:

> "One thinks that learning language consists in giving names to objects. Viz, to human beings, to shapes, to colours, to pains, to moods, to numbers, etc. To repeat - naming is something like attaching a label to a thing. One can say that this is preparatory to the use of a word. But what is it a preparation for?" (Philosophical Investigations I §26) [Wittgenstein, 1953].

NAME AND REFERENCE

The archetype of this activity is naming people when they are born. Names (*definienda*) are an anticipate preparation for referring to human beings later. This scenario makes clear that: (a) names are facilitators of this reference process; (b) why people are inclined to assign distinct names to their surrounding people; (c) names can be borrowed from one context to the other; and (d) concepts that are not likely to be referred to will improbably be labelled.

SCOPE AND CONTEXT

This also explains why encyclopedias are more likely to embody articles about those names that are more likely to be referred to by a considerable amount of people. Certainly, the number of references to a particular name grows in consonance with the pertinence of its essential characteristics. This relevance is also determined by the size of its context. For example, the kind of achievements accomplished by the referred person or organisation(e.g., the best or first in history). These are normally famous people and organisations in some prominent contexts (e.g., *biology*, *music*, *technology organisations*, and *sports*). Inversely, it is harder for people relatively well-known, minor contexts or private contexts to be included in KBs. In summary, contexts fluctuate in size, and the larger it is, the more likely their encircled meanings of terms are to be embraced by KBs.

ACCUMULATIVE MEANINGS

On a different note, a *definiendum* may not depend on whether it refers to something that actually exists (e.g., *unicorn*). In reality, if something ceases to exist, the word or name for that thing may still have meaning [Wittgenstein, 1953]. In this category, biography articles fit perfectly as they commonly describe dead people. This also helps us to understand that meanings of words or names are accumulative, and therefore are of increasing ambiguity; particularly, nowadays with the skyrocketing capacity of electronic storage. Although some meanings are not frequently used anymore in our daily lives, old connotations can still be consulted by a user. In actuality, this infrequent usage makes them more likely to be prompted to a definition QA system. In the working term, the user can enquire about some denotations of "*God*" utilised by ancient tribes and civilisations (e.g., the Golden Calf, Greek and Roman gods).

Some Characteristics of Definitions

TIMELESS CHARACTERISTICS

One characteristic of explanations of meanings is that they are tend to convey timeless properties of the *definiendum*. Since this class of attributes has a permanent relationship with the *definiendum*, they are very likely to be utilised for elucidating its meaning. The fact that a property, however, is immutable does not make it a prerequisite. The feature has to additionally characterise the *definiendum*. Consider the definition of the term "*tree*" supplied by the Oxford Dictionary:

Trees have a thick central wooden stem (the trunk) from which branches grow, usually with leaves on them.

This explication details several timeless properties of trees: "*have one stem, wooden stem, central stem, thick stem, have branches, the branches grow from the stem, may have leaves on the*

branches." However, these attributes are not only timeless, but they are normally common to the essence of all types of trees. Plainly speaking, in the past, the present, and probably in the future, trees will have these virtues.

HYPERNYM-HYPONYM

One can also argue the essentiality of these attributes. That is to say, why the trunk is a required feature of a tree, while not the fact that it is a "*living thing*". This might sound like a valid argument, but it is somewhat debatable. The elemental quality of being a "*living thing*" is subsumed in the fact that a tree is a (type of) plant. Thus, in order to imply the essential properties of a tree as a plant, it is naturally preferable to say explicitly: "*Trees are plants*", or to do this implicitly by providing a subtle or clear hint to this fact: "*the branches grow from the stem*". This sheds light into how a hypernym can influence an explanation of meaning. Now, hyponyms of trees also derive its properties, consider the "*maple*":

A tall tree with leaves that have five points and turn bright red or yellow in the autumn/fall. Maples grow in northern countries.

As the descriptions correspond to lower terms in the semantic hierarchy, they are inclined to focus their attention on more distinguishing features (e.g., "*leaves that have five points*"). Expressly, it would be unusual to enrich an abridged definition of "*maple*" with the fact that it is a "*living thing*". Ergo signifying that the definitions corresponding to preceding hypernyms in the hierarchy might be tacitly considered. In conclusion, values of properties (e.g., "*types of leaves*") that categorise a *definiendum* into its different hyponyms are more likely to be utilised for explicating the meaning of its hyponyms than the *definiendum* itself, whereas the explication of the *definiendum* would probably refer to the property (e.g., "*it has leaves*"). Similarly, "*a thick wooden stem*" distinguishes a "*tree*" from another plant like a "*bush*" or a "*shrub*". This means these qualities, along with their categorising values, encapsulate part of the essence of the hyponyms. On the other hand, leading or indispensable characteristics of the hypernym are more likely to be tacit or subtly referenced. This conjecture relies on the extension of the definition. Consider biographical articles, they are likely to be thorough, and they hence portray some attributes emanated from their hypernyms. For instance, articles on painters talk about some of their qualities as a person (e.g., birth/death place and date).

TIME-DEPENDENT PROPERTIES

Contrarily, time-dependent attributes are normally explicitly stipulated. One way of doing this is by extending the name (e.g., "*Trees in 1970*"), or by adding a temporally anchored explanation to the description "*in 1970 trees decreased their height by 80%*". Since time and space are the same concept in different dimensions, features reliant upon a physical location are conveyed in the same way (e.g., "*trees in Colombia*" and "*grow in northern countries*"). This class of properties increases the complexity of the description.

MERONYMY

Another peculiarity of definitions is the fact that a *definiendum* can be explicated with the assistance of a list of its foremost parts. These parts can be typically recognised by means of our senses (sight, hearing, smell, taste and touch). Distinctly, it can be observed in the working definition of "*tree*" that its parts: "*stem*", "*branches*" and "*leaves*", are utilised for recalling or projecting the image of an actual tree into the mind. However, listing the parts of the *definiendum* is usually not enough to unambiguously communicate its meaning. In fact, depicting a tree as a stem put together with a set of branches and leaves could cause the receptor of this explanation to give rise to a misconception, this means not necessarily the tree. It is therefore necessary to specify the condition that the branches must grow from the stem, and the leaves from the branches, and/or providing greater details about its parts (e.g., "*wooden stem*" and "*thick stem*").

PARAPHRASE

Definitions can also describe the *definiendum* by means of a paraphrase. The next are delineative samples taken from the Oxford Dictionary:

Odour ⇒ a smell, especially one that is unpleasant.

Scent ⇒ the pleasant smell that sth has.
Aroma ⇒ a pleasant, noticeable smell.
Smell ⇒ to have a particular smell.

These explanations of meanings also unveil that, in some cases, the description can use a synonym of the *definiendum*, or the *definiendum* itself (e.g., smell) [Swartz, 1997]. On the other hand, some definitions are elucidated in terms of a sequence of examples (Oxford Dictionary):

DEFINITION BY EXAMPLES

Blue ⇒ having the colour of a clear sky or the sea/ocean on a clear day.

In this type, the meaning can be inferred from the likeness between the essences of the exemplars. This sort of definition, however, incurs the risk of making the reader to deduce the wrong array of attributes that permeate the examples, and therein lies the potential mistakes when applying the derived rules. As well as that, the list of examples can be negative. Take the term "*animal*" (Oxford Dictionary):

Animal ⇒ a creature that is not a bird, a fish, a reptile, an insect or a human.

Lastly, the connection between a word and its meaning may be arbitrary. In a peculiar context, a person may arbitrarily choose to adopt the term "*cold*" to describe something which is actually warm, while the word "*warm*" to something which is cold [Wittgenstein, 1953].

ARBITRARINESS OF MEANING

1.6.1 Types of Definitions

Practically speaking, [Swartz, 1997] identified seven distinct kinds of definitions. This classification does not contemplate completeness or exhaustiveness, but it still yields a good framework for fleshing out some of their assorted characteristics:

1. **Stipulative definitions** present new terms (e.g., abbreviations), or narrow the usage of a word in a special context. The introduction of this type of definition normally causes inconsistencies as they bring about conflicts with the unstipulated use of the redefined term.

2. **Lexical definitions** are specifications of common usages of words. Dictionary definitions are the quintessence of this group as they can be utilised also for regulating and standardising the utilisation of terms. At large, [Swartz, 1997] distinguished the following subcategories of this class: Synonyms, reports on the grammatical use of words, Species-Genus, Anonyms, Implicit and Explicit cause, Functional and Circular. Notably, *Species-Genus* definitions cooperate on discovering an instance (specie) of the *definiendum* (genus), while *Circular* definitions require the use of the *definiendum* in their explanations of meanings. This is commonly utilised for describing non-visible phenomena including scents, pains, and sounds. In other cases, the circle is broken by means of pictures.

3. **Precising definitions** refine the meaning of words, whose explanation is nebulous in a context. Recurrently, this is found in the legal and medical domains, where terms are vague or incomplete, hence their definitions are constantly ameliorated.

4. **Theoretical definitions** are associations of words with a well-defined set of properties. This array of properties is predicated on well-established beliefs or theories. This can

be in the context of science or in daily life. For instance, the term "*fruit*" has some inherent attributes like its origin, being the product of something, and the time it takes to ripen. These properties can be abstract and applied to refer to other concepts bearing these similarities like "*the fruits of your labour.*"

5. **Operational definitions** are built with a specific purpose. Normally, they stipulate a condition that it is imperative to draw further inferences and/or to understand concepts. These are constructed on top of the operational description of the *definiendum*. In this kind of definition, the validity of these inferences does not necessarily uphold when accounting for alternative operational descriptions. Excellent examples are currencies, for instance: "*The Euro is the equivalent to 1.495 American Dollars.*"

6. **Recursive definitions** consist of two parts. To understand this clearly, one takes two different sentences where the first one contains the *definiendum* and its description, and the second establishes a link between a new concept and the explanation stressed by the first sentence. Consequently, a tacit connection between the new concept and the *definiendum* exists. A delineative example is the "*parent of*" relationship and the next two sentences: "*George W. Bush is the parent of Barbara Pierce Bush.*" and "*George H.W. Bush is the parent of George W. Bush.*". The induction or recursive step supplies the definition "*George W. Bush is the parent of a parent of Barbara Pierce Bush.*"

7. **Persuasive definitions** are directed at making someone to agree or believe that something is true, when its validity could be at stake. To exemplify, [Swartz, 1997] brought up the case of citations of definitions in heated arguments. Some simple cases are as follows:

   Islam ⇒ religion that teaches hatred and violence and intolerance.
   Evolution ⇒ world as created by God.
   Pre-emptive war ⇒ supreme international crime.

   These examples convey debatable information, but they might get the support of many people, ergo appearing to be definitive. However, consider the next description in juxtaposition to the previous example:

   Pre-emptive war ⇒ case when power A attacks power B because of an overriding belief that power B is certain to attack power A in the near future.

   Certainly, this sentence renders less arguable information, hence unveiling the dubious nature of its counterpart.

1.6.2 Definitions in TREC

Fundamentally, the conception of definition utilised in the TREC challenge extends the idea of essentiality to, broadly speaking, biographical knowledge. In other words, answers to definition questions are not seen solely as succinct meaning explanations that embody the essence of the *definiendum*, but rather they shift their focus of attention to find a set of essential nuggets about the *definiendum*. Take, for instance, the definition question "*Who is Flavius Josephus?*" and the following illustrative text fragments:

also known as Yosef Ben Matityahu
Jewish

historian and apologist
born AD 37
recorded the destruction of Jerusalem in AD 70
wrote the Jewish War in AD 75
wrote Antiquities of the Jews in AD 94
In 71 became Roman citizen
In 75, married for third time
father of Flavius Hyrcanus, Flavius Justus
father of Flavius Simonides Agrippa

The underlying motivation behind this change in interpretation stems from the fact that, as shown by query logs, web users can seek ample coverage and diverse nuggets about numerous concepts, persons, locations, events or things (see explicitly specified queries in section 1.5). This is principally because of the wide variety of information that can be found on the Internet. Certainly, not all titbits qualify for the final output, but rather, as [Han et al., 2006] put it, those classes of elements that can typically be found across dictionaries and encyclopedias:

> "The definition about a [*definiendum*] consists of conceptual facts or principal events that are worth being registered in a dictionary or an encyclopedia for explaining the [*definiendum*]" [Han et al., 2006].

Although this definition does not explicitly state what is "*worth*" and what is not, it establishes a link between the content of KBs and answers to definition queries. Thus, "*worth*" can be perceived in terms of the support given by KBs. One way of quantifying this support could be the frequency count of the nugget type. In a statement, answers to definition questions in TREC tend to yield biographical knowledge, in general.

1.6.3 Length of Definitions

Since it is all-important to provide enough context to ensure the readability of the final answer, definition QA systems prefer sentences to nuggets. However, some systems still output paragraphs, sentences, or provide links to the full-page. The criterion for selecting the abstraction level of the answers depends largely on the goals of the definition QA system.

1.7 Evaluation Metrics

Generally speaking, there is no single metric that supplies a definitive and complete view of the performance of definition QA systems. In part due to the fact that different systems or components are designed to meet distinct requirements, stressing different facets of their output. Ergo, several metrics have been utilised for evaluating systems and components in order to assess their different pivotal aspects. Between the most broadly used metrics, one can find: $\mathcal{F}(\beta)$-Score, precision at k, Mean Average Precision (MAP) and Accuracy.

$\mathcal{F}(\beta)$-Score has been used regularly for assessing definition QA systems in the TREC track since 2003 [Voorhees, 2003, Harman and Voorhees, 2005]. This measurement balances the precision and recall of a system by making a judgement about its output with respect to a manually generated ground truth. To exemplify, the TREC gold standard with respect to the 2004 *definiendum* "*Nirvana*" contains the following nuggets in table 1.1:

ID	Category	Nugget
1	**vital**	seminal band
2	**vital**	originated in Seattle
3	okay	Grohl was Nirvana's drummer
4	okay	Grohl later played in Foo Fighters
5	okay	Courtney Love was married to Kurt Cobain
6	**vital**	Cobain committed suicide
7	okay	Love fronted band Hole
8	okay	Cobain died at 27
9	okay	Cobain died in 1994
10	okay	Nirvana ended with Cobain's death
11	okay	Cobain played guitar
12	okay	Cobain used heroin

Table 1.1: TREC 2004 ground truth for the *definiendum* "*Nirvana*".

Basically, there is no explicit limit to the number of nuggets per *definiendum*, and each nugget is associated with a category. In this assessment, the ground truth gives a hierarchy of nuggets, which consists of "vital" nuggets (must be in the description of the concept) and "okay" nuggets (not necessary). These labels are assigned by human assessors in agreement with the requirements and the objectives of each particular definition QA system. The list of nuggets is essentially constructed in relation to the target corpus. In detail, this assessment takes into consideration the next aspects:

v = number of vital nuggets returned in a response.
o = amount of okay nuggets returned in a response.
g = number of vital nuggets in the gold standard.
h = amount of non-whitespace characters in the whole output.

LENGTH ALLOWANCE

Then, a length allowance (α) of 100 non-whitespace characters per matched nugget was imposed in order to cope efficiently with two central aspects: (a) different paraphrases of a particular nugget can be found, and hence their corresponding lengths differ from one rewriting to the other, and (b) many nuggets need their context to be readily comprehensible.

PRECISION

The allowance of the output of a system is accordingly defined as $\alpha = 100 \times (v + o)$. If the length of response exceeds this allowance, the precision (P) obtained by the system is then linearly downgraded:

$$P = \begin{cases} 1 & \text{if h} < \alpha \\ 1 - \frac{h-\alpha}{h} & \text{otherwise} \end{cases}$$

As a matter of fact, the parameter α can also vary in agreement with the intentions of the application. For instance, descriptive sentences taken from web snippets are about 110 non-whitespace characters long on average [Figueroa and Neumann, 2007], they can thus be interpreted as nuggets, and therefore, a reference value of 110 might seem to be more

RECALL

appropriate. The recall (R) of the system is subsequently calculated as follows:

$$R = \frac{v}{g}$$

This ratio implies that the recall rewards a system solely for the amount of "vital" nuggets subsumed in the output. The $\mathcal{F}(\beta)$-Score value is, eventually, computed by balancing the trade-off between precision(P) and recall(R):

$$\mathcal{F}(\beta) - Score = \frac{(\beta^2 + 1) \times P \times R}{\beta^2 \times P + R}$$

In TREC 2003, β was set to five, but since this value was heavily biased towards large responses, it was later decreased to three. Like α, the value of β also depends chiefly on the type of application. To reinforce this point, imagine a system that returns small text messages, this would prefer concise nuggets over large responses, while a web search engine would probably favour more contextual information strongly. To illustrate this measure, consider the working output in table 1.2 belonging to a hypothetical definition QA system operating on the AQUAINT corpus:

Nuggets	Sentences
6v,9o,12o	Along with Seattle's fame for high-tech innovation and cultural vibrancy, Seattle has long battled a grim notoriety for Seattle's heroin problem, one that was aggravated by the 1994 suicide of Kurt Cobain, the grunge rock star and lead singer for the group Nirvana, who had struggled with heroin addiction.
NO-NUGGETS	Former Nirvana bassist Krist Novoselic talked about Former Nirvana bassist Krist Novoselic's own political activism, and of witnessing the November riots at the World Trade Organization meeting in Seattle.
6v	But the scene was fading by 1994 after Nirvana front man Kurt Cobain killed Nirvana front man Kurt Cobain and other bands were stymied by drug problems, were breaking up or were backing away from the limelight.

Table 1.2: Sample response corresponding to the *definiendum*: "*Nirvana*" (after co-reference resolution).

Accordingly, in the working example, the set of parameters is as follows: v = 1, o = 2, g = 3, and h = 617 characters. The recall is consequently given by $R = \frac{1}{3}$ and $\alpha = 100 \times (1 + 2) = 300$. Since h=617 is greater than α=300, the precision of this output is determined by: $P = 1 - \frac{617-300}{617} = 1 - \frac{317}{617} = \frac{300}{617}$. Hence, the $\mathcal{F}(\beta)$-Score is computed as follows:

$$\mathcal{F}(\beta) - Score = \frac{(\beta^2 + 1) \times \frac{300}{617} \times \frac{1}{3}}{\beta^2 \times \frac{300}{617} + \frac{1}{3}}$$

Table 1.3 shows the values obtained for different βs:

β	1	2	3	4	5
$\mathcal{F}(\beta)$-Score	0.39552	0.35570	0.34416	0.33962	0.33741

Table 1.3: Example of $\mathcal{F}(\beta)$-Score (β=1...5).

ZERO RECALL

As [Lin and Demner-Fushman, 2006, Figueroa and Neumann, 2007, Figueroa, 2008c] duly pointed out, whenever a definition QA system does not discover at least one "vital" nugget, it finishes with a recall equal to zero, ergo bringing about a $\mathcal{F}(\beta)$-Score equal to zero. This issue grossly distorts the comparison of systems, because some systems can still output "okay" nuggets and their output lengths can dramatically differ. Furthermore, definition QA systems that do not return an answer are punished as equally as systems that render only

unrelated information as a response. Since these zero values are completely unuseful for juxtaposing systems, [Lin and Demner-Fushman, 2006] adopted a new $\mathcal{F}(\beta)$-Score to mitigate this problem. This modifies the recall to make allowances for weighted nuggets as follows:

$$R = \frac{\sum_{x \in A} z_x}{\sum_{y \in Z} z_y}$$

In this recall, A and Z are the set of all nuggets in the output and in the ground truth, respectively; while z_x and z_y denote the weights of two nuggets x and y. More specifically, [Lin and Demner-Fushman, 2006] computed these weights by averaging the opinions of several assessors regarding the text fragments in the gold standard. Precisely, the weight of each nugget is equal to the number of assessors who labelled it as "vital". These weights are subsequently normalised by dividing by the highest score.

GOLD STANDARD COVERAGE

In the TREC assessment, the ground truth is manually compiled, and it is known that systems in this evaluation were able to find pertinent nuggets, which did not make it onto the list, even as "okay" nuggets. As a means to support this observation, [Hildebrandt et al., 2004] brought out the following cases extracted from the AQUAINT corpus, which were left unconsidered from the TREC 2003 gold standard:

(a) The *definiendum* "*Alberto Tomba*" and the fact that he is Italian was not judged to be relevant.

(b) The *definiendum* "*fractals*" and the idea that they can be described by simple formulas, which is one of their most important virtues.

(c) Additional descriptions include the following:

Aga Khan is the founder and principal shareholder of the Nation Media Group.
The vagus nerve is the sometimes known as the 10th cranial nerve.
Alexander Hamilton was an author, a general, and a founding father.
Andrew Carnegie established a library system in Canada.
Angela Davis taught at UC Berkeley.

GROUND TRUTH INCONSISTENCY

Another problem with the TREC gold standard is its inconsistency; some types of nuggets are interpreted as crucial to one query, but irrelevant to other *definienda*, and hence excluded from their respective ground truth. To illustrate this, in contrast to "*Alberto Tomba*", the nugget "*Danish*" made it as "vital" onto the list of the assessors concerning "*Niels Bohr*". This sharp difference in the interpretation of attributes seriously compromises the quality of the TREC evaluation, because it lacks the necessary basic level of coherence to make the performance comparable between distinct instances of the same kind of *definiendum* (e.g., persons). Simply put, it would be unreasonable to expect that definition QA systems arbitrarily or randomly include/exclude some descriptions in/from the ground truth, therefore consistent guidelines are necessarily needed to carry out robust and meaningful evaluations.

CONTEXT QUERIES

Unfortunately, this is not an isolated case. A closer look of the TREC 2004 evaluation provides more insight into this issue. In TREC 2004, context queries were adopted in order to increase the complexity of the question answering task. A sample of context queries is shown in table 1.4:

ID	Category	Question
11.1	FACTOID	Who is the lead singer/musician in Nirvana?
11.2	LIST	Who are the band members?
11.3	FACTOID	When was the band formed?
11.4	FACTOID	What is their biggest hit?
11.5	LIST	What are their albums?
11.6	FACTOID	What style of music do they play?
11.7	OTHER/DEFINITION	Other

Table 1.4: TREC 2004 context queries about "*Nirvana*".

In this context, definition QA systems are forced to ban or exclude all descriptions that respond to the previous questions prompted in the same context. As a repercussion, making the output to definition questions and their corresponding ground truth dependent upon the respective context. This is relevant, because the array of previous queries can vary from one *definiendum* to the other, causing the gold standard for definition questions to be inconsistent and incoherent. For example, consider the juxtaposition of the context queries concerning "*The Clash*" and "*Nirvana*" (tables 1.4 and 1.5).

ID	Category	Question
39.1	FACTOID	What kind of music does the band play?
39.2	FACTOID	In what year was their first major album recorded?
39.3	LIST	Name their songs.
39.4	OTHER/DEFINITION	Other

Table 1.5: TREC 2004 context queries about "*The Clash*".

While it is true that questions 11.6 and 39.1 can be homologated (analogously to the pair 11.5 and 39.3), it is also true that the query 39.2 could also be applied to "*Nirvana*", in the same way that other queries in table 1.4 could also be prompted when dealing with "*The Clash*". But not only the preceding queries differ, tables 1.1 and 1.6 also contrast tangible differences between both ground truths.

ID	Category	Nugget
1	**vital**	mainstream success album Combat Rock
2	**vital**	Clash drummer Topper Headon
3	okay	mistakes treated as learning experiences
4	okay	Rancid plays faster than Clash
5	**vital**	Mick Jones co-founded Clash
6	okay	first U.S. tour in New York City
7	okay	richness of Clash songwriting

Table 1.6: TREC 2004 gold standard for the *definiendum* "*The Clash*".

In a sense, nuggets, such as "*Clash drummer Topper Headon*" (vital) and "*Grohl was Nirvana's drummer*" (okay), along with questions, like 11.2 in table 1.4, indicate that responses to potential preceding factoid or list queries could be incorporated into the list of nuggets corresponding to their respective definition questions, as also pinpointed by [Han et al., 2006]. In a nutshell, this inconsistency reduces the portability of the ground truth to other corpus

and to another definition queries of the same types.

However, in favour of the design of the TREC gold standard, it can be said that it depends heavily on the target corpus, hence it might be perfectly possible that this corpus provides occurrences for some nugget types for one instance of a particular kind of *definiendum*, while for another occurrence of the same sort of *definiendum* the instances for the same types of nuggets do not exist. The corpus, therefore, can be a partial justification for a purpose-built gold standard, and thus, for the internal inconsistency in the ground truth. This is reasonable, but not strictly valid, because definition QA systems should also be capable of detecting when the instance of a particular kind of nugget cannot be found in the target collection. Certainly, this seems to be a difficult achievement so far, consequently making this justification eminently acceptable.

CORPUS AND GOLD STANDARD CONSISTENCY

Another aspect is the relevancy and sufficiency of the nuggets in the gold standard. Specifically, manifold nuggets within the ground truth might seem unnecessary or uninteresting for some users (e.g., "*first U.S. tour in New York City*" or "*seminal band*"). Conversely, several supplementary and novel nuggets, or nuggets closely related to ones already embraced by the ground truth were totally ignored. As an illustration of this "idiosyncratic" issue, the reader can consider the next two nuggets which were deemed as irrelevant by the assessors: "*dead from an apparently self-inflicted gunshot wound*" and "*dead in Seattle*", along with their supporting sentence taken from the AQUAINT corpus (after co-reference resolution):

NUGGET IMPORTANCE

> Five years ago: Kurt Cobain, singer and guitarist for the grunge band Nirvana, was found dead in Seattle from an apparently self-inflicted gunshot wound; Kurt Cobain, singer and guitarist for the grunge band Nirvana, was 27 .

Inherently, there is an element of subjectivity in the list of the assessors. First, the coverage of the list must be defined, that is, which facets will be conceived as relevant in conjunction with their extent. Determining the extent of each facet demands the stipulation of the necessary attributes that must be taken into consideration. A textbook case are *definienda* aimed at historical figures. One of the facets is the event of their death, thus the assessor will have to settle whether its extension incorporates the death date or death place, or cause of death, or a combination of these into the gold standard. Furthermore, after specifying which facets and their corresponding extension, the assessors need to decide the coverage of each text fragment. In other words, they are required to stipulate how many nuggets each facet will provide, that is, the death place and date will comprise one or two distinct nuggets. What is more, when performing the semantic matching, it is indispensable to restrict the length (sufficiency) of details that an answer candidate (nugget) must embody to qualify for being a valid response (e.g., "*band*", "*grunge band*", or "*rock and grunge band*").

GROUND TRUTH COVERAGE

NUGGET COVERAGE

NUGGET SUFFICIENCY

A promising solution to the issue of subjectivity is exploring similarities across Wikipedia abstracts about distinct instances of the same kind of *definiendum*. To neatly illustrate, figures 1.3 and 1.4 yield the abstracts corresponding to the working examples. In both cases, they draw attention to facts about their line-ups, and how they changed over time, formation date, type of music, genre, origin, the message of their music (lyrics), information about their most important albums, and achievements. Under the assumption that most pertinent classes of nuggets will have a higher frequency in the context of a particular sort of *definiendum*, the gold standard and the weights z_x can be assigned in tandem with these frequencies instead of the opinion of a certain group of assessors, which can be entitled to their idiosyncrasy. Certainly, extracting a set of weighted nugget types for each plausible kind of *definiendum* would involve the construction of a full-taxonomy of definitions. In the long-term, this taxonomy can additionally help to distinguish which nuggets are more important and/or harder to

TAXONOMY OF DEFINITIONS

Nirvana was an American rock band that was formed by singer/guitarist Kurt Cobain and bassist Krist Novoselic in Aberdeen, Washington in 1987. Nirvana went through a succession of drummers, the longest-lasting being Dave Grohl, who joined the band in 1990.
With the lead single "Smells Like Teen Spirit" from the band's second album Nevermind (1991), Nirvana entered into the mainstream, bringing along with it a subgenre of alternative rock called grunge. Other Seattle grunge bands such as Alice in Chains, Pearl Jam, and Soundgarden and also the San Diego based band Stone Temple Pilots had also gained popularity, and as a result, alternative rock in general became a dominant genre on radio and music television in the United States during the early-to-mid-1990s. As Nirvana's frontman, Kurt Cobain found himself referred to in the media as the "spokesman of a generation," with Nirvana the "flagship band" of Generation X. Cobain was uncomfortable with the attention and placed his focus on the band's music, believing the band's message and artistic vision to have been misinterpreted by the public, challenging the band's audience with its third studio album In Utero (1993).
Nirvana's brief run ended with Cobain's death in April 1994, but the band's popularity continued in the years that followed. In 2002, "You Know You're Right," an unfinished demo from the band's final recording session, topped radio playlists around the world. Since their debut, the band has sold over twenty-five million albums in the US alone, and over fifty million worldwide.

Figure 1.3: Abstract extracted from Wikipedia about "*Nirvana*" (As of October 2009).

detect, and assign their weights accordingly. Special nuggets not contemplated in the taxonomy could obtain a pre-determined standard weight. Eventually, weights related to nuggets that can be applied to a *definiendum* can be normalised so that the recall can range from zero to one.

AUTOMATIC SEMANTIC MATCHING

On a different note, in order to assess the performance of a particular system, each assessor has to manually validate which nuggets within the ground truth are included in the response of the definition QA system. This manual matching process is also a demanding task. As a means to deal with that, [Lin and Demner-Fushman, 2005] proposed an adaptation of POURPRE under the hypothesis that term co-occurrence statistics can serve as a surrogate for the manual semantic matching process. More specifically, they added unigrams co-occurrences between nugget terms and words in the output. They confined this matching so that all nugget words appear within the same response string (e.g., sentence). [Lin and Demner-Fushman, 2005] imposed this restriction under the assumption that nuggets represent coherent concepts, they are thus unlikely to be spread across various answer strings. Accordingly, the recall is now computed as a ratio of the sum of the matching scores for all "vital" nuggets to the total amount of "vital" nuggets. Consequently, the length allowance is now given by 100 non-whitespace characters per nugget that obtains a matching score greater than zero. Additionally, they studied their matching strategy when enriched with features including different term weights and stemming.

However, this matching strategy is useful when text fragments in the gold standard share a substantial number of words with most of their respective paraphrases across the target corpus. In this scenario, this automatic matching is naturally preferable to a manual semantic judgement. However, in larger target collections (e.g., the Internet), there is a rise in the probability of incorporating paraphrases into the response that do not share terms with the respective entry in the ground truth. The exclusion of these paraphrases from the gold standard in conjunction with their inclusion in the final outputs actually brings about a decline in the $\mathcal{F}(\beta)$-Score, because they enlarge the response without increasing precision and contributing to the recall. In the case of web-based systems, this vital fact is more likely to happen, because definition QA systems discover manifold nuggets paraphrased with words not used by their counterparts within the gold standard.

In regard to the ground truth, the extraction and construction of this gold standard is, in general, an arduous task, because it inherently entails manually checking the target corpus.

The Clash were an English rock band that formed in 1976 as part of the original wave of British punk rock. Along with punk, they experimented with reggae, ska, dub, funk, rap and rockabilly. For most of their recording career, The Clash consisted of Joe Strummer (lead vocals, rhythm guitar), Mick Jones (lead guitar, vocals), Paul Simonon (bass, backing vocals, occasional lead vocals) and Nicky "Topper" Headon (drums, percussion). Headon left the group in 1982, and internal friction led to Jones's departure the following year. The group continued with new members, but finally disbanded in early 1986.
The Clash were a major success in the UK from the release of their debut album, The Clash, in 1977. Their third album, London Calling, released in the UK in December 1979, brought them popularity in the United States when it came out there the following month. Critically acclaimed, it was declared the best album of the 1980s a decade later by Rolling Stone magazine.
The Clash's politicised lyrics, musical experimentation and rebellious attitude had a far-reaching influence on rock, alternative rock in particular. They became widely referred to as "The Only Band That Matters", originally a promotional slogan introduced by the group's record label, CBS. In January 2003 the band-including original drummer Terry Chimes-were inducted into the Rock and Roll Hall of Fame. In 2004, Rolling Stone ranked The Clash number 30 on their list of the 100 Greatest Artists of All Time.

Figure 1.4: Abstract excerpted from Wikipedia about "*The Clash*" (As of October 2009).

As a rough rule of thumb, in the evaluations in sections 5.3 (page 128) and 6.5 (page 149), the TREC 2003 consists of 50 different *definienda*: 30 are for people (e.g., "*Alberto Tomba*"), 10 are for organisations (e.g., "*ETA*") and 10 are for other entities (e.g., "*vagus nerve*"). In this assessment, [Figueroa, 2008b, Figueroa and Atkinson, 2009] retrieved about 300 web snippets for each of *definiendum*, therefore about 15,000 web snippets must be manually inspected in order to determine the gold standard. Assuredly, this number can double to about 30,000 when a baseline system is taken into consideration.

WEB NUGGET WEIGHTS

As to web-targeted definition QA systems, [Figueroa, 2008b, Figueroa and Atkinson, 2009] preferred equally weighted nuggets, that is $z_y = 1, \forall z_y \in Z$. Their reason to use uniform weights is three-fold: (a) under the assumption that more relevant nuggets will be embraced by a larger amount of documents, they attempted to weight them according to the amount of snippets where they occur, but this caused all systems to obtain a high recall, because high in frequency nuggets are usually easier to discover, and little is gained when nuggets low in frequency are identified; (b) if a highly frequent nugget is missed by a system, it needs numerous nuggets in low frequency to recover from the loss, bringing about a gross distortion of the performance as many of these low frequent nuggets can be hard to discern; and (c) the gold standard is largely reliant on the search strategy, and therefore the distribution of weights could sharply vary from one search approach to the other, turning to be a critical factor when likening different systems. All things considered, [Figueroa, 2008b, Figueroa and Atkinson, 2009] defined a $\mathcal{F}(\beta)$-Score, where the recall is calculated as the ratio of the number of detected nuggets to the amount of nuggets in the ground truth of the respective definition question. Ideally, these weights should be given by a purpose-built standard taxonomy, as mentioned earlier.

In praxis, this changes the value of the recall obtained by the output in table 1.2 to: $R = \frac{3}{12} = \frac{1}{4}$. Correspondingly, the $\mathcal{F}(\beta)$-Score is worked out as follows:

$$\mathcal{F}(\beta) - Score = \frac{(\beta^2 + 1) \times \frac{300}{617} \times \frac{1}{4}}{\beta^2 \times \frac{300}{617} + \frac{1}{4}}$$

For values of β equal to three and four, this formula returns 0.26 and 0.25, respectively. Both values are considerably lower than their counterparts in table 1.3.

Precision at k. A disadvantage of the $\mathcal{F}(\beta)$-Score is the fact that it does not assess the ranking order of the nuggets/sentences within the response. On the other hand, precision at k measures the ratio of sentences that are actual definitions between the first k positions of the ranking. This is a key issue whenever definition QA systems output sentences as it is also all-important to determine whether the highest positions of the ranking carry actual descriptive information. In the working example outlined in table 1.2, the precision values at the three different levels are: $k = 1 \Longrightarrow 1$; $k = 2 \Longrightarrow 0.5$; and $k = 3 \Longrightarrow 0.667$.

Mean Average Precision (MAP). The "uncomfortable" aspect of the previous metric is that a different value is computed for each level k. In order to deal with this, the MAP consolidates these k precision values by averaging them as follows [Manning and Schütze, 1999]:

$$MAP(Q) = \frac{1}{\mid Q \mid} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} Precision\_at\_k$$

Here, Q is a question set (e.g., TREC 2003), and m_j is the number of ranking sentences in the j-th output. Habitually, m_j is truncated to fixed values (e.g., one or five) that are of the interest of the evaluators. This metric is selected because of its ability to show how good the outcomes are in the first positions of the ranking. Simply put, for a given question set Q, MAP-1 shows the fraction of questions that ranked a valid definition on the top. In the working example in table 1.2, MAP-3 is equal to 0.7223.

Accuracy. This measure reflects another view of the response of definition QA systems. In a statement, it is the ratio of the amount of sentences labelled as definitions that are genuine definitions plus the number of sentences labelled as non-definition that are actual non-definitions to the total number of answer candidates. In the illustrative example in table 1.2, Accuracy is equal to 0.6667, assuming that the system picked all answer candidates.

R-precision. This indicates the average precision of the top R-ranked sentences with respect to a set Q of definition queries, where R varies for each question in Q in congruence with the amount of putative answers that are genuine definitions. In the example in table 1.2, R-precision is equal to 0.5, under the assumption that the system selected all genuine answers.

1.8 Conclusions

This chapter focuses its attention on the introductory aspects of definition question answering. These aspects are the common ground to understand the problems and their corresponding solutions presented in the posterior chapters.

For starters, this chapter goes over the trade-off between coverage and trustworthiness of distinct sources of potential answers. It then sketches the modules that are amply used by definition QA systems, and it subsequently familiarises the reader with the international evaluation conferences. Next, this chapter dissects some relevant facets of the query analysis stage. It also discusses some intrinsic characteristics of definitions, and it eventually elaborates on different evaluation metrics.

Chapter 2

Crawling the Web for Definitions

> "The Internet is the world's largest library. It's just that all the books are on the floor." (John Allen Paulos)
>
> "What, exactly, is the Internet? Basically it is a global network exchanging digitised data in such a way that any computer, anywhere, that is equipped with a device called a "modem" can make a noise like a duck choking on a kazoo." (Dave Barry)

2.1 Introduction

Without a shadow of doubt, one of most the fundamental and crucial components of a definition Question Answering (QA) system is the module that trawls the Internet for definitions. Upon its performance largely depends the success of the posterior phases of the answering process. In a nutshell, definition QA systems must account for an efficient and general strategy that let them fetch a wealth of descriptive information about *definienda* of various characteristics and targeted at wide-ranging topics. An effectual search approach, for instance, should efficiently cope with *definienda* such as abbreviations, books, events, locations, organisations, personal names, and sport teams, as well as, technical terms.

In practice, a major difficulty of designing a successful search technique lies in the precise nature of the input of the user. Frequently, the only significant piece of information that the user supplies is the *definiendum*. The user seldomly provides additional relevant input like the desired sense, extra contextual hints, several spellings or aliases, and/or different morphological forms. As a matter of fact, most of the time, the user has no idea about the *definiendum*, and for this reason, he or she is soliciting its definition. In essence, the user is normally unaware of the possibility of contexts different from the one he/she is thinking about, or of the potential contribution of various spellings/aliases. This unawareness prevents the QA system from making allowances for valuable knowledge that could eventually aid in sharply increasing the precision of the search, and therefore in outputting a more accurate response to the user. Anyhow, when this sort of fine detail is entered, it is not necessarily easy to identify its role in the query.

Another decisive factor involved in the design of search techniques is the response time. Definition QA systems have a limited time window to produce the correct answer for the user. This time constraint, in consequence, militates against an exhaustive search across the entire collection; in particular, when they tackling massive collections of documents (e.g., the Internet). In this situation, a full off-line processing is implausible, and under these

circumstances, systems have to perform a zooming approach that lead to the most promising set of texts spans, such as paragraphs or sentences. In truth, during this zooming process useful data can be lost.

This chapter deals at greater length with assorted strategies that assist (web) definition QA systems to find documents carrying descriptive information about the *definiendum*. The search methods discussed in this chapter account solely for the *definiendum* as the input of the user, which is the most recurrent case across query logs. In more details, this chapter fleshes out various strategies to discover descriptions across the Web. These methodologies differ in their complexity, coverage, and reliability as well. To begin with, the next chapter brings out the concept of definition relational databases. Section 2.3 subsequently gives information about the exploitation of Knowledge Bases (KB). Next, section 2.4 examines approaches predicated on task specific keywords. Section 2.5 touches on the utilisation of lexico-syntactic constructs, and section 2.6 then extends this technique by means of mining Wikipedia knowledge. Later, section 2.7 raises the subject of searching for descriptions in other languages, section 2.8 suggests future trends, and section 2.9 concludes this chapter.

2.2 Definition Relational Databases

There are three characteristics that make the Internet highly distinctive from other collections: (a) web users are constantly adding, updating and removing documents from their web-sites, making the Web a collection of documents that it is always changing, (b) it is composed of a tremendous number of documents, which some deem to be infinite due to dynamic pages, and (c) these documents encompass a broad range of formats including plain texts, videos, images, postscripts, Portable Document Formats (PDF), Microsoft Word documents, audio and HyperText Markup Language (HTML) pages. Under these conditions, commercial search engines, such as Google and MSN Search as well as Yahoo! Search, are compelled to allocate enormous computational resources to keep an updated index of this vast and diverse collection of documents. These vanguard Information Retrieval (IR) engines serve as an interface between users and the Internet.

On the contrary, other collections of documents, namely the ones used in Text REtrieval Conference (TREC) and Cross Language Evaluation Forum (CLEF), are considerably smaller in size, their content remains mainly static and consists chiefly of plain texts. These three characteristics make it easier to index and navigate them for descriptive information. In so doing, some definition QA systems, like [Kosseim et al., 2006, Qiu et al., 2007], usually take advantage of open-source search softwares, such as Lucene. This kind of tool allows the creation of a variety of views of the collection, and ergo they assist definition QA systems in zooming in to the most propitious portions of texts by means of manifold query words matching techniques. Of course, the demand for more specific indexing strategies and search functionalities as well as for reducing the retrieval time, has led definition QA systems, like [Fernandes, 2004, Hildebrandt et al., 2004, Katz et al., 2004], to the design of purpose-built techniques. Above all, this methodology boosted the recall of descriptive phrases by automatically constructing an immense relational database embodying nuggets distilled from every article in the corpus. These nuggets are about every entity within this corpus. Some illustrative entries in this type of repository are outlined below:

DEFINITION RELATIONAL REPOSITORY

Definiendum	Nuggets
Abby Cadabby	• a fairy-in-training who first appeared in the 37th season of the children's television show "Sesame Street".
C. S. Lewis	• Clive Staples "Jack" Lewis (29 November 1898 - 22 November 1963), commonly referred to as C. S. Lewis, was an Irish author and scholar. • known for his work on medieval literature, Christian apologetics, literary criticism, and fiction. • best known today for his series "The Chronicles of Narnia".
Sabena	• a former national airline of Belgium, which mainly operated from Brussels National Airport.

Table 2.1: Examples of entries in a definition relational database (nuggets surrounded with their contexts).

PROS AND CONS OF DEFINITION DATABASES

Thus trawling the target corpus for a definition consists in looking up for the right entry in this relational database. In other words, this methodology partly or fully answers definitions questions, before they are asked. This view works well with static collections, because it yields fast access to the definition relations, when entries are alphabetically sorted and the engine account for an efficient look-up algorithm (e.g., binary search). If this database is huge, sophisticated indexing approaches must be taken into account when designing. There are, however, some essential aspects that make this class of strategy less attractive:

- If the collection does not stay totally static, the relational repository must be updated every time a document is added, removed or modified.
- The creation of this relational database implies preprocessing all documents beforehand, despite the fact that most of them will contribute with entries that are very unlikely to be accessed later.
- These first two points underline the pertinence of the definition nuggets detector. If this demands large computational resources, the plausibility of this kind of technique resides heavily in the rate of updates and the size of the collection.
- Categorically, most entities have aliases or synonyms (e.g., "*Thomas Hanks*"/"*Thomas Jeffrey Hanks*" and "*George Walker Bush*"/"*George W. Bush*"). These aliases play a pivotal role in the performance of this strategy, because different entries in this repository can coincide with distinct aliases of the same entity (see table 2.2). Hence, descriptive information subsumed in entries that do not perfectly match the *definiendum* will be missed, bringing about a decline in recall. This problem becomes graver whenever the alias formulated by the user does not exist in the database, but alternative names do exist.
- As *definienda* normally have numerous senses, there is a need to split the entries in this database into their respective different senses. A concrete example is depicted in table 2.2. The entry "*Thomas Hanks*" should contain separate references to the actor and the seismologist. Definitively, this need for sense discrimination grows in accordance with the size of the collection, and/or whether or not the collection is open or domain specific.
- Lastly, the difficulty of extracting definitional relations is dependent on the variety of document formats and on the degree of structure of the articles within the collection.

Definiendum	Nuggets
Tom Hanks	• an Academy Award-winning actor. • an American seismologist.
Thomas Jeffrey Hanks	• an actor born in 1959 in California.

Table 2.2: Examples where problematic issues regarding definition relational databases can be seen.

AMBIGUITY IN DEFINITION REPOSITORY

Preferably, entries in this relational repository should be grouped in agreement with entities and senses instead of entities only as shown in table 2.3. This grouping approach consists of two tables. Since the *definiendum* prompted by the user can be ambiguous, the first table does the mapping to its possible senses within the collection, and the second brings nuggets together from the respective aliases. From one angle, it is the underspecified entry given by the user which causes the need for disambiguation. Presumably, due to his/her lack of knowledge about the content of the collection. While the user can explicitly enter "*Thomas Jeffrey Hanks*", it is more probable that one of its underspecified variations (e.g., "*Tom Hanks*") is given to the system. On the other hand, it is the structure of the collection which raises the ambiguity as it relies largely on the amount of different descriptions within its various contexts.

Definiendum	Sense #
Thomas Jeffrey Hanks	14534
Tom Hanks	14534, 56298

Sense #	Nuggets
14534	• an Academy Award-winning actor. • an actor born in 1959 in California.
56298	• an American seismologist.

Table 2.3: Preferable structure of a relational database.

At any rate, automatically achieving this lay-out is a very complex task. In the first place, it is necessary to identify when two distinct aliases refer to the same entity. Here, the technique for learning aliases proposed by [Figueroa, 2008a] might help (see section 2.6.1). This approach searches for sentences that match some lexico-syntactic patterns that often express aliases, and creates a repository of pairs accordingly. Another interesting method was introduced by [Wu et al., 2004], this profited from synsets in WordNet to find the aliases. In the second place, once the descriptive information corresponding to all aliases of a particular *definiendum* is gathered into one single group, it must be split into groups in concert with the respective senses. This task of finding classes of similar contexts such that each class represents a single word/entity sense is known as *word sense discrimination* [Purandare and Pedersen, 2004, Lupsa and Tatar, 2005]. A quality of natural language texts that might cooperate on tackling this problem for definition questions is the One Sense Per Discourse principle [Gale et al., 1992]. This asserts that if a polysemous word appears two times in a well-written discourse, it is extremely likely that they share the same sense. In the case of definition QA systems, this claim is not straightforwardly applicable, because dictionary or disambiguation pages do not observe this principle, whereas blogs and forums are more probable to meet this criterion. Encyclopedias such as Wikipedia can, nevertheless, offer fruitful information including translations that can greatly support in discriminating senses (cf. [Brown et al., 1991, Dagan et al., 1991]). One last remark on the relational definition database is that each nugget should have a pointer to the source document, however, this pointing strategy is largely reliant on the update approach used by the repository.

GOOGLE DEFINE FEATURE

The works of [Hildebrandt et al., 2004] and [Katz et al., 2004] are aimed at discovering answers in the AQUAINT corpus, while Google offers a feature for crawling the Web for def-

Figure 2.1: Example of Google's define feature (*definiendum*:"Tom Hanks").

initions. Every time a user enters "define:*definiendum*", the search engine returns an array of glossaries that embrace definitions of the *definiendum*. To the best of our knowledge, it is hitherto unknown how Google gathers these glossaries: Which strategies are involved? What is manual or automatic? This uncertainty makes this strategy difficult to analyse. Nonetheless, [Xu et al., 2005] observed that these glossaries seem to have some common properties: the pages are titled with task specific clues including "*glossary*" and "*dictionary*", the terms in the page are alphabetically sorted and presented with the same style, for instance italics and bold print. Under this observation, this method yields wider coverage, but succinct definitions taken from different glossaries are very likely to convey redundant information, while at the same time, new concepts are rarely found in glossaries, but rather in web-sites such as blogs or forums. Along with glossies, Google collects definitions snippets from KBs, such as WordNet and Wikipedia (see figure 2.1). Google accounts for the first definition lines in these resources, and outputs them for the user. The underlying goal is to offer users a set of concise descriptions of the *definiendum*. In short, this feature aided with descriptions to [Cui et al., 2004b] for 25 out of the 64 TREC 2004 questions.

GOOGLE TIMELINE FEATURE

In addition, Google made available another experimental service (Google Timeline) that allows users to discover pertinent events associated with the *definiendum* (see figure 2.2). This class of resource is extremely beneficial as manifold nuggets are temporally anchored. However, this kind of tool is still in its first steps, and therefore they need many improvements to be considered as authoritative as KBs. Specifically, [Katz et al., 2007] realised that Google Timeline mixes references to assorted items bearing the same name. In this trend, web-sites can also be found that commemorate the most important historic events of each day. In these archives, one can find births and deaths of celebrities, independence days, dates related to world records, important achievements, etc. Some prominent archives include: www.thisdaythatyear.com, www.worldofquotes.com, and www.historyorb.com, as well as www.theday2day.com.

HISTORIC EVENTS

DEFINITION WEB QA

When a QA system is geared towards the Web, some of the subsequent aspects must be taken into consideration. Firstly, web documents are added, updated and removed constantly, and hence the index must be updated regularly. Secondly, the size of the Internet

Figure 2.2: Example of Google's timeline feature (*definiendum*:"Alan Moore").

is commonly much larger than any domain-specific collection, and thus computing a look-up table is computationally demanding. It also entails a crawling methodology, scheduling, the retrieval of documents, etc. Fortunately, this task is efficiently performed by commercial search engines such as Altavista, Excite, Google and MSN Search as well as Yahoo! Search. These state-of-the-art commercial engines make a search interface to the Web available. At any rate, commercial search engines are tuned to perform IR tasks, or to bring in substantial revenue from advertisements, ergo not to perform specific QA needs. Therefore, adapting this technology to serve QA purposes is attractive.

2.3 Using Specific Resources

The overwhelming majority of TREC definition QA systems discover descriptive text fragments on the Web by profiting from online KBs. This prominent strategy commonly involves downloading the full-document and the design of a specialised wrapper for each KB [Hildebrandt et al., 2004, Sun et al., 2005].

Table 2.4 lists the most widely used KBs and their corresponding systems. A breakdown of these techniques is as follows:

- [Jijkoun et al., 2003] capitalised on www.biography.com and WordNet. Whenever nothing was found in these KBs, this method utilised Google, with queries formed by putting together the name of the person in question with varying subsets of a predefined array of hand-crafted features, including "*born*", "*graduated*", and "*suffered*". For questions directed at organisations, an identical approach was used, but with a set of

Knowledge Base	System
WordNet glossaries	Jijkoun et al. [2003], Echihabi et al. [2003], Gaizauskas et al. [2003] Xu et al. [2003, 2004], Cui et al. [2004c], Wu et al. [2004, 2005a] Zhang et al. [2005], Zhou et al. [2006], Chali and Joty [2007]
Merriam-Webster dictionary	Xu et al. [2003, 2004], Hildebrandt et al. [2004] Katz et al. [2004], Zhang et al. [2005]
Wikipedia	Xu et al. [2003, 2004], Gaizauskas et al. [2004], Cui et al. [2004c] Ahn et al. [2005], Zhang et al. [2005], Kosseim et al. [2006] Hickl et al. [2006], Shen et al. [2006], Qiu et al. [2007] Schlaefer et al. [2007], Razmara et al. [2007], Katz et al. [2007] Shen et al. [2007]
Columbia Encyclopedia	Xu et al. [2003, 2004]
www.s9.com	Xu et al. [2003, 2004], Zhang et al. [2005], Hickl et al. [2006]
www.encyclopedia.com	Wu et al. [2004, 2005a], Zhang et al. [2005], Zhou et al. [2006]
Britannica Encyclopedia	Gaizauskas et al. [2003, 2004]
answers.com	Sun et al. [2005]
www.biography.com	Jijkoun et al. [2003], Echihabi et al. [2003] Cui et al. [2004c], Hickl et al. [2006]
Who2	Prager et al. [2003]
Google Timeline	Katz et al. [2007]
web snippets	Jijkoun et al. [2003], Xu et al. [2003, 2004] Cui et al. [2004c], Chen et al. [2006], Qiu et al. [2007]
full web pages	Gaizauskas et al. [2004], Hickl et al. [2007] Schlaefer et al. [2007]
unspecified online resources	Han et al. [2004], Wu et al. [2005a], Zhou et al. [2006]

Table 2.4: KBs utilised by TREC-oriented systems.

properties concerning organisations. As a final fallback option, they simply submitted the "*definiendum*" to Google and mined the returned surrogates afterwards.

- [Echihabi et al., 2003] sifted 14,414 biographical entries from www.biography.com. They used these entries to gather core biographical knowledge for specific people as well as identify words that are indicative of biographical information.

- [Xu et al., 2003, 2004] extracted existing definitions of the *definiendum* from: WordNet glossaries, Merriam-Webster dictionary, the Columbia Encyclopedia, Wikipedia, the biography dictionary at www.s9.com and Google snippets.

- [Han et al., 2004] took advantage of biographical resources for tackling *definienda* directed at persons. These resources contributed to their system with pieces of information about their personal identities and related events.

- [Sun et al., 2005] preferred wrappers for specific websites to general search engines, this way they obtained more precise results. Their system accumulated existing definitions from answers.com.

- [Katz et al., 2004] exploited the Merriam-Webster online dictionary for acquiring definitions. Keywords from these definitions were used in a Lucene query to download documents from the AQUAINT corpus afterwards.

- [Cui et al., 2004c] benefited from KBs (i.e., WordNet, and www.biography.com as well as Wikipedia) and 200 web snippets.

- [Zhang et al., 2005] studied the influence of several resources exerted to the answering process: www.s9.com, www.encyclopedia.com, Wikipedia, and Merriam-Webster dictionary as well as WordNet glossaries.
- [Ahn et al., 2005] profited from the online encyclopedia, and queried an integrated IR engine built on top of Wikipedia for the *definiendum*.
- [Wu et al., 2004, 2005a, Zhou et al., 2006] mined definitions of the *definiendum* from a number of online KBs: WordNet glosses and other online dictionaries such as the biography dictionary at www.encyclopedia.com.
- [Hickl et al., 2006] collected descriptive phrases from Wikipedia, www.s9.com, and www.biography.com.
- [Kosseim et al., 2006] distinguished marking terms related to the *definiendum* across the Wikipedia online dictionary. They found the proper Wikipedia article by crawling the domain using the Google Application Programming Interface (API) and the *definiendum* as query. In this method, the first Wikipedia article that satisfies the query is taken. Whenever no Wikipage satisfies the query, the query is loosened. Eventually, if still no Wikipage is discovered, the top N AQUAINT documents are then utilised for discovering the marking terms.
- [Schlaefer et al., 2007] took advantage of Wikipedia articles. For targets not found in Wikipedia, they made use of Google as a fallback solution by fetching the first 100 hits.
- [Qiu et al., 2007] acquired a corpus related to the *definiendum* from Wikipedia, and utilised the first 100 Google hits returned by submitting the *definiendum*. Analogously to this strategy, [Hickl et al., 2007] took into account the first 100 pages retrieved from Google bearing the *definiendum*.
- [Chali and Joty, 2007] utilised WordNet glossary entries, while [Razmara et al., 2007] and [Gaizauskas et al., 2004] exploited Wikipedia and the Britannica Encyclopedia, respectively, as a source of descriptive phrases.
- [Katz et al., 2007] harvested descriptive information from two distinct KBs: Wikipedia and Google Timeline.
- [Shen et al., 2006, 2007] did not consider special facilities for responding to definition queries, apart from the implementation of a basic Wikipedia-based snippet retrieval component.

KBS COVERAGE

As this breakdown reveals, most systems are inclined to capitalise on KBs instead of web snippets or full web documents. In a special manner, the two more prominent and prevalent mines of descriptive information (KBs) are: Wikipedia and WordNet. The reason for this is that the experiments have shown great improvements caused by KBs [Cui et al., 2004c]. Chiefly, they cooperate on getting more precise results [Sun et al., 2005, Zhang et al., 2005] at the expense of a detriment to coverage. Rigorously speaking, [Cui et al., 2004c] found out that Wikipedia covered 34 out of the 50 TREC–2003 definition queries, whereas 23 out of 30 questions with respect to people were covered by www.biography.com, all together providing answers to 42 queries. Further, [Hildebrandt et al., 2004] was assisted by a wrapper for the online Merriam-Webster dictionary, which retrieved about 1.5 nuggets per question.

KBS IMPACT

In their work, [Zhang et al., 2005] examined the relationships and the differences between definitions from several KBs. They analysed the coverage supplied by five distinct resources

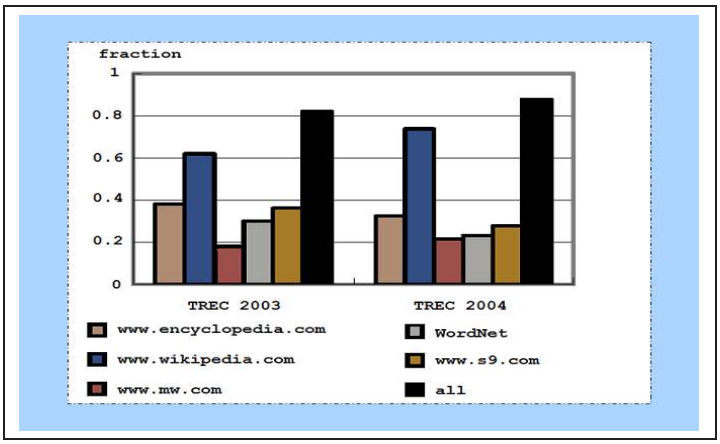

Figure 2.3: Contribution of different KBs (adapted from [Zhang et al., 2005]).

to the 50 and 64 definition questions distilled from the TREC 2003 and 2004 tracks, respectively. Their findings are shown in graph form in figure 2.3. This graph unveils that these five resources combined covered between 80% to 90% of the queries. Predominantly, Wikipedia is the largest contributor, followed by www.encyclopedia.com. However, in this kind of comparison, slight differences are not necessarily meaningful, because the distribution of types of *definienda* in the question set plays an essential role. For instance, the contribution of www.s9.com will be constrained by the number of *definienda* in the question set that are aimed at biographies. Another aspect of consideration when comparing KBs is their updating rate, that is how many articles are added in a given period of time. Ostensibly, KBs with a higher rate are naturally preferable. The amount of queries covered by the KBs is just one aspect. Another determining factor is the extent to which these resources answer a definition question. The observations of [Zhang et al., 2005] reveal that their definitions can be short or long, concise or detailed, and accordingly, yielding few or many descriptive nuggets. By and large, [Zhang et al., 2005] contrasted a TREC-oriented system that makes allowances for these five resources with another which does not. Both systems operated on the TREC 2004 data set. They showed that these five KBs substantially bettered the $\mathcal{F}(3)$-Score from 0.231 to 0.404. The reader is also encouraged to look at sections 4.8 and 4.9 on pages 91 and 93, respectively, for complementary information on this subject.

KBS VERSUS WEB-SNIPPETS AND FULL-DOCUMENTS

The reason that prevents definition QA systems from making use of web-snippets and full-documents is five-fold: (a) they are noisy, meaning they can express not only closely related descriptive sentences, but also large amounts of spurious and unrelated -or loosely connected- information about the *definiendum*, creating the need for a technique that weeds out false hits [Xu et al., 2004], (b) finding web-pages that are more likely to provide definitions inherently involves the design of an ad-hoc search procedure that sharply ameliorates the recall of promising documents, (c) the *definiendum* entered by the user can be ambiguous (e.g., "*Tom Hanks*" and "*Jim Clark*"), and it is thus difficult to foresee whether or not the desired sense will match one of its predominant senses on the Internet, or by default when this desired sense is not explicitly stipulated; it is hard to satisfactorily discriminate the senses existing across the fetched documents, while systems can retrieve information from Wikipedia unambiguously by seeking for articles on the *definiendum* [Schlaefer et al., 2007], and (d) in online encyclopedias, users make sure that the most relevant information is put into words in a concise and concentrated fashion with little noise [Kosseim et al., 2006, Schlaefer et al., 2007], whereas finding the same amount of data across non-definition full web pages would inevitably imply processing a significant amount of documents, and (e) in the case of full-documents, there is a download time involved that in some cases can be

extremely long, whereas definition QA systems can incorporate a built-in IR engine on top of Wikipedia [Ahn et al., 2005]. Another factor that makes working with full-documents less attractive is that they can be found in manifold formats, including PDFs and plain text, hence specialised text extraction techniques must be taken into consideration.

DOMAIN SPECIFIC KBS

Although some KBs supply articles on wide-ranging topics (e.g., Wikipedia), most of them are targeted at one particular type of *definiendum*, like persons and locations, or at one domain including movies, sports, and books. The former type can be conceived as *open-domain*, while the latter as *close-domain* or *domain-specific*. The reason why *close-domain* KBs are seldom utilised by definition QA systems (see table 2.4) is that they need to implement a wrapper around each of them, and whenever any of these resources changes its lay-out, the system must revise the implementation of the respective wrapper.

One potential advantage of *close-domain* over *open-domain* KBs is that they can offer a way of detecting ambiguity and discriminating some senses of the *definiendum*. That is to say, different entries in distinct *domain-specific* KBs can signal different senses. Some *definienda*, however, can still transpire several domains. For example,

Maria Gioacchina Stajano Starace (1932 -) writer, actor, journalist, painter.

WEB-SNIPPETS

With regard to web-snippets, as explained in table 2.4, a comparatively fewer number of systems take advantage of them as a source of descriptive information. One of the main disadvantages is intentional breaks inserted by search engines. These breaks truncate sentences, materialising not only the loss of critical data, but also the incorporation of misleading knowledge. To reinforce this, consider the next web snippet in relation to "*Jar Jar Binks*" returned by Yahoo! Search:

Jar - TvWiki, the free encyclopedia
A Mason **jar** is a glass jar sporting a screw-on cap or wire-sprung lid. ... **Jar Jar Binks** is a fictional character from the Star Wars universe.
www.tvwiki.tv/wiki/Jar

The source text of this illustrative surrogate is provided in order to adequately understand this problem:

The word **jar** has several meanings:
...
• A Mason **jar** is a glass jar sporting a screw-on cap or wire-sprung lid.
...
• **Jar Jar Binks** is a fictional character from the Star Wars universe.
...

NOISE IN WEB-SNIPPETS

Here, the first piece of text within the retrieved snippet renders a description of a Mason jar, while the second verbalises a definition of the *definiendum* of interest. This sort of keyword matching oriented summarisation creates, from the QA viewpoint, the retrieval of noisy information. Furthermore, web snippets can be unrelated or loosely related to the *definiendum*:

#6 Tom Hanks - Forbes.com10 Feb 2009 ...
10 Feb 2009 ... Forbes.com: #6 Tom Hanks - A look into how this famous actor is ranked next to their peers by the entertainment industry.
www.forbes.com/business/lists/2009/58/star...

Even though this surrogate outlines "*Tom Hanks*" as an actor, it primarily conveys noisy information. This class of web snippets is not considered preferable when dealing with prominent senses of the *definiendum*. What is more; attributes (e.g., the title) that are a good indicator in identifying related KBs, can turn out to be misleading when recognising web-snippets embracing descriptive information. Despite the fact that they carry unrelated content, some descriptive knowledge can still be found in the full document. An excellent example is gossip blogs:

IRRELEVANT WEB-SNIPPETS

Tom Hanks | PopSugar - Celebrity Gossip & News12 Jan 2009 ...
Oct 28, 2008 - After posing for photos on the red carpet, Julia Roberts, Casey Affleck, Bruce Willis, and Tom Hanks were just a handful of ...
www.popsugar.com/tag/Tom+Hanks

Nonetheless, web-snippets are an effective way to avoid dealing with the various document formats existing on the Web, and a convenient way to prevent diverting time away from downloading and processing full documents.

WEB-SNIPPETS ADVANTAGES

2.4 Finding Additional Resources through Specific Task Cues

For the purpose of surmounting the difficulties exhibited when dealing with web-snippets and with the narrow coverage provided by KBs, definition QA systems have attempted several strategies to ameliorate the recall of descriptive knowledge across web-snippets. An increment of descriptive sentences within web-snippets brings about a growth in the quantity of full-documents that carry descriptive phrases about the *definiendum*. In light of the results obtained by [Zhang et al., 2005], an increase in descriptive sentences would lead to an enhancement in the performance of definition QA systems. The underlying idea behind these strategies is rewriting the query, or in this case the *definiendum*, in such a way that the new query biases the search engine in favour of web-snippets that are very likely to express definitions of the *definiendum*.

KEYWORD: *biography*

The studies of [Xu et al., 2003, 2004] took advantage of Google for trawling the Web for biographies. They extended the *definiendum* with the word "*biography*". Take, for example, the search string "*George Bush, biography*". In this methodology, this task specific cue tries to boost the retrieval of biography pages. This keyword is, of course, more helpful when aiming at definitions of persons than of diseases, for instance. A simple rule-based classifier was utilised later to filter out false hits. These rules included how many times a third person pronoun (i.e., he, him, she and her) are used, whether the document contains a birth date and so forth.

LEARNING SEARCH CLUES

In their work, [Chen et al., 2006] reformulated the query by simply adding task specific keywords to the questions. More exactly, queries like "*Who is ...?*" were extended with the word "*biography*". In the event of "*What is ...?*" questions, the cues "*refers to*" and "*is usually*" were added. These clue words were learnt using a method akin to [Ravichandran and Hovy, 2002]. The new queries are sent to Google and gather the five highest terms co-occurring with the target. These five words are perceived as query expansion terms, and are later used to download the first 500 snippets returned by Google. However, what makes this technique less attractive is the fact that it is unclear how to automatically assign the right expansion keyword(s) in the first stage. Firstly, not all definition questions across search engine logs are posed in concert with the "*What/Who is ...?*" template (see sample queries in section 1.5 on page 6). Secondly, this assumes that the user knows if he is defining a person or other thing. Thirdly, it takes for granted that a *definiendum*, targeted at a person, cannot also aim at another class. In reality, this is not necessarily true, *definienda* can have numerous potential

senses, imagine a user looking for "*Calvin Klein*". This sort of query expansion is too typical to the TREC challenge, and in the same line, follows the hand-crafted rules proposed by [Jijkoun et al., 2003].

OPTIMAL NUMBER OF WEB-SNIPPETS

Another final aspect to compare is the different number of surrogates downloaded by the distinct approaches. It is unclear what the optimal amount of snippets to use actually is. On the one hand, [Cui et al., 2004c] utilised 200 Google snippets, while [Chen et al., 2006] and other systems much more. This is an open research question, and its answer seems to depend largely on the size of collection of KBs that definition QA systems benefit from.

2.5 Using Lexico-Syntactic Constructs

Most of IR engines are devised to search for crucial documents [Salton and McGill, 1983], they are thus not suitable for aiding in looking for definitions [Xu et al., 2005]. Contrary to previous approaches and apparently in congruence with the idea of [Gaizauskas et al., 2003], [Figueroa and Neumann, 2007, Figueroa et al., 2009] collected descriptive phrases from the Internet that are very likely to carry a definition by sequentially submitting ten purpose-built queries. These queries are aimed specifically at biasing the search engine in favour of web snippets that are very likely to match some lexico-syntactic constructs that often render definitions, and as a natural consequence, directed essentially at increasing the recognition of descriptive phrases within the web-snippets. In this method, the first submission conforms to the initial query:

q_1:"<definiendum>"

COPULAR SEARCH QUERIES

The remaining queries are focused on various lexico-syntactic patterns that are often used for conveying open-domain definitions. Since copular constructions are very likely to contribute a lot of descriptive knowledge, this procedure generates the following three queries by synthesising these copular constructs:

q_2:"<definiendum> is a" ∨ "<definiendum> was a" ∨ "<definiendum> were a" ∨ "<definiendum> are a"
q_3:"<definiendum> is an" ∨ "<definiendum> was an" ∨ "<definiendum> were an" ∨ "<definiendum> are an"
q_4:"<definiendum> is the" ∨ "<definiendum> was the" ∨ "<definiendum> were the" ∨ "<definiendum> are the"

In these three search queries, constructs are equally spread in consonance with their different tenses and number, that is each query has two clues: one directed at the present and another at the past tense. In addition, each query has a clue directed at both numbers: plural and singular. The tacit assumption here is that each query will be able to find descriptive information, independently as to whether or not the number of the *definiendum* is singular or plural, reducing the possibility of making fruitless submissions, and hence alleviating a possible detriment to recall. Tenses, for the same reason, were equally distributed among queries, in order to target at *definienda* in the present and the past. However, it is worth duly noting here that the relation between the location of the *definiendum* in the timeline and the tense of which its respective descriptive information is expressed is decidedly weak. This weakness is due to the fact that descriptions are occasionally not bound to the location of their context in time (i.e., past, present or future). Furthermore, the date and the style of writing of the document become very influential in this aspect. Clear cases of this are shown in the following phrases:

Aaron Copland is the youngest of five children born to Harris and Sarah Copland, ...
Aaron Copland is the most honoured of American composers.
Aaron Copland (1900-1990) was an American composer who wrote modern music.

Under these observations, [Figueroa and Neumann, 2007, Figueroa et al., 2009] distributed these clues equally. The next two submissions are as follows:

q_5:"<definiendum> has been a" ∨ "<definiendum> has been an" ∨ "<definiendum> has been the" ∨ "<definiendum> have been a" ∨ "<definiendum> have been an" ∨ "<definiendum> have been the"
q_6:"<definiendum>, a" ∨ "<definiendum>, an" ∨ "<definiendum>, the" ∨ "<definiendum>, or"

SYNONYMS, ALIASES AND APPOSITIVES

Since the search construct "<definiendum>, or" has a low occurrence across documents on the Internet [H. Joho and M. Sanderson, 2001] and often conveys a synonym (e. g., *"myopia or nearsightedness"*), these two kinds of patterns are merged into one query. Alternative names of people, organisations or abbreviations are seldomly expressed in this way, but they are likely to match the other clauses within q_6. As a result, combining both patterns allows this method to reduce the amount of web searches. As [Chen et al., 2006] also stressed, it is always crucial to seek a balance between download time and recall. The next three submissions are in congruence with:

q_7=("<definiendum>" ∨ "<definiendum> also" ∨ "<definiendum> is" ∨ "<definiendum> are") ∧ (called ∨ nicknamed ∨ "known as")
q_8="<definiendum> became" ∨ "<definiendum> become" ∨ "<definiendum> becomes"
q_9="<definiendum> which" ∨ "<definiendum> that" ∨ "<definiendum> who"

WAS-BORN QUERY

Eventually, the tenth query tries to retrieve snippets that match "<definiendum> was born" and "(<definiendum>)". Homologously to q_6, both patterns are fused into one query on the grounds that the former deals with *definienda* concerning persons and the latter focuses essentially on acronyms [H. Joho and M. Sanderson, 2000]. Hence, this technique avoids an unproductive retrieval without lessening the number of retrieved descriptive sentences:

q_{10}="<definiendum> was born" ∨ "(<definiendum>)"

BIRTHDATE AND SENSE DISCRIMINATION

Here, it is worth remarking that the pattern "(<definiendum>)" would work depending on the interface offered by the search engine. An appealing aspect of this last lexico-syntactic regularity is that it can produce valuable data about the different senses:

Jim Clark was born in 1944, in Texas, USA.
Jim Clark was born in Kilmany, (in the county of Fife), to a Scottish farming family.
Jim Clark was born in Byrdstown, Tennessee.

Glaring inconstancies or discrepancies are good indicators of ambiguity (i.e., *born in Texas* or *Kilmany*). In the opposite way, minor discrepancies between dates and places might trigger that there is uncertainty about the factual accounts. In juxtaposition, consider the following examples with respect to *"Alexander Hamilton"*:

Alexander Hamilton was born in Charlestown, Nevis, in the West Indies on January 11, 1757 (or 1755)*, to James Hamilton, a Scottish merchant of St...
Alexander Hamilton was born on the island of Nevis in the British West Indies on January 11, 1757.
Alexander Hamilton was born as a British subject on the island of Nevis in the West Indies on the 11th of January 1755.

Corpus		Baseline	LS-Search
	Total Number of Questions	Answered Questions	Answered Questions
TREC 2001	133	81	**133**
TREC 2003	50	38	**50**
CLEF 2004	86	67	78
CLEF 2005	185	160	173
CLEF 2006	152	102	136

Table 2.5: Performance of lexico-Syntactic search (source [Figueroa and Neumann, 2007]).

However, to fully take advantage of this pattern for discriminating senses, it would automatically imply the design of a strategy capable of homologating place names and standardising dates as well as dealing with underspecifications in terms of dates and places.

In their study, [Figueroa and Neumann, 2007] compared the coverage provided to a definition QA system by this strategy (LS-SEARCH) with a BASELINE (see table 2.5). In their experimental settings, LS-SEARCH fetched a maximum of thirty web snippets per query, altogether supplying a maximum of 300 surrogates. Accordingly, [Figueroa and Neumann, 2007] implemented a BASELINE that also downloaded 300 hundred snippets by submitting "<definiendum>" to the Internet, resembling the approach by [Cui et al., 2004c, Qiu et al., 2007]. The main differences are that LS-SEARCH and BASELINE acquired 300 surrogates and utilised MSN Search as an interface to/with the Web, while [Cui et al., 2004c] gathered 200 and [Qiu et al., 2007] 100 web snippets, and both fetched the web snippets by means of Google.

COVERAGE OF LEXICO-SYNTACTIC CLUES

In table 2.5, the number of answered questions is the amount of responses that embraced at least one correct nugget. These "Answered Questions" values were collected by manually verifying the output. Here, CLEF data-sets consider all English translations from all languages. To put it more exactly, LS-SEARCH cooperated on discovering nuggets for all questions in (2), in contrast to [Cui et al., 2004c], who was aided in finding nuggets for solely 42 questions by using 200 web snippets along with KBs. Overall, LS-SEARCH covered at least 94% of the questions, whereas BASELINE EN-I covered at least 74%. This outcome achieved by this query rewriting puts forward the idea of using web snippets as a productive source of descriptive phrases. Precisely, assisting definition QA system in biasing search engines in favour of surrogates from any type of KB. These snippets also offer the advantage that localised pieces of texts that match well-known lexico-syntactic constructs that are likely to put into words definitions. Other constructs , including "*stands for*" or "*was grounded*", can still be utilised for increasing the recall of definitions.

ANSWER LENGTH

Another beneficial aspect of this rewriting approach is that it fetches longer sentences. Explicitly, the length of the retrieved sentences was 125.70 ± 44.21 and 109.74 ± 42.15 with and without white spaces, respectively. By sending the *definiendum*, the achieved lengths are 118.168 ± 50.20 and 97.81 ± 41.80 with and without white spaces, respectively. A side effect is that these longer sentences mitigate the impact of truncations on web-snippets.

Lastly, it is worth underlining that [Gaizauskas et al., 2003] capitalised on fifty patterns to locate unique relevant documents on the Internet. Unfortunately, they did not stipulate which clues and how they are utilised, and the benefit in coverage they bring forth. It is also worth stressing that they made use of this procedure for finding descriptive sentences within the most propitious documents

2.5.1 The Grammatical Number of the *Definiendum* Guessed Beforehand?

The prior technique suffers from the following drawback: Principally, [Figueroa, 2008c] detected that the static nature of this query rewriting results in a drop in recall. They observed, more precisely, that clauses such as "*Allen Iverson were a*" and "*Allen Iverson are a*" bring about misleading sentences, when the grammatical number of the *definiendum* is singular. In like manner, this phenomenon also emerges when definition QA systems send to the search engine constructs including "*Caribbean islands is a*" and "*Caribbean islands is the*" and they are dealing with *definienda* plural in grammatical number. As examples, consider the next surrogates retrieved by Yahoo! Search:

USATODAY.com - Breaking down the categories: Best of the best
Fan view: Cheers for visiting **Allen Iverson were a** slap in the face to the Clippers. ... The team ran out of $5 programs before tip-off, a slap in the face to ...
usatoday.com/sports/basketball/nba/2005-04-13-arenas-breakdown_x.htm

N.B.A. ROUNDUP; Arenas Leads The Wizards To Victory - New York Times
... 10. Carmelo Anthony and **Allen Iverson were a** combined 1 for 10 in the third, when Denver ... Iverson, who had missed 9 of the past 10 games with an ...
query.nytimes.com/gst/fullpage.html?.../index.html

Map Caribbean Islands is a Free online Map for most all the Caribbean ...
Map of **Caribbean Islands is a** Free online map of the Caribbean Islands including detailed road map and useful city travel information maps.
map-caribbean-islands.com

These three illustrative web snippets signal the chief obstacle here. A disagreement between the *definiendum* and the grammatical number of the matching lexico-syntactic clue can indicate that the subject of the sentence was shifted to another topic. As a logical consequence, the main focus of attention might be unrelated or loosely related to the *definiendum*, causing the sentences to convey indirect or non-definitional knowledge about it. In the previous three cases, phrases serving as focus shifters are:

TOPIC SHIFT

10. Carmelo Anthony and <definiendum> were a
Fan view: Cheers for visiting <definiendum> were a
Map of <definiendum> is a

This linguistic phenomena is specially relevant to definition QA systems that are crawling the Web for descriptive information, because most commercial search engines, at the time of writing, do not supply a way to incorporate this type of restrictions into the search query. For instance, filtering out some words placed between the beginning of a sentence and the *definiendum*, or allowing a maximum amount of words between the beginning of the sentence and the *definiendum*. It is, nevertheless, still unclear how this sort of attribute could be profitably exploited to boost the recall of promising descriptive sentences. Further, the application of these kinds of features depends heavily on the amount of knowledge existing on the Web about the *definiendum*, because if there is not a lot of knowledge, then snippets as the following becomes very important to recognise:

Caribbean
Here's a link to some of out Most Popular Posts ... Jamaica, the third largest of the **Caribbean islands, is a** beautiful, wild, and diverse place.
kathika.com/category/destinations/caribbean

COLLECTIVE NOUNS

In this web snippet, the reader gets informed about the fact that the "*Caribbean islands*" comprise at least three islands and that one of them is named Jamaica. Simply put, the importance of detecting this piece of information lies in the amount of pieces of text found about the "*Caribbean islands*", because a larger number increases the probability of finding the same knowledge paraphrased in a way that it is easier to distinguish. But things are not black or white, collective nouns or instances of collective nouns that refer to groups of people, such as sport teams, lie in the middle ground between both groups. The next two snippets verbalising definitions regarding "*Manchester United*" exemplify this:

Manchester United fan site
Manchester United is the reigning champion of the Premier League ...
In May 2008, the total value of Manchester United is about 897 million pounds.
manchester-united.gemzies.com

Manchester United F.C. - Wikipedia, the free encyclopedia
For other uses, see MUFC (disambiguation). "Manchester United" redirects here.
... **Manchester United are the** reigning English, European, and World Champions having ...
en.wikipedia.org/wiki/Manchester_United_F.C.

This "number mismatch" is actually a natural and logical property of the human language, and it is a subtle shift in the thought of the underlying words.

GRAMMATICAL NUMBER DEDUCTION

Incidentally, an array of unpromising lexico-syntactic patterns can be set in the same query and hence, bring forth an unproductive retrieval, diminishing the number of descriptive utterances. Nonetheless, these clauses observe a local lexico-syntactic dependency with the *definiendum*. Specifically, they are unlikely to embody accompanying words in between them. This is an important fact because off-line n-gram counts supplied by Google can be used to transform this static query construction into a more dynamic one. In the working example regarding "*Allen Iverson*", an excerpt of Google 4-grams counts is as follows:

Search Clauses concerning "*Allen Iverson*"	Frequency
Allen Iverson **is a**	209
Allen Iverson **is an**	68
Allen Iverson **is the**	425
Allen Iverson **was a**	57
Allen Iverson **was the**	101

Table 2.6: Examples of search clauses regarding "*Allen Iverson*".

The first beneficial aspect of Google n-grams is that, in some cases, the grammatical number can be inferred. In particular, in the case of "*Allen Iverson*", singular lexico-syntactic clues are most propitious. However, it is not always possible to draw a clear distinction. A delineative example is "*fractals*":

fractals are a 176 (e.g., "Fractals are a powerful tool...")
fractals are an 86 (e.g., "Fractals are an exquisite...")
fractals are the 215 (e.g., "Fractals are the place...")
fractals is a 124 (e.g., "Fractals is a new branch of...")
fractals is the 148 (e.g., "Fractals is an innovative...")

A dynamic strategy was then designed which selects a grammatical number whenever more than three keywords conforming to one grammatical number exist, and zero to the

another. The second favourable aspect is that the frequencies give hints about the hierarchy within the lexico-syntactic patterns. This dynamic method takes advantage of this hierarchy for configuring the ten search strings. First, the queries q_7 and q_{10} are intermixed into one $q_7^{'}$. This search query is composed of the following clues:

"<definiendum> also called ", "<definiendum> also nicknamed", "<definiendum> also known",
"<definiendum> is called", "<definiendum> stands for", "<definiendum> are called",
"<definiendum> is known",
"<definiendum> are nicknamed", "<definiendum> are known", "<definiendum> was born",
"<definiendum> was founded", "<definiendum> was founded", "<definiendum> is nicknamed"

Accordingly, $q_7^{'}$ consists merely of the clauses that can be found in Google n-grams. If any construct cannot be found, $q_7^{'}$ is set to $\emptyset$. In every case, $q_{10}^{'}$ remains as $\emptyset$. It is worth pointing out that, the term *"stands for"* replaces the parentheses in q_{10}. Second, $q_5^{'} = q_5$, $q_6^{'} = q_6$ and $q_8^{'} = q_8$ as well as $q_9^{'} = q_9$. Additionally, the $q_1^{'}$ is set to $\emptyset$. Third, the clauses included in the queries q_2 and q_3, as well as q_4, are dynamically sorted across the available queries, as highlighted in table 2.7.

$q_7^{'} = \emptyset$	$q_7^{'} \neq \emptyset$
$q_1^{'}$:"<definiendum> I_1" $q_2^{'}$:"<definiendum> I_2"	$q_1^{'}$:"<definiendum> I_1" $q_2^{'}$:"<definiendum> I_2"
$q_3^{'}$:"<definiendum> I_3" $q_4^{'}$:"<definiendum> I_4"	$q_3^{'}$:"<definiendum> I_3" $q_4^{'}$:"<definiendum> I_4"
$q_5^{'}$:"<definiendum> I_5" $q_7^{'}$:"<definiendum> I_6"	$q_5^{'}$:"<definiendum> I_5" $\vee$ "<definiendum> I_6"

Table 2.7: Dynamic queries (grammatical number known).

Where I_1 and I_6 coincide with the highest and lowest frequent lexico-syntactic regularities in accordance with Google frequency counts, respectively. In the event that the grammatical number cannot be distinguished, the queries are as follows:

$q_1^{'}$:"<definiendum> is a" $\vee$ "<definiendum> were an" $\vee$ "<definiendum> was the"
$q_2^{'}$:"<definiendum> was a" $\vee$ "<definiendum> are an"
$q_3^{'}$:"<definiendum> are a" $\vee$ "<definiendum> was an" $\vee$ "<definiendum> were the"
$q_4^{'}$:"<definiendum> were a" $\vee$ "<definiendum> is an"
$q_{10}^{'}$:"<definiendum> is the" $\vee$ "<definiendum> are the"

In the case $q_{10}^{'} = \emptyset$, the following queries are reformulated:

$q_1^{'}$:"<definiendum> is a" $\vee$ "<definiendum> were an"
$q_3^{'}$:"<definiendum> are a" $\vee$ "<definiendum> was an"
$q_7^{'}$:"<definiendum> was the" $\vee$ "<definiendum> were the"

In summary, the goal of [Figueroa, 2008c] is to enhance the efficiency of the former search methodology, whilst keeping the same amount of queries and, whenever possible, without missing essential nuggets all the time.

GRAMMATICAL NUMBER EFFECT

The work of [Figueroa, 2008c] compared both query reformulation procedures by means of the fifty definition questions supplied by TREC 2003. As a means of assessing both construction methods, a prior definition QA system was used and it was fed with sentences found by each technique. The underlying idea behind this approach is that whenever a significant difference between both strategies exists, this definition QA system would experiment some tangible improvement in its performance. As a matter of fact, an alternative

way of evaluating would be likening the number of distinct nuggets in each retrieval, that is, juxtaposing the recall of both retrieval. At any rate, this sort of assessment does not take into consideration all the misleading information that one technique could have not fetched along with the possible reduction in redundancy of the retrieved sentences. Precisely, it ameliorated the performance for 29 queries while deteriorating it for 17 questions. Figure 2.4 highlights the $\mathcal{F}(5)$ score per question for both strategies.

In general, the static query rewriting backs the definition QA system by finishing with an average $\mathcal{F}(5)$ score of 0.5472, while the dynamic query reformulation helped to better the average value to 0.5792 (5.8%). As well as that, it is worth remarking that this improvement was reached without raising the number of submitted queries.

In a nutshell, results prove that prior knowledge of the grammatical number of the *definiendum* can lead to an enhancement in the overall performance of a definition QA system.

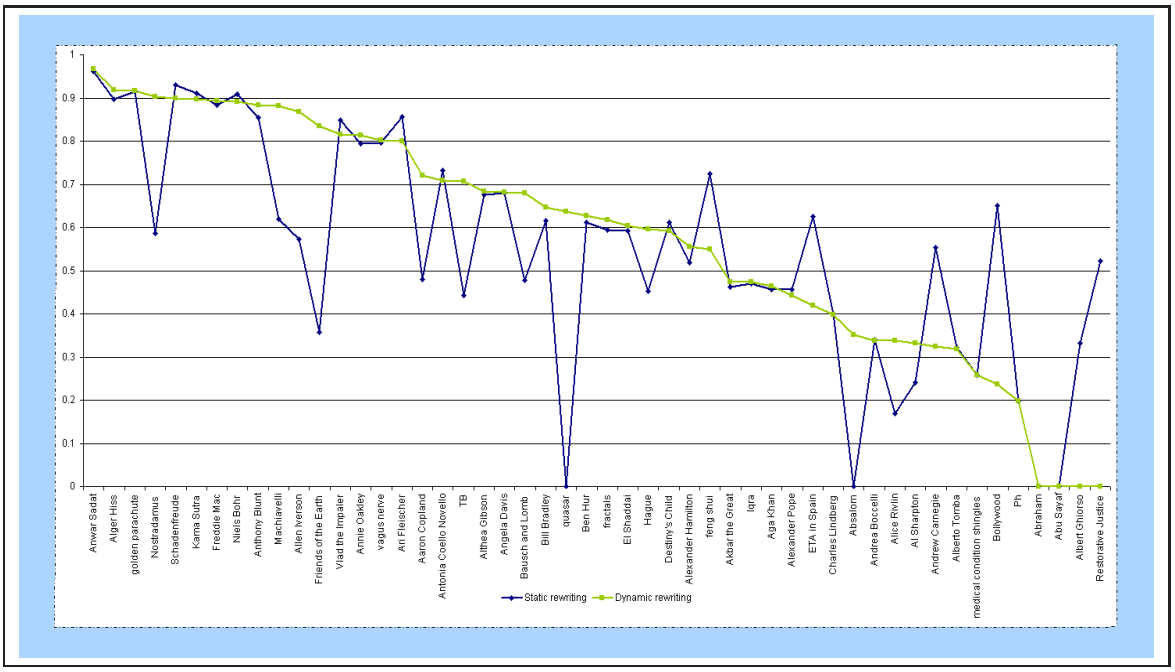

Figure 2.4: Comparison between $\mathcal{F}(5)$ scores achieved for each *definiendum* in the TREC 2003 question-set by the static and dynamic query rewriting.

More Search Engines?

MULTIPLE SEARCH ENGINES

Following the suggestion of [Chen et al., 2006], [Figueroa, 2008c] tested the influence of using several search engines on the performance. They sent, for this purpose, a copy of the queries generated by the dynamic query rewriting to Yahoo! Search. This kind of idea makes perfect sense, because distinct search engines index different parts of the Internet, and accordingly, their indexes could substantially differ. Since [Figueroa, 2008c] sifts nuggets directly from web snippets, the differences in the algorithms that compute the surrogates, namely regarding truncations, turned out to be critical. Paradoxically, this extra search engine slightly boosted dynamic query rewriting by modestly improving the average $\mathcal{F}(5)$ value from 0.5792 to 0.5842 (0.8%). To be more specific, it enhanced the performance for 27 questions, whereas worsening it for 20 queries. Figure 2.5 contrasts the $\mathcal{F}(5)$ score per question for both approaches.

All in all, a marginal rise was achieved at the expense of sending ten auxiliary queries

to the additional search engine, stressing the 5.8% obtained by guessing the grammatical number in conjunction with ten queries and one search engine.

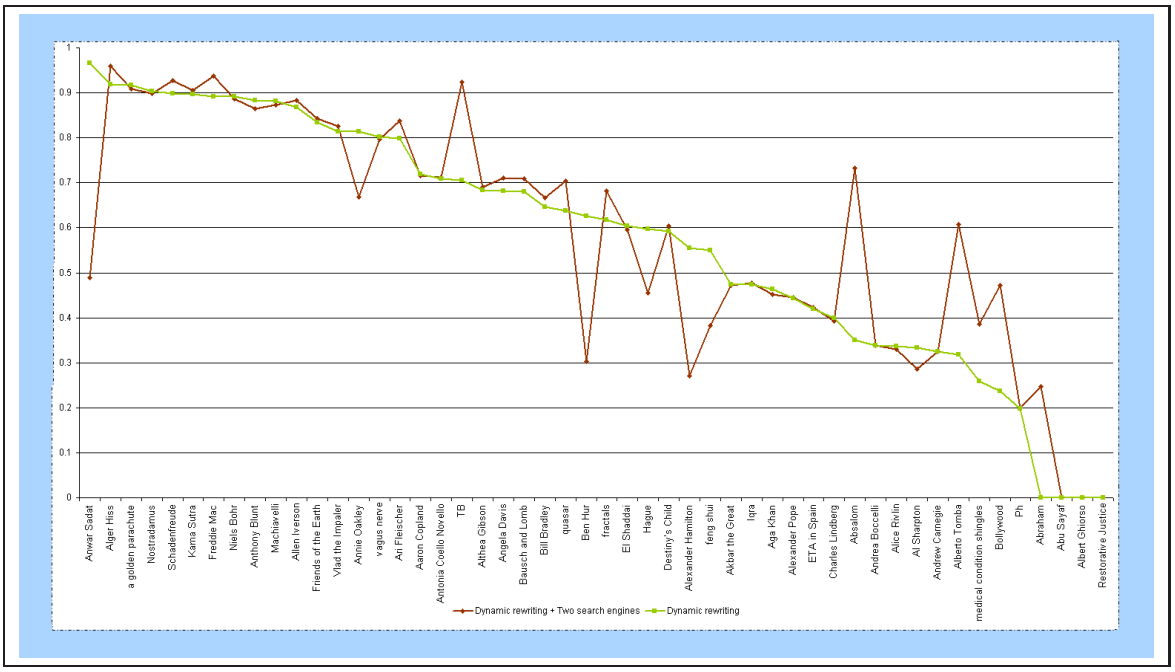

Figure 2.5: Comparison between $\mathcal{F}(5)$ scores achieved by the incorporation of an extra search engine (TREC 2003 question set).

Additional Remarks on the Assessment

The static and dynamic query rewritings scored zero for four distinct *definienda*, despite the "okay" nuggets found by both systems. As a matter of fact, if a system does not discover any nugget assessed as "vital", it finishes with a $\mathcal{F}(5)$ value equal to zero. The dynamic reformulation, for instance, scored zero for three questions; in particular, for the following output in relation to "*Albert Ghiorso*":

- said **Albert Ghiorso, a** veteran Berkeley researcher, who holds the Guinness world record.
- **Albert Ghiorso is a** nuclear scientist at Lawrence Berkeley National Laboratory in Berkeley, Calif.
- That's what Berkeley Lab's **Albert Ghiorso, a** man who has participated in the discovery of more atomic elements than any living person, told the students and teachers who packed.
- **Albert Ghiorso is an** American nuclear scientist who helped discover several elements on the periodic table.

The "okay" nugget is underlined that matches the list of the assessors provided by TREC 2003:

vital designed and built cyclotron accelator
okay nuclear physicists/experimentalist
vital co-creator of 12 artificial elements
vital co-discovered element 106

Like [Hildebrandt et al., 2004] also noticed, "okay" nuggets, like *nuclear physicists/experimentalist* can easily be perceived as "vital". As a support for this notion, consider abstracts supplied by Wikipedia as a third-party judgement, at the time of writing, one finds:

- **Albert Ghiorso** (b. 15 July 1915) is an American nuclear scientist who helped discover numerous chemical elements on the periodic table.

NUGGETS IMPORTANCE

Furthermore, some pertinent text fragments, including *veteran Berkeley researcher*, are unconsidered, enlarging the response, and thus decreasing the $\mathcal{F}(5)$ score. Then, [Figueroa, 2008c] hypothesised that a nugget can be seen as "vital" or "okay" in accordance with how often its **type** (birthplace, birthdate, occupation, outstanding achievement) occurs across abstracts and/or bodies of online encyclopedias, such as Encarta or Wikpedia. Accordingly, [Figueroa, 2008c] deemed this sort of **type-oriented** evaluation would be more appropriate to web-based definition QA systems, because it is impossible to make allowances for an always updating gold standard for the Web. Only in one *definiendum* were all strategies unable to discover any nugget in the list of the assessor: "*Abu Sayaf*". The reason is uncovered when the following frequencies on Google n-grams are checked:

Abu Sayyaf 96204
Abu Sayyafs 89
Abu Sayaf 1156
Abu Sayaff 3205

In this situation, the spelling of the *definiendum* in the query is unlikely to occur in the Internet, causing an $\mathcal{F}(5)$ equals to zero (the reader can also check the case of "*Andrea Boccelli*" and "*Andrea Bocelli*", and the reference [Blair-Goldensohn et al., 2003]). In the opposite way, when the methods process "*Abu Sayyaf*", the scores reached by each technique are depicted in table 2.8.

Strategy	"Abu Sayyaf"	Average
static	0.844	0.5472 → 0.564
dynamic	0.8794	0.5792 → 0.5967
dynamic + 2 engines	0.8959	0.5842 → 0.602

Table 2.8: Impact of misspellings on the $\mathcal{F}(5)$ scores (*definiendum*:"*Abu Sayyaf*").

ADDITIONAL SENSES IMPACT

Another complicated problem is that the list of the assessor is aimed predominantly at one possible sense of the *definiendum*. Ergo, discovered descriptive knowledge concerning ancillary senses, akin to the unconsidered nuggets, bring about a diminution in the $\mathcal{F}(5)$ value. To exemplify this, a descriptive sentence found by one of the methods regarding "*Nostradamus*":

- **Nostradamus is a** neural network-based, short-term demand and price forecasting system, utilized by electric and gas utilities, system operators and power pools, electric...

In deed, it is highly frequent to find ambiguous terms. For example, Wikipedia compounds more than 19,000 different disambiguation pages. In this case, the list of the assessor solely accounts for the reference to the French astrologer/prophet. When sentences respecting other senses are manually removed, the $\mathcal{F}(5)$ values for this concept grow as sketched in table 2.9. Obviously, a more noticeable difference is due to *definienda* with more senses such as "*Absalom*".

Strategy	*"Nostradamus"*
static	0.5871 → 0.5936
dynamic	0.9028 → 0.9182
dynamic + 2 engines	0.8977 → 0.9167

Table 2.9: Impact of ancillary senses on the $\mathcal{F}(5)$ scores (*definiendum*:"*Nostradamus*").

2.6 Enriching Query Rewriting with Wikipedia Knowledge

The prior sections went over strategies geared towards boosting the recall of descriptive phrases within web snippets, and accordingly, the recall of documents containing definitions. So far, the described techniques accounted for Google n-grams as a supporter for reformulating the query. This section, conversely, deals at length with the query rewriting approaches adopted by [Figueroa, 2008a]. These techniques capitalise on Wikipedia resources for coping with some of the issues presented in the preceding section: misspellings, alias resolution and focused query rewriting. Other techniques that might be instrumental include [Bilenko et al., 2003, Cohen et al., 2003].

2.6.1 Alias Resolution & Misspellings

Generally speaking, Wikipedia consists of various sorts of pages including redirection, disambiguation, definition, list, and categories. Distinctly, redirection pages embody no descriptive content, but they link an input string with the respective definition page. [Figueroa, 2008a] perceived these input strings as rewritings of the main concept. To neatly illustrate, the redirection page of "*Clive S Lewis*" connects this name variation to the definition page of "*C. S. Lewis*". These mappings were used essentially for building an **off-line** repository of name rewritings (aliases): REDIRECTIONS

<C. S. Lewis, Clive Staples Lewis>
<C. S. Lewis, C.S. Lewis>
<C. S. Lewis, Clive S Lewis>

With regard to the working examples given in the prior section, the next pairs were discovered in the redirection pages:

<Andrea Bocelli, Andrea Boccelli>
<Abu Sayyaf, Abu Sayaf>
<Abu Sayyaf, Bearer of the Sword>

These pairs underline the utility of redirection pages for tackling misspellings and alternative names head-on. This database is additionally enriched with the alternative name rewritings conveyed in first definition sentences. Take the following case corresponding to "*C. S. Lewis*": FIRST LINE

"**Clive Staples 'Jack' Lewis**" (29 November 1898 - 22 November 1963), commonly referred to as "**C. S. Lewis**", was an Irish author and scholar.

Sentences bearing variations of names are discriminated **off-line** on the grounds of predefined lexico-syntactic clues. These clauses were determined by inspecting recurrent n-grams within these sentences that trigger name aliases. Table 2.10 emphasises the constructs identified by applying these n-gram patterns at the surface level. In the working example, the next mapping was obtained:

Lexico-Syntactic clues		
a.k.a.	colloquially known as	or more commonly
aka	commonly abbreviated	or more precisely
nicknamed	commonly called	or simply
, called	commonly known as	otherwise known as
, known as	commonly referred to as	previously called
, or the	commonly written as	previously known as
abbreviation for	formely known as	previously written as
abbreviation of	generally called	referred to as
also called	generally known as	referred to simply as
also known as	generally written as	sometimes called
also spelled	informely known as	sometimes known as
also written	initially known as	sometimes spelled
also written as	officially called	sometimes spelt
an acronym for	officially known as	sometimes written as
best known as	officially written as	still known as
better known as	often abbreviated	widely known as

Table 2.10: Highest frequent n-grams, in Wikipedia, that signal alternative names.

<C. S. Lewis, Clive Staples 'Jack' Lewis>

TRANSLATIONS

Here, the underlying assumption is that the first line renders details of the corresponding main concept. Definition pages also provide another contributory source of rewritings: translations. For example, the following are variations extracted from the translations of C.S. Lewis's book "*Mere Christianity*": "*Christentum schlechthin*" into German, and "*Mero Cristianismo*" into Spanish:

<Mere Christianity, Christentum schlechthin>
<Mere Christianity, Mero Cristianismo>

To exemplify, these translations cooperate on inferring that the following two names refer to the same entity in the next surrogate:

0061140015: **Mero Cristianismo** by CS Lewis, Veronica Fernandez Muro ...
Mere Christianity is a book that uncovers common ground upon which all those who have ...
Mero Cristianismo. by Lewis, CS, and Muro, Veronica Fernandez ...
www.alibris.com/search/books/isbn/0061140015

Simply stated, this **off-line** repository comprises of 2,418,886 rewritings taken from redirection pages, while 1,797,492 pairs from first line definitions, and 3,353,663 from translations. Then, finding out the aliases of a particular *definiendum* consists of looking for the right entry in this database.

Selecting an alternative *definiendum*

WORDNET SYNSETS

In a specific manner, web definition QA systems are occasionally unable to find descriptive information because users misspell the *definiendum*, or this was correctly entered, but the input is improbable to occur on the Web. Correspondingly, [Wu et al., 2004] dealt with this by making use of synsets in WordNet for turning the *definiendum* into an array of *definienda*.

They illustrated this with the *definiendum*: "*Khmer Rouge*". This can be expanded into: "*Khmer Rouge*", "*KR*", "*Party of Democratic Kampuchea*", and "*Communist Party of Kampuchea*". For their particular purposes, this expansion benefits the retrieval from both the AQUAINT corpus and KBs. In any event, what makes this procedure less attractive is the narrow coverage given by WordNet. Figure 2.3 juxtaposes this with the coverage supplied by Wikipedia. This substantial difference fostered [Figueroa, 2008a] to examine candidate aliases within the aliases repository.

Since all resources utilised for extracting aliases are not perfectly and equally trustworthy, [Figueroa, 2008a] singled out candidates discovered in the first definition lines, and whenever nothing was found there, their approach exploited variations extracted from redirections, thereby ensuring that most dependable aliases are considered first. Due to the query length restrictions imposed by search engines, only aliases candidates written with two or three words were chosen. The more promising candidates aliases are then selected as follows: ALIAS SELECTION

1. If the submitted *definiendum* is formed of three words, aliases that account for the removal of one term are picked. For instance, "*Angela Merkel*" would be weighed if the input is "*Angela Dorothea Merkel*".

2. Aliases bearing the same amount of words, such as "*Nicolas Sarkozy*"⇔"*Nicolas Sarcozy*", are contemplated.

3. If the alternative name resolves or corresponds to an acronym.

4. Only aliases that contain letters and numbers, a hyphen, spaces and/or an ampersand, are taken into account.

Whenever a candidate was found, for the purpose of selecting the right replacement, [Figueroa, 2008a] sent the search engine five search queries per alias candidate. These five purpose-built queries were targeted at the copular lexico-syntactic clues detailed in the preceding section, fetching a maximum of thirty snippets per submission. The underlying assumption here is that clearer evidence (more descriptive phrases) will come into light in the case of the most propitious alias. Ideally, in this step, the user can provide the best variations, but this would inevitable entail an intermediate step, where the user is asked accordingly.

Lastly, it is worth remarking that [Schlaefer et al., 2007] tackled variations of names of organisations by generating queries with and without determiners, and producing an acronym for its name. Some organisation names and their respective acronyms can certainly be found in the alias repository. In addition, the acronyms for some organisations cannot be straightforwardly guessed from its name because they have their origin in a foreign language, or they are irregularly extracted. For instance, the pair : "*Stabilisation Force*" and "*SFOR*", or may be more meaningful to point out: "*Text REtrieval Conference*" (*TREC*).

2.6.2 Definition Focused Search

As a means of enhancing the precision of the hits fetched by the search engine, [Figueroa, 2008a] harvested search clauses from Google 5-grams by examining n-grams starting with the *definiendum*. 5-grams are then seen as search cues, which are filtered by checking as to whether or not the extensions are recurrent across Wikipedia abstracts. Search constructs are hence ranked in agreement with their Google 5-grams frequency. Some examples are the search clauses with respect to "*Angela Merkel*":

Search Clauses concerning *"Angela Merkel"*	Frequency
Angela Merkel , **the conservative**	112
Angela Merkel , **the leader**	319
Angela Merkel , **the opposition**	53
Angela Merkel , **who makes**	57
Angela Merkel , **who took**	48

Table 2.11: Examples of search clauses regarding *"Angela Merkel"*.

By default, based on the spirit of the query reformulation presented in section 2.5, this focused search boosts the retrieval of snippets carrying descriptive phrases by replacing the clues of the following queries:

q_6:"δ, a" $\vee$ "δ, an" $\vee$ "δ, the" $\vee$ "δ or"
q_8:"δ becomes" $\vee$ "δ become" $\vee$ "δ became"
q_9:"δ which" $\vee$ "δ who" $\vee$ "δ that"
q_{10}:"δ was founded" $\vee$ "δ was born" $\vee$ "δ was grounded" $\vee$ "δ stands for"

The first step in the substitution is verifying whether or not more specific constructs for q_6 and q_8 exist. If no clue is found, these queries are sent as they are. For instance,

q_6:"Angela Merkel, the conservative" $\vee$ "Angela Merkel, the leader" $\vee$ "Angela Merkel, the opposition"

Contrarily, q_9 and q_{10} are modified with the highest clauses that were not subsumed in the previous replacements, but substitutions geared towards the original clauses are preferred. Consider the following case:

q_9:"Angela Merkel, who makes" $\vee$ "Angela Merkel, who took"
q_{10}:"Alexander Hamilton was born in" $\vee$ "Alexander Hamilton was born on" $\vee$ "Alexander Hamilton , a founding" $\vee$ "Alexander Hamilton who served at"

The idea behind this modification is trying to focus directly on more specific clues that are very likely to verbalise definitions. Evidently, queries are constrained by the length imposed by the search engine.

2.6.3 Experiments

Figure 2.6 highlights the ratio of the number of nuggets fetched by this technique to the nuggets retrieved by the method of the preceding section. For the TREC 2003 question set, the average value of this ratio was 1.15 ± 0.46 (1.14 ± 0.34 and 1.28 ± 0.72, for TREC 2004 and 2005, respectively). This enhancement was due to 29 questions (58%), for which this strategy retrieved a higher number of different nuggets, whereas in twelve cases (24%) fetched fewer nuggets. In nine (18%) of the questions, there was not a tangible improvement or deprovement. The interesting point in figure 2.6 is that the three more remarkable enhancements stem from the methodology of finding alternative aliases. Given these outcomes, it can be concluded that the repository of aliases is especially helpful for the robustness of this class of systems. However, this enhancement is at the expense of sending auxiliary queries to the Internet, and at the same time, delaying the response time to the user, and in some cases, a misleading candidate can bring about snippets corresponding to another senses. One extreme case is the *definiendum "Artificial Intelligence"*. When consulting the alias repository, one

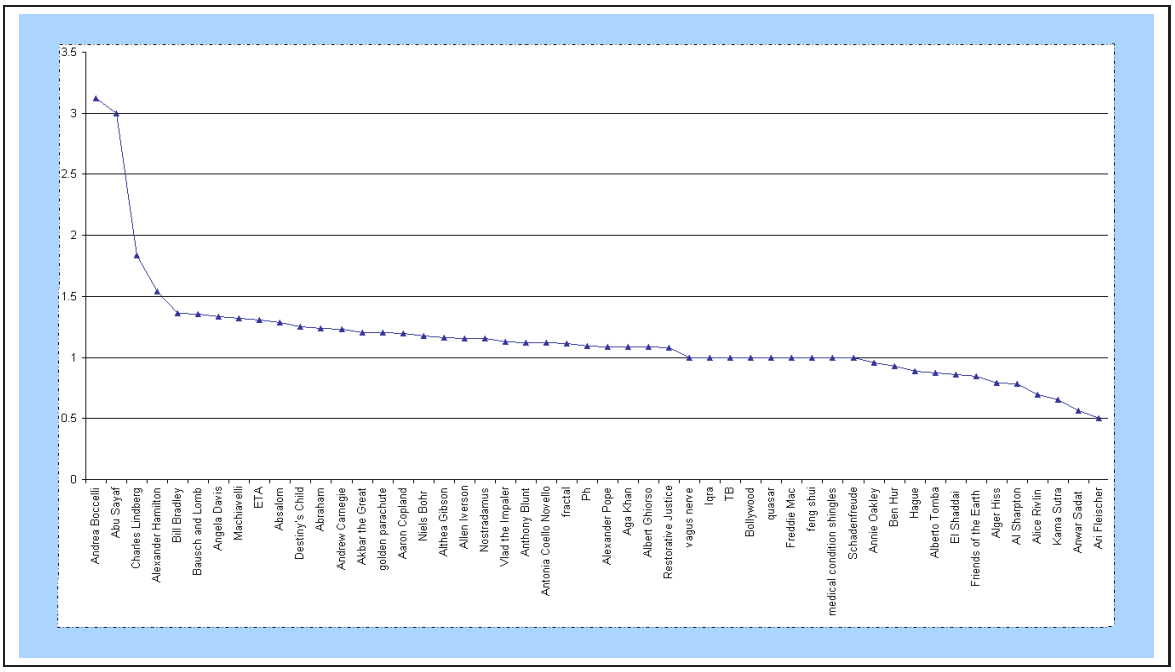

Figure 2.6: Comparison between $\mathcal{F}(5)$ scores reached by each strategy for each *definiendum* in the TREC 2003 question-set.

can get "*AI*" as candidate, which is utilised for numerous purposes, creating the mistaken impression that it is the right candidate. Nevertheless, it can be envisioned that, whenever it is possible, an intermediate phase consisting of requesting at the user for the validation of the alternatives in place of querying the Internet, would be more appropriate.

In the fourth top *definiendum*, the following two queries boosted the recall of descriptive phrases:

q_9:"Alexander Hamilton, who wrote" $\vee$ "Alexander Hamilton that resulted in" $\vee$ "Alexander Hamilton, who favored" $\vee$ "Alexander Hamilton who served at"
q_{10}:"Alexander Hamilton was born in" $\vee$ "Alexander Hamilton was born on" $\vee$ "Alexander Hamilton , a founding" $\vee$ "Alexander Hamilton who served at"

Some representative examples of the array of web snippets fetched by these two queries are as follows:

The American Experience | The Duel | People & Events | The Republican Party
... Federalists such as **Alexander Hamilton, who favored** a strong central government. They feared that the concentration of federal power under George ...
www.pbs.org/wgbh/amex/duel/peopleevents/pande09.html

NIAHD Journals
Aaron Burr is best known for his duel with **Alexander Hamilton that resulted in** Hamilton's death while Vice President of the United States to Jefferson's ...
niahd.wm.edu/index.php?browse=date&id=531

Washington, George - FREE Washington, George Information ...
Washington proved unable to heal the divisions between Thomas Jefferson and **Alexander Hamilton that resulted in** the creation of the Federalist Party and the ...
www.encyclopedia.com/doc/1O142-WashingtonGeorge.html

Amazon.com: Alexander Hamilton: Founding Father and Statesman ...
A biography profiling the life of **Alexander Hamilton, a founding** father of the United States and the first Secretary of the Treasury. ...
www.amazon.com/Alexander-Hamilton-Statesman-Signature-Revolutionary/dp/0756510732

The American Experience | The Duel | People & Events | Alexander ...
Alexander Hamilton was born on January 11, 1757, in Nevis, British West Indies. His father, James Hamilton, was a Scottish trader. ...
www.pbs.org/wgbh/amex/duel/peopleevents/pande06.html

On the contrary, in the case of "*Ari Fleischer*", the diminution in the performance resulted from selected clauses that were semantically similar, and consequently, they brought about the retrieval of descriptive phrases that convey similar nuggets. Particularly, a quite fruitless retrieval eventuated from the following query:

q_6:"Ari Fleischer , the president" $\vee$ "Ari Fleischer , the press" $\vee$ "Ari Fleischer , a spokesman"

Basically, the next two delineative snippets summarise the retrieved set:

CNN.com - Transcripts
THIS COPY MAY NOT BE IN ITS FINAL FORM AND MAY BE UPDATED. ... This is **Ari Fleischer, the president**'s spokesman with his afternoon briefing. ...
transcripts.cnn.com/TRANSCRIPTS/0109/24/se.21.html

1110cona
Whenever the press attempts to define a new Presidency, ... **Ari Fleischer, the press** secretary who announced it, told the press ...
www.bartcop.com/1112cona.htm

On the whole, extending the query by profiting from Wikipedia resources is a promising idea. The drop to recall shown here can be allayed by enlarging instead of replacing the number of queries. This depends, however, on the sort of the application and the amount of time the user is likely to wait.

2.7 Searching in other Languages: Spanish

Fundamentally, surface patterns for English have been studied broadly, whereas regularities for other languages have been systematically explored only in the context of the CLEF campaigns. Until 2005, CLEF focused exclusively on definition questions targeted at abbreviations and the position of persons [Vallin et al., 2005, Magnini et al., 2006]. These surface patterns were therefore specialised for recognising this specific sort of descriptive knowledge. In contrast, systems in TREC are encouraged to extract as much useful descriptive information as possible about *definiendum* [Hildebrandt et al., 2004]. Ergo, these surface patterns provide a comparatively ampler coverage than the constructs known for other languages.

For the case of surface patterns for Spanish, two additional issues complicate the identification of descriptive content on the Web. Firstly, the regularities are premised largely on punctuation signs [Montes-y-Gómez et al., 2005] and closed class words [Denicia-Carral et al., 2006], which are usually ignored by some search engines. Secondly, these punctuation signs and closed class words tend to be separated by a large span of text. To reinforce this point, consider the following two examples:

Queries
q_1:"<definiendum>"
q_2:"<definiendum>, fue un" ∨ "<definiendum> son lo" ∨ "<definiendum>, la"
q_3:"<definiendum> fue la" ∨ "<definiendum> es el" ∨ "<definiendum> son el"
q_4:"<definiendum> que" ∨ "<definiendum> son las" ∨ "<definiendum>, lo"
q_5:"<definiendum> es un" ∨ "<definiendum> ha llegado a ser" ∨ "<definiendum> son la" ∨ "<definiendum> fueron las"
q_6:"<definiendum> fue el" ∨ "<definiendum> son unas" ∨ "<definiendum>, uno" ∨ "<definiendum> ha sido la"
q_7:"<definiendum> quien" ∨ "<definiendum> los cuales" ∨ "<definiendum>, un" ∨ "<definiendum> son una"
q_8:"<definiendum> se ha transformado" ∨ "<definiendum> es lo" ∨ "<definiendum> fue fundado"
q_9:"<definiendum>, el" ∨ "<definiendum> son unos" ∨ "<definiendum> fue una" ∨ "<definiendum> fue fundada"
q_{10}:"<definiendum> es la" ∨ "<definiendum> llego a ser" ∨ "<definiendum> ha sido el" ∨ "<definiendum> son un"
q_{11}:"<definiendum> es una" ∨ "<definiendum> fue lo" ∨ "<definiendum> ha sido un"
q_{12}:"<definiendum> se transformo" ∨ "<definiendum> fue uno" ∨ "<definiendum> , las"
q_{13}:"<definiendum> la cual" ∨ "<definiendum>, una" ∨ "<definiendum> ha sido una"
q_{14}:"<definiendum> es uno" ∨ "<definiendum> nacio" ∨ "<definiendum> el cual" ∨ "<definiendum>, los"

Table 2.12: Queries for searching definitions in Spanish.

, el <description>, <definiendum>, dijo
y el <description>, <definiendum>.
El <description>, <definiendum>, se

The last pattern, for instance, matches sentences such as *"El presidente de España, Jose Luis Zapatero, se..."* The snippets acquired by the respective query rewriting *"El"* ∧ *", <definiendum>, se"* are unlikely to yield definitions, and additionally, portions of the large span of text between the *definiendum* and the closed class word *"El"* can be replaced with an intentional break (often denoted by "...") by the search engine.

TRANSLATIONS

All things considered, the study of [Figueroa and Neumann, 2007] extended this set of regularities by taking into account translations of the English lexico-syntactic constructs into Spanish. These translations are outlined in table 2.12, and they cooperate on overcoming the special difficulties exhibited when taking advantage of the patterns by [Juárez-Gonzalez et al., 2006, Denicia-Carral et al., 2006]. To be more precise, they conform a compact block of consecutive words that disallows this problematic span of text. These fourteen queries bias the retrieval in favour of web snippets that match these definition lexico-syntactic regularities that often render descriptions in Spanish.

COVERAGE

Plainly speaking, the more successful this query reformulation is, the larger the recall of web snippets, and hence, documents containing definitions, is. This query rewriting aided in answering 32 and 22 out of the CLEF 2005 and 2006 questions, respectively. However, the runs submitted by the best two systems in CLEF 2005 answered 40 out of the 50 definition questions [Vallin et al., 2005, Montes-y-Gómez et al., 2005]. Nonetheless, the third-best system solely responded to 26 questions. It is fair to highlight here that these answers were

distinguished on the Web, in juxtaposition of CLEF systems, this also makes the nuggets in the ground truth dependent on the EFE corpus. For this reason, results are positively encouraging.

Additionally, the best system in CLEF 2006 responded to 35 out of the 42 definition questions, whereby this method supported in finding answers to 22 out of the 35 questions answered by this best system. Unfortunately, CLEF 2006 gold standard provides only one nugget for solely these 35 questions. Furthermore, in [Figueroa, 2008b], this procedure cooperated on discovering nuggets for 17 out of 19 concepts distilled from the CLEF 2008 question set. It is also worth remarking that this query rewriting technique fetched longer sentences, [Figueroa and Neumann, 2007] compared the length of the sentences obtained by this strategy: 135.78 $\pm$ 45.21 and 113.70 $\pm$ 37.97 with and without white spaces, respectively, with the length of the phrases gathered by sending the *definiendum*: 104.98 $\pm$ 36.43 and 85.88 $\pm$ 29.87 with and without white spaces, respectively. This improvement in terms of length helps to cushion the effects of truncations on web snippets.

DESCRIPTION LENGTH IN SPANISH

The reason why a noticeable growth in the recall of descriptive information enhances the chance of correctly answering a definition question is two-fold: (a) it raises the probability of matching the context of a model previously learnt from annotated examples, including those taken from online encyclopedias and dictionaries, and consequently (b) it facilitates the selection of the most relevant and reliable, as well as descriptive answers.

In actuality, the success of this approach lies in the size of the target corpus, in this situation, the Spanish web. A larger corpus tends to support with broader coverage, and therefore, likely to assist QA systems in leaving less unanswered questions. But more important a considerably larger corpus yields a large-scale redundancy. It is worth duly pointing out that, by redundancy it is not meant duplicate information, but rather distinct paraphrases of the same underlying ideas. QA systems can undoubtedly benefit from paraphrases, because they markedly increase the probability of matching query terms and purpose-built patterns. As a consequence, they considerably boost the chance of finding more and fuller answers.

Unfortunately, there is a big difference between the number of web documents in English and Spanish. As a very rough rule of thumb, this difference can be estimated approximately by submitting some lexico-syntactic clues to the Internet in order to get their web frequency counts. Table 2.13 emphasises this difference.

English		Spanish	
is the	6,840,000,000	es un	323,000,000
is a	8,720,000,000	es una	172,000,000
is an	2,440,000,000	es la	162,000,000
		es el	150,000,000
		es uno	36,700,000
	+	es las	+1,710,000
	18,000,000,000		845,410,000

Table 2.13: Comparison between frequencies of definition clues.

CORPUS SIZE

These rough estimates indicate that the size of this corpus falls into a drastic decline from about 18 billion in English to 1 billion in Spanish. This comparatively small number of matches enforces definition QA systems to divert time and effort away from extracting answers to perform an exhaustive search for propitious documents. In general, when answering definition questions in English, few queries directed at a small array of lexico-syntactic constructs suffice to obtain a high recall of web snippets, and hence, documents that carry descriptive information about the *definiendum*. Since the amount of Spanish web documents

is much smaller, the probability of matching these lexico-syntactic constructs dramatically decreases. The system is, for this reason, compelled to submit a larger amount of queries to the search engine, as a means to sharply increase the probability of obtaining diverse and sufficient knowledge to satisfactorily answer the question.

MORPHOLOGICAL COMPLEXITY

There are also linguistic aspects that make the search process more demanding. Most nouns in modern English lack grammatical gender. This gender is triggered by two indefinite articles: "*a*" and "*an*", and one definite article: "*the*", which is also used for indicating plural forms. Therefore, a query as follows would be enough to retrieve many descriptive nouns:

q_*:"<definiendum> is a" ∨ "<definiendum> is an" ∨ "<definiendum> is the"

Inversely, Spanish uses three grammatical genders: feminine, masculine, and neuter, which are signalled by six definite and indefinite articles. Furthermore, Spanish utilises four additional morphological forms of these articles for agreement with the number of the noun phrase they modify, that is indicating plural nouns. All together, this increases the number of articles from three to ten. The reader can verify this increase by inspecting the queries depicted in table 2.12.

This growth in morphological complexity demands an extra effort that goes into the search process. More specifically, a richer noun morphology leads to more lexico-syntactic clues, which means more search clauses, and by the same token, a longer retrieval time.

Lastly, contrary to English, amalgamating this strategy with the utilisation of Google n-grams and an alias repository is not as fruitful for Spanish as for English. Firstly, there are no Google n-grams available for Spanish, and thus the same approach for guessing the grammatical number or rearranging the clauses across the purpose-built search queries is, for the moment, implausible. Secondly, it is worth underlining that the coverage offered by Wikipedia, widely varies from English to another language. At the time of writing, Wikipedia supplied about 2,500,000 English articles, whereas only about 450,000 articles in Spanish. Wikipedia, consequently, yields a really limited number of entries for the alias repository in Spanish.

More languages?

GERMAN

Adapting the methods to cope with other languages, including German, brings about four challenges: (a) discriminate descriptive phrases in the present tense from sentences in perfect tense with "*sein*", (b) German has a richer morphology, causing a growth in the amount of search clues, (c) relative clauses in German have a more complex structure, and (d) coping with the orthographical variations caused by umlauts and compounds such as "*Koeln*" and "*Köln*". A deeper analysis of these phenomena is beyond the scope of this work. Nevertheless, the reader can go over [Sacaleanu et al., 2007, 2008] for some passage retrieval strategies exploited in the context of definition QA systems directed at the German language.

2.8 Future Trends

Looking for descriptions in external KBs seems to be the path of least resistance, due to their reliability. Definition QA systems make use of these reliable descriptions in different ways. One can envision, for this reason, systems with access to a larger amount of KBs, this way obtaining a wider coverage of reliable descriptions. In order to explore KBs supplementary to the ones presented in table 2.4, a web-based definition QA system operating on web-snippets was utilised for extracting 1,829,889 descriptive phrases of about 291,026 different *definienda*. These *definienda* were distilled from Wikipedia, and they regard wide-ranging domains. As a

result, this system sifted these presumed definitions from 511,044 different hosts. As a means of knowing the most noticeable places, the frequencies of the hosts running websites were counted. Table 2.8 underlines the most conspicuous hosts, excluding those already depicted in table 2.4:

Website	Hits	Website	Hits
www.absoluteastronomy.com	22363	www.probertencyclopaedia.com	2487
www.amazon.com	14098	www.usatoday.com	2457
dbpedia.org	6607	www.myspace.com	2339
dbpedia.openlinksw.com:8890	6601	www.time.com	2091
duckduckgo.com	6424	www.mahalo.com	2070
infao5501.ag5.mpi-sb.mpg.de:8080	6071	www.washingtonpost.com	2043
www.guardian.co.uk	5008	www.rootsweb.ancestry.com	2017
www.thefreelibrary.com	4436	www.imdb.com	1938
www.spock.com	4400	sports.espn.go.com	1885
www.reference.com	4188	www.bbc.co.uk	1866
query.nytimes.com	3760	ezinearticles.com	1757
www.freepatentsonline.com	3629	testserver.semantic-mediawiki.org	1735
en.allexperts.com	3121	www.pbs.org	1714
www.geocities.com	2971	findarticles.com	1711
articles.latimes.com	2855	encycl.opentopia.com	1698
www.youtube.com	2799	www.powerset.com	1638
news.bbc.co.uk	2774	www.sfgate.com	1630
profile.myspace.com	2740	www.boston.com	1557

Table 2.14: Websites providers of descriptive sentences.

The idea behind this table is that internet websites which are more probable to render reliable descriptive sentences will often come up across various *definienda*. This idea is somehow similar, but not identical, to the Wisdom Of The Crowds principle suggested by [Surowiecki, 2004]. Of course, the reliability of this table depends largely on the system used for mining the phrases. Obtaining authoritative top entries, is nevertheless, expected. Table 2.8 shows these frequency counts and remarks the need for making allowances for *close-domain* KBs, such as www.imdb.com and sports.espn.go.com. Some illustrative definitions extracted from these resources:

Alan Grant is a former NFL defensive back who played collegiately at Stanford.
sports.espn.go.com/espnmag/story?id=3649085

Mini Biography: **Aaryn Doyle** is a Canadian actress, singer, songwriter, and model.
www.imdb.com/name/nm1401989/

Since these KBs are *close-domain*, there is a chance to design procedures that discriminate descriptions on grounds of domain particularities, and therefore, to achieve a high precision. Table 2.8, on the other hand, highlights trustworthy *open-domain* KBs, including www.absoluteastronomy.com , and not so dependable sources of definitions:

Fast Glass is a feature film starring the Pilots, planes and aerial camera work of Afterburner-Mach One Aviation.
www.youtube.com/watch?v=PmGSYM6btRs&fmt=18

John H. Farr is a writer, photographer, and Web designer in Taos, New Mexico.
www.youtube.com/profile?user=jhfarr

Umberto Bindi is a great personnage in the modern song history.
www.youtube.com/watch?v=w4lyysUo_WI

In light of these results, it can be concluded that this class of websites can provide definition QA systems with valuable pieces of information, like descriptions in www.youtube.com videos. Another trend is capitalising on *pseudo relevance feedback*, to enhance the recall of descriptive knowledge. This could be done analogously to [Chen et al., 2006]. This kind of method is an arable field to work in conjunction with the strategies explicated in sections 2.5 and 2.6.

2.9 Conclusions

Conventionally, most definition QA systems take advantage of on-line and/or off-line KBs as reliable answer sources. The quintessence of each class are Wikipedia and WordNet, respectively. Broadly speaking, these kinds of resources have proven to better the performance of TREC-oriented systems. But on the other hand, they have also shown to suffer from narrow coverage. Since TREC-oriented systems project trustworthy nuggets found across these dependable resources into the AQUAINT corpus, they overcome this coverage problem by devising query rewriting techniques that allow them to ameliorate the recall of promising web-snippets and full-documents.

This chapter discusses various approaches to discover propitious sources of answers on the Internet at length. These strategies are basically directed at: (a) disregarding information about the domain, that is they are utterly open-domain; and (b) boosting the probability of carrying the descriptive knowledge within the document surrogate returned by the search engine, this way definition QA systems can avoid fetching full-documents. In conclusion, results indicate that Google n-grams and Wikipedia resources are particularly useful for optimising the retrieval of descriptive knowledge within web snippets, and as a logical consequence, they can also back QA systems in fetching a larger amount of promising full-documents. Further, given the obtained outcomes, it can also be concluded that an alias resolution phase leads to marked enhancement in terms of recall and performance. Furthermore, results additionally reveal that the utilisation of multiple search engines brings about a negligible improvement only.

In addition, this chapter dissects the advantages and disadvantages of the design and use of definition relational databases, and it also looks into the obstacles encountered when trawling the Internet for descriptive knowledge in another languages (i.e., Spanish). In essence, some of these issues encompass the marked difference in the size of the Spanish and English Web, and the increase in morphological complexity of Spanish in comparison to English. A final observation is due to the fact that different crawling techniques fetch different pieces of descriptive information, and thus bring to mind the idea of amalgamating them as a means to boost recall at the expense of download time. This chapter also envisages an increment in the exploitation of *close-domain* KBs as sources of authoritative descriptive knowledge.

Chapter 3

What Does this Piece of Text Talk about?

> "We are drowning in information but starved for knowledge." (John Naisbitt)

> "As we know, there are known knowns. There are things we know we know. We also know there are known unknowns. That is to say we know there are some things we do not know. But there are also unknown unknowns, the ones we don't know we don't know." (Donald Rumsfeld)

3.1 Introduction

By and large, definition Question Answering (QA) systems take advantage of Information Retrieval (IR) engines for finding passages and/or documents relevant to the question asked by the user. In general, each IR engine has a set of views of the collection reflecting a combination of global (all documents) and local (each document) statistics. These representations are utilised typically for indexing the collection, and accordingly, for locating the most propitious documents given a search string.

For the most part, IR engines return top-ranked documents in accordance with a given purpose-built similarity measure. Broadly speaking, this similarity measure takes terms within the search string, and scores each document containing any of these terms in concert with how representative each term of the document and the collection is. Most delineative passages are ordinarily returned as document surrogates along with a link to the respective resource. Assuredly, a myriad of diverse approaches to compute this likeness do exist.

There are several reasons why this abstraction and ranking can be problematic for definition QA systems. For starters, the index of the collection and the similarity function are rarely optimised for privileging descriptive content about the search string, thus causing a detriment to recall. It is important to underline here that systems typically take into consideration solely the top N hits (normally, N=100...1000), since there is always a trade-off between quality and response time. Secondly, a disagreement between the writing of the *definiendum* formulated by the user and the alias found in the collection exists, forcing to slack the matching, which is materialised in the form of disperse matches that influence the recall of descriptive knowledge, and from the standpoint of definition QA systems, the precision of the set of results.

In practice, the array of features that each similarity measure prioritises goes hand in hand with the intention behind the application. A commercial search engine, for instance, may prefer documents that raise their revenue as opposed to more informative hits. In other

words, the input of the user might be only one of the key factors when ranking and retrieving the most promising documents.

This chapter is aimed specifically at fleshing out some of the complicated matters found across passages obtained from the target collection of documents, when querying a traditional IR engine for the *definiendum*. The organisation of this chapter is as follows: section 3.2 elaborates on the issue of the alignment of terms embodied in the *definiendum* with words in retrieved excerpts, section 3.3 discusses some of the relations and sources of disagreement between the topic(s) of the fetched passages and the topic(s) stipulated by the *definiendum*, section 3.4 describes some approaches to ensure that some classes of excerpts (i.e., passages that observe some widely known definition patterns) will talk about the *definiendum*, or that is to say, to ensure the agreement between the topic(s) of the *definiendum* and the topic(s) of particular kinds of passages, section 3.5 goes over some aspects related to co-reference resolution and predicts future trends, and finally, section 3.6 highlights the main conclusions.

3.2 *Scatter Matches* of the *Definiendum*

DISPERSE MATCHES

To begin with, *scatter matches* is the first phenomenon that can be encountered across surrogates and documents returned by IR engines. Precisely, some of the top-ranked results can bear only disperse matches of the words within the query (*definiendum*). These *scatter matches* can have both a positive and negative impact on the recall of descriptive knowledge. On the positive side, a disperse match allows insertions of terms between and/or around *definiendum* words, thus boosting the chance of discerning a greater amount of potential descriptions. Inversely, *scatter matches* escalate the probability of retrieving more spurious and misleading contexts. By *scatter matches*, it is also understood as a partial matching of the *definiendum*. To reinforce this point, consider the next surrogate retrieved by Yahoo! Search when searching for "*Barack Obama*":

President Barack Obama
THE ADMINISTRATION • PRESIDENT **BARACK OBAMA** ... PRESIDENT **BARACK OBAMA**. **Barack** H. **Obama** is the 44th President of the United States. ...
www.whitehouse.gov/administration/president_obama/

In this fragment, the insertion of the token "*H.*" makes the recognition of the definition pattern "<*definiendum*> is the <description>" harder. Note that this pattern is widely utilised for singling out promising answer candidates. The reader can look at a list of widespread definition patterns in section 3.4 on page 67. As a means of tackling this matter, [Xu et al., 2003] allowed segmenting *definienda* targeted at person names by a sequence of at most three terms "F w_1 w_2 w_3 L", where F and L denote the first and last names of the *definiendum*, respectively. On the other hand, they required the exact match for any other kind of *definiendum* after converting plurals to their respective singular forms. This technique assists in matching "*Barack H. Obama*", when coping with "*Barack Obama*". By any means, this sort of procedure can provide false positives (e.g., "*George Bush*" can overmatch "*George H. W. Bush*" or "*George W. Bush*"). Another linguistic phenomenon observed within descriptive phrases is the deletion (partial matching) of *definiendum* words. To exemplify, consider the next surrogate:

TOKEN INSERTION

TOKEN DELETION

Barack Obama: Biography from Answers.com
Obama was elected to the Illinois state senate in 1997, where he served as ...
Obama was elected to the Illinois Senate in 1996, and then to the U.S. Senate in ...
www.answers.com/topic/barack-obama

Fundamentally, [Katz et al., 2004] coped with this phenomenon by splitting the *definiendum* and checking the matches afterwards. Even so, not all *definiendum* terms deserve the same consideration, as many of them carry few meaning (e.g., "*The Fool and His Money*", "*Hatfield and the North*" and "*For Love or Money*"). Further, in many cases a full match must be enforced (e.g., "*For Love or Money*") because the *definiendum* is compounded from words that have the potential of matching extremely diverse topics. The archetypes of this kind of term are stop-words. This approach seems to be, nevertheless, the path of least resistance when overcoming this drawback. Yet, this problem becomes more difficult when deletions and insertions happen simultaneously:

INSERTIONS AND DELETIONS

Five Leadership Lessons From Obama's First Month - Forbes.com
The president has done three things that all CEOs should ... more than a month since **President Obama** was sworn in as the 44th President of the United States.
forbes.com/.../25/obama-lessons-president-leadership-managing_ceo.html

Research into definition QA systems has not directly addressed this issue yet. One reason for this might be the fact that redundancy mitigates the consequences implied by this phenomenon. This redundancy comes from amalgamating evidence from the target corpus (i.e., AQUAINT) and Knowledge Bases (KB) as well as the Web. That is, descriptions found in sentences with a disperse match of the *definiendum* can also be discovered across sentences that provide more precise matches. On the same token, the evidence yielded by contexts carrying *scatter matches* can be corroborated by the context taken from the corresponding entry in KBs and web sentences containing the exact *definiendum*. These solutions, nonetheless, work for *definienda* high in frequency or amply covered by knowledge bases, and they rule out the fact that senses are not specified by KBs, but rather by the entire collection of target documents.

Anyway, [Soubbotin, 2001] pioneered efforts in tackling this head-on by accounting for a pack of complex rules that recognise changes, such as word permutations and the insertion of punctuation and/or titles (e.g., "*His Excellency*"). They noticed that definition patterns, like "<*definiendum*> is the <description>", with more sophisticated internal structures are more indicative of the answer. Another approach is due to [Wu et al., 2004], who took advantage of synsets in WordNet for dealing with alias resolution. Certainly, this method relies heavily on the coverage of WordNet. On a different note, the repository of aliases presented in section 2.6.1 on page 45 can also offer some help.

PERMUTATION RULES

WORDNET SYNSETS

3.3 Topic Shift in Descriptive Sentences

From another perspective, passages returned by the IR engine can subsume descriptions, but of topics different from the *definiendum*. This phenomenon can be the result of two distinct crucial factors: (a) as aforementioned, *scatter matches* can produce a false positive that materialises a shift in the focus of the definition towards another topic; and (b) even though a text passage can exactly match the *definiendum*, its context can signify the shift. The following surrogates illustrate how *scatter matches* can bring about a topic shift:

TOPIC SHIFT

American Presidents: Life Portraits
George Bush is the second president whose son became president. ...
Dec 12, 11:59 AM: Interview with President **George Bush** recorded November 19th at the ...
www.americanpresidents.org/presidents/president.asp?PresidentNumber=40

George Walker Bush - Infoplease.com
George Walker Bush was born on July 6, 1946, in New Haven, Conn., the first child of future president **George H. W. Bush**.
In 1948, the family moved to Odessa ...
www.infoplease.com/ipa/A0878291.html

George Herbert Walker Bush - Infoplease.com
George Herbert Walker Bush was born June 12, 1924, in Milton, Mass., to Prescott and Dorothy Bush.
The family later moved to Connecticut. ...
www.infoplease.com/ipa/A0760625.html

DISPERSE MATCHES

In these working snippets, the addition of the tokens "*Walker*" and/or "*Herbert*" might induce the detection of descriptions about two different but strongly related presidents. In actuality, this is the crystallisation of a much deeper problem in definition QA systems: *sense discrimination*. Both descriptions could have been equally verbalised by referring only to the name "*George Bush*", entailing consequently the need for a deeper analysis to detect the ambiguity, and accordingly, the necessity for grouping answer candidates in congruence with their senses. It is crystal clear, nevertheless, that this kind of *scatter matches* aggravates the possibility of perceiving spurious putative answers as genuine, because misleading answers are given higher chances of being selected as answer candidates. Like in the working examples, in many cases, the degree of relatedness of the distinct concepts makes the discrimination the different senses harder (e.g., both presidents of the same country, and the same bloodline). Further, [Figueroa and Neumann, 2007] observed that some shifts can bring forth interesting descriptive phrases. An excelent case is the disease "*neuropathy*", which can be shifted to the next closely related disorders:

HYPERNYMY TOPIC SHIFT

- **Peripheral neuropathy** is a common nerve disorder caused by damage to nerves and nerve processes outside the brain and spinal cord.
- **Diabetic neuropathy** is a progressive disease that is most probably caused by the effects of a chronic deficiency of prostacyclin and prostaglandins.
- **Alcoholic neuropathy** is a disorder involving decreased nerve functioning caused by damage that results from excessive drinking of alcohol.
- **Auditory neuropathy** is a hearing disorder in which sound enters the inner ear normally but the transmission of signals from the inner ear to the brain is impaired.

These four related diseases can be seen as hyponyms of the same hypernym, that is they pertain to the putative input of the user "*neuropathy*". It is notable that some semantic similarities used for these four explanations offer insight into the meaning of the *definiendum* (i.e., *disorder/disease*, *nerve*, *signal* and *caused by damage/effects of*):

- **Neuropathy** is a disease of the nervous system.

While showing these four explications to the user is disputable and depends on the goals of the application, they help to understand the underlying linguistic phenomenon behind the recognition of definitions (see the discussion in sectio 1.6 on page 10). On the contrary, some shifts can generate -what arguably are- loosely related sentences: "*G7*" to "*Powershot G7*".

- The **G7** is an informal group of Finance Ministers and Central Bank Governors from the United Kingdom, the United States, France, Canada, Italy, Japan and Germany.
- The **Powershot G7** is the successor to the popular G6 model from 2004 and both share a resemblance to that of a 35mm rangefinder camera.

At any rate, insertions can also be neuter, that is they do not necessarily need to add information about the *definiendum* and they do not shift the topic:

The Bourbon Asylum: A vote for Count Chocula is a vote for freedom
Your choices are **Count Chocula**, the wussy whine boy vampire from Twilight (retch), some banal vampire from an obscure HBO series, and a Muppet.
bourbonasylum.blogspot.com/

On The Colorado - Articles
Whereas removing **Glen Canyon Dam**, which is not essential water infrastructure, would increase river habitat by 500 miles on the Colorado, San Juan, and all the convergent.
www.onthecolorado.org/articles.cfm?mode=detail&id=1230799075869

In summary, what is at stake is a trade-off between precision and recall. While a flexible matching of the *definiendum* increases the recall, it can also overmatch, seeing some misleading sentences as genuine definitions. A strict alignment can, in contrast, leave some definition questions unanswered, or at least, it can exclude some pertinent facets of the *definiendum* from the output of the user. It is possible, nonetheless, that some spurious sentences can be filtered out in posterior steps of the answering process, namely ranking.

CONTEXT SHIFT

This last kind lies in the middle ground between the two causing factors of topic shift, since the extra terms can be interpreted as the context in which the -full- matching of the *definiendum* occurs. Generally speaking, the second factor of topic shift encompasses a variety of texts, including opinions. Relaxing the matching of the *definiendum* can match some structures that are commonly used for expressing opinions (e.g., "*I think/believe that definiendum is the* ..."). This class of shift situates the explanation in a private/personal context, making it very subjective, and ergo untrustworthy. Another representative case of context shift stems from some matches of the *definiendum* when contained in prepositional phrases. A straightforward example of this second sort of factor is as follows:

Texas Academy of Science Conference Success for TAMIU, Laredo
The 105th Annual meeting of the Texas **Academy of Sciences** was an enormous success thanks to the hard work and dedication of a number of people.
www.tamiu.edu/newsinfo/3-21-02/article12.htm

GRAMMATICAL NUMBER

Previously, section 2.5.1 briefly touched on the subject of topic shifts. Distinctively, this section focused on the disagreement between the grammatical number of the subject of the sentence and the grammatical number of the *definiendum*. This discrepancy, by all means, can be utilised as one latent feature. That is to say, it can sometimes cooperate on filtering out some misleading hits.

3.4 Strategies to Match the *Definiendum* when Aligning Definition Patterns

In recent years, the problem of aligning the *definiendum* entered by the user with an occurrence within a candidate descriptive sentence has not been the focus of attention of definition QA systems. As mentioned earlier, the reason for this is that most systems allay this problem by relying largely on the redundancy supplied by knowledge bases, the Internet and/or the target collection of documents. The growing need, however, for outputting more diverse output to the user stimulates the demand for more efficient matching techniques.

ANTECEDENT NOUN PHRASE

BASE NOUN PHRASES

Predominantly, aligning methods capitalise on shallow syntactic information, namely chunking. For instance, [Hildebrandt et al., 2004] made sure that the noun phrase preceding the verb phrase embodying the definition verb (e.g., "*is*" and "*was*" as well as "*became*") bears the *definiendum* (see section 3.4). In the same spirit, [Xu et al., 2005] checked the first two noun phrases respecting to the first sentence of a paragraph. In order to be conceived as a candidate answer, they imposed the restriction that both noun phrases must be joined with one of the next two prepositions: "*by*" and "*for*". This extension allowed [Xu et al., 2005] to capture sentences defining more complex *definienda*, such as "*Perl for ISAPI*":

Perl for ISAPI is a plug-in designed to run Perl scripts...

Certainly, the imposition of constrains is aimed specifically at boosting the accuracy of the definition pattern matching at the expense of worsening the recall, and consequently, a diminishment of the diversity of the output. Nevertheless, the degree of diversity is dependent on the type of application. While a web user can be expecting a rich variety of information, a user in front of a mobile phone can be waiting for a concise, but quick and precise response. One kind of information that is very likely to be missed when placing the restriction of [Xu et al., 2005], is historical events such as the following:

Montenegro Adventures | History
In 1077 **Duklja** became a kingdom named Zeta, under the rule of Mihailo Vojisavljevic who was acknowledged by Pope Gregory VII as Sclavorum Regi - King of the Slavs.
montenegro-adventures.com/montenegro-History-f-60x83

As a matter of fact, every time a noun phrase containing information about the location in the timeline of an historical event related to the *definiendum* is located in the beginning of the sentence, this sentence will fail to meet this restriction, and therefore increase the probability of missing the descriptive information about the corresponding historical event. The reader can also equate this line of reasoning to domain-specific definitions, like those starting: "*In algebra, ...*" and "*In statistics, ...*". Assuredly, the redundancy supplied by some massive collections lessens this adverse effect. On the contrary, verifying only the antecedent noun phrase can eventuate in the assimilation of definitions loosely or closely related to other *definienda* or non-definitions (e.g., opinions and advertisements) into the final output. The next two surrogates are delineative of the former ramification:

Hologram
Ford status: The P2000 Hologram that was displayed at the **1999 North American International Auto Show** was the world's largest full color, full parallax, digital hologram produced to date.
media.ford.com/print_doc.cfm?article_id=2606

Nathan Twining
Nathan Twining, graduate of the **U. S. Naval Academy** was an ensign when he was assigned to the battleship U. S. S. Iowa just prior to the war.
www.uwm.edu/Library/arch/Warletters/spanam/twining.htm

Undoubtedly, the first snippet suggests that this phenomenon can come to pass, whenever a location is being defined. By the same token, manifold events are also directly connected with a particular location during a limited window of time, and accordingly, they can also materialise a shift in the focus or in the topic of the definition. But, still yet, this can happen when tackling other sorts of *definienda* including institutions and diseases.

Topic LM

In general, as noted by [Han et al., 2006], a sentence can be a definition and embody the *definiendum*, but its descriptive content is not necessarily about the *definiendum*. For this reason, [Han et al., 2006] made use of language models for estimating the probability that a given sentence talks about the topic or context of the *definiendum*, this way they attempted to boost the accuracy of the syntactic definition patterns presented in [Han et al., 2005]. The underlying idea behind their strategy is that definition sentences talk about the *definiendum* when they are in one of the contexts of its potential senses. Essentially, these language models are built by weighing evidence taken from (see greater details in section 6.4 on page 146): (a) top-ranked web pages downloaded by submitting the *definiendum*, (b) the respective entry to knowledge bases, and (c) top-ranked documents extracted from the target collection. They concluded that knowledge bases render the most authoritative source of evidence, whereas web pages the most noisy. Simply put, their results show that the evidence extracted from knowledge bases was useful for discarding some definitions in relation to other concepts, because these extracted contexts act as filters of sentences belonging to contexts irrelevant to the various plausible senses of the *definiendum*. It is, by all means, hitherto unknown how this method performs when filtering definitions originated from contexts similar to the *definiendum*. By similarity, it is meant coming from overlapping contexts like the ones sketched by the *definiendum* "*Farouk of Egypt*" and the next web snippet:

The best-laid plans - Haaretz - Israel News
The accord with King **Farouk of Egypt**, the strongest Arab country, which the IDF did not succeed in removing from its strongholds in Palestine (namely, the Gaza Strip), was fragile.
www.haaretz.com/hasen/spages/1054000.html

Language Portability

There are two additional facets that need to be balanced when coming to the decision of which strategy to follow: language portability and the computational resources demanded. Both approaches make use of linguistic processing, this means both systems must reallocate computational resources into parsing or chunking sentences. The technique of [Han et al., 2006] additionally demands the computational resources required for downloading the respective articles from the KBs and for building the language models thereof, as well as the resources needed for rating the candidate sentences. In terms of language portability, both methodologies request linguistic tools dependent on the target language. Ergo, their portability of both strategies is predicated on the portability of the linguistic tools and the inherent properties of the target language. An attractive aspect of the procedure of [Han et al., 2006] is that this can still be applied to other languages by matching language-specific definition patterns at the surface level (see for example tables 3.1 and 3.2). This is without the mandatory use of the linguistic tool that outputs more precise syntactic dependencies in the target language. This is relevant because of the fact that the performance of these linguistic tools widely varies from one language to another. Therefore, in this envisioned scenario, diverting computational resources from parsing and chunking to building language models will be capitalised in terms of a betterment in portability, and in a presumable enhancement with respect to an approach based exclusively on noun phrase analysis. But, as [Han et al., 2006] pointed out, this method will still be, nonetheless, heavily dependent on the coverage of the knowledge bases in the target language, especially on the number of senses. Certainly, this coverage radically changes from one language to another [Figueroa, 2009] (see section 6.5 on 149).

While it is true that the strategy adopted by [Han et al., 2006] has interesting features, things need to be put into perspective. Matching patterns, and thus the *definiendum*, is just one early step in the answering process. Many misleading definitions can still be discarded in posterior steps such as ranking and summarisation. For instance, some approaches, such

as [Figueroa and Atkinson, 2009] (see section 6.5 on page 149), bias the ranking of answer candidates in favour of predominant potential senses. This way they expunge some of the spurious definitions at the expense of a reduction in the diversity of the output. Independently on how these relevant contexts are determined, by means of knowledge bases or web redundancy, both approaches are inclined to discard putative answers that do not significantly match these pre-determined pertinent contexts. This approach, nevertheless, has the advantage that these prominent contexts are inferred from the same set of answer candidates, cushioning the problem of the narrow coverage supplied by the knowledge bases for some *definienda*. On the other hand, it is worth noting that since this method also accounts for linguistic information, namely dependency trees, it is hence less portable to other languages.

OPINIONS AND ADVERTISEMENTS

The second repercussion coming to pass when slackening the pattern alignment, is the assimilation of opinions and advertisements into the final output. Some opinions and advertisements are also very probable to observe some widely used definition patterns. In any event, opinion sentences matching these regularities frequently yield evidence that helps to recognise their opinioned nature. This evidence is in the form of some constructs that can normally be located at the beginning of the sentence. Some of these structures include: "*But I don't believe/think*" and "*But I must say that*". The following web snippets exemplify this:

Daily Kos: Clinton going on Bill O'Reilly
But I don't believe that **ABC news** was founded and operated with the primary goal in mind of Electin Republicans.
www.dailykos.com/story/2008/4/29/123511/859/794/50562

Vendetta Online - Vendetta Message Board
But I don't think **breaking the game** is the way to fix things.
www.endetta-online.com/x/msgboard/3/15317?page=4

Ten Favorite Bands of All Time [Archive] - RockMyMonkey Forums
But I must say that **Agents Of Oblivion** was the closest thing to Acid Bath godliness to come out later.
www.endetta-online.com/x/msgboard/3/15317?page=4

Without a shadow of doubt, a battery of these expressions can be exploited to filter out some opinions. At any rate, this problem is more difficult, because an opinion can still convey factual information, independent from the usage of phrases such as "*But I don't believe/think*". In this case, these clues indicate the disbelief of the author of the actual fact. For instance,

(But) I don't believe that **Linda Lou Taylor** has married 23 times.

Interestingly enough, this obstacle becomes more problematic when opinions presented as actual facts are kept in mind. This is the case of advertisements, lies and some opinions. For example,

Barack Obama is the Antichrist
Barack Obama is the anti-chirst.
www.sodahead.com/blog/69195/barack-obama-is-the-antichrist/

All in all, an efficient ranking methodology should be able to eliminate most of these misleading sentences. Another solution introduced by [Figueroa and Neumann, 2007] resides in capitalising on the *Jaccard Measure* for distinguishing more reliable descriptive sentences. The *Jaccard Measure*, J, of two terms w_i, w_j, is the ratio between the number of distinct *uni-grams* that they share and the total amount of different *uni-grams*:

Pattern	Threshold
$\delta^{'}$ [is|are|has been|have been|was|were] [a|the|an]	0.33
[$\delta^{'}$|$\eta^{'}$], [a|an|the] [$\eta^{'}$|$\delta^{'}$] [,|.]	0.25
$\delta^{'}$ [become|became|becomes] $\eta^{'}$	0.25
$\delta^{'}$ [,|] [which|that|who] $\eta^{'}$	0.25
$\delta^{'}$ [was born] $\eta^{'}$	0.5
[$\delta^{'}$|$\eta^{'}$], or [$\eta^{'}$|$\delta^{'}$]	0.25
[$\delta^{'}$|$\eta^{'}$][,|][|also|is|are] [called|named|nicknamed|known as] [$\eta^{'}$|$\delta^{'}$]	0.25
[$\delta^{'}$|$\eta^{'}$] ([$\eta^{'}$|$\delta^{'}$])	0.25

Table 3.1: Some surface patterns for English (source [Figueroa and Neumann, 2007]).

$$J(w_i, w_j) = \frac{\mid w_i \cap w_j \mid}{\mid w_i \cup w_j \mid} \tag{3.1}$$

Jaccard Measure

In accordance, they compared the subject of the candidate sentence with the *definiendum*. For example, consider the *definiendum* δ^*="*John Kennedy*", which might also be expressed as $\delta_1^{'*}$="*John Fitzgerald Kennedy*" or $\delta_2^{'*}$="*Former US President Kennedy*". The values for $J(\delta^*, \delta_1^{'*})$ and $J(\delta^*, \delta_2^{'*})$ are $\frac{2}{3}$ and $\frac{1}{5}$, respectively. This methodology filters trustworthy descriptive knowledge by means of a pattern specific threshold, avoiding additional purpose-built hand-crafted insertion/deletion rules and ad-hoc linguistic processing.

Table 3.1 shows eight surface definition patterns and their respective thresholds. These values were determined after empirically testing various thresholds from 0.2 to 0.7, and thus manually counting the corresponding number of non-descriptive or spurious selected sentences. Of course, some sentences embracing useful nuggets will be eliminated, but these discarded nuggets can also be found in other retrieved phrases, e.g., "*Former US President Kennedy*" in "*John Fitzgerald Kennedy was a former US President.*" In short, this approach trusts implicitly in the redundancy of the Internet for discovering numerous paraphrases.

Other languages: Spanish

Jaccard Measure: Spanish

Unquestionably, the experimental thresholds utilised by [Figueroa and Neumann, 2007] do not supply a pinpoint accuracy, failing in some of the same aspects than other approaches. Nonetheless, the interesting facets of this strategy are its speed and potential for an easy language-portability as well as its KB-contexts independency. In theory, whenever a definition pattern exists in the target language, porting this procedure involves experimentally setting its respective threshold. In practical terms, there is no drastic change between porting to another language and adding a new definition pattern to a previous existing set (e.g., English).

In order to go over language portability, consider Spanish as target language. To begin with, let us examine the following example that was fetched when searching for "*Hugo Chávez es la*":

Hugo Chávez y las huelgas sindicales y universitarias en Francia y ...
Una de las reformas más importantes de **Hugo Chávez** es la constitución que un Fondo Monetario Latinoamericano, que llaman Banco del Sur y que precisamente en ...
usuarios.lycos.es/euroim/huelgasyEstadistas.htm

Pattern	Threshold
δ' [es|son|fueron|fue|ha sido|han sido][la|lo|el|un|una|uno|unos|unas|las|los]η'	0.33
δ' [,|;] [un|una|uno|la|lo|el|los|las] η' [,|;|.]	0.25
δ' [ha llegado a ser|llego a ser|se transformo|se ha transformado] η'	0.25
δ' [,|] [el cual|la cual|los cuales|quien|que] η'	0.25
δ' [nacio|fue fundado|fue fundada] η'	0.4

Table 3.2: Surface patterns for identifying definitions in Spanish (source [Figueroa and Neumann, 2007]).

This illustrative web snippet communicates information about a reform advocated by *Hugo Chávez*, but not about himself. In this working example, the *Jaccard Measure* between "*Hugo Chávez*" and "*Una de las reformas más importantes de Hugo Chávez*" is $\frac{2}{8} = 0.25$. Therefore, according to the definition patterns for Spanish sketched in table 3.2 and their respective thresholds, this technique filters out this working sentence.

The special advantage of this word overlapping methodology is that it can be applied to different languages indistinctly, which is vitally important in designing multilingual definition QA systems. However, applying this technique to a new language inevitably implies computing new experimental thresholds. Still yet, there are two additional difficulties that arise when applying this strategy to Spanish: (a) the discarded sentences can possibly embrace descriptive information that is not present in the group of sentences seen as dependable, and (b) sentences in Spanish do not necessarily need to carry an explicit subject. In the case of English, the former is tangibly alleviated by the amount of redundancy provided by the Internet (see section 2.7 on page 50). For the purpose of explaining the latter, let us consider the next first four sentences taken from the Wikipedia article regarding "*Genovevo Rivas Guillén*":

TACIT SUBJECT

(1) Gral. **Genovevo Rivas Guillén** (1886-1947) fue un militar y Gobernador provisional de San Luis Potosí mexicano.
(2) Nacio en Rayon, San Luis Potosí, en 1886.
(3) Lucho como maderista desde 1910, bajo las ordenes del Gral. Alberto Carrera Torres.
(4) Durante la Expedición Punitiva se distinguió en la Batalla de Carrizal, que fue un enfrentamiento contra tropas norteamericanas que perseguían Francisco Villa en el año 1924, concediéndosele la condecoración del Valor Heroico.

In this paragraph, sentences 2-4 omit all explicit references to "*Genovevo Rivas*", but they nevertheless put into words factual information about him. Sentence 4 especially serves to highlight the case of the passive "*se*" construction, which is chiefly used in the third person, increasing its probability of being utilised for defining concepts.

The absence of references forces definition QA systems to process the entire paragraph, as a means of determining to whom each sentence refers. While it is arguable that, in the case of biographical sources, the title and the position of the sentences are good features to solve this problem, it is also true that numerous other classes of documents do not observe these patterns, and they still verbalise descriptive information. Taking into account all sorts of documents is particularly important for languages where a small redundancy is provided. For instance, consider the following blog entry:

(1) La persona a quien más admiro es **Ricky Martin**.
(2) Es cantante.

(3) Nació en Puerto Rico en 1971.
(4) A la edad de seis años apareció en anuncios en la televisión.
(5) Fue selecionado para el grupo "Menudo" a los doce años.
(6) Con su primer álbum obtuvo ocho discos de oro en México, Chile, Argentina, Puerto Rico y Estados Unidos.

In this blog entry, the *definiendum* is the direct object of first sentence, and all the posterior sentences talk about this object without being explicitly referenced. Conversely, descriptive sentences in English usually convey an explicit subject, making it easier to disambiguate about who/what they are talking. In the particular case of definitions, the subject can refer to the *definiendum* by means of pronouns (e.g., he, it and her), orthographical variations, aliases, synonyms or constructions such as "*the album*", "*the 1935 film*", "*the place*", "*the French writer*" and "*the president*".

All things considered, recognising implicit subject pronouns is particularly important to maximise the chances of identifying descriptive information from assorted documents as much as possible. This linguistic phenomenon reaffirms the need for a higher level of redundancy, and from an alternative standpoint, it stresses the necessity for deeper linguistic processing at the paragraph level.

Enriching the *Jaccard* Ratio with Definition Knowledge

It is evident clear that redundancy backs definition QA systems in reducing the amount of descriptions that are missed specifically due to pattern mismatching. However, it is not the magic bullet that completely remedies this difficulty. Let us take a closer look at some of the plausible root causes of this obstacle. As aforementioned, sometimes noun phrases are inserted at the beginning of the sentence. These noun phrases are frequently directed at verbalising and contextualising definitions (e.g., the location in the timeline of an event). Take the working example yielded by the following web snippet:

Open Mic: No Money in Retirement | Bleacher Report
At 47 years old, Atlanta Falcons' kicker **Morten Andersen** is the oldest active player in the NFL... The Atlanta Falcons' 20 Greatest Games of All Time ...
bleacherreport.com/articles/40260-open-mic-no-money-in-retirement

In fact, the phrase "*At 47 years old*" will induce a misalignment regardless of the utilisation of a strategy premised on exact chunking matching or on *Jaccard* ratios, and thus the loss of the description "*oldest active player in the NFL*". To be more precise, the *Jaccard* score for this sentence is $\frac{2}{9} = 0.22$, which make it ineligible to be an answer candidate.

DEFINITION MARKERS

A possible way of attenuating this effect is by slackening the threshold restriction whenever there is evidence suggesting that it is trustworthy to do so. This reliable evidence can be discovered in the form of *definition markers*. These markers are phrases that are very likely to come out at the beginning of descriptive sentences (ahead of the *definiendum*), especially within those sentences matching a pre-determined array of definition patterns. The dependability of *definition markers* can be stipulated in terms of whether or not: (a) they preserve the topic established by the *definiendum*; (b) they do not embody pronouns; and (c) they are likely to start descriptive sentences.

More exactly, *definition markers* can be collected by traversing sentences aligning definition patterns across Wikipedia articles, and subsequently, by extracting sequences of words encompassing the beginning of the description until the first comma. The benefit of this surface-motivated approach versus using chunking is that the former makes it possible to process a large-scale corpus faster, while a frequency count along with a threshold can cut

Definition Markers	Definition Markers	Definition Markers
About CD years later,	Fundamentally,	So far,
Also in that year,	In Greek Mythology,	The founder,
At CD years old,	In ancient times,	The stadium,
During the CDth century,	More precisely,	Traditionally,
During the war,	On MONTH CD,	Under the law,
For many people,	Several years later,	Until the CDth Century,

Table 3.3: Highly frequent *definition markers* discovered across WIKIPEDIA abstracts.

off most spurious findings. As a natural advantage, it is also plausible to employ this procedure to recognise markers in languages differ from English (e.g., Spanish).

Table 3.3 underlines some highly frequent *definition markers* harvested from Wikipedia abstracts. In this selection, *definition markers* carrying pronouns and words starting with a capital letter (excluding the first word) were not taken into consideration, and numbers were replaced with the placeholder "CD". Then, whenever any of the gathered markers is found at the beginning of the candidate sentence, the sequence of words or the respective chunk can be removed, and next, the definition QA system proceeds to test the set of definition patterns. In the illustrative surrogate, the *Jaccard* ratio increases from $\frac{2}{9} = 0.22$ to $\frac{2}{5} = 0.4$. This increment turns this sentence into a qualifier for the succeeding steps of the answering process. Lastly, it is worth pointing out that these *definition markers* can also serve as extra evidence of descriptive content when coupled with definition patterns. In other words, they can be utilised as a salient feature for scoring answer candidates.

Another way of diminishing the reliance on redundancy, while at the same time boosting the chances of distinguishing genuine descriptions, is detecting common *definition phrases* that are very likely to antecede the *definiendum*. To aptly illustrate, consider the following web snippet on "*Jagdish Bhagwati*":

Jagdish Bhagwati - IndiaOn.com
Economics professor at Columbia University, **Jagdish Bhagwati** is an expert at trade, WTO and multilateralism.
www.indiaon.com/education/articles/1129-jagdish-bhagwati

DEFINITION PHRASES

This working surrogate does not reach the succeeding steps of the answering process independent from the usage of a strict pattern matching or the *Jaccard* ratio. In particular, this sentence finishes with a *Jaccard* score of $\frac{2}{7} = 0.29$. One of the reasons for its disqualification is that this sentence carries descriptive content anteceding the *definiendum*. More specifically, the mismatch eventuated from the initial *definition phrase*: "*Economic professor*".

In like manner, *definition phrases* can be collected by inspecting the first definition line supplied by Wikipedia abstracts. Lines matching the first definition pattern in table 3.1 are highly probable to convey these *definition phrases*, which can be extracted by trimming the respective description at the first comma or verb. Here, as a means of achieving reliable counts, part-of-speech tagging and a frequency threshold can be utilised. Correspondingly, table 3.4 shows some recurrent definition phrases across Wikipedia first definition sentences.

Consequently, *definition phrases* can be exploited in a similar way to the *definition markers*. In the illustrative sentence, the new value of the *Jaccard* score increases to $\frac{2}{5} = 0.4$, hence qualifying for the posterior steps of the answering process.

Definition Phrases	Definition Phrases	Definition Phrases	Definition Phrases
American actor	U.S. Representative	fourth album	second single
American author	census town	high school	second son
American lawyer	census-designated place	live album	secondary school
American politician	compilation album	multi-use stadium	small town
American writer	county seat	plant pathogen	small village
Economic professor	debut album	political party	state highway
English footballer	eldest son	public high school	studio album
French commune	fictional character	radio station	television station
Italian painter	football club	railway station	trade union
Norwegian politician	founding member	second album	train station

Table 3.4: Highly frequent *definition phrases* extracted from WIKIPEDIA first lines.

3.5 Co-reference & the Next Sentence

Most definition QA systems aim at finding as much descriptive information as possible. In achieving this, they are also compelled to analyse sentences that do not explicitly bear the *definiendum*. The first and natural approach is interpreting as most promising sentence candidates: (1) the ones close to an explicit mention of the *definiendum*, chiefly the subsequent sentences, and (2) sentences containing pronouns. For instance, [Xu et al., 2003] made use of tools for resolving co-references, and accordingly, they took into consideration sentences carrying noun phrases that either match the *definiendum* directly (via string comparison) or indirectly (through co-reference). However, depending on the algorithm, the number of answer candidates, and the length and/or grammaticality of each sentence, this type of linguistic processing can demand a considerable amount of computational resources.

Some definition QA systems, therefore, focus primarily on immediately succeeding sentences. Principally, sentences in which the head word is represented as an anaphora. More crucial, [Han et al., 2004, 2005] observed that an anaphora refers to the *definiendum* if the anaphora is used as a subject and the *definiendum* is also used as a subject in the previous sentence. In particular, [Chen et al., 2006, Figueroa and Neumann, 2007] assumed that if any successive sentence starts with a pronoun, this pronoun is deemed to conform to the *definiendum*. This procedure assists in dealing with web snippets such as:

NEXT SUBJECT

Californians teeming
Adderley was born in Houston but grew up in New York and Los Angeles.
She and her husband, Dallas native Todd Jenkins, were living in LA when they looked at ...
www.statesman.com/news/content/business/stories/statesmanhomes/09/28/0928californians.html

This method will then overwrite the pronoun "*She*" with "*Adderley*", whenever the *definiendum* typed by the user is "*Adderley*". A more aggressive strategy can also predicate on the replacements adopted by [Keselj and Cox, 2004, Abou-Assaleh et al., 2005], that is additionally resolving possessive pronouns, cooperating on linking the *definiendum* with interesting entities or events (see some examples in tables 3.6 and 3.7). This sort of substitution can connect the *definiendum* with some of its plausible hyponomic relations such as author-books and singer-songs [Figueroa and Neumann, 2008]. This class of relation can be of the interest to the user when reading a definition about the corresponding hypernym (an example can be seen in table 3.7). The replacements are accordingly detailed in table 3.5.

SUBSEQUENT PRONOUNS

On the one hand, this approach aids in recognising more descriptive information, in particular some descriptions low in frequency, ergo providing wider coverage and a more di-

Co-reference	Replacement
/ it / / he / / she / / they / /^It / /^He / /^She / /^They /	/ <definiendum> /
/ its / / his / / her / / their / / theirs / /^Its / /^His / /^Her / /^Their / /^Theirs /	/ <definiendum>'s /

Table 3.5: Substitutions utilised for shallow co-reference resolution (source [Keselj and Cox, 2004, Abou-Assaleh et al., 2005]).

verse output to the user. On the other hand, this kind of technique entails incurring higher risks in diminishing accuracy by identifying some misleading nuggets that have a low frequency, and hence they can be hard to validate. Certainly, some systems can rely on the posterior steps of the answering process (i.e., ranking) for lessening the impact of misleading reference resolutions. In the specific case of web snippets, truncations makes the degree of uncertainty higher, making harder, even for human readers, the accurate resolution of pronouns.

Broadly speaking, there are two key aspects that must be kept in mind when resolving co-references. In the first place, definition QA systems do no necessarily need to resolve all co-references existing in a paragraph or a document. Actually, only co-references within descriptive contexts are needed. Resolving only these co-references can effectuate a reduction in the amount of noise in the posterior steps of the answering process. Doing this however, inherently involves a prior recognition of these descriptive contexts, which is not a straightforward task. In the second place, few pronouns referring to other entities or concepts are indispensable. In essence, only those correlating to the *definiendum* or with one of its references within descriptive contexts are necessary. In fact, the same descriptive information can sometimes be discovered in another document with explicit references, but the appropriate analogies are difficult to draw.

TASK-FOCUSED RESOLUTION

A practical way of easily finding propitious descriptive contexts embodying pronouns is by inspecting definitions across Wikipedia abstracts. Here, the main focus of attention will be pronouns within the first chunk of descriptive phrases. The reason to pay attention only to this chunk is two-fold: (a) this co-reference is more likely to point to an entity or concept mentioned in the preceding sentence, while at the same time, (b) it allows the extension of the method used by [Chen et al., 2006, Figueroa and Neumann, 2007] from accounting solely for pronouns located at the first word (subject) to pronouns within first chunks. These templates were manually checked, thereby ensuring these contexts are probable to refer to a *definiendum* stipulated in the previous sentence. It is worth noting that these templates can be utilised along with replacing an initial pronoun with the *definiendum*.

LEARNING NEXT SENTENCE DESCRIPTIVE CONTEXTS

Tables 3.6 and 3.7 exemplify some interesting findings. In these tables, `CD`, `PRP` and `PRP2` denote a number, a pronoun and a possessive pronoun, respectively. The placeholder `NNP` stands for a sequence of words tagged as proper noun. At answering time, whenever speed is a determining factor, these syntactic categories can be discriminated on the ground of regular expressions and a list of lexical items. In this case, `NNP` will correspond to a sequence of words, where each one starts with a capital letter. In a statement, the poor linguistic knowledge needed to match these templates brings about the advantage in speed.

There is an extra, but contributing, factor behind these templates. They are normally utilised for conveying various kinds of descriptions. For instance, templates bearing the placeholder `CD` are more likely to express temporal information. These templates can be viewed as a sister methodology or a potential extension of the approach proposed by [Paşca, 2008]; that is, they can be useful for presenting a timeline of events related to a particular

Template	Example
ABOUT CD **PRP**	James HOGAN was born in the Maury Co. area about 1822, the son of John HOGAN and Elizabeth PAYNE. **About 1845 he** married Elizabeth FRY, who was born about 1825 in TN.
ACTUALLY PRP	Grimbergen is an old rural town, originated in the 8th century. **Actually it** counts 32,746 inhabitants and is situated at exit 7 of the mainway Ring round Brussels.
AFTER PRP2 **FATHER**	Jonas Lie was born in Oslo, the son of a Norwegian civil engineer and an American mother. **After his father** died he went to Paris in 1892 to live with his uncle, the noted...
ALTHOUGH PRP	Cecrops was a mythical king of ancient Athens. **Although he** was not the first king Cecrops was viewed by the Athenian s as their city's ancestor, indeed Athens was sometimes...
AT AGE CD PRP	Robert Taylor was born in Los Alamos, New Mexico and, following the end of WWII, his family moved to rural South Texas. **At age 14 he** decided on a career as a veterinarian and had ...
BEFORE PRP2 **DEATH**	Frances Hamerstrom was an Aldo Leopold graduate student at the University of Wisconsin. **Before her death** (1998) Dr. Hamerstrom was a distinguished ornithologist and writer.
BEFORE PRP	Brian Aldiss was born in East Dereham, Norfolk, England, son of a department store manager. **Before he** became a full time professional writer, he served in Burma and Indonesia ...
BETWEEN CD **AND** CD PRP	Lars Eriksson is a former soccer goalkeeper from Sweden. **Between 1988 and 1995 he** played 17 matches for the Sweden national football team, but was often used as bench cover for ...
BOTH OF PRP2 **PARENTS**	Dr. Krakowiak was born in Poland, but has been in the US since she was 11 years old. **Both of her parents** are chemists and she has "inherited" the passion for science from them.
CD **YEARS LATER** PRP	One of them was Grzegorz Gajdus, who was 13, clocked 3:50:22 and finished 1024th. **23 years later he** was to become a national record holder in the marathon.
CURRENTLY PRP	Ahmose is a former charter member of Vatra Gitana and a student of Europamoon of Wilmington. **Currently she** is the leader of kenansville's tribal troupe The Barefoot Gypsies.

Table 3.6: Examples of template co-reference resolution. (Part I)

definiendum to the user. However, as a means of making this sort of strategy efficient, this will require a special index of the collection that makes it possible to substantially boost the recall of temporary anchored definition snippets or sentences. Another type of templates link the *definiendum* with assorted entities such as parents (e.g.,"*BOTH OF* PRP2 *PARENTS*"), events (e.g.,"*DURING* NNP PRP" and "*AFTER* PRP2 *MARRIAGE*"), add some clarification (e.g.,"*ACTUALLY* PRP" and "*DESPITE* PRP2 *NAME*"), and causation (e.g.,"*BECAUSE* PRP2 *LOCATION*" and "*DUE TO* PRP2 *POPULARITY*").

NEXT SENTENCE: SORTS OF DESCRIPTIVE CONTEXTS

These templates were extracted under the assumption that they are uttered in a descriptive context. Hence, a higher accuracy is achieved when the system already knows that the previous sentence is a definition.

3.6 Conclusions

To sum up, this chapter dealt at length with some of the phenomena emerging when matching words carried by the *definiendum* with document extracts obtained from the target col-

Template	Example
DESPITE PRP2 **NAME**	The African goose is a remarkably massive bird which has a heavy body and a thick neck. **Despite its name**, the African goose is a relative of the Chinese goose; both of them have ...
DURING NNP PRP	RAF Chilbolton was a World War II airfield in England located 4 miles SE of Andover in Hampshire. **During World War II it** was used by the Royal Air Force and the United States Army Air ...
FOR CD **YEARS** PRP	Brent Wilson is a Penn State emeritus professor of art education. **For 33 years he** has studied childrenfs graphic narratives in diverse cultures and has published the results of ...
FROM CD **TO** CD PRP	General Ruben Fulgencio Batista was a Cuban military officer, politician and military leader from 1933 to 1940. **From 1940 to 1944 he** was the president of Cuba.
FROM THERE PRP	Andrew Coyle is a writer/director who has trained as an actor at the National Theater. **From there he** studied film at the Royal Melbourne Institute of ...
IN CD PRP	Attila was born near Budapest in Central Europe to the royal family of Huns. **In 433 he** became king of the Huns and began the process of turning his tribesmen into a powerful ...
IN PRP2	Annie Dillard is a Christian poet and author who lived on a creek in Pennsylvania for a year. **In her** book Pilgrim at Tinker Creek, she wrote about her experience of living in the ...
OFTEN PRP	Bresegard bei Picher is a municipality in the German state of Mecklenburg-Vorpommern. **Often it** is referred to simply as Bresegard.
SINCE CD PRP	Coca Cola GM is the top football division in Greenland and it was created in 1958. **Since 1971 it** is organised by the Football Association of Greenland.
THEN PRP	Chris Roberts was a fighter pilot, a flying instructor and a test pilot. He served a tour with the Arrows. **Then he** was a test pilot at Dunsfold on Harriers, finally becoming a pilot ...
TODAY PRP	Abdul Latif Abdul Hamid was born in 1954 and graduated in Cinema from the Moscow Higher Institute in 1984. **Today he** works at the Damascus General Film Institute.
UPON PRP2 **ARRIVAL**	Yves Brayer was born in Versailles in 1907, but spent most of his childhood in Bourges. **Upon his arrival** in Paris in 1924, he set out immediately...

Table 3.7: Examples of template co-reference resolution. (Part II)

lection. It then dissects some of the reasons why the topic(s) specified by the *definiendum* do not necessarily coincide with the topic(s) verbalised by the document excerpts retrieved by the IR engine. It next presents and contrasts some strategies to cope with these two issues utilised by current definition QA systems. In general, both problems are critical to the performance of definition QA systems, and ultimately, current evaluations have not settled ways to quantitively compare different systems in these facets.

In addition, this chapter discusses some methodologies employed in anaphora resolution by definition QA systems, and it provides insight into the use of evidence mined from Wikipedia for aiding in this task.

Generally speaking, these challenges have to do with recognising and selecting the most promising answer candidates that will be processed in the succeeding steps of the answering pipeline. Since this is an early and intermediate phase in the entire process, definition QA systems are unlikely to divert computational resources from other vital stages to these tasks, entrusting crucial succeeding tasks and redundancy with improving the outcome of these

two challenges.

To be more specific, definition QA systems can efficiently extract descriptive sentences about the *definiendum* by resolving anaphoras focusing primarily on the *definiendum*, that is without full anaphora resolution for all anaphors in all documents. At any rate, it is unclear whether or not the usage of full co-reference resolution will result in a better overall performance at the expense of processing time.

Chapter 4

Heuristics for Definition Question Answering used by TREC systems

> "1) Numbers are tools, not rules. 2) Numbers are symbols for things; the number and the thing are not the same. 3) Skill in manipulating numbers is a talent, not evidence of divine guidance. 4) Like other occult techniques of divination, the statistical method has a private jargon deliberately contrived to obscure its methods from non-practitioners. 5) The product of an arithmetical computation is the answer to an equation; it is not the solution to a problem. 6) Arithmetical proofs of theorems that do not have arithmetical bases prove nothing." (Ashley-Perry Statistical Axioms)
>
> "Definition of Statistics: The science of producing unreliable facts from reliable figures." (Evan Esar).

4.1 Introduction

In the last decade, a myriad of techniques have been developed as a means of discovering answers to definition questions in, predominantly, English. Most of these approaches have been developed in the context of the Question Answering (QA) track of the Text REtrieval Conference (TREC) challenge. In this track, QA systems are encouraged to seek answers across a collection of news documents known as the AQUAINT corpus. Accordingly, this chapter builds on some notable strategies designed by these systems:

- Section 4.2 dissects the relevant aspects regarding the utilisation of definition patterns for discerning answers to definition questions in natural language texts.
- Section 4.3 focuses on the exploitation of Knowledge Bases (KB) for definition QA.
- Section 4.4 investigates into the use of superlatives and numerals as indicators of promising answer candidates to definition questions.
- Section 4.5 lays out attributes that cooperate on differentiating descriptions from other sorts of texts.
- Section 4.6 goes over the benefit provided by triplets consisting of the *definiendum*, named entities and some prominent and characterising nouns and verbs.

- Section 4.7 deals at length with the amalgamation of positive and negative evidence.
- Section 4.8 presents the centroid vector.
- Section 4.9 fleshes out the soft matching approach to aligning definition patterns.
- Section 4.10 details the ranking of definition patterns in conformity with the ground truth given by TREC.
- Section 4.11 specifies a strategy predicated on trigram Language Models (LM).
- Section 4.12 elaborates on the influence of web frequency counts on the rating of putative answers.
- Section 4.13 discusses a technique for extracting definition patterns premised on Part-of-Speech (POS) tags and named entities.
- Section 4.14 stresses the pertinence of the head of noun and verb phrases for distinguishing answer candidates.
- Section 4.15 underscores the importance of predicates and the layout of KBs web pages in the ranking of putative answers.
- Section 4.16 outlines the construction of the profile of the *definiendum*, and the extraction of propositions.
- Section 4.17 digs deeper into the use of BLEU metric and the utilisation of relational databases as a source of answers.

Lastly, section 4.18 abridges and reiterates the key aspects laid out in this chapter, and it brings this to an end.

4.2 Definition Patterns

In their early work on definition QA, [H. Joho and M. Sanderson, 2000] carried out experiments designed to assess the performance of a lightweight definition QA system, which selected the top five and ten sentences rated in concert with a combination of the following three criteria:

1. There are some lexico-syntactic regularities that frequently signify descriptive content. These constructs are typified in the form of patterns. To be more precise, the ranking function of [H. Joho and M. Sanderson, 2000] checks as to whether or not the candidate sentence observes any of the following patterns on the surface level:

   (a) (<description> such | such <description>) as <definiendum>
   ⇒ "*... diseases such as **mad cow***"

   (b) <definiendum> (and | or) other <description>
   ⇒ "***mad cow** and other diseases ...*"

   (c) <description> especially <definiendum>
   ⇒ "*diseases especially **mad cow**...*"

   (d) <description> including <definiendum>
   ⇒ "*diseases including **mad cow**...*"

(e) <definiendum> (<description>) or (<description>) <definiendum>
⇒ "***TB*** *(Tuberculosis) ...*"

(f) <definiendum> (is | was | are | were) (a | an | the) <description>
⇒ "***Gordon Brown*** *is a politician...*"

(g) <definiendum>, (a | an | the) <description>
⇒ "***Gordon Brown***, *the politician,...*"

(h) <definiendum>, which (is | was | are | were) <description>,
⇒ "*That's the* ***Gordon Brown***, *which is pro-nuclear, pro-airport expansion, pro-Iraq war and pro-unlimited capitalism.*"

(i) <definiendum>, <description>, (is | was | are | were)
⇒ "***Gordon Brown***, *British Prime Minister, is preparing to use a speech to the US ...*"

As a matter of fact, [H. Joho and M. Sanderson, 2000] associated each of these clues with an experimental accuracy, since they noticed that some lexico-syntactic constructs are better than others at discovering descriptions. This assessment was conducted on a small training set of sentences embodying the *definiendum*, and table 4.1 accordingly highlights the experimental accuracies that they obtained. Interestingly, all rules are relatively rare in comparison with the number of identified descriptions that do not observe one of their pre-determined regularities. Some clauses, on the one hand, seem to be much more accurate than others, but they, on the other hand, are less frequent in natural language texts. Thus accounting solely for the most precise rules can bring about a detriment to the diversity of the final output, and to the number of answered questions. In conclusion, this difference in accuracy and frequency stresses the necessity for increasing the precision of all patterns, because less frequent and more specific lexico-syntactic clauses are likely to convey similar descriptions. All in all, figures in table 4.1 were employed to assign higher scores to answer candidates matching more reliable rules.

PATTERN ACCURACY

Pattern	Not Relevant	Relevant	Total	Accuracy
No Pattern	6424	872	7296	12.0%
especially	0	0	0	0.0%
<definiendum>, <description>,	89	63	152	41.4%
is a	23	18	41	43.9%
including	20	17	37	45.9%
or other	1	1	2	50.0%
such as	59	59	118	50.0%
acronym	14	23	37	62.2%
and other	9	23	32	71.9%

Table 4.1: Accuracy of definition patterns (source [H. Joho and M. Sanderson, 2000]).

2. In effect, [H. Joho and M. Sanderson, 2000] noticed that if a *definiendum* is outlined in one document, it is then very likely to be described in other documents. Under this assumption, they reasonably determined that some of these descriptions will share some words, hence creating redundancy that can be exploited to recognise trustworthy descriptive knowledge. First, they examined each document that contains the *definiendum*, and took only the first sentence where it was found. Second, the case of the words in these sentences was normalised, stop words were removed, and a stemmer

TERM CO-OCCURRENCE REDUNDANCY

was applied. Third, they kept the twenty most frequent terms. Each candidate answer was later assigned a score based on the amount of these top twenty words. Ergo, those candidates embracing more of these common terms were given a higher score.

3. The ordinal position of the sentence embracing the *definiendum* was incorporated into the ranking. In this order, earlier sentences obtain a higher score.

SENTENCE POSITION

One of the aims of [H. Joho and M. Sanderson, 2000] was to study the effectiveness of applying their rules at the word level. In practical terms, an efficient procedure of this nature would be, in all likelihood, preferable to a method predicated on parsing because of its speed, simplicity, and its potential application to wide-domains and large collections. Eventually, in order to rate a candidate sentence A_i, these three criteria were synthesised as follows:

$$Score(A_i) = 2000 * KPW + 1 * WC + 75 * (500 - SN)$$

In this ranking function, KPW is the key phrase accuracy, while WC denotes co-occurring word count, and SN coheres with the sentence number. This approach takes advantage of the second and third criteria as a fallback, when no pattern matches a candidate sentence. Experimentally, they tested their ranking methodology in a collection of documents that is part of the TREC data, and it comprised 475mb of articles distilled from the *LA Times* between 1989 and 1990.

FIRST MENTION IN NEWS ARTICLES

Interestingly enough, a reinterpretation of this scoring function leads to thinking that the inception of a relevant concept, that is its first mention, in a piece of news usually goes together with a brief description. Additionally, it can also be conceived that, at some point in the news article, the more a concept has been previously mentioned, the less likely that it will be described. This conjecture, of course, does not hold when the concept is the topic of the news article. Simply put, crucial people, events, locations and things taking part in a piece of news are very likely to be defined the first time they are mentioned in the article, this way the news reader can get a clear picture of the reported event. As a logical consequence, these introductory sentences can be rendered as good answer candidates, when the introduced object matches the *definiendum*. However, every rose has its thorn: this type of feature can make the definition QA system less portable to other kinds of collections.

PATTERNS VS. COLLECTION SIZE

In their experiments, fifty *definienda* were utilised as a test set. In substance, each sentence was assumed to be pertinent whenever it aided in understanding more about the respective *definiendum*. Above all, table 4.2 emphasises one interesting finding: the performance of their definition QA system declined as long as the size of the collection was reduced, due to a diminution in the chance of aligning a definition pattern.

Precision at k	100%	75%	50%	25%	10%
1	0.75	0.78	0.69	0.63	0.62
5	0.51	0.51	0.46	0.38	0.32
10	0.42	0.40	0.35	0.28	0.24

Table 4.2: Precision at k versus collection size (source [H. Joho and M. Sanderson, 2000]).

For the best performing method, 90% of the queries had a correct answer in the top five, whereas 94% in the top ten. Further, [H. Joho and M. Sanderson, 2000] dissected their results by juxtaposing the performance accomplished by each separate ranking criterion, and the tree merged. Table 4.3 draws attention to the fact that the integration of ranking criteria outperforms each individual score, and definition patterns showed to be a great contributor to the performance of their system. Furthermore, [H. Joho and M. Sanderson, 2001] tested this

procedure on sentences containing the *definiendum* taken from the top 600 web documents returned by Google. As a result of this extra experiment, their definition QA system sharply boosted its performance; more precisely, it got an answer ranked in the top five for all 96 testing queries. Given these outcomes, they speculated that their system was able to collect more reliable cross document term occurrence statistics, and the size of the corpus raised the probability of discovering an answer candidate that matches a purpose-built definition patterns.

Precision at k	Combined	Key Phrases	Word Co-occurrence	Sentence Number
1	0.76	0.75	0.37	0.25
5	0.57	0.51	0.35	0.27
10	0.46	0.42	0.35	0.27
15	0.42	0.36	0.33	0.24
20	0.38	0.32	0.32	0.23
30	0.32	0.26	0.28	0.22
100	0.17	0.15	0.16	0.15

Table 4.3: Precision at k of each rating strategy (source [H. Joho and M. Sanderson, 2000]).

Contrary to [H. Joho and M. Sanderson, 2000], [Hildebrandt et al., 2004] benefited from an array of clauses that operated at the word and part-of-speech level. They grouped words in consonance with their Part-of-Speech (POS) tags, this way they roughly identify boundaries of noun phrases. They applied eleven rules to the entire AQUAINT corpus as a means of building a searchable database of definitions (see also section 2.1 on page 25). The following is the list[1] of patterns presented by [Hildebrandt et al., 2004]:

POS-BASED PATTERNS

(a) NP_d be NP_n
⇒ "*A **rabona** is a move that involves a player shooting or crossing a ball after bringing his kicking foot from behind his non-kicking foot.*"

(b) NP_d become NP_n
⇒ "***Psammetichus** I became independent only after the end of Assurbanipal's reign (625 BCE), and ruled until 610 BCE to see Assyria disappear, and Babylonia rise to power in Asia.*"

(c) NP_d *v* NP_n
⇒ "***A Corner of the Universe** won a Newbery Honor in 2003.*"

(d) NP_d, NP_n // NP_n, NP_d
⇒ "***Charles Higham**, British archaeologist.*"

(e) NP_n NP_d
⇒ "*MLB outfielder **Tom Brown***"

(f) NP_d (NP_n)
⇒ "***ACM** (Advanced Cruise Missile)*"

(g) NP_d, (also) known as NP_n // NP_n, (also) known as NP_d
⇒ "***ARP spoofing**, also known as ARP poisoning, is a technique used to attack an Ethernet network which may allow an attacker to sniff data frames on a switched local area network (LAN) or stop the traffic altogether (known as a denial of service attack).*"

[1] In this list, // signals an alternative use of the construct. This normally means an inversion or a swap between the *definiendum* and its description.

(h) NP_n, (also) called NP_d
⇒ "*Giovanni Battista Cima, also called* ***Cima da Conegliano*** *(c. 1459 - c. 1517) was an Italian Renaissance painter.*"

(i) NP_d, or NP_n
⇒ "***myopia****, or nearsightedness....*"

(j) NP_n (such as | like) NP_d
⇒ "*...former Confederates such as* ***Thomas Jefferson Foster*** *were denied ...*"

(k) NP_d (which | that) VP_n
⇒ "***Creeping elegance****, which is related to creeping featurism and second-system effect is the tendency of programmers to disproportionately emphasize elegance in software at the expense of other requirements such as functionality, shipping schedule, and usability.*"

In this pattern representation, NP_d and NP_n stand for the noun phrase pertaining to the *definiendum* and the description, respectively. Some remarks on these regularities[2] are: (1) construct (a) enforces a determiner at the beginning of the sentence, this way discarding some spurious matches, (2) the verb "*v*" in rule (c) represents a commonly-used biographical verb gathered from biographies of people, (3) pattern (e) checks whether or not a noun denoting an occupation (e.g., "*actor*", "*skier*" and "*writer*") is the head of NP_n. Table 4.4 underscores the accuracy reported by [Hildebrandt et al., 2004] for each pattern. This accuracy was obtained by means of a test set of 160 definition questions.

Pattern	Accuracy
(a)	0.3537
(b)	0.2500
(c)	0.2609
(d)	0.3040 / 0.6000
(e)	0.6935
(f)	0.3491
(g)	0.8571
(h)	0.5000 / 0.8000
(i)	0.6774
(j)	0.6460
(k)	0.6667

Table 4.4: Accuracy of definition patterns (source [Hildebrandt et al., 2004]).

EXTRA PATTERNS

In a related manner, various additional lexico-syntactic clauses can be found across the literature on definition QA. Take for instance, some of the patterns taken into consideration by [Xu et al., 2005, Cui et al., 2007] in conjunction with regularities previously sketched in table 3.1 on page 65:

(a) <definiendum> is one <description>
⇒ "***Hemochromatosis*** *is one of the most common genetic disorders in the United States.*"

(b) <description>, (a | an) <definiendum>
⇒ "*... in* ***Berlin****, a cosmopolitan city...*"

[2] An extensive study of this rule matching approach can be found in [Fernandes, 2004].

(c) <definiendum> (is | are) (usually | generally | normally)* (being used to | used to | referred to | employed to | defined as | formalized as | described as | concerned with | called) <description>
⇒ "***Lymphangiectasis*** *is defined as a congenital or acquired dilation of the lymphatic vessels.*"

(d) <definiendum> (refer to | refers to | satisfies | satisfy) <description>
⇒ "***Spinal Stenosis*** *refers to a narrowing of the canal surrounding the spinal cord.*"

4.3 Knowledge Bases

Originally, [Wu et al., 2004] implemented a definition module geared towards discovering definitions across the AQUAINT corpus. In detail, this component selected putative answers from this collection of documents, and rated them thereafter in accordance with their similarities to an array of trustworthy definitions. This dependable descriptive knowledge was gathered from several Knowledge Bases (KB) such as www.encyclopedia.com and WordNet glosses.

Since they noticed that KBs differ from each other in several aspects including: coverage, reliability and relevance, this ranking technique makes use of experimental weights w_d for counterbalancing/balancing these intrinsic differences. Consequently, in their model, the score of an answer candidate A_i is determined by the weighted sum of its similarity to its corresponding definition extracted from each knowledge base D_d:

$$Score(A_i) = \sum_{\forall d} w_d * sim(A_i, D_d) \tag{4.1}$$

This similarity score is given by the Term Frequency-Inverse Document Frequency (TF-IDF) with D_d and A_i treated as a bag of words. Here, they employed the TF-IDF formula outlined in [Salton and Buckley, 1988]. Incidentally, the empirical weights w_d were normalised and assigned in agreement with the authoritativeness of the KB[Zhang et al., 2005], and the output of their system comprised top-ranked sentences. In order to test the influence of these experimental weights, they attempted two slight different configurations in the context of the TREC 2004 challenge. These configurations finished with two quite dissimilar average $\mathcal{F}(3)$-Scores: 0.404 and 0.389. This definition QA module acquired, nevertheless, the second best run in this track, in which the best response across all participants accomplished an average $\mathcal{F}(3)$-Score of 0.460 [Voorhees, 2004]. KBS IMPACT

According to [Wu et al., 2004], definitions from KBs share some common pieces of descriptive information, while other pieces can be irrelevant. For this reason, they selected an "*Essential Definition Set*" by means of a purpose-built summarisation algorithm. Answers were then chosen by their resemblance to each definition in this essential set in combination with a threshold. At each selection, the putative answer with the higher similarity is picked until there is no further selection possible (no similarity surpasses the threshold or no answer candidate is left) or a maximum character-length is reached. As a result, the performance of their system declined from 0.404/0.389 to 0.367. ESSENTIAL DEFINITION SET

In the next TREC assessment, [Wu et al., 2005a] extended and enhanced this system by making allowances for several new attributes. First, they classified the *definiendum* into three pre-defined sorts including person, organisation and thing. This classification is performed via heuristics rules, and the obtained type is utilised for deciding which KBs will be mined. Later, the ranking score of putative answers conform to equation 4.1. As a fallback strategy, they rated answer candidates in consonance with a vector compounded of words along with their respective frequencies. Therefore, [Wu et al., 2005a] implicitly postulated that words highly correlated with the *definiendum* in the target corpus are more fundamental *definiendum*-TYPE KB SELECTION TERM CO-OCCURRENCE

to characterise it, analogously to [H. Joho and M. Sanderson, 2000] (see section 4.2). Eventually, this improvement cooperated on obtaining a $\mathcal{F}(3)$-Score of 0.231 in the TREC 2005 challenge.

On a different note, [Cui et al., 2004c] observed that definitional patterns can filter out statistically highly-ranked sentences that do not express a definition. They also noticed that these regularities can bring the definition sentences that are written in certain styles for definitions, but are not statistically significant, into the answer set. In light of this observation, [Wu et al., 2005a] harvested *definiendum*-nugget pairs from the two previous TREC definition QA challenges, in order to generate a set of rules for each of their *definiendum* classes. Subsequently, they formed a training set encompassing the answer sentences pertaining to this training set grouped into their *definiendum*-types. Afterwards, they extracted windows of five words[3] centred in/on the *definiendum*. They next rated these regularities in tandem with their frequency in the training set. Some top-ranked constructs acquired for the type person are listed in table 4.5.

definiendum-TYPE PATTERNS

Structure Pattern	Weight
<*definiendum*>, the	0.094
<*definiendum*>, a	0.042
<*definiendum*>, who	0.030
<*definiendum*>, was a	0.012
known as <*definiendum*> , is	0.012

Table 4.5: Top-ranked definition patterns for the type “Person" (source [Wu et al., 2005a]).

Thus, [Wu et al., 2005a] made use of hard or strict pattern matching for rating an answer candidate A_i, and both scores were accordingly weighted as follows:

$$Score'(A_i) = \sum_{\forall d} 0.7 * Score(A_i) + 0.3 * PatternWeight(A_i)$$

Here, 0.7 and 0.3 are normalised empirical weights, and their definition QA system returned the top-ranked sentences. This enhancement enabled their system to finish with an average $\mathcal{F}(3)$-Score of 0.232 in the TREC 2005 challenge. This result corresponded to the second best response in this track, where the best run achieved an average $\mathcal{F}(3)$-Score of 0.248 [Voorhees and Dang, 2005].

In the succeeding TREC definition track, [Zhou et al., 2006] reformulated their ranking strategy to the following procedure:

$$InitScore(A_i) = 0.8 * DefiniendumScore(A_i) + 0.2 * DocumentScore(A_i) \qquad (4.2)$$

This ranking function weights the evidence regarding the *definiendum* as more instrumental than the information taken from documents. First of all, the evidence concerning the *definiendum* is derived as follows:

EVIDENCE FROM *definiendum*

$$DefiniendumScore(A_i) = 0.3 * \frac{c(w)}{n_w} + 0.3 * \frac{c(p)}{n_p} + 0.4 * \frac{c(e)}{n_e}$$

In this mathematical relation, n_w, n_p, n_e are the number of words, phrases and named entities within the putative answer (A_i), whereas $c(w)$, $c(p)$, $c(e)$ stand for the amount of the

[3] In deed, patterns in table 4.5 show that they can be shorter and not necessarily centred on the *definiendum*, this can be interpreted as the result of some posterior processing, namely trimming.

words, phrases and named entities that are in both A_i and the *definiendum*. Secondly, the evidence coming from the documents is gathered as follows:

EVIDENCE FROM TARGET DOCUMENTS

$$DocumentScore(A_i) = Maxdocw(A_i) * \left(2 - \frac{2 * docn(A_i)}{1 + docn(A_i)^2}\right)$$

Specifically, $docn(A_i)$ is the number of returned documents (e.g., by the Information Retrieval (IR) engine) subsuming the answer candidate A_i, while $Maxdocw(A_i)$ stands for the max score of these documents.

As a means of tackling data sparseness and the underspecification of the *definiendum*, [Zhou et al., 2006] sifted an array of related (expansion) terms T (i.e., words, phrases and entities) from the set of candidate sentences A. To put it more precisely, [Zhou et al., 2006] computed the $Relativity(t_i)$ between a plausible expansion term $t_i \in T$ and the *definiendum* in sympathy with the following equation:

definiendum EXPANSION TERMS

$$Relativity(t_i) = \sum_{\forall A_i \in A} E(t_i, A_i) * InitScore(A_i)$$

Where $E(t_i, A_i)$ is equal to one whenever $t_i \in A_i$, otherwise is zero. Next, [Zhou et al., 2006] chose the top fifteen terms as expansion words, phrases and entities. Subsequently, the relative term score $RWScore(A_i)$ of a candidate sentence A_i is stipulated as below:

$$RWScore(A_i) = \tag{4.3}$$

$$0.3 * \left(\sum_{\forall w_i \in r_w} \frac{Relativity(w_i)}{n_w}\right) + 0.3 * \left(\sum_{\forall p_p \in r_p} \frac{Relativity(p_p)}{n_p}\right) + 0.4 * \left(\sum_{\forall e_e \in r_e} \frac{Relativity(e_e)}{n_e}\right)$$

The similarity between *definiendum* and relative words, phrases and entities are denoted by $Relativity(w_i)$, $Relativity(p_i)$ and $Relativity(e_e)$, respectively. The respective weight values suggest that named entities are more significant to this query expansion technique than words and phrases. The parameters n_w, n_p, n_e cohere with the amount of words, phrases and named entities, respectively, within the answer candidate A_i. All these words, phrases, and entities embracing an amount of r_w relative words, r_p phrases and r_e entities. The final value is a linear combination[4] of the web score (equation 4.1), the initial score (equation 4.2) and the related terms score (equation 4.3).

ENTITIES RELEVANCE

After ranking, redundant sentences are removed and their definition QA system outputs the twenty highest ranked sentences. They tested this new technique in the context of TREC 2006, reaching an average $\mathcal{F}(3)$-Score of 0.222, in contrast to the system utilised in the previous year, which finished in this challenge with an average $\mathcal{F}(3)$-Score 0.159, and an integration of both systems obtained an average score of 0.223. All runs ranked amongst the top systems, especially their best response, was the second best across all participants in this challenge [Dang et al., 2006].

4.4 Definition Markers: Indicators of Potential Descriptive Knowledge

Inherently, the methodology introduced by [Kosseim et al., 2006] rates a candidate sentence A_i as the sum of the weights of the interesting terms T_j it contains.

[4] Apparently, the complete specification of this combination of factors is not meticulously detailed by [Zhou et al., 2006].

$$Score(A_i) = \sum_{\forall T_j \in A_i} Weight(T_j) \tag{4.4}$$

These interesting marking terms are mined from Wikipedia articles related to the *definiendum*. Words with a frequency higher than one are deemed to be specific and crucial to the topic. In essence, [Razmara and Kosseim, 2007] considered as interesting-markers only named entities coinciding with locations, dates, person names and organisations (e.g., *Titanic* ⇒ *White Star Line* and *J.F. Kennedy* ⇒ *Lee Harvey Osward*). They discarded all entities, whenever they obtained more than twenty instances of each type, because they assume this to be an indicator of an article biased towards a particular point of view. These interesting terms are exploited for fetching documents from the target collection, and rated in tandem with their frequency as follows:

INTERESTING MARKERS

$$Weight(T_j) = Log(Frequency(T_j)) \tag{4.5}$$

UNIVERSAL MARKERS

Additionally, [Razmara et al., 2007] distinguished universal markers, which are important terms regardless of the class of *definiendum*. These words were determined empirically by examining TREC data respective to previous challenges. Particularly, they took into consideration two different kinds of universal markers: superlatives and numerical.

SUPERLATIVES

In truth, superlative adjectives and adverbs are critical because of the observation that people are interested in knowing about the best, the first, and the most wonderful more than discovering normal or average facts [Razmara et al., 2007]. Numerals are also very likely to supply useful descriptive information. Good examples are the sentences "*Lufthansa flights to 200 different cities around the world.*" and "*Jesus started his ministry at age 30*".

NUMERALS

As a matter of fact, [Razmara and Kosseim, 2007] verified this hypothesis by finding the percentage of superlatives in "vital", "okay" and "uninteresting" sentences across the TREC 2004 challenge data. They found that superlatives and numbers are more likely to belong to sentences perceived as vital by the assessors.

	Number Of Words	Superlatives (%)	Numerals (%)
vital	49,102	0.52	2.46
okay	56,729	0.44	2.26
irrelevant	2,002,525	0.26	1.68

Table 4.6: Proportion of superlatives and numerals versus sort of sentence (adapted from [Razmara and Kosseim, 2007]).

As a means of accounting for this, the score of sentences embodying numerals and superlatives was raised by 20% for each of their instances they bear. However, numerals that are part of a date expression are excluded as they are already included in the interesting terms acquired from the Wikipedia entry. In the context of the TREC 2006 track, slightly different configurations of this strategy achieved average $\mathcal{F}(3)$-Score values from 0.180 to 0.199, while accomplishing average $\mathcal{F}(3)$-Score values from 0.275 to 0.281 in the TREC 2007 challenge. In both TREC tracks, this definition QA system submitted the third best response. In TREC 2006, the first and the second runs scored 0.250 and 0.233 [Dang et al., 2006], respectively, whereas 0.329 and 0.299 in TREC 2007[Dang et al., 2007].

In sympathy with superlatives and numerals, [Razmara and Kosseim, 2007] also explored interesting terms corresponding to three different classes of *definiendum*: thing, person and organisation. Like superlatives and numerals, they inspected TREC 2004 data for this purpose. They scored each stemmed keyword K_i in consonance with the following formula:

definiendum-TYPE MARKERS

$$Score(K_i) = Frequency(K_i) * Distrib(K_i)^2 \tag{4.6}$$

where $Frequency(K_i)$ is the frequency of the keyword K_i and $Distrib(K_i)$ stands for the number of *definienda* whose sources embrace K_i. The underlying idea here is preferring keywords that are referred to in a higher number of *definienda*. To the detriment of those cues that appear frequently, but only for fewer *definienda*, and ergo the former are viewed as more critical. In order to compare terms carried by uninteresting sentences versus those keywords within interesting sentences, [Razmara and Kosseim, 2007] used the ratio of both $Score(K_i)$ values. In accordance with this ratio, table 4.7 highlights the fifteen highest rated keywords per *definiendum*-type. This table also supports the importance of stressing superlatives. Consequently, [Razmara and Kosseim, 2007] favoured sentences embodying these words by increasing their score by 20% per matching term.

	Markers
all types	found, die, associ, life, begin, publish, first, public, servic, group, death, see, countri, old, most
thing	kind, fall, public, found, countri, offici, field, program, develop, director, begin, discov, particl, power, figur
person	born, servic, serv, become, film, general, old, movi, chairman, place, receiv, begin, win, life, intern
organisation	chang, publish, establish, first, leader, associ, larg, found, releas, project, group, lead, organ, begin, provid

Table 4.7: Top interest-marking keywords versus class of *definiendum* (adapted from [Razmara and Kosseim, 2007]).

The contribution of each sort of marker was assessed by means of the definition question set produced by the TREC 2005 challenge. In their evaluation, [Razmara and Kosseim, 2007] tested the performance of their system using all types of markers, and leaving out one of them at a time.

Markers	$\mathcal{F}(\beta)$-Score
including all types	0.265
excluding superlatives	0.255
disregarding numerals	0.257
ignoring marking keywords	0.266

Table 4.8: Results obtained when dealing with the TREC 2005 definition questions (adapted from [Razmara and Kosseim, 2007]).

The outcomes in table 4.8 show that numeral and superlative markers improved the performance, while keyword markers are inclined to diminish it. The reason for this deprovement is that the TREC 2004 set is too small for training (64 questions solely). Distinctly, the TREC 2005 challenge enriched the *definiendum* types with events, whereas TREC 2004 did not contemplate this kind of *definiendum*. A breakdown of the achieved average $\mathcal{F}(3)$-Score per class is as follows: person 0.300, thing 0.277, organisation 0.268 and event 0.210. Since the TREC 2005 question set was composed of nineteen questions per type, with the exception of only eighteen questions targeted at events, the steep decline in the case of events reaffirms the data sparseness of the the TREC 2004 training material.

Along the same lines, [Kaisser et al., 2006] investigated the nuggets subsumed in the TREC gold standards. To begin with, [Kaisser et al., 2006] created a word frequency list from all nuggets judged as vital in the TREC 2004 and 2005 ground truths. The assumption here is that frequently occurring words serve as importance indicators in answering definition questions. The following table shows the twenty highest frequent words:

Rank	Word	Frequency	Rank	Word	Frequency
1	of	1,262	11	from	215
2	in	1,255	12	**first**	198
3	the	895	13	**largest**	187
4	to	755	14	**million**	181
5	and	498	15	at	175
6	for	37	16	on	174
7	a	411	17	with	164
8	is	341	18	as	157
9	was	265	19	has	153
10	by	254	20	**most**	139

Table 4.9: Twenty highest frequent terms across TREC gold standard (source [Kaisser et al., 2006]).

A key finding of [Kaisser et al., 2006] points to the fact that all but one of the instances of "*most*" assessed as "vital" are part of a superlative construction: *most* + adjective/adverb.

SUPERLATIVES

More crucially, [Kaisser et al., 2006] also discovered that, on average, at least half of the TREC *definienda* have at least one superlative nugget. In essence, 32 out of 69 superlative nuggets were judged as "vital" by more than nine assessors. Principally, their findings are in line with and corroborate the observations of [Razmara and Kosseim, 2007].

Further, [Scheible, 2007] conducted extra studies on the nuggets containing superlatives.

TYPES OF SUPERLATIVES

These studies suggest that distinct sorts of superlatives can be discriminated on the grounds of the relationship between the *definiendum* and the superlative target:

1. both match, e.g., "*AARP*" ↔ "*Largest seniors organization*".
2. or the target is part of the *definiendum*, closely related to, or part of the comparison set, e.g., "*Florence Nightingale*" ↔ "*Medal Nightingale highest international nurses award*".
3. Lastly, the *definiendum* is unrelated or only loosely related to the target of the superlative, e.g., "*Kurds*" ↔ "*Irbil largest city controlled by Kurds*".

Furthermore, [Kaisser et al., 2006] discovered that 46 out of the 69 TREC 2004/05 superlative nuggets fall in the first group, whereas fifteen and eight in the second and third categories, respectively. The distribution of judgements given by the assessors showed that 87% of the 46 nuggets with respect to the first class were interpreted as "vital", while only 59% of the fifteen and 37% of the eight in relation to the second and third groups, respectively. In TREC 2006, sixty nuggets embodying superlatives were perceived "vital" or "okay". What is more, [Kaisser et al., 2006] noticed that these text fragments show a similar distribution: 91% of the first class superlatives were judged as "vital", but solely 54% and 20% of the second and third categories, respectively.

One must bear in mind, however, that superlatives are not the magic bullet, that will solve the definition ranking problem. In reality, the sole presence of a superlative does not make a sentence a definition. Take, for instance, the sentence "*definiendum is the most arrogant*

INSUFFICIENCY OF SUPERLATIVES

man in the world." Additional properties must be, therefore, taken into consideration when rating sentences carrying superlatives.

4.5 Negative Evidence: Markers of Potential Non-Descriptions

Exceptionally, an interesting aspect of the ranking strategy adopted by [Kil et al., 2005] is due to the notion of negative evidence. They conceived negative evidence as those elements that make an answer candidate unlikely to be a genuine answer. That is to say, strong indicators of content different from descriptive knowledge. The deliberate intention behind making allowances for negative evidence is to diminish the ranking of candidate sentences embodying these attributes, or alternatively, simply discarding these putative answers.

PRONOUNS IN SUBJECTS

More specifically, [Kil et al., 2005] considered the presence of a first or second person pronoun in the subject of the sentence, namely "*I*", "*We*" or "*You*" as negative evidence, because these candidate sentences are highly likely to only render a subjective opinion. It is unclear, however, how they validate this, and one should be aware of *definienda* embracing these pronouns when checking. For instance, royal names such as "*King Henry I*" or names of books or songs including Michael Jackson's "*I Wanna Be Where You Are*". Nonetheless, some factual descriptive content can still be put into words in combination with subjective opinions. The next three web snippets about "*US President Barack Obama*" exemplify this ambivalency:

OPINIONS AND FACTS

Obama Wont Be First Black President ...
I think **Barack Obama is the** ONLY Black President. Until the people in this world find out that Barack Obama is not the first Black President ...
www.diversityinc.com/public/1461.cfm

Barack Obama and His Childhood in Indonesia ...
I think **Barack Obama is the** best president ever. Barack Obama RULES and ROCKS!!!! ojitoz says: 3 months ago. barack obama is the best president so far bout ...
hubpages.com/hub/barrack-obama-indonesia

Power for You ...!!!!!!!!!! > Jobs
I believe **US President Barack Obama** will make some efforts to bring back US to normal Situation. In addition to this, the four-week moving average was 6, ...
powerfm904.com/category/jobs/

Therefore, depending on the amount of redundancy and the degree of diversity required in the final output, it is necessary to employ some linguistic processing and alias resolution technique in order to prevent the definition QA system from discarding crucial information all the time.

RANK INCREMENT

In this methodology, all answer candidates start with the same initial score, and the values are thereafter boosted or decreased in accordance with the evidence each candidate offers. The score is augmented whenever sentences fully bear the *definiendum*, otherwise it is abated. Additionally, the ranking value was reinforced whenever the answer candidate observed some widespread syntactic definition patterns (see greater details on definition patterns in section 4.2):

definiendum (is|was|who|which|that)
definiendum(s|es) (are|were|who|which|that)

On the other hand, [Kil et al., 2005] worsened the ranking score of sentences that are too long or short with respect to the allowance of one hundred non-whitespace characters

RANK DECREASE

imposed by the TREC assessment (see section 1.7). Further, they also downrated the ranking of answer candidates which contain unrelated words including "*say*", "*ask*", "*report*", "*If*", "*Unless*", interrogatives, and subjective pronouns. Also, [Kil et al., 2005] attenuated the score of sentences that give the impression of being just an enumeration of nouns such as names and places. These misleading candidates were discerned by comparing the amount of non-trivial words or proper nouns with the number of other sorts of words.

ENUMERATION OF PROPER NOUNS

In TREC 2005, exploiting negative evidence for ranking helped their strategy to accomplish an average $\mathcal{F}(3)$-Score between 0.179 and 0.196, reaching the ninth place. The median of this track was 0.156, while the best run scored 0.248.

Following the same trend, [Schlaefer et al., 2006] discarded document surrogates that are part of enumerations of proper names. Specifically, they noticed that these enumerations are very likely to be a list of stock prices when the *definiendum* is an organisation, while a list of tracks whenever a song or a band is being defined. This filter benefited from the observation that all words were part of a proper name, thus they were capitalised, and they consequently discarded snippets carrying more than their half of non-stop-words capitalised. This ad-hoc filter, however, might be good for the TREC challenge, in which this kind of descriptive information is usually not included in the ground truth, and it hence makes the output noisier and larger, causing a diminishment in the final $\mathcal{F}(3)$-Score. This class of nugget, nevertheless, can still be useful for some users. In particular, a list of albums or tracks is commonly found in the biographical content (e.g., Wikipedia) of an musical artist.

ENUMERATION OF ENTITIES

Furthermore, when the type of *definiendum* is person, [Schlaefer et al., 2006] eliminated sentences that contain both the first and the last name of *definienda* in two distinct named entities. This methodology has difficulties when dealing with long *definienda*, that is, the ones composed of several phrases. These kind of *definienda* has been outlined by [Xu et al., 2005] (see section 3.4 on page 61). Like [Kil et al., 2005], they also reduced the rating score of sentences that do not carry the *definiendum* as a whole, because these sentences can be very noisy, but they can still supply interesting nuggets. In TREC 2006, [Schlaefer et al., 2006] obtained an average $\mathcal{F}(3)$-Score between 0.143 to 0.150, capturing tenth place. The median for this track was 0.125, while the best run reaped an average $\mathcal{F}(3)$-Score of 0.250.

PARTIAL *definiendum* MATCHES

Also, [Schlaefer et al., 2007] additionally enhanced their system by filtering out (a) statements of persons in indirect speech, and (b) direct speech formulations citing the statement of a person. They observed that the latter can still render descriptions, but in general, they are more likely to render opinions than facts. Their definition QA system capitalised on a refinement of the scoring methodology proposed by [Kaisser et al., 2006] (see details in section 4.12), achieving an average $\mathcal{F}(3)$-Score between 0.156 to 0.189 in the TREC 2007 challenge. The median of this track was 0.118, while the best run finished with an average $\mathcal{F}(3)$-Score of 0.329.

INDIRECT AND DIRECT SPEECH

4.6 Triplets Redundancy

Fundamentally, the goal of [Roussinov et al., 2004, 2005] is to build a QA system based largely on redundancy. In the event of definition questions, [Roussinov et al., 2005] took advantage of the redundancy that it is brought forth by multiple co-occurrences of the *definiendum* with named entities at the sentence level. They hypothesised that this redundancy could potentially capture the most interesting attributes of the *definiendum*.

Mainly, they gather statistics about word triplet co-occurrences embodied in the top fifty documents retrieved from the target collection. Afterwards, they extract and rate text surrogates pertaining to the most frequent word triplets.

definiendum & NAMED ENTITIES CO-OCCURRENCE

Firstly, their definition QA system discerns all nouns and verbs along with name entities across these top fifty fetched documents. As to named entities, they made allowances for the following classes: date, locations, organisation, person, and time. They created a list comprising the top ten highest frequent nouns and verbs in addition to all detected name entities.

Secondly, they preserved only those tuples of named entities that at least one of the two elements was a substring of or equal to the entire *definiendum* (e.g., for "*George Walker Bush*," it was enough to contain "*Bush*."). For each *definiendum*, they acquired a list of triplets, where each triplet has two named entities (or frequent nouns) and a verb or a noun that appeared between both. For each triplet on the list, they counted the number of occurrences across the top fifty documents for the *definiendum*, and eventually, triplets with a frequency higher than one were selected. The table below presents two triplets extracted for the target "*OPEC*".

definiendum & NOUN/VERB CO-OCCURRENCE

Frequency Count	Triplet
21	price/[HFNN] - barrel/NN - OPEC/[ORGANIZATION]
11	OPEC/[ORGANIZATION] - world/NN - oil/[HFNN]
7	City/[HFNN] - host/VB - Games/[HFNN]
5	Nagano/[LOCATION] - get/VB - games/[HFNN]

Table 4.10: Sample of word triplets regarding "OPEC" and "1998 Nagano Olympic Games" (source [Roussinov et al., 2005]). In this illustrative samples, HFNN signals a high frequency noun.

On the whole, [Roussinov et al., 2005] theorised that this technique achieved a high recall because critical information is typically repeated across the collection of documents. Overall, their system accomplished an average $\mathcal{F}(3)$-Score of 0.171, which was above the median average (0.156) of the TREC 2005 track. It is worth mentioning that the best run reached an average $\mathcal{F}(3)$-Score of 0.248 [Voorhees and Dang, 2005]. Given this result, one can conclude that redundancy is a key and promising feature for definition question answering.

Contrarily, [Roussinov et al., 2005] observed that their technique is very likely to pick text snippets embracing similar descriptions (normally paraphrases), instigating an increase in redundancy in the output to the user. For instance, their system singled out the next two sentences:

ANALOGIES AND REDUNDANCY

Japan's most famous film director, Akira Kurosawa, died at his home Sunday at the age of 88, Kyodo news agency reported.
Japan's internationally renowned film director Akira Kurosawa died Sunday at age 88.

4.7 Combining Positive and Negative Evidence

An interesting facet of the approach introduced by [Yang et al., 2003] is that their definition QA system takes sentences from documents related to the *definiendum*, and clusters them into two categories afterwards: positive and negative. Members of the former encompassed sentences carrying any part of the *definiendum* and their context, specifically one preceding and succeeding sentence, whereas members of the latter are the remaining sentences within the document. They thereafter discriminate new sentences on the grounds of these two groups.

POSITIVE AND NEGATIVE SAMPLES

One fundamental aspect to bear in mind about this classification method is its underlying assumption of relevance. On the one hand, this technique interprets only sentences that are close to the *definiendum* as members of the positive group. This can work when tackling

POSITIVE SAMPLES

news documents because news articles are normally not biographical in their entirety [5], and central descriptions are more likely to be expressed in contexts that include the *definiendum* (see also the discussion in section 4.2). However, these localised contexts do not necessarily need to supply descriptive content, and they can consequently cause noise in the positive class. On the other hand, the negative set can still have definitions, because some sentences can still elucidate interesting facets of the *definiendum* by means of closely related events or entities. A textbook case is the pilot "*Jim Clark*" who died in a car accident, and given the fact that the description of this accident can delineate some critical facts about his death (e.g., *death date*, *death place*). Furthermore, depending on the learning strategy, the selection procedure of the negative category can also cause noise because descriptions about other concepts might be subsumed in this set, namely concepts of the same type (e.g., persons, locations and organisations) whose descriptions could possibly share similarities as they are involved in the same context (e.g., "*President*" and "*Former President*" or "*Vice President*").

NEGATIVE SAMPLES

LANGUAGE DEPENDENCY

Another aspect to keep in mind is the language dependency of this criterion. Some languages, including Spanish, are more unlikely to explicitly convey the *definiendum* or a co-reference, making it harder to make this categorisation. On the whole, under the principle of Large Numbers, this method supports capturing the most crucial aspects of the *definiendum*, while being heavily reliant on the quality and the coverage of the recognised localised contexts surrounding the *definiendum*.

CORPUS RANK

For the purpose of scoring sentences, [Yang et al., 2003] merged evidence extracted from the Internet and the target collection (i.e., AQUAINT corpus). A candidate sentence A_i is ranked in a TF-IDF-like fashion in sympathy with the relevance of its words. The relevance of a word w is measured by counting the number of sentences containing the word in both classes. On this account, they deemed words that have a high frequency in the positive examples as more essential, while at the same time, they rarely occur in the negative training material. The following is the formula utilised by [Yang et al., 2003]:

$$Rank_{Corpus}(A_i) = log\left(1 + \sum_{w \in A_i} SentenceCount_{Pos}(w) * log\left(1 + \frac{NumberOfNegativeSentences}{SentenceCount_{Neg}(w)}\right)\right)$$

An aspect that makes this function less attractive is its reliance on the coverage and the quality of the contexts supplied by the corpus from which it is calculated. In many cases, the corpus offers limited coverage to draw accurate inferences about the pertinence of each particular word, especially, when the definition QA system is attempting to distinguish descriptions across the same corpus where these estimates are deduced. For this reason, in order to tackle the sparseness of the AQUAINT corpus, [Yang et al., 2003] sifted through additional evidence from the Internet. As a means of downloading documents related to the *definiendum*, they submitted queries composed of terms within sentences belonging to the positive set. These terms were chosen in concert with the next equation:

QUERY EXPANSION

$$Weight_{Expansion}(w) = \frac{SentenceCount(w)}{NumberOfSentences} * log\left(1 + \frac{Correlation(definiendum, w)}{Freq(w) + Freq(definiendum)}\right)$$

The first factor implies the relevance of the word w to the positive class, in other words, how representative of this group it is, and the second factor signifies the strength of the semantic relationship between w and the *definiendum* in congruence with positive examples.

WEB RANK

These two criteria serve as predictors of descriptive content about the *definiendum*, and they are, therefore, applied for ameliorating the recall of definitions within the web snippets

[5] In fact, some news articles can be biographical. In particular, when a famous person passes away.

returned by the search engine. A candidate sentence is then weighed in consonance with the evidence discovered on the Internet as follows:

$$Rank_{Web}(A_i) = \sum_{w \in A_i} log(1 + WebSentenceCount(w)) * log\left(1 + \frac{NumberOfPositiveSentences}{SentenceCount_{Pos}(w)}\right)$$

Eventually, candidate answers, emanated from the AQUAINT corpus, are scored by linearly fusing both rank values:

$$Score(A_i) = \lambda * Rank_{Corpus}(A_i) + (1 - \lambda) * Rank_{Web}(A_i) \tag{4.7}$$

In this rating function, λ is an empirical factor that balances the lexical knowledge discovered on the Internet and the evidence found in the target corpus. In this procedure, the selection strategy of the query expansion terms plays an essential role in focusing the retrieval on the right sense of the *definiendum*, which also highlights its sensitivity to the coverage given by the corpus and the Web.

Later, sentences with a higher score are ranked in the top, and in order to reduce redundancy, a variation of the Maximal Marginal Relevance [Carbonell and Goldstein, 1998, Goldstein et al., 2000] was employed to select a subset of the sentences on this list. Slightly different configurations of their systems achieved average $\mathcal{F}(5)$-Score values from 0.432 to 0.473 in the TREC 2003 challenge. These differences mainly reflect slight variations in the summarisation technique. This definition QA system finished with the second best run in this challenge (the best system scored an average $\mathcal{F}(5)$-Score value of 0.555) [Voorhees, 2003].

REDUNDANCY REMOVAL

4.8 Centroid Vector

In TREC 2004, [Cui et al., 2004b] also amalgamated evidence taken from the Web and the AQUAINT corpus. From the latter, they fetched the top 800 documents retrieved from the collection by querying the *definiendum*, and from the former, they took descriptive content from six KBs by means of pre-defined wrappers. Table 4.11 parallels the coverage provided by each of these KBs for the 64/65 TREC 2004 definition questions (one question was dropped).

Resource Name	Topic Coverage (65 max.)
biography.com	19
S9	15
Wikipedia	63
bartleby.com	37
Google glossaries	25
WordNet glossaries	13

Table 4.11: Web KBs and their TREC 2004 respective coverage (source [Cui et al., 2004b]).

As for ranking, [Cui et al., 2004b] exploited the centroid vector (see also [Chen et al., 2006]), previously exploited by [Xu et al., 2003]. This centroid vector is usually made up of words selected in agreement with the mutual information measure. More precisely, equation 4.8 shows how this weight is specified for a word w:

CENTROID VECTOR

$$Weight_{Centroid}(w) = \frac{log(Correlation(w, definiendum) + 1) * IDF(w)}{log(SentenceCount(w) + 1) + log(SentenceCount(definiendum) + 1)} \quad (4.8)$$

In the above formula, IDF stands for Inverse Document Frequency, and [Cui et al., 2004a,b] made use of the statistics produced by Web Term Document Frequency and Rank site[6] to approximate the IDF values. Posteriorly, words surpassing the average weight score plus a standard deviation are selected as centroid words. This strategy additionally augments the weight of words that also appear within the definitions found across KBs (see for instance section 4.9). The resulting centroid vector is eventually employed for rating candidate sentences by computing the cosine similarity to this vector. The underlying idea behind this similarity metric is modeling the context of the *definiendum* predicating on the Distributional Hypothesis [Harris, 1954, Firth, 1957]. This means they rank putative answers in conformity with the degree in which their respective words characterise the *definiendum*, where the degree of characterisation of each term is grounded on the mutual information measure. In deed, the augmentation of the weights corresponding to centroid words that overlap definitions across KBs tends to enhance the performance because these overlapping words are leaning to yield a better characterisation, and by the same token, to boost the likelihood of recognising descriptions in the target corpus that are very likely to define the most pertinent facets of the *definiendum*. It should not be forgotten, however, that this augmentation approach relies largely on the procedure applied for finding the right articles in these six KBs, and the coverage supplied by these six authoritative resources (see table 4.11). In a special manner, it counts on the fact that there is no great disparity between the senses in the target collection and the senses encountered in the KBs.

TERM CO-OCCURRENCE

At any rate, taking into consideration only metrics grounded solely on word correlations for rating candidate sentences does not ensure pinpoint accuracy. Some misleading and spurious candidate sentences can still obtain a high ranking score, thus being included in the final output. Let us consider the following illustrative case regarding "*British Prime Minister Gordon Brown*":

INSUFFICIENCY OF WORD CORRELATIONS

According to the Iraqi **Prime Minister**'s office, **Gordon Brown** was reluctant to signal the withdrawal of **British** troops.

Even though this sentence carries words such as "*British*", "*Prime*" and "*Minister*", that are very likely to be found across KBs articles regarding this *definiendum*, it does not convey descriptive information about "*Gordon Brown*". Ergo, [Cui et al., 2004b] capitalised on definition patterns for filtering out some unreliable hits. They made use of two kinds of rules: hard and soft definition patterns (explained in sections 4.2 and 4.9, respectively) [Cui et al., 2004a]. Their array of hard patterns consisted of manual rules that are well-know in definition QA systems, such as copulas and appositives (see lists of these patterns in sections 3.4 and 4.2). This hard matching aid in detecting definitions that are missed by their soft matching strategy or centroid vector. Their rating strategy reached average $\mathcal{F}(3)$-Scores between 0.379 to 0.460 in the TREC 2004 challenge, where the variation in values depends on the pre-determined length allowance of the final output [Cui et al., 2004b]. This definition QA system culminated with the best run in this track and all its responses were amongst the best systems [Voorhees, 2004].

SOFT & HARD PATTERNS

[6] http://elib.cs.berkeley.edu/docfreq/

4.9 Soft Pattern Matching

The technique presented in the previous section integrates the application of two distinct kinds of rule matching approaches: hard and soft. According to [Cui et al., 2007], hard patterns, that is, manually constructed lexico-syntactic rules, are too rigid to cover all plausible ways of conveying descriptive content, since sentences exhibit a variety of lexico-syntactic clauses when delineating concepts, in particular regularities bearing the same meaning. That is to say, the implementation of definition QA systems predicated on hard patterns normally requires the manual codification of each lexico-syntactic rule. In all respects, this is an undesirable task as it inherently demands considerable and sustained human efforts, while collecting and coding, as well as extending the array of clues.

Under the observation that strict rules fail to match definition sentences due to some inserted and/or deleted tokens such as adverbs or adjectives, soft patterns treat definition lexico-syntactic constructs as sequences of lexical and syntactic tokens [Cui et al., 2007]. Simply stated, pattern matching can be, therefore, conceived as the probabilistic generation of these test sequences premised on training sequences.

According to [Cui et al., 2004a, 2005, 2007], soft patterns outperform hard patterns because they model these types of language variations probabilistically. Still yet, [Cui et al., 2004b] also realised that soft patterns can miss some definitions detected by aligning hard patterns.

In praxis, [Cui et al., 2004a] abstracted or generalised local contextual regularities from a set of training sentences bearing the *definiendum*. This rule induction process accounts for contexts enriched with POS taggings and chunking information. The initial step of this process consists in overwriting some noun phrases and syntactic categories with placeholders. Table 4.12 details these selective replacements. This selective substitution increases the chance of matching new sentences, generating representative patterns and countering overfitting [Cui et al., 2007].

SELECTIVE SUBSTITUTIONS

Token	Substitution
Any word in the *definiendum*	<Definiendum>
Centroid Words	Syntactic Class
Noun Phrases	NP
Adjective and adverbs	deleted
is, am, are, was, were	BE$
a, an, the	DT$
all numeric values	CD$

Table 4.12: Selective substitutions (source [Cui et al., 2004a]).

Likewise, [Cui et al., 2004a, 2007] recognised chunks in the training sentences, and replaced noun phrases with a placeholder afterwards. This specific overwriting is aimed essentially at similar scenarios that usually do not share the same instance of a noun phrase. Another fundamental aspect of this substitution strategy is replacing of sequences of the same placeholder by one instance only.

WINDOW SIZE

The subsequent step is then deducing the soft patterns. For this purpose, contexts are modelled as windows of $2 * L + 1$ words centred on the placeholder corresponding to the *definiendum*. These text fragments are aligned and counted from the array of training sentences. In [Cui et al., 2004a, 2007], the value of L was set to two, that is these training windows encompassed two terms to the left and to the right of the *definiendum*. The outcome of this process is a vector representing soft definition patterns. In this vector representation, a

VECTOR REPRESENTATION

pattern P_a is denoted as follows:

$$P_a :< slot_{-L}, \ldots, slot_{-1}, < definiendum >, slot_1, \ldots, slot_L >$$

Where each $slot_l$ is a vector of pairs of tokens and their respective probabilities:

$$slot_l :< (token_{l1}, weight_{l1}), \ldots, (token_{lm}, weight_{lm}) >$$

As to the stipulation of a token, [Cui et al., 2004a] interpreted any word, punctuation or syntactic tag in a slot as tokens (see table 4.12). Thus, weights $weight_{lm}$ are stipulated as the conditional probability of a token occurring in a slot:

$$weight_{lm} = Pr(token_{lm} \mid slot_l) = \frac{Freq(token_{lm})}{\sum_{l=1}^{s} Freq(token_{ls})}$$

As a means to neatly illustrate this method, let us consider the following six definitions:

1. In **2004** , <definiendum> **was the** top scoring player (five goals , tied with Ali Karimi) in the Asian Cup 2004 .
2. Initially appointed in **1997** , <definiendum> **is the** first IPC of Ontario to be re-appointed for a second term (until 2009) .
3. Along with fellow CNN reporter Jacki **Schechner** , <definiendum> **is one** of the first " Internet reporters " in mainstream television news .
4. Around 980 to **985** , <definiendum> **wrote a** commentary on the " Calculus " of Victorius of Aquitaine , before the introduction of Arabic numerals , when calculations were often quite complex .
5. On April 1 **1814** , <definiendum> **was awarded** UK patent number 3 , 799 for his steam engine design .
6. Robert Baillie (**known as** <definiendum> ; **c.1634** December 24 , 1684) was a Scottish conspirator implicated in the Rye House Plot against King Charles I I .

REPLACEMENT ORDER

Appropriately, the next six pieces of text underline the five-word contextual fragments along with their respective selective substitutions obtained from the working example, in which it is assumed that replacing noun phrases is preferred to overwriting cardinals. For the sake of simplicity, centroid words were also omitted in the examples. Some dates and the lemma of words, including *wrote* and/or *awarded*, could eventually be part of a plausible centroid vector[7].

1. **NP** , <definiendum> **BE$ DT$**
2. **NP** , <definiendum> **BE$ DT$**
3. **NP** , <definiendum> **BE$ DT$**
4. **NP** , <definiendum> **wrote DT$**
5. **NP** , <definiendum> **BE$ awarded**

[7]To the best of our knowledge, it is unclear the order or hierarchy of these substitutions. A noun phrase, for instance, can consist solely of a year, which can be seen as NP$ or CD$. To exemplify, take the case of "*1859*" in the next definition: "*A tale of Two Cities (1859) is the second historical novel by Charles Dickens*." Different orderings would bring about distinct abstractions, and thus diverse probability models.

6. **known as** <definiendum> ; **c.1634**

Definitely, the identification of tokens such as "*c.1634*" depends on the tokeniser utilised for this task. Ergo, the four slot vectors $slot_l$ are as follows:

- $slot_{-2}$:<(NP , $\frac{5}{6}$), (known , $\frac{1}{6}$)>
- $slot_{-1}$:<(, , $\frac{5}{6}$), (as , $\frac{1}{6}$)>
- $slot_{+1}$:<(BE$, $\frac{4}{6}$), (wrote , $\frac{1}{6}$), (; , $\frac{1}{6}$)>
- $slot_{+2}$:<(DT$, $\frac{4}{6}$), (awarded , $\frac{1}{6}$), (c.1634 , $\frac{1}{6}$)>

It is remarkable that [Cui et al., 2004a] used an empirical factor (0.1) to cushion the bias in favour of syntactic classes and punctuation. Then, the likeness of a candidate sentence A_i to the soft pattern vector P_a is estimated in two steps. In the first place, they calculate the similarity assuming that all slots are independent:

$$Pr(A_i \mid P_a) = \prod_{l=-L}^{L} Pr(token_{lm} \mid slot_l)$$

According to [Cui et al., 2004a], one advantage of using this Naïve Bayes rule is that it can still be determined, although some slots are missing. Here are some illustrative examples (A_1 and A_2) of ranked sentences in accordance with the working models:

1. In the field of intellectual property licensing , **an** <definiendum> **is a** payment made by the licensee to the licensor at the start of the period of licensing (usually immediately upon contract , or on delivery of the property being licensed) which is to be offset against future royalty payments . This test sentence provides the following soft pattern: **, DT$** <definiendum> **BE$ DT$** $\Longrightarrow Pr(A_1 \mid P_a) = \frac{4}{6} * \frac{4}{6} = \frac{16}{36} = 0.44$.
2. In **1985** , <definiendum> **joined Newsday** as a general assignment reporter ; currently , <definiendum> is a staff columnist . This test sentence yields the next two soft patterns: **NP ,** <definiendum> **joined NP** $\Longrightarrow Pr(A_2 \mid P_a) = \frac{5}{6} * \frac{5}{6} = \frac{25}{36} = 0.69$, and **; ,** <definiendum> **BE$ DT$** $\Longrightarrow Pr(A_2 \mid P_a) = \frac{5}{6} * \frac{4}{6} * \frac{4}{6} = \frac{80}{216} = 0.37$.

In the second place, in order to filter out unlikely sequences of tokens, and consequently to ameliorate precision, [Cui et al., 2004a] modelled how likely a sequence of tokens occur in congruence with the underlying soft pattern. This model comprises probabilities independently specified for the right and left context of the *definiendum*. The right context model is given by:

$$Pr(seq_{right} \mid P_a) = P(token_1) * P(token_2 \mid token_1) * \ldots * P(token_L \mid token_{L-1})$$

where $P(token_l \mid token_{l-1})$ is the ratio of the frequency of the bigram $< token_{l-1} token_l >$ to the frequency of the unigram $token_{l-1}$. Accordingly, the probabilities of the left sequences are calculated analogously. The unigrams probabilities $P(token_1)$ and $P(token_{-1})$ are worked out by counting the occurrences of the token to the right and to left of the *definiendum*, respectively. Both contexts are weighed as follows:

$$Weight_{P_a Seq}(A_i, P_a) = 0.3 * Pr(seq_{left} \mid P_a) + 0.7 * Pr(seq_{right} \mid P_a)$$

LEFT & RIGHT CONTEXTS

The weight 0.7 empirically shows that the right context is more influential than the left context. This finding, however, might be directly connected with the English language. Eventually, the weight of a pattern P_a is given by:

$$Weight_{pattern}(A_i, P_a) = \frac{Pr(A_i \mid P_a) * Weight_{P_a Seq}(A_i, P_a)}{fragment\ length}$$

In the formula above, *fragment length* is a normalising factor, and these pattern weights are inferred from an array of training sentences (cf. [Cui et al., 2004a]). The final score of a candidate sentence A_i balances both its centroid based and the soft pattern matching weights:

$$Score(A_i, P_a) = (1 - \delta) * Weight_{centroid}(A_i) + \delta * Weight_{pattern}(A_i, P_a)$$

SEMANTICS VS. SYNTAX

The experimental parameter δ favours either the centroid or the soft pattern weight. In their experiments, [Cui et al., 2004a] set the value of this parameter to 0.6, this way they biased this score in favour of pattern rules. Their outcome showed that the amalgamation of statistics and soft patterns is much more effective than using only methods based on word statistics (e.g., word correlations). As a natural consequence, a combination of semantics and syntactic information is required to properly score candidates to definitions. Further, the empirical value of δ underlines the pertinence of syntactic structures when ranking.

Using Knowledge Bases to enhance the ranking score

CENTROID VECTOR AUGMENTATION

Since semantic evidence plays a pivotal role in rating answer candidates, especially the role of correlated words that typify the *definiendum*, [Cui et al., 2004c] enriched the centroid vector with information emanated from KBs such as WordNet, Wikipedia and biography.com (see a comparison amongst the resources most amply exploited by some definition QA systems in table 2.4 on page 31). In short, the idea behind this enrichment is boosting the rate of putative answers whenever they match the contexts yielded by these three authoritative resources. However, these three KBs occasionally supply narrow coverage or no coverage at all, and therein lies the fact that the contribution of these three resources must be weighted with the evidence of other documents accordingly. In [Cui et al., 2004c], the augmentation of weights in relation to words embodied in texts originated from these three KBs is given by:

$$Weight'_{Centroid}(w) = \begin{cases} Weight_{Centroid}(w) * (1 + log(1 + SF(w))) & \text{if } w \text{ in web snippets;} \\ Weight_{Centroid}(w) * (1 + \gamma) & \text{if } w \text{ in KBs.} \end{cases}$$

It is worth recalling here that these centroid words are identified across the candidate sentences, which are distilled from the target collection, namely the AQUAINT corpus, and $Weight_{Centroid}(w)$ denotes centroid weights in sympathy with equation 4.8. In this formula, $SF(w)$ gives the number of snippets that contain the word w, while γ is a constant factor. As a means of setting the value of this constant, [Cui et al., 2004c] tried values from 0.2 to 1.0 in order to optimise the performance of their system, and its value was eventually set to 0.6 grounded on these preliminary experiments. As a finding, [Cui et al., 2004c] observed that the centroid vector enhances its recall as a result of adding the information coming from the KBs, conversely, the use of web snippets reflected only a minute improvement.

Using syntactic information to enhance the ranking score

As a means to intermix more syntax with their soft pattern matching methodology, [Cui et al., 2005, 2007] took advantage of bigram Language Models (LM) for representing

pattern instances. To be more specific, they made use of the linear interpolation of unigrams and bigrams for modelling the likelihood of bigrams. According to [Cui et al., 2007], the reason for this is two-fold: (1) to smooth probability distributions in order to generate more accurate statistics for unseen data, and (2) to incorporate the conditional probability of individual tokens appearing in specific slots. In this method, unigrams and bigrams are interpolated exactly as follows:

BIGRAM LMS

INTERPOLATION

$$P(t_1 \dots t_L) = P(t_1 \mid S_1) * \prod_{l=2}^{L} (\lambda * P(t_i \mid t_{i-1}) + (1 - \lambda) * P(t_i \mid S_i))$$

In this formula, $P(t_i \mid S_i)$ stands for the conditional probability of token t_i occupying the slot S_i, and λ is a mixture weight that integrates both models: unigrams and bigrams. These models utilise conditional probabilities of unigrams located in each particular slot to represent unigram probabilities, due to the fact that [Cui et al., 2007] stipulated that token positions are instrumental. For instance, a comma always appears in the first slot right of the target in an appositive expression [Cui et al., 2007]:

Dr. <definiendum> , a strong believer in Gandhian principles of non-violence , rural development and self-sacrifice , has shaped the Sarvodaya Movement in ways that forged a significant link between secular principles of development and Buddhist ideals of selflessness and compassion .

PARTIAL MATCHING

The incorporation of individual slot probabilities assists the bigram model in allowing partial matching, which is a characteristic of soft pattern matching [Cui et al., 2007]. Essentially, every time some slots are matched, the bigram model can still yield a high matching score by merging the probabilities corresponding to the matching unigram slots.

NORMALISATION

Since test instances frequently differ in length, the log-likelihood of $P(t_1 \dots t_L)$ is normalised by the number of tokens l of the respective test instance:

$$P_{norm}(t_1 \dots t_L) = \frac{log(P(t_1 \mid S_1))}{l} + \frac{1}{l} * \sum_{i=2}^{L} log(\lambda * P(t_i \mid t_{i-1}) + (1 - \lambda) P(t_i \mid S_i))$$

PROBABILITY ESTIMATES

Subsequently, [Cui et al., 2005, 2007] approximated unigram and bigram probabilities by their maximum likelihood (ML) estimates:

$$P_{ML}(t_i \mid S_i) = \frac{\mid t_i(S_i) \mid}{\sum_k \mid t_k(S_i) \mid}$$

$$P_{ML}(t_i \mid t_{t-1}) = \frac{\mid t_i(S_i) t_{i-1}(S_{i-1}) \mid}{\mid t_i(S_i) \mid}$$

where $t_i(S_i)$ denotes that token t_i appears in slot S_i and $|t|$ coincides with the frequency of the token t. Given the fact that taking into consideration token counts with respect to slot positions effectuates larger data sparseness, [Cui et al., 2005, 2007] made use of Laplace smoothing on unigram probabilities:

$$P(t_i \mid S_i) = \frac{\mid t_i(S_{(i)} \mid + \delta}{\sum_k \mid t_k(S_j) \mid + \delta * \mid N(t) \mid}$$

SYNTACTIC TAGS DISTRIBUTIONS

In this smoothing formula, $|N(t)|$ coheres with the total number of unique tokens in the training data, and $\delta = 2$ is a smoothing constant. Due to the fact that tags normally applied in selective substitutions have a considerable higher frequency in comparison to individual

lexical items, [Cui et al., 2007] made allowances for frequencies of words and general syntactic tags separately. When both are put together, the distribution would be strongly biased towards the syntactic categories, causing distortions of the word distributions. All things considered, the probability of unigrams are consequently estimated in consonance with its own set.

Next, [Cui et al., 2005, 2007] worked out the optimal value of λ=0.3 by the Expectation Maximisation (EM) algorithm [Dempster et al., 1977]:

1. Initialize λ to a random value between 0 and 1, e.g., 0.5.
2. Update λ using:

$$\lambda^{'} = \frac{1}{\mid INS \mid} * \sum_{j=1}^{|INS|} \frac{1}{l_j - 1} * \sum_{i=2}^{l_j} \frac{\lambda * P(t_i^{(j)} \mid t_{i-1}^{(j)})}{\lambda * P(t_i^{(j)} \mid t_{i-1}^{(j)}) + (1-\lambda) * P(t_i^{(j)} \mid S_i^{(j)})} \tag{4.9}$$

3. Repeat 2 until the algorithm converges.

In this equation, $|INS|$ is the number of sentences subsumed in the training/development set.

Making the syntactic Matching more flexible

According to [Cui et al., 2005], the previous approach falls short of coping with gaps. Consider, for instance, the following training sentence outlining "*Akira Sakamoto*":

<definiendum> **is known for** clean and modern design.

PATTERN DERIVATIONS

In effect, the construct "*<definiendum> is known for*" is more likely to be found than some of its derivations. More often than not, there are numerous plausible derivations for a particular regularity. Table 4.13 likens the frequencies provided by Google n-grams for some of the most frequent cases pertaining to "*is known for*". In substance, [Cui et al., 2005, 2007] noticed that the training material usually does not offer enough coverage to develop systems that can potentially match all possible derivations, in particular those originated by inserting and deleting tokens. It is important for definition QA systems to have the flexibility of recognising these sorts of variations, this way they can distinguish more potential answers; and presenting thus a more trustworthy and diverse output to the user. For this reason, [Cui et al., 2005, 2007] coped with this by means of Profile Hidden Markov Models (PHMM), which make it possible to account for deletions and insertions while matching.

INSERTIONS AND DELETIONS

In their strategy, PHMMs comprise a sequence of L, match states M_i; each of them is related to a slot in pattern instances. Each state M_i can emit a token t with a probability $P(t|M_i)$. These tokens belong to the set of tokens within training sentences. Each state is additionally enriched with a deletion state D_i utilised for skipping the respective match state. Insertion states take place after a match or a deletion state and they allow self-loops, this way multiple insertions can occur. The probability of a sequence of tokens t_1,...,t_N generated by moving through the states S_0 (start state),...,S_{L+1} (end state) is as follows:

$$P(t_1, \ldots, t_N \mid S_0, \ldots, S_{L+1}, \mu) = P(S_{L+1} \mid S_L) \prod_{i=1}^{L} P(t_{n(i)} \mid S_i) P(S_i \mid S_{i-1}) \tag{4.10}$$

In this formula, μ stands for the model and $P(t_{n(i)} \mid S_i)$ is equalised to one for all deletion states. For the purpose of ranking candidate sentences, [Cui et al., 2005, 2007] selected the

Pattern Derivations			
is known for	1018523	is especially known for	5614
is additionally known for	42	is internationally known for	13406
is also known for	73843	is mainly known for	3164
is best known for	300966	is more known for	2185
is better known for	15303	is most known for	8583
is chiefly known for	1354	is mostly known for	5319
is nationally known for	5576	is wellknown for its	260
is particularly known for	5849	is widely known for	17879
is primarily known for	6549	is especially well known for	2229
is well known for	236287	is most well known for	6427
is particularly well known for	3641	is very well known for	4039

Table 4.13: Some derivations of the pattern "*is known for*".

most probable state path via equation 4.10 coupled with the Viterbi algorithm. This way they approximate the likelihood of the sequence being given all possible state paths. Incidentally, the probability of the observed sequence $P(t_1, \ldots, t_N \mid S_0, \ldots, S_{L+1}, \mu)$ was estimated by the forward-backward algorithm [Manning and Schütze, 1999].

With respect to the transition and emission probabilities, these were approximated by means of the standard Baum-Welch algorithm [Manning and Schütze, 1999]. Exceptionally, the calculation of the sequence probability conforms to the path with the highest probability determined by the Viterbi procedure during the re-estimation process, and not all possible paths as specified in the traditional Baum-Welch algorithm.

Transition and Emission Probabilities

In PHMMs, probabilities can be automatically induced by making use of an iterative EM algorithm. This procedure can start with random or uniform likelihood estimations. However, the re-estimation process only guarantees local convergence, which in some cases can be the global optima. According to [Cui et al., 2007], definition patterns are diverse and sparse in terms of both lexical tokens and POS tags. Therefore, initialising the EM procedure with random or uniform approximations can bring about a suboptimal model that is unable to discriminate between different sequences. For this reason, together with the fact that [Cui et al., 2007] accounted for a small training set, they worked on the assumption that the most probable state path for a sequence should go through as many match states as possible.

Furthermore, insertion and deletion states add flexibility, but they can adversely impact the generalisation of the underlying definition patterns whenever they obtain a high probability. Consequently, [Cui et al., 2007] smoothed the emission probabilities for each match and insertion state through the maximum likelihood estimate of the emission probabilities.

The initial state transition probability $P(t|I_i)$ for a state was arbitrarily set to $\frac{1}{n}$, where n is the amount of transition links that lead from the state. This way the likelihood of emitting a token from matching states is always higher than from insertion states.

Initial Probabilities

PHMM versus the Bigram Soft Pattern Model

Since both models conceive definition patterns as token sequences, there is an intrinsic relationship between them. More precisely, [Cui et al., 2007] noticed that the Bigram Soft Pattern Model can be seen as the PHMM with one state per token. The difference between them lies in the fact that the bigram model takes into consideration unigram probabilities, while PHMMs make use of emission likelihood estimates for representing the independent probability of a particular token occupying each particular position.

Further, [Cui et al., 2007] also pointed out that PHMMs are more robust in terms of model settings, they need more training data, and have a more complex topology that aggregates token sequence probabilities into state transition probabilities. Furthermore, in linguistic terms, both models make allowance for some shallow syntactic regularities, namely the sequential order of tokens. Since the position of these tokens is usually close to the *definiendum*, it can be alleged that they are very likely to share some context and syntactic relation with the *definiendum*. According to [Cui et al., 2007], this shallow syntactic information is captured by bigram likelihood estimations and the state transition probabilities in the bigram and PHMM model, respectively.

In their experimental settings, [Cui et al., 2005] made use of 761 sentences taken from the TREC-12 data set for training their soft pattern matching models, while they took advantage of the fifty questions submitted in the TREC-13 challenge for testing their models.

As an outcome of their experiments, both models were found to ameliorate the performance of the original technique (cf. [Cui et al., 2005]). The Bigram Soft Patterns and PHMM Soft Patterns improved the $\mathcal{F}(3)$-Score by 7.36% and 5.00%, respectively. Two interesting findings are: (a) PHMM Soft Pattern is more sensitive to the amount of training data than Bigram Soft Pattern, and (b) as long as the training material increased the difference in performance between both models narrowed.

In addition, [Sun et al., 2005] tested the Bigram Pattern Model in the TREC-2005 challenge. Their system produced fourteen definition sentences as a final output, finishing with an average $\mathcal{F}(3)$-Score value of 0.195, which is better than the median average obtained by the 71 submitted runs (0.156). This methodology achieved a score of 0.211 when combined with hard pattern matching in the same challenge. This response positioned this system as the fifth best system (the best run scored 0.248) [Voorhees and Dang, 2005].

4.10 Ranking Patterns in Concert with the TREC Gold Standard

Another way of judging the relevance of definition patterns was adopted by [Wu et al., 2005b]. Their approach learnt and ranked regularities, that are often employed to elucidate different facets of the *definiendum*, in congruence with a training material comprising:

1. An array of *definienda*.
2. A list of sentences carrying answering nuggets is supplied for each *definiendum*.
3. Each nugget is labelled as "*vital*" or "*okay*".

SYNTACTIC PATTERNS

In more detail, they capitalised on the TREC 2004 definition corpus, which consists of 64 definition queries. Their definition QA system parsed answer sentences, and walked through the trees in a bottom-up fashion. In this walking process, they detected the parts of the tree that match the answer nuggets, ergo preserving the syntactic structures conforming these matchings.

GROUND TRUTH SCORING

FEATURES

Subsequently, [Wu et al., 2005b] rated these preserved patterns in accordance with the labels (i.e., "*vital*" or "*okay*") assigned by the TREC 2004 assessors. For examples [Wu et al., 2005b] indicated that this procedure discovers constructs such as: "`NP VP`", "`NP NP`", "`NP PP`" as well. These generalised regularities, nonetheless, can still overmatch, and thus the aligned sentences must be reexamined. As a means of doing this, [Wu et al., 2005b] enriched their syntactic rules with semantic attributes, such as comparative adjectives, digits, topic related verbs and phrases. Next, patterns are re-evaluated, and the constructs that are shown to be more likely to recognise "*vital*" and "*okay*" nuggets are preferred over those that are

more inclined to match irrelevant text fragments. Numerous low ranked rules are expunged at this point. As an outcome of this process, they ended up with 34 rules such as:

VBD PP PP_t PP_d
NP JJS NN NN_t
NP JJS NN NNS_t

When ranking answer candidates, their definition QA system follows the same procedure. As a means of dealing with partial matches, their system rates putative answers in consonance with how well their semantic features align. The final value is hence the product of both the pattern and matching score. In the TREC 2005 challenge, this method helped their definition QA system to occupy the sixth place with an average $\mathcal{F}(3)$-Score between 0.205 and 0.207. It is worth remarking here that the median in this track was 0.156 and the best run obtained an average $\mathcal{F}(3)$-Score of 0.248. In their error analysis, [Wu et al., 2005b] mainly alleged that some deterioration in performance was caused by inexact answers eventuated from inaccuracies of the Named Entity Recogniser (NER) tagger, namely comma-separated names of people.

In the TREC 2007 challenge, [Wu et al., 2007] extended this definition QA system by integrating the analysis of relative words. More precisely, they incorporated highly frequent terms within sixty-word windows around the *definiendum* in conjunction with delineative words collected from the Web.

Analogously to [Wu et al., 2005b], this definition QA system parses and walks through the syntactic structure of the candidate sentences. At each level of the parse tree, the content is evaluated, and a score is assigned in sympathy with the next two equations[8]:

$$S^{'} = \sum \frac{S_{topic} + S_{digit} + S_{rep} + S_{adj}}{4}$$

$$S = \frac{\alpha * S^{'}}{L_{pattern} + 1} + \frac{(1 - \alpha) * 64}{L_{content}}$$

In these formulae, $S^{'}$ is the sum of the score of every syntactic unit at each different level of the parse tree. To make this point clearer, [Wu et al., 2005b] mentioned that the value of $S^{'}$ for the level "`NP JJS NN NNS`" would be the sum of the individual scores for "`NP`", "`JJS`" and "`NN`" as well as "`NNS`". In these equations, $L_{pattern}$ is the number of syntactic units, while $L_{content}$ the amount of words in text snippets, and α is an empirical weighing factor. The remaining parameters of these mathematical relations are not clearly specified in their work.

Eventually, [Wu et al., 2007] singled out the top thirty nuggets for the final output, achieving an average $\mathcal{F}(3)$-Score between 0.216 to 0.242 in the TREC 2007 definition QA subtask, holding sixth place. It is worth noting here that the best response across all participants accomplished an average $\mathcal{F}(3)$-Score of 0.329, while the median was 0.118.

4.11 Combining Trigram Language Models with TF-IDF

In general terms, the definition QA system implemented by [Whittaker et al., 2005] is predicated on a variation of a speech summarisation approach proposed by [Kikuchi et al., 2003]. First, this method discards high frequent words that are unlikely to belong to an answer nugget. Secondly, this system chooses the top 500 sentences carrying between 40 and 220

[8]These formulae are formed of factors that are obscurely explained in [Wu et al., 2007]. It is very interesting, nevertheless, to keep in mind what features they use and how they are intermixed as a means of rating answer candidates.

bytes long. These sentences are then rated in tandem with the amount of topic words they bear. To be more precise, the larger the number of terms (w_i) related to the *definiendum* they embody, the higher their position in the rank is. The scores of these words were based on Inverse Document Frequency (IDF) values acquired from the AQUAINT corpus. Thirdly, they selected up to 175 sentences in agreement with a combination of a linguistic score $L(w_i)$ (trigram language models) and a significance score $I(w_i)$ (measured by a TF-IDF score) as well as the following equation:

TRIGRAM LMS

$$S(A_i) = \frac{1}{N} * \sum_{i=1}^{N} (L(w_i) + \alpha * I(w_i))$$

In this equation, N stands for the number of words in the putative answer A_i. In TREC 2005, this approach assisted in reaching an averaged $\mathcal{F}(3)$-Score between 0.091 to 0.138, while it supported in getting an averaged $\mathcal{F}(3)$-Score between 0.60 to 0.64 in the TREC 2006 challenge[Whittaker et al., 2006]. The best run in TREC 2005 scored 0.248 [Voorhees and Dang, 2005], whereas 0.250 in TREC 2006 [Dang et al., 2006]. In TREC 2007, a slightly different system finished with an averaged $\mathcal{F}(3)$-Score between 0.110 to 0.118 (best 0.329)[Dang et al., 2007]. The difference focussed on the sentence retrieval and selection modules [Whittaker et al., 2007].

4.12 Web Frequency Counts

In TREC 2006, [Kaisser et al., 2006] outputted the best response ($\mathcal{F}(3)$-Score = 0.250) by enhancing their system presented in [Kaisser and Becker, 2004]. The first step in their answering strategy is collecting term frequency counts from the Web. These counts are distilled from the top fifty web snippets returned by a search engine[9]. Like other strategies also do, stop-words are left unconsidered when counting. As a means to illustrate, [Kaisser et al., 2006] listed the frequency count of the words produced for the *definiendum "Warren Moon"*:

148: "moon"
145: "warren"
30: "football"
27: "nfl"
20: "houston" "oilers"
18: "autographed"
11: "quarterback", "jerseys"
10: "hall", "time", "18","throwback"
9: "player", "born", "only", "1956", "pro", "november", "sports"
8: "jersey", "1"
7: "los", "angeles", "team", "authentic", "career", "fame", "free"
...

Candidate sentences are later scored in conformity with the weights of their terms. In other words, they are rated by summing their respective term weights deduced from the Web, and divided by their length in non-white space characters afterwards. They iteratively singled out the highest scored sentences and removed them from the candidate set after their selection. At each iteration, the weights of words belonging to the chosen sentence is divided by five. In this decay method, terms belonging to the *definiendum* are divided by two.

DECAY FACTORS

[9] Actually, [Kaisser et al., 2006] did not deal at great length with their search strategy.

Sentences are then re-rated and this process iterates until the length of all selected answers surpasses an experimental threshold.

	Run I	Run II	Run III
length	1400	850	650
result	0.250	0.229	0.203

Table 4.14: Length of the output versus final score (source [Kaisser et al., 2006]).

INTERESTING MARKERS

An interesting aspect of the third run is that their system took advantage of a re-ranking procedure grounded on *importance indicators* (e.g., superlatives). Intrinsically, this re-scoring method is in the same spirit as the works of [Kosseim et al., 2006, Razmara and Kosseim, 2007] (see section 4.4 for more details). As table 4.14 shows, the performance of this third run dropped. At any rate, this diminution can stem from shorter outputs (a lower threshold was enforced) in conjunction with the fact that the TREC evaluation is biased in favour of larger responses.

TERM CO-OCCURRENCE

To a great extent, this technique relies largely on the Distributional Hypothesis [Harris, 1954, Firth, 1957], this means it finds high frequent terms within the context of the *definiendum* and makes use of these words for rating answer candidates accordingly. Certainly, the degree to which these terms typify the *definiendum* depends heavily on the amount of redundancy downloaded from the Internet. This methodology is thus sensitive to this factor when determining trustworthy web frequency counts. Nevertheless, this strategy mitigates this drawback by solely utilising the words within putative answers taken from the AQUAINT corpus when ranking. Regularly, this ensures that some misleading terms extracted from the Web will be useless, because they will be unlikely to appear within the array of candidate sentences. Put differently, the answer candidates also act as a filter of some spurious words gathered from the Web. Inversely, banking exclusively on these web frequency counts can induce the loss of some good answers that are not covered by these web terms. Conspicuously, but not in every respect, when there is a sharp dissonance between the predominant contexts -or senses- across web snippets and across the target collection.

In TREC 2007, [Schlaefer et al., 2007] achieved an average $\mathcal{F}(3)$-Score between 0.156 to 0.189 (the median of this track was 0.118, and the best run reached a value of 0.329) by making use of this score computation strategy plus some enhancements:

(a) A parameter that signals how the decay factor of a word is devalued after contributing to the score of a selected answer.

(b) In order to lower the bias in favour of longer document surrogates, they normalised the score by utilising the logarithm of the number of terms it contains.

WEIGHTS BALANCE

(c) They capitalised on an online dictionary for acquiring global frequency counts of words. The logarithm of these global counts was applied specifically for normalising raw counts harvested from the Internet, to state it more precisely, for compensating the overweight of common terms, while at the same time, for counteracting the underweight of more specific terms.

Furthermore, it is worth stressing here that [Schlaefer et al., 2007] considered Wikipedia as their primary source of web terms. When this is unsuccessful, they get the top 100 snippets from Google for a pack of generated queries, download the documents, extract all words and count their frequencies.

4.13 Part-of-Speech and Entity Patterns

Basically, [Gaizauskas et al., 2004] continued the trend of TREC definition QA systems and learnt words that co-occur with the *definiendum* across three distinct resources: web pages, Wikipedia and Britannica Encyclopedia articles. They linked each term to a weight in accordance with these co-occurrence frequencies. In a special manner, this list of words was extended by adding normalisations and morphological variations of these terms. To illustrate, [Gaizauskas et al., 2004] listed some of the terms regarding "*Horus*" and "*Crips*":

Definiendum	Terms
Horus	falcon-headed, god, solar, Egyptian, deity, ...
Crips	graffiti, art, gangs, gang, los, angeles, ...

Table 4.15: Sample terms (source [Gaizauskas et al., 2004]).

Subsequently, their definition QA system takes into consideration the next three factors[10] for ranking a candidate sentence A_i:

1. A function $Main(A_i)$ that returns one whenever the *definiendum* exactly matches A_i, while 0.5 if A_i contains an alias, otherwise zero.

2. The sum $Rel_{Term}(A_i)$ of the weights of the terms that overlap with the characterising words harvested from the three external sources.

3. If the answer candidate aligns any of the two kinds of definition patterns: (1) lexical rules, and (2) POS/named entity patterns, then:

   (a) Another score $Def_{Patterns}(A_i)$ regards a boolean value symbolising whether or not the answer candidate observes a lexical definition pattern.

   (b) The next ranking value $POS_{Patterns}(A_i)$ sums the weights of the POS/entity patterns that match the answer candidate.

POS/ENTITY PATTERNS INDUCTION

With regards to POS/entity patterns, they were inferred by exploiting data-sets supplied by prior definition QA tracks, namely TREC 2003. This class of regularity was derived for each type of *definiendum* separatively, and it has two forms: (a) <*definiendum*> X_2 X_3 X_4, or (b) X_1 X_2 X_3 <*definiendum*>. In these pattern structures, the slots X_i can be occupied by a date, POS tag, punctuation mark, or title, whereas <*definiendum*> can also be an alias of the concept being described. In deed, this sort of syntactic construct shows a slight resemblance to the *definiendum*-type oriented regularities depicted in table 4.5. One distinction between both strategies is the overwriting of some tokens with their POS tags. In this respect, the strategy of [Gaizauskas et al., 2004] is akin to the sequences of tokens utilised for learning soft patterns (see section 4.9) and the syntactic patterns induced by the technique presented in section 4.10.

In the first place, sentences embodying the *definiendum* were fetched from the AQUAINT corpus and automatically marked with the *definiendum*, POS information,dates, and titles[11]. Only the top ten ranked sentences make it to the next step. In the second place, co-references

[10] Actually, [Gaizauskas et al., 2004] also consider an additional score $Exclude_{Terms}(A_i)$ aimed specifically at removing the facts presented in previously answered questions pertaining to the same *definiendum*. This section does not deal with this factor because it turns out to be a property too specific to this challenge, and not a determining factor for definition ranking.

[11] For this purpose, they took advantage of GATE tools: http://gate.ac.uk

Definiendum	Descriptions
Horus	Osiris, the god of the underworld, his wife, Isis, the goddess of fertility, and their son, **Horus**, were worshiped by ancient Egyptians.
Crips	to the FlyPlayaWeb site will see the words “C’z Up," a greeting used by the **Crips**, an infamous street gang

Table 4.16: Some outputted examples provided by this system (source [Gaizauskas et al., 2004]).

are resolved, and instances of both sequences introduced earlier are gathered. In the third place, the score associated with each sequence was given by the reciprocal value of their ranking position. Therefore, regularities found in the most crucial sentences obtain a score of 1.0, whereas those found in the least important sentences get a score of 0.1. Lastly, scores for each pattern were summed for each instance.

As a means of singling out answers for producing the final output, each sentence is sorted in descending order by (in order): $Main(A_i)$, $Rel_{Term}(A_i)$, $Def_{Patterns}(A_i)$, $POS_{Patterns}(A_i)$ and in ascending order by $Exclude_{Terms}(A_i)$. Sentences are then aggregated until they met a length allowance. In TREC 2004, their definition QA system finished with an average $\mathcal{F}(3)$-Score between 0.317 to 0.321, securing fourth place. The best system in this track reaped an average $\mathcal{F}(3)$-Score of 0.460. Table 4.16 sketches some nuggets found by this rating procedure.

4.14 Phrases, Head Words and Local Terms Statistics

Notably, [Han et al., 2004] interpreted noun and verb phrases as answer candidates instead of sentences. They picked these putative answers by checking as to whether or not the syntactic trees of their respective sentences observe some regularities. These regularities encompassed noun phrases that: (a) directly modify the *definiendum*, and (b) are used as a complement in copulas. Further, they extracted verb phrases that: (a) a nominative or possessive relative pronoun directly changes the *definiendum*, (b) are particle phrases, and (c) whose head is not a stop verb. Interestingly enough, this conception of stop verbs goes hand in hand with the verbs perceived as negative evidence by [Kil et al., 2005] (see section 4.5). Furthermore, for the purpose of tackling misparsing, [Han et al., 2004] implemented some POS-based heuristics such as attaching isolated determiners, adjectives and prepositions when they are left outside of the boundaries of the succeeding phrase. They also trimmed incomplete noun phrases (i.e., those ending with a conjunction or a relative pronoun).

STOP VERBS

According to [Han et al., 2004], the head word is essentially the gist of each answer candidate. To understand this more clearly, “*player*" is the most important part of the noun phrase “*a tennis player*". For this reason, one of the factors $rdd(A_i)$ they weighted when ranking a putative answer A_i was the frequency of its head word as head word of all sentences fetched from the corpus. This frequency is then divided by the amount of answer candidates of its corresponding type (noun or verb phrase).

HEAD WORDS REDUNDANCY

A second factor $loc(A_i)$ they took into consideration when scoring putative answers was term statistics across the passages retrieved from the AQUAINT collection:

LOCAL STATISTICS

$$loc(A_i) = \frac{\sum_{t_i \in A_i} \frac{sf(t_i)}{max\_sf}}{|A_i|}$$

In this equation, $sf(t_i)$ denotes the number of retrieved sentences carrying the term t_i, max_sf is a normalising factor and stands for the highest value of $sf(t_i)$, and $|A_i|$ is the amount of content words in A_i.

BIOGRAPHICAL WEIGHT

In addition, [Han et al., 2004] calculated a specific weight for *definienda* of type person. In so doing, they capitalised on KBs. The underlying idea behind this is their observation that frequent words across a restricted set of KB articles about people will also be more salient in other biographical definitions regarding people. Ergo, they defined the probability of a term t with respect to an array of training biographical articles of people:

$$P_{person}(t) = \frac{Frequency\_in\_Knowledge\_Bases(t)}{\sum_{\forall t} Frequency\_in\_Knowledge\_Bases(t)}$$

The key and attractive aspect of this method is that, homologously to $P_{person}(t)$, they also stipulate $P_{text}(t)$ as the likelihood of finding the term t in general texts, and the biographical weight $weight(t)$ of a word t then becomes the ratio of $P_{person}(t)$ to $P_{text}(t)$. This term probability ratio assigns higher weights to words embodied more frequently in the encyclopedia, while at the same time, they are embraced a few times in general texts. In TREC 2004, they made use of two different biographical factors for rating an answer candidate, A_i. Both synthesise the previously presented weight and probabilities differently:

$$bio_1(A_i) = \begin{cases} \frac{\sum_{\forall t_i \in A_i} log_2(P_{person}(t_i) * weight(t_i)+1)}{|A_i|} & \text{if } A_i \text{ is a noun phrase;} \\ \frac{\sum_{\forall t_i \in A_i} log_{10}(P_{person}(t_i)+1)}{|A_i|} & \text{if } A_i \text{ is a verb phrase.} \end{cases}$$

$$bio_2(A_i) = \frac{\sum_{\forall t_i \in A_i} P_{person}(t_i)}{|A_i|}$$

These factors are eventually linearly interpolated as follows:

$$Score(A_i) = \alpha * rdd(A_i) + \beta * loc(A_i) + \gamma * bio(A_i)$$

PHRASES REDUNDANCY

Where the empirical parameters α, β and γ must add one. Redundant answers were expunged by verifying the overlap between their words and the semantic classes of their heads. They eliminated the lower-ranked answer whenever the result of the comparison with a higher-scored answer indicated that their term overlap was equal or greater than 70%, or they share the same synset in WordNet together with a word overlap equal or greater than 30%. In TREC 2004, [Han et al., 2004] generated three runs, the first one utilised $bio_1(A_i)$, achieving an average $\mathcal{F}$(3)-Score of 0.246, while the second response took advantage of $bio_2(A_i)$, accomplishing an average $\mathcal{F}$(3)-Score of 0.229, and the third run did not benefit from the biographical term weight. This last configuration reached an average $\mathcal{F}$(3)-Score of 0.247, occupying seventh place. The best response across all participants finished with an average $\mathcal{F}$(3)-Score of 0.460.

4.15 Predicates as Nuggets and Page Layout as Ranking Attribute

Primarily, [Ahn et al., 2004] noticed that most descriptive nuggets can be verbalised with simple predicates, i.e., normally the verb with all its arguments and modifiers. Therefore, their definition QA system additionally balances the structure of predicates when rating answer candidates. For starters, predicates are identified by means of Minipar[12][Lin, 1994]. Secondly, each nugget t_i starts with an initial score $I_{initial}(t_i)$ assigned by the retrieval engine in sympathy with the document rank where the nugget t_i was found. Precisely, they

[12] Minipar is available at webdocs.cs.ualberta.ca/∼lindek/minipar.htm.

gathered these nuggets from the top twenty documents fetched by the retrieval engine from the target (AQUAINT) collection.

Thirdly, similarly to the target collection, [Ahn et al., 2004] harvested descriptive information from a KB article related to the *definiendum*. Most remarkably, these mined nuggets were rated in congruence with heuristics predicated on the layout of the document. As for features, they considered their order of occurrence in the document. Put differently, like [H. Joho and M. Sanderson, 2000] (see section 4.2), [Ahn et al., 2004] made allowances for the proximity to the beginning of the article as more pertinent nuggets are typically expressed immediately. As well as that, they realised that data in tables is usually critical.

PAGE LAYOUT

As a mean to better the accuracy of the initial score given to each nugget emanated from the target collection, [Ahn et al., 2004] made use of two sorts of sentence-level similarity measures: lexical and semantic. The former is determined conforming to the degree of word overlap with the sentences collected from the KB article. In this comparison, they utilised a stemmed version of the sentences, and they also removed stop-words, this way they strengthened the morphological resemblance of the sentences. Thus, [Ahn et al., 2004] applied the *Jaccard* measure [Jaccard, 1912], which aided in calculating the degree of overlap and to normalise over the length of sentences.

Jaccard MEASURE

As to semantic similarity between sentences, two sorts of metrics were employed: (a) the total WordNet distance of words appearing within the sentences[Boni and Manandhar, 2003], and (b) the similarity scores between pairs of words derived from proximities and co-occurrence in large corpora[Lin and Pantel, 2001], and sum the total proximity measure for the words in the two segments. Then, the refined estimates of answer candidates $I_{refined}(t_i)$ are given by the following equation:

WORDNET DISTANCE

WORD PROXIMITY

$$I_{refined}(t_i) = I_{initial}(t_i) * max_j(I(r_j) * sim(t_i, r_j))$$

In this formula, r_j stands for an authoritative nugget from the KB, and the similarity between two nuggets is denoted as $sim(t_i, r_j)$, and the importance of the dependable nugget as $I(r_j)$. Eventually, [Ahn et al., 2004] returned the top ranked nuggets. On a final note, as an outcome to their experiments, they found out that their definition QA system missed many nuggets because they were not subsumed in the top twenty documents retrieved from the collection. Another source of error was the normalised word overlap. Because of its sparsity, the decision about a match was done chiefly on the basis of only one overlapping word, despite the fact that they stemmed sentences prior to comparison.

4.16 Propositions, Relations and *Definiendum* Profile

To begin with, the definition QA system devised by [Xu et al., 2003] differentiates two classes of definition queries: *who* and *what*. They then retrieved the top 1,000 documents from the collection, and separated those sentences that explicitly or implicitly (co-references) bear the *definiendum* afterwards. Posteriorly, [Xu et al., 2003] exploited the following five strategies for extracting putative answers:

1. The first source of answer candidates regards appositive and copula constructions (for examples, see patterns (f) and (g) on the first listing, and rules (a) and (d) on the second listing in section 4.2). These constructs were identified from parse trees by means of simple rules.

COPULA

2. They capitalised on a rough approximation of predicate-argument structures (called *propositions*), which were discerned from the parse tree. They generated a list of propositions that were very likely to be used for delineating an entity (e.g., "<*PERSON*> *was*

PROPOSITIONS

born on <DATE>"). Propositions that matched one of some pre-defined structures were seen as special, while others were interpreted as ordinary. Special propositions focus chiefly on a particular *definiendum* (e.g., person).

3. In a spirit similar to [Harabagiu et al., 2003, Blair-Goldensohn et al., 2003], they implemented more than forty handcrafted structured rules that typically signal descriptions. Basically, these patterns are derived from the parse trees and are aimed essentially at the same structures listed in section 4.2. At any rate, the reader can still check the details on their implementation provided by [Xu et al., 2003] on his/her own.

4. They normalised propositions in agreement with relations that can be found in an ontology. The deliberate intention is grouping propositions that describe the same relationship.

5. As a fallback mechanism, they utilised sentences containing the *definiendum*.

definiendum PROFILE

Subsequently, [Xu et al., 2003] rated the obtained answer candidates in concert with two factors: their class and their likeness to the profile of the *definiendum*. To be more precise, putative answers are put in the following order: appositives and copulas at the top (1), then privileges structured patterns (3), next prioritises special propositions(2), later relations(4), and eventually ordinary propositions (2) and sentences (5). The answer candidates of each kind are then ranked in congruence with their similarity to the profile. This likeness is given in terms of the TF-IDF score and the bag of words representation.

In regard to the profile of the *definiendum*, [Xu et al., 2003] outlined three complementary ways of construction. In the first place, they looked for articles about the *definiendum* across KBs (see table 2.4 in section 2.3). Additionally, they took advantage of Google for mining the Web for descriptive knowledge (see section 2.4 for details). The discovered descriptions were utilised thereafter for forming a centroid vector comprising words and their respective frequencies. This vector was accordingly used as the profile of the *definiendum*.

CENTROID VECTOR FOR PERSONS

Two complementary options were considered as a result of the fact that their definition QA system could not download relevant articles from the KBs for all *definienda*: (a) in the case of *who* queries, or in other words, in the case that the *definiendum* is a person, they learnt the centroid vector from a collection of 17,000 short biographies harvested from www.s9.com, and (b) in the case of *what* questions, they learnt the centroid from all putative answers. The assumption here is that the highest co-occurring words with the *definiendum* are its most typifying words. This is, of course, an assumption that permeates most of the techniques presented in this chapter. There is, nevertheless, one key extra aspect of this fallback strategy which must be cogitated as an alternative way of tackling definition queries. The idea of building the centroid vector with biographies connected with several persons, and rating answer candidates in sympathy with this vector afterwards is interesting.

The underlying idea is that biographies of people share some common features and attributes, or to put it another way, kinds of nuggets (e.g., birthday, birthplace, relevant achievements, and names of parents). Correspondingly, these commonalities can be exploited to recognise descriptive content regarding *definienda* of people, for which KBs do not supply coverage. Another positive advantage of this idea is that it can assist in coping with the data sparseness that characterises strategies that learnt features (terms) from KBs and project them into the candidate sentences afterwards. In particular, this can cooperate on distinguishing those nuggets conveyed with words that do not appear in articles about the *definiendum*, but overlap with several articles about people.

The disadvantage is, however, that this method pre-assumes a kind of *definiendum* before constructing the centroid vector, that is, it needs an extra step that makes this distinction in

the beginning of the answering process. Depending on the output of this initial phase, the system could be dealing with potential senses different from the ones existing in the target collection. This is a critical issue because it is well-know that *definienda* are very likely to bear several senses. For instance, many places or companies are named after people, small companies can also be named after places, some movies, books and songs can easily bear the same name (e.g., "*Ben-Hur*"). In the specific case of the TREC 2003 challenge, due to its size, the target (AQUAINT) corpus does not yield the level of ambiguity that can be found in more massive collections, such as the Internet, which can be aimed at a larger amount of wide-ranging topics. Consequently, this sort of abstraction should be used in conjunction with some efficient potential sense detection strategies when tackling larger collections.

	Run I	Run II	Run III
m	0	0	5
n	5	20	10
$\mathcal{F}(5)$-Score	0.521	0.520	**0.555**

Table 4.17: Results obtained by [Xu et al., 2003] in TREC 2003.

In TREC 2003, [Xu et al., 2003] generated three runs by adjusting some parameters in their definition QA system. Table 4.17 presents the achievements of this system. The length allowance of this system was 4,000 bytes; m coheres with the number of answer candidates accepted ignoring their type (only in terms of the profile); n stands for the amount of sentences and ordinary propositions in the final output. Table 4.18 highlights, more interestingly, a breakdown of the response that achieved the best average $\mathcal{F}(5)$-Score across all participants in the TREC 2003 challenge. Overall, the system shows a good performance, despite the class of *definiendum*.

Type	Number of Questions	$\mathcal{F}(5)$-Score
Who	30	0.577
What	20	0.522
Total	50	0.555

Table 4.18: Breakdown of the results achieved by the best run in TREC 2003 (source [Xu et al., 2003]).

4.17 The Definition Database and the BLEU Metric

In TREC 2005, [Katz et al., 2005] extracted answer nuggets from their pre-compiled definition repository. This database was built from the AQUAINT corpus by means of the strategy adopted by [Fernandes, 2004, Hildebrandt et al., 2004] (see section 2.2 on page 26 for more details). This array of putative answers was also extended by searching for short text fragments across the target collection that overlap with the definition of the *definiendum* in the Merriam-Webster dictionary. They also considered the top-ranked sentences fetched from the corpus as answer candidates.

For the purpose of rating answer candidates, [Katz et al., 2005] weighted the outcome of two scores: topic accuracy, and the product of the IDF and the term frequency of non-topic terms within putative answers. The latter factor follows the idea of local statistics exploited by manifold definition QA systems including [H. Joho and M. Sanderson, 2000, 2001,

LOCAL STATISTICS

Han et al., 2004] (see sections 4.2 and 4.14). Above all, topic accuracy is stipulated as an $\mathcal{F}$-measure based on word-based precision and recall of the topic. Their definition QA system singled out at most the top 24 sentences. This method assisted them in reaping an average $\mathcal{F}(3)$-Score of 0.1557.

The novel aspect of their second and third runs is not the harvest of the first paragraph of the best Wikipedia article about the *definiendum*, but the fact that they computed the likeness between a putative answer and this paragraph in congruence with the BLEU metric [Papineni et al., 2002]. This enrichment helped their system to reach an average $\mathcal{F}(3)$-Score of 0.1606. Eventually, the third response extends the second by accounting for anaphora resolution. This last run accomplished an average $\mathcal{F}(3)$-Score of 0.1602.

BLEU SIMILARITY

As a means of taking part in the TREC 2007 QA subtask, [Katz et al., 2007] sought to substantially enhance their definition QA system [Katz et al., 2005, 2006]. An interesting aspect of this system is that it makes allowances for passages and sentences gathered from the collection as answer candidates. Like their old system, this also interprets hits from their definition repository as putative answers. In achieving this, [Katz et al., 2007] enlarged their definition database by approximately 1.25 million definitional snippets pre-extracted from the AQUAINT2 corpus.

PARAGRAPHS AS ANSWERS

As for ranking answer candidates, [Katz et al., 2007] benefited from the next four[13] attributes:

definiendum MATCHING

1. $\mathcal{F}_{topic}$ captures the overlap between the *definiendum* (or synonyms) and the words within the response in accordance with $\mathcal{F}(P, R, \beta = 2)$:

   Q = the set of unique terms in the *definiendum* (or an alias).
   R = the words in the best named entity and keyword matches in the response.
   M = the exact matching set between R and Q.

   The precision P and recall R are then given by:

$$P = \frac{\sum_{w \in M} \mathrm{IDF}(w)}{\sum_{w \in R} \mathrm{IDF}(w)}$$

$$R = \frac{\sum_{w \in M} \mathrm{IDF}(w)}{\sum_{w \in Q} \mathrm{IDF}(w)}$$

   At large, $\beta = 2$ signifies that recall is more critical than precision. As a matter of fact, this strategy can be conceived as a plausible alternative to the *Jaccard* Measure explicated in section 3.4 on page 61. In practical terms, it separately models this matching from two standpoints: the *definiendum* and the sentence, contrary to the *Jaccard Measure*, which computes a sole global score.

   Both measures remark the need for considering the accuracy of matching the *definiendum* as an indicator of the likelihood or fitness of an answer candidate to be a genuine answer. On the other hand, this score assigns a weigh to each word, whereas the *Jaccard Measure* interprets each word as equally weighted. Although the contribution of some terms is nullified by the enrichment elucidated in section 3.4 on 67, and the *Jaccard Measure* measure is also directed at verifying whether or not there is a shift in the topic of a sentence that observes some regularities. Lastly, it is unclear why [Katz et al., 2005]

[13] In effect, they incorporated an extra attribute in order to cope with specialities of the TREC challenge.

amalgamated both recall and precision into one score, since they could have capitalised on the three factors separately as features.

2. F_{inform} approximates the "*informativeness*" of a particular response R by contrasting their combined IDF score with the corpus average IDF, IDF_{avg}:

$$F_{inform} = \frac{\sum_{w \in R} \text{IDF}(w)}{\text{IDF}_{avg} \mid w \in R \mid}$$

The idea behind this ingredient is detecting responses (e.g., sentences, chunks or paragraphs) embracing words that have some special distribution in the corpus. By special, it is meant terms that differ from the behaviour of average words within the collection. This factor bears some resemblance to the general text probabilities put into use by [Han et al., 2004] (see section 4.14 on page 105).

3. F_{source} is a list of fifteen plausible sources, e.g., appositive pattern, and whether it is a paragraph or sentence.

4. $F_{projection}$ quantifies the n-gram distributional similarity grounded on the BLEU metric [Papineni et al., 2002], between a response and an array of sentences extracted from articles regarding the *definiendum* across KBs (see table 2.4 on page 31).

These properties are fused into a score function that returns a boolean value. This function was trained on the basis of the TREC 2006 data and four distinct binary classifiers: Support Vector Machine (SVM), logistic regression, radial basis function, and decision trees. These outcomes were synthesised by a logistic function ranging between zero and one. They additionally used their TREC 2006 system to break ties.

In the TREC 2007, they submitted two runs. Both responses harvested Wikipedia and Google Timeline as KBs. The difference between both runs is that the second takes advantage of the new scoring function, while the first does not. The first response finished with an average $\mathcal{F}(3)$-Score of 0.198, and the second run 0.235, securing the seventh place. In this track, the best run scored 0.329 and the median of all systems was 0.118.

In addition, [Katz et al., 2007] carried out experiments to test several classifiers and KBs. As a finding, they realised that using Wikipedia alone yields a larger betterment of the performance than making use solely of Google Timeline. They noticed that Google Timeline occasionally mixes references to various items with the same name, whereas Wikipedia articles are either right on target or very noisy, which makes it possible for the projection phase to filter out all inconsistencies.

4.18 Conclusions

To recapitulate, this chapter describes the most interesting aspects of distinct ranking functions implemented by numerous definition QA systems in the TREC challenge. One of the strategies that transpires most systems is benefiting from definition patterns, which are regularities commonly found across sentences conveying descriptive content. These constructs are typically a couple of tokens preceding and/or succeeding the *definiendum*. Definition patterns can be both independent or dependent on the kind of *definiendum*, and an instance of a pattern can slightly vary. These variations imply the insertions and/or deletions of tokens, and they can be modelled by means of soft patterns. Categorically, definition patterns can overmatch, since these clauses are not solely used for definitions, but also for expressing opinions, advertising, and writing general texts. Broadly speaking, more accurate rules have

low frequency, or are used for outlining a specific class of nugget, whereas less precise cues supply a wider diversity of descriptions and they can hence have a high frequency.

In order to boost the accuracy of the recognition of descriptive phrases observing -and also not observing- definition patterns, QA systems take advantage of linguistic processing such POS tagging, syntactic parsing and morphological analysis. For instance, several systems profit from selective substitutions as a method for abstracting and deducing as well as aligning clues. At any rate, none of these tools have been the panacea for this classification problem, so far.

Another prevalent technique is predicated on learning specific regularities across articles about the *definiendum* gathered from KBs. These regularities normally include lexical items and entities. More often than not, this kind of definition QA system is capable of detecting descriptive phrases that have a significant word overlap with the respective KB articles, while at the same time, they tend to miss many essential nuggets that have few or no term overlap with these articles. In addition, these systems single out many misleading sentences for the final output. Two ways of coping with this obstacle include: (a) enriching this projection with syntactic information, and (b) growing the number of KBs exploited by this class of QA system. However, this sort of strategy starts with the wrong assumption that the context/senses are ruled by the KBs, not the target corpus or sentence, and benefiting from descriptions obtained from a larger amount of KBs has not shown to be the final solution.

As for features, definition QA systems account for diverse attributes. In the first place, the first mention of an entity (potential *definiendum*) in a piece of news is likely to be accompanied by an introductory description. In the second place, superlative adverbs and adjectives are instrumental, but not unmistakable signs of essential characteristics. In the third place, there are also negative indicators of definitions: some verbs and enumeration of named entities. In fourth place, the layout of pages, the ranking assigned by the IR engine and the length of the answer candidate are utilised as features. Lastly, lexical items are the most widely used property.

With respect to machine learning approaches, several methodologies have been tried including SVM and loglinear models. However, the vital issue is the acquisition of negative samples. On the one hand, massive and reliable positive samples are distilled from KBs; it is hard to obtain, on the other hand, a large-scale and balanced negative set of training examples without involving manual annotations. Therefore, strategies that learn from positive examples, such as LMs, are naturally preferred.

By and large, definition QA systems acquire highly correlated words with the *definiendum* as the most prominent and discriminative features. Ergo, these systems require a considerable amount of contexts carrying *definiendum* in order to infer these characterising terms. In so doing, systems take advantage of extra contexts fetched from the Internet. In many cases, simple frequency counts learnt from web snippets have shown to be a decisive factor in the enhancement of definition QA systems.

On a final note, systems typically rate sentences as putative answers in concert with an array of pre-defined features (i.e., lexical items). Conventionally, definition QA systems inspect the set of answer candidates as a means to derive some predominant values for some attributes, and accordingly rank putative answers afterwards. However, for the purpose of dealing with the inherent data sparseness eventuated from this array of sentences, systems boost the ranking score of candidates that bear some similarity to articles about the *definiendum* taken from KBs. This boosting can be done in terms of augmenting the weights of words collected from the set of answer candidates, or by straightforwardly computing the resemblance of each putative answer to the articles harvested from KBs. In a special manner, the most widespread ranking method is the centroid vector.

In closing, this chapter dealt at greater length with assorted approaches to discover an-

swers to definition questions. Chiefly, methods designed in the context of the TREC challenge, that are therefore directed at collections of news documents, namely the AQUAINT corpus. On the whole, these definition QA systems in conjunction with the results accomplished in the different versions of this challenge show that the phenomenon behind this task is still scientifically unexplained. For this reason, along with the fact that definition questions vastly requested by Internet users, it remains an interesting research area.

Chapter 5

Extracting Answers to Multilingual Definition Questions from Web Snippets

> "Thousands of people were producing new Web sites every day. We were just trying to take all that stuff and organize it to make it useful." (David Filo)
>
> "The beginning of knowledge is the discovery of something we do not understand." (Frank Herbert)

5.1 Introduction

Generally speaking, the definition Question Answering (QA) systems boiled down in chapter 4 geared towards discovering answers across the AQUAINT corpus, which is the official corpus of the QA subtask of the Text REtrieval Conference (TREC) challenge. Conspicuously, the trend of these systems reflects the following prominent characteristics:

- The goal of these systems is discerning answers across news articles in English. Evidently, this inherently means a large reliance on the lay-out and language as well as syntactic structures of this kind of document.
- Under the framework of evaluation specified by TREC, it is an open question which modules (strategies) are necessary to revisit in order to develop definition QA systems capable of dealing with queries in several languages. The motivation and hope here is the creation of algorithms that can port, at least, to languages akin to English.
- By and large, most of the best procedures harvest articles about the *definiendum* across Knowledge Bases (KB) as a source of descriptive knowledge, which is projected into the candidate set of answers afterwards. The overall performance of systems that capitalise on this technique normally fell into a step decline when these resources offer narrow coverage. It is thus encouraging to seek ways of widening this coverage and benefiting from other types of sources such as web snippets.
- An essential characteristic transpiring most of the best systems is the utilisation of definition patterns. There are assorted approaches to employ these rules; they differentiate in their strictness, level of abstraction and linguistic knowledge (e.g., parse trees, Part-of-Speech (POS) tagging).

All things considered, it is especially interesting to design methodologies that can cope with several languages under the assessment framework stipulated by TREC (see section 1.7 on page 15). To be more specific, the study of the key components of a definition QA system that are required to be language dependent so that it enhances its performance, and which modules can be transparent to the language. A study in this direction would also disclose diverse advantages and disadvantages pertaining to each language, and expectedly, it would also help to begin to see the light at the end of the tunnel.

Another desirable property of definition QA system is less reliance on the coverage given by Knowledge Bases (KB) to each particular *definiendum*. Particularly, when contemplating target languages different from English, as this coverage considerably narrows for other languages. This excessive reliance can be alleviated, or hopefully eliminated, by understanding the linguistic phenomena that typifies definitions in general. It is, thus, the hope to be able to learn these characteristics and apply them to recognise further descriptions afterwards.

This chapter introduces two different strategies for multilingual definition QA directed at web snippets. A practical use of a system operating on web snippets can be its exploitation as a supplementary or fallback procedure whenever a TREC system fails to find descriptions across KBs, or simply, it might be used to straightforwardly query the Web for descriptive information.

This chapter is organised as follows: The next section outlines methods that are predicated upon redundancy for distinguishing the most statistically relevant answers in English and Spanish, and section 5.3 expands on the utilisation of multilingual KBs for rating answer candidates in both languages, in a way that it is independent from the *definiendum* entered by the user.

5.2 Multi-Linguality, Web Snippets and Redundancy

Mainly, TREC definition QA systems profit from manifold techniques when recognising answers across the whole AQUAINT corpus. To begin with, some systems account for a prior off-line processing of this target collection of documents [Fernandes, 2004, Hildebrandt et al., 2004, Katz et al., 2005, 2006, 2007], because once this pre-processing is computed, this cooperates on accelerating the answering process and normally ameliorating the performance. In a sense, this is a plausible and reasonable alternative when coping with static collections. This pre-processing is, however, implausible when tackling collections that are constantly changing and growing, or which their size is too large to be entirely processed with Natural Language Processing (NLP) tools.

For the most part, the variety of NLP tools exploited by definition QA systems fluctuates from Part-of-Speech (POS) tagging (see section 4.9 on page 93) to parsing (e.g., [Xu et al., 2003] introduced in section 4.16 on page 107). These strategies are usually devised to deal with a specific language, and their performance also depends on the tool, the language and the target corpus. This dependency along with the speed required for applying NLP tools to each document and the uncertainty about the boost in performance that this might cause, encourage researchers to explore the limits of techniques that make allowances solely for surface information.

NLP-TOOLS

Certainly, the fact that one of the fundamental ranking ingredients exploited, for example by the best system in TREC 2006 (see section 4.12 on page 102), is web frequency counts indicates that much can be gained in terms of performance without deep linguistic processing. In particular, counts learnt from web snippets. This type of strategy is premised on the principle known as *Distributional Hypothesis*. This principle states that highly correlated words in the same context (e.g., sentence, shingle and paragraph) are very likely to be strongly se-

WEB FREQUENCY COUNTS

TERM CO-OCCURRENCE

mantically related [Harris, 1954, Firth, 1957], and ergo, in the case of definition QA systems, to characterise the *definiendum*.

KNOWLEDGE BASES

This principle is also the building block of numerous techniques including those which collect statistics from articles about the *definiendum* harvested from KBs: [Xu et al., 2003] (section 4.16 on page 107), [Han et al., 2004] (section 4.14 on page 105), [Wu et al., 2004, 2005a, Zhang et al., 2005, Zhou et al., 2006] (section 4.3 on page 81), [Cui et al., 2004b,a, 2005, Sun et al., 2005] (section 4.8 on page 91). As [Han et al., 2004] stressed in their ranking procedure, the major advantage of utilising KBs instead of general texts is the reliability of the extracted highly frequent terms. These words are very likely to be semantically connected, and at the same time, highly probable to be descriptive. At any rate, the restricted coverage yielded by these resources is what makes them less attractive. More precisely, this narrowness or data sparseness hurts the performance [Zhang et al., 2005, Han et al., 2006]. Particularly, the sharp contrast in their coverage for various languages. On top of that, the potential dissonance between the senses/contexts outlined in KBs and the target array of answer candidates is a problematic issue for this sort of methodology.

On a different note, [H. Joho and M. Sanderson, 2000, 2001] demonstrated that a marked enhancement in performance can be achieved when adding syntactic information at the lexical and surface level. They showed that the usage of some definition patterns in amalgamation with frequency counts of words correlated with the *definiendum* can bring about a good performance (see section 4.2 on page 76). They also experimentally revealed that the performance goes hand in hand with the size of the collection, because it sharply increases the likelihood of matching a pre-determined set of rules, thus detecting the most promising descriptive terms. This finding makes the use of definition patterns propitious when coping with massive collections such as the Internet.

REDUNDANCY

Another advantage of massive collections is their amount of redundancy, which make it possible to find numerous paraphrases of the same underlying ideas, therefore increasing the chances of finding a rewriting that observe a purpose-built pattern, while at the same time, attenuating the number of missed descriptions. In substance, [Roussinov et al., 2004, 2005] noticed that redundancy is very likely to produce a high recall of descriptions (see section 4.6 on page 88), but a low precision, because it is very likely that the definition QA system will get manifold paraphrases of the most essential descriptions.

Contrary to the trend of definition QA systems presented so far, [Figueroa and Neumann, 2007, Figueroa et al., 2009] went beyond KBs, and designed a framework for discovering answers to definition queries within web snippets. In this way, they challenged the observation of [Cui et al., 2004c] about the fruitfulness of web-snippets (see details in section 4.9 on page 96). This framework is essentially aimed at multi-linguality, and it is thus compelled to diminish its dependence on the coverage of KBs and NLP processing. More precisely, this definition QA system (M-DEFWEBQA) deals specifically with English and Spanish as target languages. The intention is also to build algorithms that do not rely on the information supplied by KBs about the *definiendum* as this is not necessarily trustworthy all the time, and moreover, the objective is capitalising on web document surrogates to cushion the already mentioned problem of coverage. The general framework is sketched in figure 5.1.

As a means of reaching a large degree of language independency, this system makes use of poor linguistic knowledge, basically a stop-list and an array of definition lexico-syntactic regularities at the word level. In detail, the pack of patterns utilised for English and Spanish are listed on tables 3.1 (page 65) and 3.2 (page 66), respectively. In this framework, the DEFINITION MINER module takes advantage of the corresponding rules for generating the query rewritings explicated in sections 2.5 on page 36 (English) and 2.7 on page 50 (Spanish). In a nutshell, this query rewriting strategy assists in biasing the search engine in favour of web snippets that are likely to put into words descriptive information about the *definiendum* in the

WEB SEARCH respective language. This imposed bias essentially boosts the redundancy, strengthens the recall of sentences matching definition patterns, and consequently makes it possible to produce the final output to the user accounting chiefly for statistical processing, and lessening the reliance on statistics taken exclusively from KBs. Later, the PATTERN MATCHER component checks which of the fetched sentences match the pre-determined set of constructs by means of the *Jaccard Measure* as underlined in section 3.4 (page 61). Analogously to the technique outlined in section 4.12 (page 102), this selection process is predicated on term co-occurrences across web snippets, that is, the array of answer candidates.

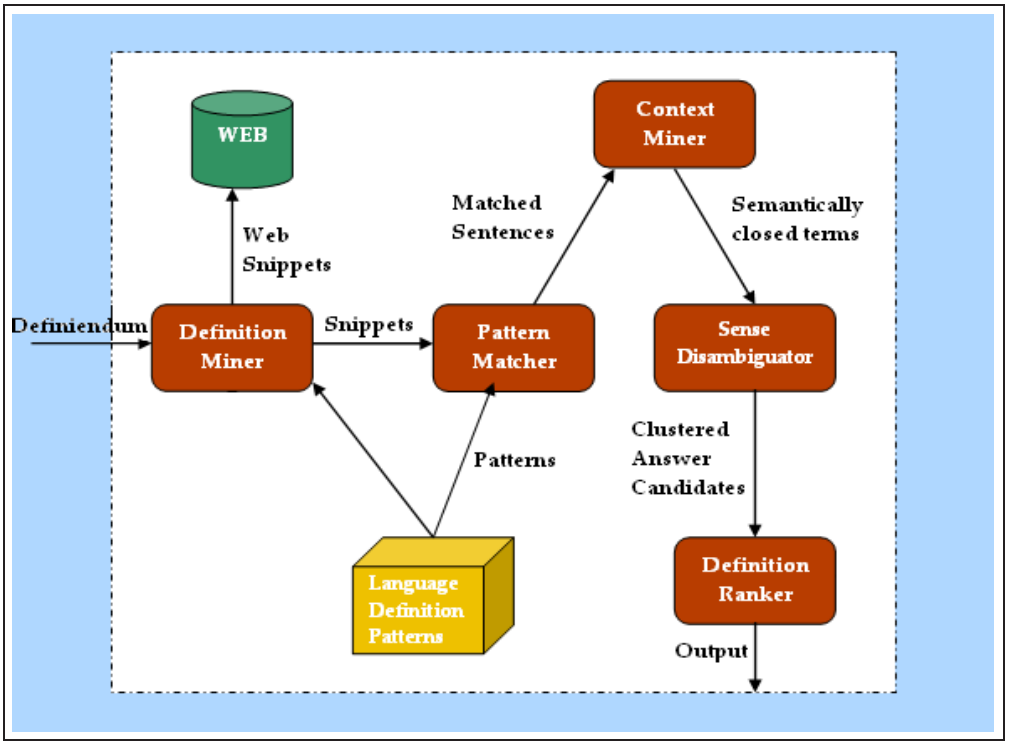

Figure 5.1: Framework for answering definition questions in several languages.

A novel and interesting attribute of the approach adopted by [Figueroa and Neumann, 2007, Figueroa et al., 2009] is the SENSE DISAMBIGUATOR component, which also takes advantage of the obtained redundancy for clustering matched sentences in agreement with the potential senses of the *definiendum*. Retrieved sentences that aligned the array of definition patterns are grouped and selected by the next three components of the flow depicted in figure 5.1: CONTEXT MINER, SENSE DISAMBIGUATOR and DEFINITION RANKER.

5.2.1 Snippet Context Miner

There are many-to-many mappings between names and their concepts. The same name or word can refer to several meanings or entities. For instance, places and companies are named after famous people and owners/locations, respectively. On the other hand, distinct names can indicate the same meaning or entity. As a means of tackling this, a (snippet) context miner sifts semantically connected terms from sentences. To illustrate, take the next set A of descriptive sentences (answer candidates) recognised by the DEFINITION MATCHER:

1. Tom Hanks is an Academy Award-winning actor.
2. Thomas Jeffrey Hanks is an actor born in 1959 in California.

3. Tom Hanks is an American seismologist.

In these sentences, "*American actor Tom Hanks*" is referred to as "*Thomas Jeffrey Hanks*" and "*Tom Hanks*", whereas "*Tom Hanks*" also signals an American seismologist. In this method, a *sense* is a meaning of a word or one possible reference to a real-world entity.

LSA

The context miner is directed at extracting the different senses of the *definiendum* by observing the correlation of its neighbors in the reliable semantic space supplied by Latent Semantic Analysis (LSA) [Deerwester et al., 1990]. This multi-dimensional semantic space is built from a term-sentence matrix (M) which renders the *definiendum* as a *pseudo-sentence* and weighted in conformity to the traditional Term Frequency-Inverse Document Frequency (TF-IDF) metrics. The SNIPPET CONTEXT MINER distinguishes all the possible different *n-grams* (W) in A and their frequencies. The size of W, and hence M, is then reduced by removing n-grams, which are substrings of other equally frequent terms. In addition, [Figueroa and Neumann, 2007, Figueroa et al., 2009] discarded HyperText Markup Language (HTML)-tags, isolated punctuation marks, and term frequency counts were consolidated in congruence with their uppercase variation. This removal step allows the system to speed up the computation of the semantic vectors (M as $UDV^{'}$) when using Singular Value Decomposition (SVD) in LSA. It is also worth noting here that the absence of syntactical information of LSA is slightly mitigated by considering strong local syntactic dependencies (n-grams).

CORPUS SEMANTICS

While running time is not a key issue in state-of-the-art QA systems (effectiveness metrics such as $\mathcal{F}(\beta)$-Score are those often used to compare methods), [Figueroa et al., 2009] still ran several setting experiments in order to obtain a taste of the running times taken by this critical task. Overall, the SVD task cannot be pulled apart from the computation of semantic vectors by LSA, otherwise, [Figueroa et al., 2009] would need a completely different kind of corpus-based semantic analysis method. Nevertheless, the obtained running times were 0.07 seconds (standard deviation=0.12) for SVD and 1.13 seconds (standard deviation=0.86) for LSA. This considered an average of 186 dimensions for the dictionary (W) and 88 dimensions for the sentences premised on the semantic vectors built by LSA. All in all, setting tests showed that the running time for SVD and dimensional reduction tasks are not meaningful for this definition QA system.

SIMILARITY MEASURE

Accordingly, [Figueroa and Neumann, 2007, Figueroa et al., 2009] make use of $\hat{D}$, the greatest three eigenvalues of D, and the corresponding three vectors $\hat{U}$ and $\hat{V}$ for constructing the semantic space as $R = \hat{U}\hat{D}^2\hat{U}^{'}$. Then, [Figueroa and Neumann, 2007, Figueroa et al., 2009] preferred the dot product to the traditional cosine for measuring of the semantic relatedness $R(w_i, w_j) = \hat{u_i}\hat{D}^2\hat{u_j}^{'}$ ($\hat{u_i}, \hat{u_j} \in \hat{U}$) of two terms $w_i, w_j \in W$. The major reasons are (a) it was noticed empirically that, because of the size of web snippets (texts shorter than 200 words), the cosine draws an unclear distinction of the semantic neighbourhood of *definiendum*, bringing about spurious inferences [Wiemer-Hastings and Zipitria, 2001], and (b) the length of vectors was found to draw a clearer distinction of the semantic neighbourhood of *definiendum* as this biases R in favour of contextual terms, which LSA knows better [Deerwester et al., 1990].

SEMANTIC NEIGHBOURHOOD

In the multi-dimensional semantic space generated by LSA, the neighbourhood of a particular word w_i provides its context [Deerwester et al., 1990, Kintsch, 1998], and consequently its correct meaning by pruning, for instance, inappropriate senses [Kintsch, 1998]. Similarly, the *definiendum* is also a term defined by its neighbourhood in this semantic space. Hence the web context miner singles out an array of the highest closely connected words to the *definiendum*, that is, terms that are likely to specify its meaning. Different experiments were set in order to adjust the optimum size of the set of linked words from five to fifty. Best results in terms of semantic closeness and inter-cluster distance were observed for groups of forty

words.

ALIAS MATCHING

As an outcome of the relaxed pattern matching performed by the PATTERN MATCHER, the system also accounts for all *n-grams* that are substrings of the *definiendum* δ as some internal *n-grams* $\delta^+ \in W$ are more likely to occur within descriptive sentences (i.e., names or surnames are more frequent than their corresponding full names).

In the previous example, "*Hanks*" has a higher frequency than "*Tom Hanks*" so the snippet CONTEXT MINER considers an array $\bar{W}$ compounded of the forty highest pairs $\{w_i, R_{max}(\delta, w_i)\}$, where $R_{max}(\delta, w_i) = \max_{\delta^+ \in W} R(\delta^+, w_i)$. The miner eventually normalises terms in $\bar{W}$ based on the following computation:

$$\hat{R}(\delta, w_i) = \frac{R_{max}(\delta, w_i)}{\sum_{\forall w_j \in \bar{W}} R_{max}(\delta, w_j)} \quad \forall w_j \in \bar{W}$$

5.2.2 Sense Disambiguator

One of the difficulties with targeting massive collections of documents is term ambiguity. In the case of definition QA systems, this obstacle translates into the need for clustering matched sentences in accordance to their respective potential senses. For the purpose of resolving this ambiguity problem so as to detect the correct sense, [Figueroa and Neumann, 2007, Figueroa et al., 2009] devised a sense disambiguation component.

TERM CORRELATIONS

This module is geared towards resolving λ distinct senses of the *definiendum* existing in A by discovering a set of uncorrelated words. Let Φ be the term-sentence matrix, where a cell $\Phi_{is} = 1$, if the term $w_i \in \bar{W}$ occurs within the descriptive phrase A_s (zero, otherwise). The correlation amongst words is then given by the dot product $\hat{\Phi} = \Phi\Phi'$. This dot product between two row vectors of Φ reflects the extent to which two terms have a similar pattern of occurrence across the array of sentences. To exemplify, for the words in $\bar{W}$: w_1= "*Academy*", w_2= "*actor*" and w_3= "*seismologist*", the computed values of Φ and $\hat{\Phi}$ are shown below:

$$\Phi = \begin{pmatrix} & A_1 & A_2 & A_3 \\ w_1 & 1 & 0 & 0 \\ w_2 & 1 & 1 & 0 \\ w_3 & 0 & 0 & 1 \end{pmatrix} \quad \hat{\Phi} = \begin{pmatrix} & w_1 & w_2 & w_3 \\ w_1 & 1 & 1 & 0 \\ w_2 & 1 & 2 & 0 \\ w_3 & 0 & 0 & 1 \end{pmatrix}$$

POTENTIAL SENSE MARKERS

Let W^λ be the set of *sense markers*, that is the set of terms in $\bar{W}$ that signal a potential sense of the *definiendum*. This array W^λ is built iteratively by choosing the word with the highest correlation count with terms in $\bar{W} - W^\lambda$, regardless of the frequency. Initially, W^λ starts as empty, and candidates in $\bar{W} - W^\lambda$ are forced to fulfil the constraint of no-linkage or correlation with an already selected term in W^λ. The iterations finish when there is no candidate in $\bar{W}$ that has a correlation count higher than zero.

In the example, "*Academy*" and "*actor*" co-occur in one sentence, and "*seismologist*" does not, then the corresponding values of the correlation counts become: two for w_1 and w_2, one for w_3. As a consequence, the SENSE DISAMBIGUATOR adds w_2="*actor*" to $\bar{W}^\lambda$, because a random term is selected whenever it is necessary to break ties.

In the next step, due to the fact that w_2 was randomly picked, the correlation count is now equal to one for the three words in the next cycle. Nonetheless, w_3="*seismologist*" is chosen, because the other two terms are correlated with the element in W^λ. Eventually, $\bar{W}^\lambda$ includes "*actor*", "*seismologist*".

ORTHONORMAL BASIS

Since words that indicate the same sense co-occur, term vectors build up an orthonormal basis in which each direction becomes a distinct *potential sense*. Thus, each sentence is associated with its corresponding *sense marker* and assigned to one cluster C_λ. Sentences, that were not directly or indirectly (via correlation) connected with an element in $\bar{W}^\lambda$, were grouped

in a special cluster C_0. Accordingly, the outcome for the working example are: C_0=∅, C_1= A_1, A_2 and C_2= A_3.

Furthermore, the SENSE DISAMBIGUATOR will attempt to reassign each sentence in C_0 by searching for the strongest correlation between its entities and the entities[1] of a cluster C_λ. For the above example, the sentence "*Thomas "Tom" Jeffrey Hanks was a school actor in the Skyline High School in Oakland, California.*" would be attached to C_1. In a statement, this discrimination strategy is simple, fast, and works under the assumption that distinct potential senses, or at least the most dissonant and predominant ones, will be typified by a limited array of words that can form separate partitions in the semantic space.

ENTITIES CORRELATION

5.2.3 Definition Ranker

A DEFINITION RANKER produces an ordered sequence of extracted definitions. Let $N(A_s)$ be a function that returns the normalised nuggets connected with A_s, and W_N the array of terms of all normalised nuggets. Ergo, P_i is defined as the probability of finding a word $w_i \in W_N$, and is arbitrarily set to zero for all stop-words, so $W_{N(A_s)}$ is the group of terms in $N(A_s)$. For our working example, the set of ranked words becomes:

- $W_{N(A_1)}$= [ACADEMY, AWARD, WINNING, ACTOR]
- $W_{N(A_2)}$= [ACTOR, BORN, IN, 1959, CALIFORNIA]
- $W_{N(A_3)}$= [AMERICAN, SEISMOLOGIST]

In total, this illustrative array of candidates encircles eleven distinct tokens, in which the stop-word "*in*" occurs two times in the second example. The values of P_i for each w_i are as follows:

[ACADEMY,$\frac{1}{12}$], [AWARD,$\frac{1}{12}$], [WINNING,$\frac{1}{12}$], [ACTOR,$\frac{2}{12}$], [BORN,$\frac{1}{12}$], [IN,0], [1959,$\frac{1}{12}$], [CALIFORNIA,$\frac{1}{12}$], [AMERICAN,$\frac{1}{12}$], [SEISMOLOGIST,$\frac{1}{12}$]

For each cluster C_λ, the definition ranker incrementally computes a set of its sentences A_λ that maximises the relative novelty premised on their coverage and content of the nuggets as:

$$\mathbf{max}_{\forall \mathbf{A_s} \in \mathbf{C}_\lambda} \quad coverage(A_s) + content(A_s)$$

In this equation, *coverage* models the likelihood that novel terms within a normalised sentence $N(A_s)$ belong to a description. As a means of only considering novel words, the iterative process accounts for a cache of terms, which keeps track of all words with respect to previously selected sentences in the cluster C_λ. This measure shares the same spirit with other scoring methods including the strategy of the best system in TREC 2006 [Kaisser et al., 2006], and the definition QA system proposed by [Schlaefer et al., 2007] (for details, see section 4.12 on page 102). The crucial difference is due to the decay factors. [Kaisser et al., 2006] systematically lowered the contribution of a term to the score of an answer candidate. This decrease was in tandem with the number of times the word had already been included in selected answers, whereas [Figueroa and Neumann, 2007, Figueroa et al., 2009] suppressed the contribution of these terms. Of course, the latter is a special case of the

TERM COVERAGE

[1] At this step, numbers and words that start with a capital letter were interpreted as entities. This way, the system was prevented from using ad-hoc linguistics tools.

methodology of [Kaisser et al., 2006] when utilising decay factors of 100%. The idea behind all these rating methods is that they all are in concert with the second criterion of [H. Joho and M. Sanderson, 2000, 2001] (see section 4.2 on page 76).

DECAY FACTORS

The decay factors play the role of redundancy controllers, that is the slower the contribution is diminished, the higher the degree of redundancy the output will contain. In this statement, redundancy is understood at the word level. In the case of the Web, a faster decay is therefore indispensable, because the corpus already supplies redundant descriptions in terms of paraphrases and almost duplicate sentences. Inversely, in a considerably smaller corpus, such as the AQUAINT (the target corpus of the system designed by [Kaisser et al., 2006]), a slower rate is needed, because a pair of answer candidates that share a substantial amount of words can still elucidate different facets of the *definiendum*. A faster decay rate would increase the probability of missing some of these novel facets, whenever they are not included in another relatively novel putative answer.

TERM AND ENTITY CONTENT

On the other hand, *content* discriminates the degree to which $N(A_s)$ expresses definition facets of the *definiendum* on the grounds of highly close semantic terms and entities. This is calculated by summing up the semantic relationship between terms within the corresponding nuggets and the essentiality of novel entities. Each novel entity (e) is weighed in congruence with its probability P_e^{λ} of being in the normalised nuggets of C_{λ}. *Content* stresses the relevance of entities when scoring candidate sentences to a definition question, due to the likelihood that they are signalling a pertinent relationship with the *definiendum*. Other techniques also postulate the importance of entities (and also the underlying relations). For example, the best system in the TREC 2003 (section 3.2 on page 58), other systems are: [Roussinov et al., 2004, 2005] (section 4.6 on page 88), [Gaizauskas et al., 2004] (section 4.13 on 104).

Eventually, [Figueroa and Neumann, 2007, Figueroa et al., 2009] iteratively incorporated a new answer into the final output until the ranking score was lower than an experimental threshold (0.1). On the whole, sentences are rated in consonance with the order they are inserted. This means higher ranked sentences are more diverse, less redundant, and probable to embrace entities together with terms that describe aspects of the *definiendum*. In juxtaposition, other techniques eliminate a random candidate whenever a pair share a pre-determined amount of words [Hildebrandt et al., 2004]. This sort of method incurs the risk of discarding crucial descriptive knowledge expressed in quite similar constructions. Take as an example:

1. <definiendum> is an Academy Award-winning actor, who was born in Germany in 1890.

2. <definiendum> is an Academy Award-winning actor, who died in 1976 in Texas.

5.2.4 Experiments and Results

For the purpose of assessing[2] the performance of this multilingual definition QA system, two kinds of criteria were used: a BASELINE-oriented comparison which uses results from various TREC and Cross Language Evaluation Forum (CLEF) evaluations, and a comparison with other existing approaches to definition QA in TREC and CLEF. For the first assessment, five question sets where used from (1) TREC 2001, (2) TREC 2003, (3) CLEF 2004, (4) CLEF 2005, and (5) CLEF 2006.

[2]Throughout this section, $\pm$ stands for standard deviation, and CLEF data-sets include all English translations from all languages.

Experiments - English

A BASELINE was implemented for which 300 surrogates were downloaded by submitting the quoted *definiendum*. The baseline splits snippets into sentences and accounts for the same battery of constructs exploited by the PATTERN MATCHER in conjunction with a strict matching of the *definiendum*. In addition, a random sentence from a pair that shares more than 60% of its terms, and sentences that are a substring of another sentence were expunged [Hildebrandt et al., 2004].

Corpus	Baseline				M-DefWebQA			
	TQ	NAQ	NS	Accuracy	NAQ	NS	Accuracy	AS (%)
(1)	133	81	7.35 ± 6.89	0.87 ± 0.2	**133**	18.98 ± 5.17	0.94 ± 0.07	16 ± 20
(2)	50	38	7.7 ± 7.0	0.74 ± 0.2	**50**	14.14 ± 5.3	0.78 ± 0.16	5 ± 9
(3)	86	67	5.47 ± 4.24	0.83 ± 0.19	78	13.91 ± 6.25	0.85 ± 0.14	5 ± 9
(4)	185	160	11.08 ± 13.28	0.84 ± 0.2	173	13.86 ± 7.24	0.89 ± 0.15	4 ± 11
(5)	152	102	5.43 ± 5.85	0.85 ± 0.22	136	13.13 ± 6.56	0.86 ± 0.16	8 ± 14

Table 5.1: Results overview (source [Figueroa and Neumann, 2007, Figueroa et al., 2009]).

WEB SNIPPETS COVERAGE

The coverage of BASELINE and this definition QA system can be seen in table 5.1. In these figures, NAQ stands for the number of queries for which the answers embodied at least one nugget, and TQ is the total amount of questions within the question set. In light of the outcomes underscored in this table, [Figueroa and Neumann, 2007, Figueroa et al., 2009] concluded two things:

1. Contrary to the observation of [Cui et al., 2004c], empirical results support the fact that web snippets can be a fruitful source of descriptive information. Specifically, [Cui et al., 2004c] found nuggets for only 42 queries by using external dictionaries in combination with these surrogates (see details in section 4.9 on page 96), whereas M-DEFWEBQA descriptive content for all questions in (2). To be more exact, [Cui et al., 2004c] claimed that web snippets minutely enhance the performance. Although this conclusion relies on different assessments, the figures in table 5.4 yield collaborative evidence of this finding.

2. Further, M-DEFWEBQA discovered nuggets within surrogates for the 133 queries in (1), in contrast to [Miliaraki and Androutsopoulos, 2004], who found a top five ranked snippet that verbalises a definition solely for 116 questions within top 50 retrieved documents.

SENTENCE LENGTH

Furthermore, this definition QA system extracted short sentences in terms of character length (125.7 ± 44.21 considering white spaces; BASELINE: 118.168 ± 50.2), whereby [Miliaraki and Androutsopoulos, 2004, Androutsopoulos and Galanis, 2005] handled fixed windows of 250 characters. This type of fixed window can trim descriptive content or they can include too much unnecessary text. In the opposite way, sentences found by this definition QA system are 109.74 ± 42.15 (BASELINE: 97.81 ± 41.8) characters long without considering white spaces, which is comparatively longer than the 100 characters text fragments of [Hildebrandt et al., 2004]. This ratifies the competitiveness of sentences harvested from web snippets as answer units. Table 5.2 depicts a snapshot of the output of this definition QA system.

Overall, the system covered 94% of the queries, whereas BASELINE did it with 74%. This difference may be due mainly to the query re-writing step and the flexible matching of the

Sense Markers	Outputted Answers
STRANGE	• In epilepsy, the normal pattern of neuronal activity becomes disturbed, causing strange.
SEIZURES	• Epilepsy, which is found in the Alaskan malamute, is the occurrence of repeated seizures. • Epilepsy is a disorder characterized by recurring seizures, which are caused by electrical disturbances in the nerve cells in a section of the brain. • Temporal lobe epilepsy is a form of epilepsy, a chronic neurological condition characterized by recurrent seizures.
ORGANIZATION	• The Epilepsy Foundation is a national, charitable organization, founded in 1968 as the Epilepsy Foundation of America.
NERVOUS	• Epilepsy is an ongoing disorder of the nervous system that produces sudden, intense bursts of electrical activity in the brain.

Table 5.2: Sample output for the query "*What is epilepsy?*" (adapted from [Figueroa et al., 2009]).

definiendum. For all the questions in which this system and the BASELINE extracted at least one nugget, the accuracy and the average number of sentences (NS) was computed. In this aspect, the system doubled the amount of sentences and improved the performance with respect to accuracy. Interestingly enough, this observed growth in accuracy is in agreement with the finding of [H. Joho and M. Sanderson, 2000, 2001] (see section 4.2 on page 76), that is, redundancy aids in deriving better statistics and thus in determining the most dependable sentences that align a pack of definition patterns.

ENHANCEMENT IN ACCURACY

TOPIC SHIFT

The assessment in table 5.1 also took into account the proportion of sentences within NS for which the relaxed matching shifted the *definiendum* to another concept which brought about interesting descriptive sentences (AS). For example, "*neuropathy*" was shifted to "*peripheral neuropathy*" and "*diabetic neuropathy*" (refer to section 3.3 on page 59 for more details on this issue). In juxtaposition, unrelated sentences eventuated from some shifts (e.g., "*G7*" to "*Powershot G7*").

PRECISION AND RECALL

As to the evaluation of this system against a gold standard, [Figueroa and Neumann, 2007, Figueroa et al., 2009] benefited from the list of the assessors produced for the TREC 2003 data. Following the approach by [Voorhees, 2003], this system finished with 0.61 ± 0.33 for *recall* and 0.18 ± 0.13 for *precision*, whereas for the BASELINE, this was 0.35 ± 0.34 and 0.30 ± 0.26, respectively. A higher recall of 0.61 ± 0.33 suggests that the additional sentences selected by this method bore more nuggets that are seen as key ones on the list of the assessors. The high recall also stresses the essential role that web snippets can play for a definition QA system that discovers answers in the AQUAINT corpus.

It is well known that systems in TREC are capable of finding valid nuggets which may not be judged as pertinent in the list [Hildebrandt et al., 2004]. This is even more probable for Web-based systems as these discover many extra text fragments that are regarded as relevant by a user, but excluded from the list of the assessor. In the definition QA system of this section, this is a crucial issue as this significantly raises the number of selected descriptive sentences per question (see table 5.1) and so the length of the response. On the other hand, there are text fragments neither included in the list of the assessor nor in the final output, but they can still be seen as relevant by a particular user. These nuggets should be detected, and these losses should thus be taken into consideration when calculating the $\mathcal{F}(\beta)$-Score. Solving these problems, however, involves designing an entirely new ground truth (refer to

section 1.7 on page 15 for a discussion in detail).

β	1	2	3	4	5
M-DEFWEBQA	0.26	0.37	0.45	0.50	0.53
BASELINE	0.26	0.30	0.32	0.32	0.34

Table 5.3: TREC 2003 average $\mathcal{F}(\beta)$-Scores (source [Figueroa and Neumann, 2007]).

Table 5.3 parallels the outcomes by computing the $\mathcal{F}(\beta)$-Scores with different values of β, that is in other words, by weighing *recall* and *precision* differently. The outcomes of running this system contrasted to the best seven definitional QA systems in TREC 2003 can be seen in table 5.4.

Definition QA System	$\mathcal{F}(5)$	Average Length
BBN (see section 4.16 on page 107)	0.555	2059.20
M-DEFWEBQA	**0.53**	**1878**
National University of Singapore	0.473	1478.74
University of Southern California, ISI	0.461	1404.78
Language Computer Corp	0.442	1407.82
BASELINE	**0.34**	**583**
University of Colorado/Columbia University	0.338	1685.60
ITC-irst	0.318	431.26

Table 5.4: Average $\mathcal{F}(\beta)$-Scores for the TREC 2003 definition queries subtask for the best systems (source [Figueroa et al., 2009]).

In contrast to other definitional QA systems, this approach -would have- accomplished a $\mathcal{F}(\beta)$-Score of 0.53 (2nd place) which is very competitive with the best systems that achieved a value between 0.33 and 0.55 (see a snapshot of responses to a TREC 2003 question in table 5.5). Although the approaches are not directly comparable as they extracted answers from the AQUAINT corpus, whereas this system did so from the Internet, the difference in performance is still very fair. Remarkably, these figures show that web snippets offer coverage to a good amount of nuggets subsumed in this ground truth.

Sense Markers	Outputted Answers
SMITH	• Smith Akbar, the Great Mogul (1542-1605) , Clarendon Press, 1919.
KING	• Akbar the great was the next king of from Mughals (1556-1605).
EMPIRE	• A royal chronicle tells how Akbar the Great, who ruled India's Mogul Empire in the A. D. 1500's, captured at least 9,000 cheetahs during his 49-year reign to aid him in hunting deer. • Akbar the Great was a 16 th Century ruler of the Mogul Empire.
EMPEROR	• 1556 Akbar the Great becomes Mogul; emperor of India, conquers Afghanistan (1581), continues wars of conquest (until 1605)

Table 5.5: Sample output for the query "*Who is Akbar the Great?*" (source [Figueroa et al., 2009]).

SENSE DISCRIMINATION

It is also notable that for some *definienda*, the BASELINE obtained a better $\mathcal{F}(\beta)$-Score: "*Akbar the Great*", "*Albert Ghiorso*", "*Niels Bohr*" as well. This means that it extracted an an-

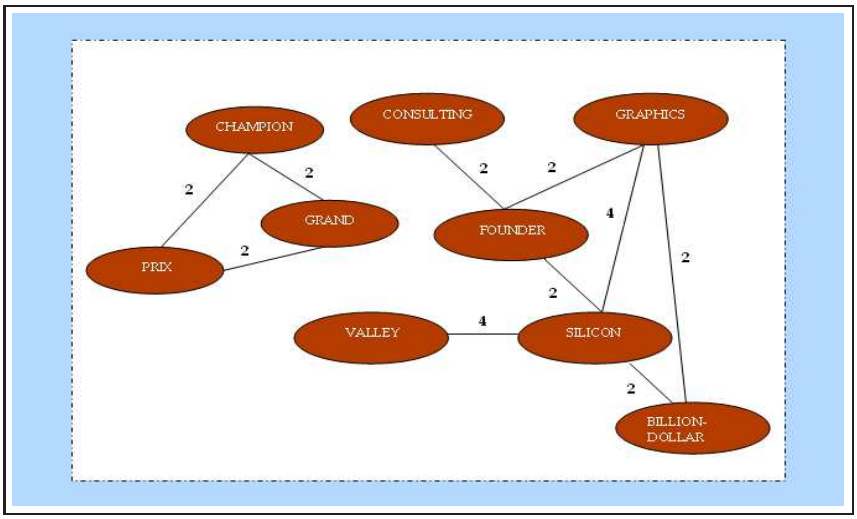

Figure 5.2: $\hat{\Phi}_{ij} > 1$ for "*Jim Clark*" (source [Figueroa and Neumann, 2007, Figueroa et al., 2009]).

swer closer to the list of the assessors. Nevertheless, a distinctive novel component of this definitional model is the incorporation of a sense disambiguation procedure. This was capable of distinguishing (and resolving) different potential senses for some *definienda* (e.g., for "*atom*", the particle–sense and the format–sense). Some snapshots of the output of this sense disambiguation model can be seen in tables 5.2 and 5.5.

On the other hand, some senses were observed as split into two separate senses, e.g., "*Akbar the Great*", where "*emperor*" and "*empire*" indicated different senses. This misinterpretation may be due to an independent co-occurrence of "*emperor*" and "*empire*" with the *definiendum*. As a means of improving this, external sources of knowledge may be necessary. Nonetheless, this may become a very hard task [Chen et al., 2006], as some *definienda* can be extremely ambiguous (e.g., "*Jim Clark*" refers to several real-world entities). While this sense disambiguator can differentiate between the photographer, the pilot, the Netscape creator (see figure 5.2), numerous executives named "*Jim Clark*" are grouped in the same cluster. In addition, entities and the correlation of highly closed terms in the semantic space supplied by LSA can be two building blocks of a more sophisticated methodology for the disambiguation of the *definiendum*.

Another distinctive attribute of this approach is the fact that the DEFINITIONAL MINER component avoids coping with specialised wrappers or downloading full documents.

Experiments - Spanish

Two baselines were designed for assessing the performance for definition queries in Spanish. Both, BASELINE ES-I and BASELINE ES-II do the same processing as the BASELINE implemented for English, but they download 420 surrogates[3]. These two baselines also differ from the baseline utilised for English in the number of terms that two sentences must share to be counted as redundant. Similarly to the criteria to set the decay factors, they account for a threshold of 90% instead of 60%, because the coverage of web space for Spanish is smaller than English and some pertinent nuggets are missed in conjunction with the redundant content (see also section 2.7 on page 50). The difference between the Spanish baselines is that

[3]The amount of retrieved snippets is balanced in order to make all systems fetch the same maximum number of hits (see also section 2.7 on page 50).

	with white spaces	without white spaces
Baseline ES-I	98.11 ± 44.90	81.06 ± 37.69
Baseline ES-II	104.98 ± 36.43	85.88 ± 29.87
M-DefWebQA	135.78 ± 45.21	113.70 ± 37.97

Table 5.6: Length of output sentences (Spanish) (source [Figueroa and Neumann, 2007]).

BASELINE ES-I is directed at the regularities in table 3.2 (page 66) whereas BASELINE ES-II is aimed at the patterns in [Montes-y-Gómez et al., 2005].

ANSWER LENGTH

Given the lengths of the outputs of BASELINE ES-I and this multilingual system (see table 5.6), it can be concluded that the increment indicates that this system outputs more complete sentences, lessening the effects of intentional breaks in web snippets. Due to the acceptable length of descriptive sentences and the fact that many nuggets seem odd without their context [Hildebrandt et al., 2004], this multilingual definition QA system outputs sentences in place of only nuggets.

As a very rough rule of thumb, the degree of redundancy of a sentence A_s was approximated at the word level by looking for a sentence $A_{s'}$ in the same response that shares the maximum number of terms with A_s:

$$redundancy(A_s) = \max_{A_{s'} \neq A_s} \frac{ns(A_s \cap A_s^{'})}{ns(A_s)}$$

where $ns(A_s)$ is the number of words in A_s, excluding stop-words. As a result, BASELINE ES-II generates an output, at least two times as redundant as this definition QA system, which supplies longer sentences (see table below). By and large, this multilingual system outputs comparatively fuller and less redundant sentences.

	(1)	(2)	(3)	(4)	(5)
BASELINE ES-I				0.32 ± 0.16	0.38 ± 0.25
BASELINE ES-II				0.54 ± 0.24	0.64 ± 0.39
M-DEFWEBQA				0.25 ± 0.17	0.25 ± 0.16
BASELINE EN-I	0.58 ± 0.26	0.61 ± 0.26	0.57 ± 0.25	0.62 ± 0.25	0.53 ± 0.23
M-DEFWEBQA	0.47 ± 0.18	0.50 ± 0.20	0.45 ± 0.18	0.45 ± 0.17	0.45 ± 0.19

Table 5.7: Redundancy overview (source [Figueroa and Neumann, 2007]).

The coverage of surface patterns for English has been studied widely [Hildebrandt et al., 2004, H. Joho and M. Sanderson, 2001, 2000] (see section 4.2 on page 76). By the same token, table 5.8 shows the amount of descriptive sentences in the final output that align each regularity in table 3.2 (page 66). Each cell represents the amount of matches for the CLEF 2005/2006 corpus respectively. In this battery of rules, the first construct provides wider coverage, while the third the most limited. Given the marked growth in the amount of recognised descriptive utterances in the final output, it can be concluded that the query rewriting of M-DEFWEBQA strongly biases the search engines, not only in favour of redundant descriptive sentences, but also in favour of diverse utterances. On the one hand, redundant sentences are undesirable in the final output, they are, on the other hand, useful for distinguishing more critical and trustworthy descriptive utterances.

In a special manner, [Figueroa and Neumann, 2007] contemplated an entirely different evaluation for each language for the following reasons: (a) the way the performance of defi-

	(1)	(2)	(3)	(4)	(5)
BASELINE ES-I	78/37	17/10	00/00	13/10	05/03
M-DEFWEBQA	470/254	168/95	03/01	59/58	54/36

Table 5.8: Coverage of patterns (source [Figueroa and Neumann, 2007]). Note that the numbers coincide with the order of presentation in table 3.2.

nition QA systems is measured differs between TREC and CLEF, and (b) CLEF gold standards for definition questions supply only one nugget regarding abbreviations or position of persons, whereas TREC 2003 provides a group of relevant nuggets.

	Baseline ES-I	Baseline ES-II	M-DefWebQA
(4)	11	**33**	32
(5)	9	12	**22**

Table 5.9: Ground truths (source [Figueroa and Neumann, 2007]).

WEB SNIPPETS COVERAGE

As for Spanish, this multilingual definition QA system responded to 32 and 22 out of the CLEF 2005 and 2006 queries, respectively (see table 5.9). However, the runs submitted by the best two systems in CLEF 2005 answered 40 out of the 50 definition questions [Vallin et al., 2005, Montes-y-Gómez et al., 2005]. Nevertheless, the third-best system only responded 26 queries. Further, the best system in CLEF 2006 answered 35 out of the 42 definition questions, whereby this system found answers for 22 out of the 35 queries responded by this best system. Unfortunately, CLEF 2006 gold standard supplies only one nugget for these 35 questions.

Since the coverage of the ground truths focuses solely on abbreviations and positions of people, together with the fact that responses to seven of the CLEF 2006 queries were missed, [Figueroa and Neumann, 2007] assigned three out of five different assessors to each data-set. Each assessor judged whether or not each output sentence rendered descriptive information. A sentence was counted as descriptive if, and only if, at least two out of the three assessors agreed (results in table 5.10). In both data-sets, this multilingual definition QA system outperformed both baselines. To be more specific, it discovered descriptive phrases for 47 out of the 50 CLEF 2005 questions. Further, this system returned more descriptive utterances (NS) with a lower level of redundancy. At any rate, the accuracy of the output sentences worsened in comparison to the English results. As a matter of fact, [Figueroa and Neumann, 2007] conceived this as a consequence of the lower amount of web redundancy for Spanish, which affects the quality of identifying the most pertinent and dependable phrases. Finally, table 5.9 shows that the performance of this system can be ameliorated by aligning patterns in [Montes-y-Gómez et al., 2005] without necessarily considering them in the query reformulation process.

5.3 Web Snippets and Mining Multilingual Wikipedia Resources

The system outlined in the preceding section takes advantage predominantly of redundancy for finding answers to definition questions within web snippets. This process proved to be competitive in terms of finding the most critical descriptions, while at the same time, it supplies a framework for building multilingual systems. This method, however, can fail to find nuggets conveyed a few times, deteriorating the diversity of the output, although these nuggets might be expressed in terms commonly found across articles in KBs.

Corpus	Baseline ES-I				Baseline ES-II		
	TQ	AQ	NS	Accuracy	AQ	NS	Accuracy
(4)	50	26	2.59 ± 2.45	0.85 ± 0.23	39	10.13 ± 10.66	0.67 ± 0.31
(5)	42	10	3.00 ± 3.13	0.61 ± 0.31	15	3.4 ± 3.31	0.65 ± 0.26

	TQ	AQ	NS	Accuracy
(4)	50	**47**	8.6 ± 4.85	0.63 ± 0.19
(5)	42	**30**	7.27 ± 6.76	0.67 ± 0.25

Table 5.10: Results overview. (TQ = Total number of questions in the question-set) (source [Figueroa and Neumann, 2007])

An alternative approach is profiting from multilingual KBs, like Wikipedia, in such a way that it is transparent or almost transparent to the system. In so doing, [Figueroa, 2008b] extended M-DefWebQA by enriching two modules with evidence supplied by Wikipedia: Definition Miner and Definition Ranker. In a way that it is independent from the *definiendum* when rating, and ergo it remains less sensitive to the sharp variations in KB coverage. This is a substantial difference to the trend of definition QA systems described in chapter 4. Figure 5.3 illustrates the new architecture.

Multilingual KBs

In the first place, the difference between both Definition Miners is that this new one accounts for the focused search presented in section 2.6.2 (page 47) instead of the methods in section 2.5 on page 36 (English) and 2.7 on page 50 (Spanish). It is worth recalling that only Google n-grams counts for English are available to this system, it therefore proceeds, in the case of Spanish, as when there is no n-grams evidence for English.

In the second place, the real difference between both systems stems from the Definition Ranker. While the old strategy discriminates answers on the grounds of redundancy, this new scoring function additionally rates answers candidates in congruence with templates and tuples (pairs and triplets) learnt from Wikipedia. It is worth duly pointing out here that some systems in CLEF have already been systematically using Wikipedia for answering queries in Spanish: [de Pablo-Sánchez et al., 2006, 2007, Martínez-González et al., 2008]. The CLEF challenge, however, does not share the same view as the TREC standard. In the context of CLEF, a short response suffices, whereas in TREC a more complex response is required (see section 1.7 on page 15). For this reason, they capitalise on other types of strategies. Nonetheless, the succeeding assessments stick to the TREC viewpoint of this task.

5.3.1 Learning Templates and Tuples from Wikipedia

To begin with, [Figueroa, 2008b] extracted sentences that match the regularities in table 3.1 on page 65 (English) and 3.2 on page 66 (Spanish) from the abstracts of Wikipedia. Secondly, entities are replaced with a placeholder (#). These entities are discriminated on the grounds of word sequences that begin with a capital letter, and a name entity recognizer [4]. In brief, this process resulted in 1,900,642 different abstractions for the English language, whereas in 527,185 for Spanish.

Thirdly, bigrams to decagrams are obtained from the definition part of these modified first sentences. These resulting n-grams are called templates, and only templates that start at any of the first four words are considered. Lastly, an histogram of templates is built (see table 5.11), and templates with a frequency lower than six are eliminated. Accordingly, the

[4]For this purpose, Stanford Named Entity Recogniser (NER) was used, which is available at nlp.stanford.edu/software/CRF-NER.shtml. In the case of Spanish, only sequences of capital letters were used as representation of Entities.

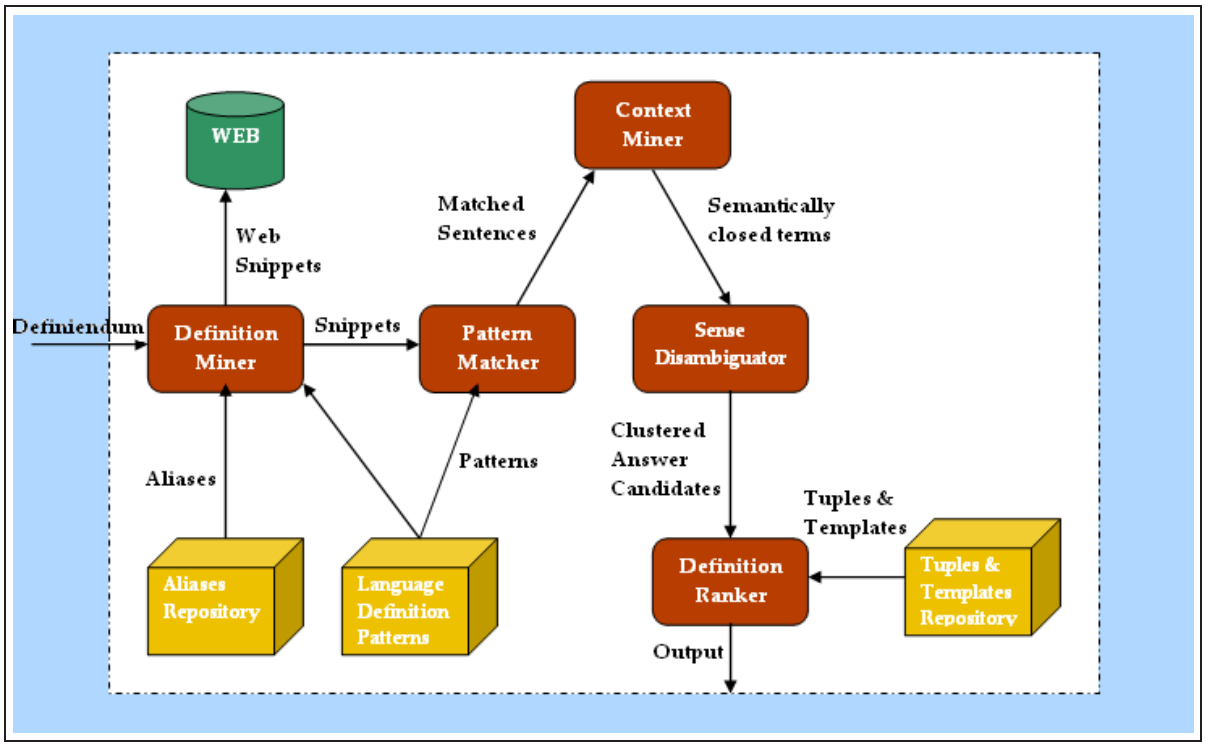

Figure 5.3: Framework for answering definition questions in several languages using Wikipedia resources.

initial diminution in variation stems from the replacement of name entities by a placeholder, aids in obtaining more reliable template counts.

English			Spanish		
Template	$len(t)$	$freq(t)$	Template	$len(t$	$freq(t)$
is a species of	4	34878	y comuna francesa en la region de #	8	3330
member of the #	4	23351	es una comuna y poblacion de # en la region	10	3310
a # politician	3	13422	es un municipio de la	5	2976
a municipality in the district of #	7	8922	es un politico	3	1471
is a # politician	4	4776	un club de futbol	4	1452
is a # politician and the	6	162			
is a # politician who is currently	7	18			

Table 5.11: Sample interesting templates (source [Figueroa, 2008b]).

TEMPLATE REPOSITORY

The basic idea behind this off-line repository is that these templates are not only highly likely to indicate definitions, but also to start these descriptions. Take, for instance, the following two definitions gathered from web snippets:

Daniel Hannan **is a** British **politician** who is currently..
Angela Dorothea Merkel (born July 17, 1954 in Hamburg) **is a** German **politician** and the conservative opposition's..

In these examples, "*is a British politician*" and "*is a German politician*" match the relatively high in frequency template "*is a # politician*" (see table 5.11), and it consequently supports in

distinguishing these descriptive phrases without needing to check whether or not an entry in a specific resource exists.

5.3.2 Definition Tuples Repository

Fundamentally, [Church and Hanks, 1990] inferred *word association norms* directly from unstructured natural language text. They proposed a measurement, named *association ratio*, predicated on the idea of mutual information. The *association ratio* (I_2) between two words w_1 and w_2 is defined as: WORD NORMS

$$I_2(w_1, w_2) = log_2 \frac{P(w_1, w_2)}{P(w_1)P(w_2)} \tag{5.1}$$

This ratio juxtaposes the likelihood of observing w_2 followed by w_1 within a fixed window of k words with the probabilities of observing w_1 and w_2 independently. This ratio differs from mutual information in the encoded linear precedence, and captures some lexico-syntactic regularities in the target corpus [Church and Hanks, 1990].

For the remainder of this section, this ratio is worked out on the descriptions by making allowances for a window size of ten, and the probabilities are estimated as explicated in [Church and Hanks, 1990]. Since this ratio becomes unstable when counts are very small, like [Church and Hanks, 1990], word pairs with a frequency lower than six were expunged. In addition, pairs solely comprising stop-words[5] were also filtered out.

Under the underpinning assumption that relevant pairs will exhibit a joint probability larger than the product of the probability of finding them by chance, this word association ratio is extended to triples as follows: TRIPLETS

$$I_3(w_1, w_2, w_3) = log_2 \frac{P(w_1, w_2, w_3)}{P(w_1)P(w_2)P(w_3)} \tag{5.2}$$

Like [Church and Hanks, 1990], [Figueroa, 2008b] noticed the larger the ratio is, the more credible results it computes. Inversely, the values become less interesting while the ratio approaches zero. Negative ratios are rare, but possible, and [Church and Hanks, 1990] suggest that it indicates a complementary relationship. Simply put, this ratio supplies an efficient way of identifying some semantic and lexico-syntactic relations.

To exemplify, table 5.12 emphasises some interesting tuples pertaining to the word w_*= "*politician*" ($freq(w_*)$ = 32,306). Some of these tuples can cooperate on identifying working descriptive phrases. In short, these tuples and their respective norms distill from the previous 1,900,642 generalisations.

$\vec{w} =< w_1, w_2 >$	$I_2(\vec{w})$	$\vec{w} =< w_1, w_2, w_3 >$	$I_3(\vec{w})$
$< w_*$, served>	7.07	$< w_*$, served, #>	33.09
$< w_*$, diplomat>	7.06	$<$ a, w_*, currently>	7.41
$< w_*$, currently>	4.33	$< w_*$, who, currently	7.14
$< w_*$, opposition>	4.15	$<$ a, w_*, conservative	2.93
$< w_*$, conservative	3.44	$<$ a, w_*, opposition>	2.71
$< w_*$, coach>	-0.30	$< w_*$, the, junior>	-5.08

Table 5.12: Some associations with "*politician*" (source [Figueroa, 2008b]).

[5] We use the 319 highly frequent close class forms encompassed in:
http://www.dcs.gla.ac.uk/idom/ir_resources/linguistic_utils/stop_words.

More to the point, table 5.13 stresses some associations with w_*="*politico*" discovered in Spanish. This table highlights a beneficial aspect of these *association ratios*: they can be deduced for several languages.

$\vec{w} =< w_1, w_2 >$	$I_2(\vec{w})$	$\vec{w} =< w_1, w_2 >$	$I_2(\vec{w})$	$\vec{w} =< w_1, w_2, w_3 >$	$I_3(\vec{w})$
<cientista, w_* >	10.54	< w_*, analista >	8.70	< w_*, democristiano, de>	28.16
< w_*, democristiano>	9.75	< w_*, derechista>	8.66	< w_*, catalanista, en>	27.72
< w_*, socialdemocrata>	9.52	< w_*, federalista>	8.64	< w_*, centrista, de>	27.16
< w_*, afiliado>	9.43	<partido, w_* >	8.57	< w_*, centro-izquierda, de>	26.26
< w_*, trotskista>	9.34	< w_*, salvadoreño>	8.50	< w_*, afiliado, al>	26.05
< w_*, catalanista>	9.30	< w_*, marxista-leninista>	8.49	< w_*, centrista, fundado>	25.99
< w_*, centro-derecha>	9.17	< w_*. intendente>	8.47	< w_*, socialdemocrata, de>	25,92
< w_*, centro-izquierda>	9.09	< w_*. sindicalista>	8.43	< w_*, centro-derecha, de>	25.75
< w_*, galleguista>	9.00	<activismo. w_* >	8.43	< w_*, sindicalista, español>	23.43
< w_*, peronista>	8.98	< w_*, militante>	8.32	<profesor, universitario, w_*>	23.29
< w_*, nacionalista>	8.85	< w_*, diplomatico>	8.30	<abogado, universitario, w_*>	22.36
< w_*, conservador>	8.74	< w_*, populista>	8.20	< w_*, guerrillero, integro>	22.02
< w_*, socialista>	8.71	< w_*, comunista>	8.19	< w_*, guerrillero, boliviano>	22.02

Table 5.13: Some strong *word association norms* with w_*="*politico*" (source [Figueroa, 2009]).

5.3.3 Ranking Definitions

First, this component computes a template representation of each answer candidate by replacing sequences of words that start with a capital letter with a placeholder. This representation helps to tackle data-sparseness, that is to say, to boost the chances of matching the content learnt from Wikipedia. From now on in this section, in order to avoid confusion, these templates will be referred to as answer candidates. Second, this ranker sifts all templates that align these putative answers from the repository. This definition ranker clusters these templates into groups in accordance with their lengths. Let Θ_l be the group containing matched templates of length l, and $fmax_{\Theta_l}$ the frequency of its highest frequent element. This module subsequently rates an answer candidate A_s in concert with:

$$R_{\Theta}(A_s) = \sum_{l=2}^{10} \xi_l \sum_{\forall t \in \Theta_l^{A_s}} \frac{freq(t)}{fmax_{\Theta_l}}$$

DESCRIPTIVE TEMPLATES

In plain words, each answer candidate A_s is ranked in agreement with its matching templates ($\Theta_l^{A_s} \subseteq \Theta_l$). This ranking value consists solely of the sum of the respective normalised frequencies (divided by $fmax_{\Theta_l}$) and a weight ξ_l. This weight factor favours definitions that match longer templates. Third, this system ranks definitions in congruence with their entities. Taking entities into consideration is vital because entities are defined by their relations with other entities. Here, this DEFINITION RANKER builds a frequency histogram of numbers and tokens that begin with a capital letter. Each definition is then rated by adding the frequencies of the entities it carries. These ranking values are thereafter normalised by dividing by the highest value. Let $R_E(A_s)$ be the normalised value in relation to the definition A_s.

ENTITIES

The reason to avoid Named Entity Recognisers (NER) is two-fold: (a) they perform poorly on web snippets, due to truncations, and (b) it is the aim to exploit as few linguistic tools and knowledge at the time of extracting answers as possible, while at the same time, systematically increasing the off-line linguistic processing while building the models. This way this

system could deal, in the future, with additional languages by only changing the content in the repositories. It is worth reiterating that other strategies also reflect the importance of entities: the best system in the TREC 2003 (section 4.16 on page 107), [Roussinov et al., 2004, 2005] (section 4.6 on page 88), [Gaizauskas et al., 2004] (section 4.13 on page 104) as well.

Fourth, this DEFINITION RANKER constructs an histogram H of pairs and triples $\vec{w}$ from the array of putative answers. Then, it sifts the respective word association ratios from the repository (I_2 and I_3), and normalises these ratios by dividing by the ratio with respect to the highest pair and triple afterwards ($\bar{I}_2$ and $\bar{I}_3$, respectively). Later, pairs and triples $\vec{w}$ with a frequency equal to one are removed from the histogram H, and this histogram is normalised similarly to the association ratios ($\bar{H}$). Each definition (answer candidate) A_s is subsequently rated in accordance with the tuples in the repository as follows: TUPLES

$$R_I(A_s) = \sum_{\forall \vec{w} \in \vec{W}^d - \tilde{W}} \bar{I}_2'(\vec{w}) + \bar{I}_3(\vec{w})$$

Where $\tilde{W}$ includes all tuples belonging to previously selected phrases, and $\vec{W}^{A_s}$ are all the tuples extracted from the definition $A_s \in A$. This $\tilde{W}$ assists in rating definitions in conformity with their novelty respecting the already selected phrases. $\bar{I}_2'(\vec{w})$ is stipulated as follows:

$$\bar{I}_2'(\vec{w}) = \begin{cases} \bar{I}_2(\vec{w}) & \text{if } \bar{I}_2(\vec{w}) \neq 0 \\ \bar{H}(\vec{w}) & \text{otherwise} \end{cases}$$

This factor $\bar{I}_2'(\vec{w})$ is aimed at harmonising evidence supplied by tuples seen in Wikipedia with some prominent regularities found in the answer candidate set, that is it is a mixture of corpus and redundancy based measure. Eventually, a sentence is ranked as follows:

$$R(A_s) = (1 + R_\Theta(A_s) + R_E(A_s)) * R_I(A_s)$$

In almost the same way as the original system elucidated in section 5.2, the higher rated sentence is chosen and its corresponding tuples are added to a cache. In the next cycles, tuples encircled by this cache are not considered when working out the ranking values of the remaining putative answers; this way sentences carrying novel and promising tuples are preferred over more redundant sentences whose scores tend to drop as long as more phrases are singled out. Sentences that obtain a rank value lower than an experimental threshold (0.1) are unconsidered. Several values were tried (0 to 0.3) to optimise this definition ranker by profiting from a subset of development queries. As a rule of thumb, values higher than 0.3 can miss many novel nuggets.

On purpose, the raking function $R(A_s)$ relies stronger on the tuples mined from Wikipedia than on tuple counts within downloaded web snippets. This way it can be checked whether or not they are efficient in distinguishing descriptive expressions.

5.3.4 Experiments

In the experiments in section 5.2.4, systems were compared on the basis of the gold standard produced by TREC 2003. This ground truth was configured in order to assess systems targeted at the AQUAINT corpus. Many nuggets in the gold standard, for this reason, are not necessarily subsumed in the retrieved web snippets. Systems, hence, would never be able to get the reward with respect to these matches, materialising a distortion in the evaluation. Systems can, additionally, recognise several nuggets excluded in the TREC ground truth, enlarging the response, and consequently causing a diminishment in terms of precision, and by the same token, $\mathcal{F}(\beta)$-Score. On the other hand, several nuggets can still exist

across fetched web snippets that were undetected by the system, and since these are ignored in the gold standard, the system cannot be accordingly punished, substantially ballooning its performance.

All things considered, a ground truth was created by manually inspecting the web snippets with respect to 189 test queries[6] supplied by the TREC 2003/2004/2005 tracks. Table 5.14 outlines the gold standard framed for the TREC 2005 *definiendum*: "*NATO*".

ID	Nugget
1	in 1949, 12 members
2	in 1994, 16 members
3	in 2005, 26 members
4	military alliance
5	political organisation
6	founded 1949
7	founded by Belgium, Britain, Canada, Denmark, France, Iceland, Italy, Luxembourg, Netherlands, Norway
8	North Atlantic Treaty Organisation
9	never conducted a military operation
10	no partner relationships

Table 5.14: Ground truth for the *definiendum*: "*NATO*".

As a means of making a fair assessment, the evaluation stuck to the most recent standard by using uniform weights for the nuggets [Lin and Demner-Fushman, 2006] (see more details in section 1.7 on page 15). It is important to note here that there was no descriptive information for eleven questions in relation to the TREC 2005 data set. In order to test the efficiency of this new method, the original system (M-DEFWEBQA) was utilised as BASELINE. For the sake of clarity, this new approach will be referred to as MKB-DWQA.

ξ_2	ξ_3	ξ_4	ξ_5	ξ_6
0.0528	0.0708	0.0861	0.09407	0.09759
ξ_7	ξ_8	ξ_9	ξ_{10}	
0.09916	0.09955	0.09983	0.09989	

Table 5.15: Definition ranker parameters (source [Figueroa, 2008b]).

Table 5.15 clarifies the value of parameters utilised in the experiments. ξ_l were fixed coinciding with the number of matching templates across a subset of development definition queries. Longer templates are certainly more trustworthy and harder to match, and ergo they are weighted more heavily.

GROUND TRUTH IMPACT

Table 5.16 parallels the outcomes achieved by each system and data-set. The first interesting observation regards the deprovement obtained by the BASELINE, when the gold standard was changed. This system previously achieved an average $\mathcal{F}(3)$-Score and $\mathcal{F}(5)$-Score of 0.45 and 0.53 (see table 5.3), respectively. This system, conversely, now reached values of 0.45 and 0.46 for the same dataset. This new $\mathcal{F}(5)$-Score value of 0.46 corroborates the good performance of the system (see also table 5.4). Furthermore, the difference in performance arising from the outcomes for the TREC 2005 data set: 2.13% ($\beta = 1$), 8.51% ($\beta = 2$), 14.89% ($\beta = 3$), 14.58% ($\beta = 4$), 16.67% ($\beta = 5$), signals the relevance of the models inferred from the training

[6] The repository of tuples was built under the exclusion of articles concerning these 189 *definienda*.

β	TREC 2003		TREC 2004		TREC 2005	
	Baseline	MKB-DWQA	Baseline	MKB-DWQA	Baseline	MKB-DWQA
1	0.44 ± 0.16	0.42 ± 0.14	0.48 ± 0.13	0.46 ± 0.18	0.47 ± 0.13	0.48 ± 0.2
2	0.44 ± 0.16	0.48 ± 0.13	0.49 ± 0.12	0.51 ± 0.13	0.47 ± 0.17	0.51 ± 0.16
3	0.45 ± 0.17	0.51 ± 0.14	0.5 ± 0.13	0.53 ± 0.13	0.47 ± 0.18	0.54 ± 0.15
4	0.45 ± 0.17	0.53 ± 0.15	0.5 ± 0.13	0.54 ± 0.13	0.48 ± 0.18	0.55 ± 0.16
5	0.46 ± 0.18	0.54 ± 0.16	0.5 ± 0.14	0.55 ± 0.13	0.48 ± 0.18	0.56 ± 0.16

Table 5.16: TREC 2003-2005 results ($\mathcal{F}(\beta)$-Score) (source [Figueroa, 2008b]).

material. This increasing difference is rendered in the detection of nuggets low in frequency as recall is weighted heavier in tandem with the value of β.

In the case of the TREC 2003 data-set, this MKB-DWQA outperformed BASELINE in 34 questions (68%), whereas the BASELINE accomplished a higher score for 16 queries (32%). First of all, there was no profound difference in the results per question between $\beta = 3$ and $\beta = 5$ as shown in figure 5.4. In other words, no sharp variation emerges from this comparison. In 13 questions, MKB-DWQA achieved more than 50% improvement, while in 17 more than 30% and in 27 more than 20%. On the other hand, the performance was considerably decreased in ten cases (20%). In the TREC 2004 question set, MKB-DWQA ameliorated the performance for 41 (64%) out of 64 queries, whereas in TREC 2005, MKB-DWQA reaped better results in 37 (49%) out of 75 questions. Given these figures, it can be concluded that the presented methods cooperates on distinguishing more nuggets low in frequency. PERFORMANCE

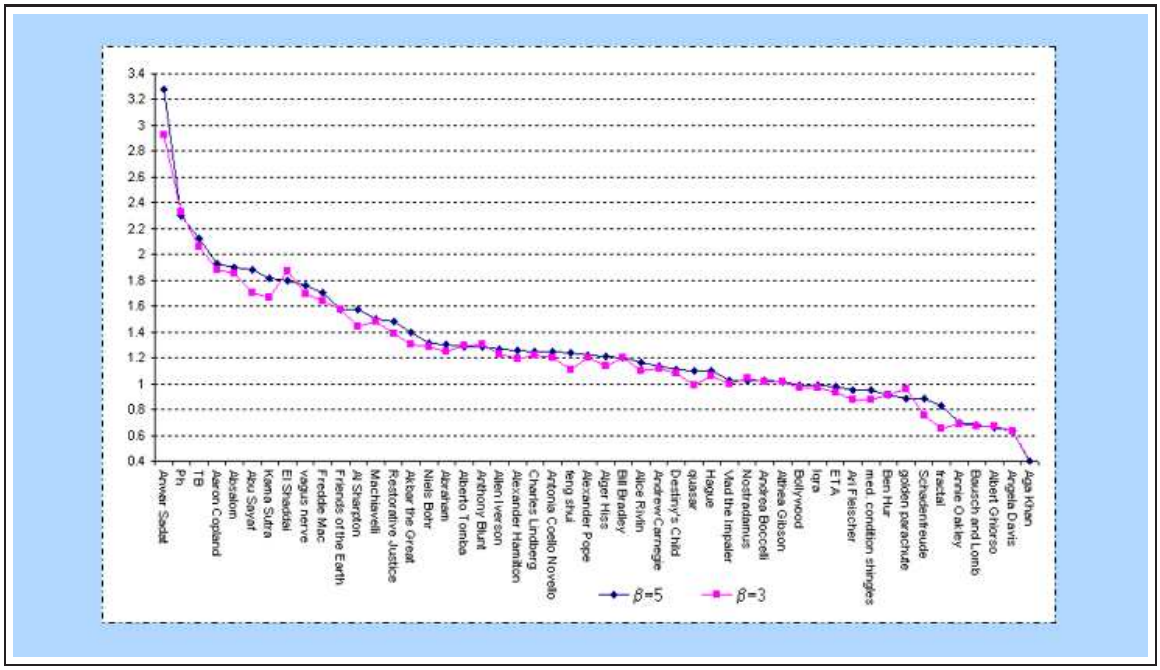

Figure 5.4: $\frac{\mathcal{F}_{\text{MKB-DWQA}}}{\mathcal{F}_{\text{BASELINE}}}$ vs. *definiendum* (source [Figueroa, 2008b]).

There are two decisive factors which worsen the performance. First, paraphrases that do not share a substantial number of words but basically put into words (almost) the same descriptive information: PARAPHRASES

Schadenfreude is a German word that refers to the guilty joy people sometimes feel at the

misfortune of others.
Schadenfreude is a German word meaning to take pleasure at the misfortune of others.

The second determining factor derives from the first: two sentences that share many words, but the few changed terms bring about several meaningful tuples, and the system, therefore, interpret this sentence as bearing significant novel definition information. To reinforce this point, consider the following two selected definitions:

Schadenfreude is a German word that means "pleasure derived from the misfortunes of others".
Schadenfreude is a German word meaning to take pleasure at the misfortune of others.

The change, here, of "*that means*" to "*meaning*" effectuates the matching of the tuples <meaning, taken, 3.93> and <means, taken, 3.86>. Both carry the same meaning, but they are seen differently by MKB-DWQA.

In the light of the achieved results, it can be concluded that the repository of templates and tuples[7] can assist in bettering the efficiency and robustness of definition QA systems in English. However, these outcomes cannot be equally extended to Spanish (see the figures in tables 5.19 and 5.18). The reason for this is that Wikipedia supplies about 2,000,000 definition pages in English, while only about 200,000 in Spanish. Therefore, the association ratios derived for Spanish were not as reliable as for English. Additionally, the number of tuples in English distilled from Wikipedia is (at least) three times larger than in Spanish. Therefore, it is harder to find matches within web snippets. In the next section, a further study on Spanish is conducted. For the sake of clarity, greater details on this evaluation are provided together with the results of the next study.

5.3.5 An extra try for Spanish: Extracting tuples from Dependency Trees

The previous technique, on the one hand, combines evidence yielded by candidate sentences with knowledge supplied by descriptive sentences across Wikipedia articles. There is still, on the other hand, a big question mark about the *word association norms* sketched in the prior section: extracting pairs and triplets from windows of ten consecutive words starts from the tacit linguistic assumption that lexical dependencies cannot occur between larger spans of words. Intuitively, this problem could be solved by accounting for larger windows, but unfortunately, this would bring out a sharper growth in the amount of tuples; to be more exact, in the number pairs and triplets pertaining to loosely related words. In reality, [Figueroa, 2009] conjectured that this increment would be more prominent than in the amount of tuples of largely related words.

WINDOW SIZE

Another valid assumption made by these norms is that a relation between all words within a given window exists. This seems to be utterly reasonable when weakly related tuples are discarded by means of an empirical threshold. At any rate, there are also manifold meaningful relationships low in frequency that would be filtered out along with these spurious tuples. This is a burning issue when dealing with a training corpus limited in size, because many essential tuples will obtain a low frequency, and ergo look irrelevant.

RELATION RELEVANCY

For the purpose of surmounting these difficulties, a dependency parser is exploited as an oracle[8] that supplies the lexical dependencies in a given descriptive sentence. This dependency parser assists in removing the window size and lowering the experimental threshold from six to two. The *word association norms* are hence computed as pairs and triplets of consecutive words in the dependency paths. Some illustrative examples taken from the dependency trees depicted in figure 5.5 are:

ORACLE OF WORD RELATIONS

[7] Available under http://www.dfki.de/~figueroa/
[8] FreeLing 2.1 was utilised as a dependency parser for Spanish.

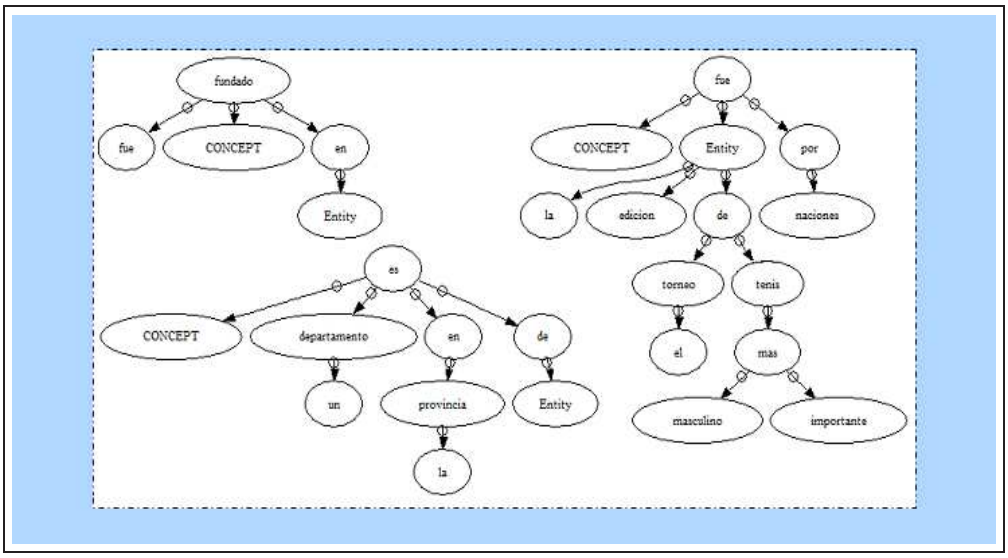

Figure 5.5: Samples of dependency trees obtained for Spanish (source [Figueroa, 2009]).

es→departamento→un
fundado→en→Entity
por→naciones

It is worth noting that dependency paths encapsulate grammatical information about word orderings. However, contrary to the tuples in the previous method, these orderings are not necessarily linear. Since only specific links are taken into account now, the number of tuples declines with respect to the previous model. Table 5.17 lays out this drop.

	Tuples from Fixed Window	Tuples from Dependency Paths	%
Pairs	719,510	243,286	33,81
Triplets	1,161,743	215,119	18,52

Table 5.17: Difference in the number of tuples (source [Figueroa, 2009]).

This method utilises the same scoring function, but the *word association norms* need to be redefined in order to account for tuples originated from the dependency paths:

$$I_2^*(w_1, w_2) = log_2 \frac{P_{link}(w_1, w_2)}{P_g(w_1)P_d(w_2)}$$

Where $P_g(w_1)$ and $P_d(w_2)$ are the probabilities that the word w_1 and w_2 are independently the head and the dependent, respectively. Further, P_{link} is the likelihood of finding the word w_1 as the head of w_2. Homologous with previous word norms, the number of links in the corpus is conceived as the corpus size, when calculating the probabilities. Analogously, $I_3^*(w_1, w_2, w_3)$ is stipulated as:

$$I_3^*(w_1, w_2, w_3) = log_2 \frac{2 * P_{link}(w_1, w_2, w_3)}{P_g(w_1)(P_g(w_2) + P_d(w_2))P_g(w_3)}$$

In a triple, the middle node serves as the dependent and the head of another node, and therein lies the average of both likelihoods in the formula.

Experiments

In the assessment, [Figueroa, 2009] utilised the 53 definition queries corresponding to the CLEF 2007-2008 Spanish QA tracks. Since they could not account for the official records, they took into consideration the 19 and 34 questions[9] in relation to the queries recognised as definitions by the best (INAOE) team in CLEF 2008 and 2007, respectively. Even though CLEF data-sets were considered, a TREC-style evaluation was performed. Consequently, for each question, the retrieved surrogates were manually inspected as a means of creating a gold standard, like [Voorhees, 2003], and nuggets in the ground truth were also equally weighed. As an example, this gold standard embodies the next three nuggets for the *definiendum* "*odometro*" included in CLEF 2007: "*indica el valor y la condicion mecanica de un auto*", "*indica la distancia recorrida*", and "*se coloca en la rueda*".

β	Baseline	MKB-DWQA	MKB-DWQA*
1	0.40 ± 0.24	0.46 ± 0.22	0.39 ± 0.19
2	0.44 ± 0.25	0.48 ± 0.22	0.42 ± 0.20
3	0.46 ± 0.26	$\mathbf{0.49} \pm 0.23$	0.44 ± 0.22
4	0.47 ± 0.27	0.50 ± 0.24	0.45 ± 0.22
5	0.47 ± 0.27	0.50 ± 0.24	0.46 ± 0.23
Precision	0.43 ± 0.31	$\mathbf{0.52} \pm 0.34$	0.41 ± 0.27
Recall	0.48 ± 0.28	$\mathbf{0.51} \pm 0.25$	0.47 ± 0.24

Table 5.18: CLEF 2007 results ($\mathcal{F}(\beta)$-Score) (source [Figueroa, 2009]).

Tables[10] 5.18 and 5.19 stress the outcomes accomplished for each question set. In both cases, MKB-DWQA reached the highest recall. This means tuples emanated from Wikipedia abstracts contributed to identifying additional descriptive information low in frequency, ratifying the finding for English. It is crystal clear that this enhancement was modest, but it is nevertheless mildly encouraging. The achievements are motivating, due to the next two reasons: (a) the number of descriptive sentences utilised for learning tuples is small, and (b) the frequent use of both genders (masculine and feminine) adversely affects the learning models. In English, most nouns have only one neuter form: "*singer*", "*president*" and "*writer*", while few nouns still bear the gender (e.g., "*congressman/congresswoman*"). In Spanish, however, most nouns usually indicate the gender: "*presidente/presidenta*" and "*escritor/escritora*", whereas few are neuter (e.g., "*cantante*"). This difference in noun forms is vital when having few training examples, because adjectives must agree with the number and the gender of the noun:

IMPACT OF MORPHOLOGY

... intelectual y escritora francesa autora de ...
... filosofo y escritor frances ...
... primera mujer elegida ...
... el primer hombre japones en ...

Incidentally, learning these *word association norms* raises the issue of orthographical variations in Spanish. The meaning of words can substantially change if they are written with their respective orthographic accents or not. Some simple cases are "*corte*" and "*rio*" as well as "*ejercito*". Spanish speakers, however, are likely to omit the orthographical accent when

ORTHOGRAPHICAL VARIATIONS

[9] We thank the INAOE team for supplying these question sets.

[10] A star is used for distinguishing the system that benefits from the repository of tuples distilled from the dependency trees.

they write on blogs, web documents, or Wikipedia articles. The reason why they leave out this accent is that they are normally unnecessary, because the context usually yields enough information to readily disambiguate the correct meaning. For this reason, together with the fact that the tuples represent contextual relations of words, tuples were inferred omitting the accent. Another final aspect regarding orthographical variations is misspellings; interchanging "*c*" with "*s*", or "*v*" with "*b*" is very common in Spanish. But unfortunately, this sort of variation is harder to correct, and has an effect on the norms, because the "new" term can already exist in the Spanish lexicon.

β	Baseline	MKB-DWQA	MKB-DWQA*
1	0.37 ± 0.18	0.32 ± 0.20	0.38 ± 0.29
2	0.43 ± 0.20	0.33 ± 0.18	0.40 ± 0.22
3	$\mathbf{0.47} \pm 0.22$	0.34 ± 0.18	0.41 ± 0.22
4	0.48 ± 0.23	0.35 ± 0.18	0.42 ± 0.22
5	0.50 ± 0.24	0.35 ± 0.18	0.43 ± 0.22
Precision	$\mathbf{0.55} \pm 0.23$	0.35 ± 0.18	0.44 ± 0.23
Recall	0.37 ± 0.21	$\mathbf{0.40} \pm 0.33$	0.38 ± 0.29

Table 5.19: CLEF 2008 Results ($\mathcal{F}(\beta)$-Score) (source [Figueroa, 2009]).

To a limited extent, this problem can be lessened by means of a morphological analyser, such as FreeLing. However, it is worth remarking that FreeLing does not offer a mapping to a "standard" form for all words (the reader can verify this by trying the given examples). In light of this observation, it can be reasonably deemed that boosting the performance will demand considerable efforts. These efforts will go into deeper linguistic processing, and at the same time, collecting a larger set of descriptive sentences. Contrary to English, where Wikipedia supplies a considerably larger collection and only one gender is predominately used.

Results obtained by MKB-DWQA* do not reflect a definite improvement in terms of recall. The reason for this is two-fold: (a) important tuples were eliminated, when reducing the models, and (b) the dependency paths computed from the candidate set of sentences did not align the paths in the models. As a means of verifying this conclusion, the average amount of matching tuples between the second and third methods were observed: 124 pairs and 57 triplets for MKB-DWQA, while 20 pairs and 2 triplets for MKB-DWQA*. One can consider two reasons for this mismatch. Firstly, errors in the output of the parser. Sentences gathered from Wikipedia are much more well-formed than occasionally truncated phrases within web snippets. Secondly, longer dependency paths might be indispensable to model the lexical relationships necessary to characterise definitions. However, dealing with these two issues would bring about a significant growth in the retrieval and processing times. PERFORMANCE

With regards to precision, results markedly varied from CLEF 2008 to CLEF 2007 and they are thus not conclusive. As a means of drawing interesting conclusions concerning precision, the Mean Average Precision (MAP) (see section 1.7 on page 15) of the top ranked and the top three ranked sentences were computed (accounting for "*Precision at one and at three*", respectively). Table 5.20 highlights these achievements. MAP

The obtained MAP scores show that using the tuples effectively contributes to ameliorating the ranking of the sentences. Essentially, they help to bias the ranking in favour of descriptive sentences that have some lexico-syntactic similarities to sentences in Wikipedia abstracts. A positive aspect of this enhancement in ranking is that the methods are aimed at selecting sentences that yield the more novel and representative content. That is, these three selected sentences are very likely to convey different information, or in the worst case,

	Baseline	MKB-DWQA	MKB-DWQA*
MAP-1	0.62	0.69	0.65
MAP-3	0.58	0.66	0.62

Table 5.20: MEAN AVERAGE PRECISION (source [Figueroa, 2009]).

different paraphrases of the same underlying ideas. This difference can also include several senses. The achieved results hold promise because of the small amount of training sentences.

Interestingly enough, the MAP values obtained by MKB-DWQA* lie in the middle between BASELINE and MKB-DWQA. This pattern in the outcomes is produced by the crystallisation of a gradual reduction in the number of matching tuples between the models and the testing sentences. Put differently, the comparatively fewer tuples entailed from the dependency trees generate a relatively modest enhancement in comparison to the larger amount of tuples derived by means of traditional *word association norms*. Nonetheless, these few matches materialised a tangible improvement, which also suggests that traditional norms are better to tackle data sparseness.

One critical facet of definition QA systems, in particular search engines, is the MAP for the top ranked sentence. In this aspect, MKB-DWQA outperformed the other two strategies, ranking a valid definition on the top in 69% of the cases. To conclude, a list of top-ranked definitions found by MKB-DWQA is presented:

- **Le Corbusier** fue uno de los miembros fundadores del Congreso Internacional de Arquitectura Moderna e hizo famoso el llamado estilo arquitectonico internacional.
  copypeist.com/2005/08/28/33/

- Gustave **Flaubert** nacio el 12 de diciembre de 1821, en Ruan, Normandia, y murio el 8 de mayo de 1880, en Croisset.
  es.answers.yahoo.com/question/index?qid=20090105120718AABaKHd

- La **revolucion de los claveles** es el nombre dado al levantamiento militar del 25 de abril de 1974 que provoco la caida en Portugal de la dictadura salazarista que dominaba ..
  www.estrelladigital.es/ED/diario/51162.asp

- **Marco Pantani** nacio en Cesena, Italia, el 13 de enero de 1970 y debuto como profesional en el Gran Premio de Camaiore para fichar con el equipo Carrera en el que milito desde...
  www.esmas.com/deportes/otrosdeportes/343766.html

- La **tarantela** es un baile popular del sur de Italia y, por lo tanto, posiblemente de las regiones italianas de Apulia, Basilicata, Calabria, Molise, Campania o Sicilia.
  video.aol.com/video-detail/jascha-heifetz-scherzo-tarantella/1919925688/?icid=VIDURVHOV07

- **INTASAT** es el primer satelite artificial cientifico español.
  valija-viaje.boonic.com/

- En la **escala de Mohs**, que indica la dureza de los materiales de 1 a 10, el zafiro ocupa la novena posicion por dureza (el diamante tiene 10).
  www.sobrerelojes.com/TECNICA/relojes-elcristal.htm

- **Leica** es una casa alemana dedicada a la fabricacion de instrumentos opticos de precision.
  es.wordpress.com/tag/leica/

- **Odessa** es la tercera ciudad mas grande de Ucrania despues de Kiev y Kharkov, un importante industrial, cultural, cientifico y recurrir centro en el norte de la region del Mar.
  www.articleset.com/Recorrido-y-ocio_articles-277_es.htm

- La **vexilologia** es la ciencia que se encarga del estudio de las banderas en todas sus variantes: guiones, estandartes, banderines, vexiloides, etc.
  vial.jean.free.fr/new_npi/revues_npi/17_2000/npi_1700/17_spai_vexillo.htm

- Los **pellets** son un nuevo tipo de combustible fabricado de una forma similar a las briquetas de madera, de los desechos de la madera y por medio del prensado...
  www.atmos.cz/spanish/paliva-energie

- **Rafael Azcona**, que fallecio el pasado 25 de marzo a los 81 años de edad a causa de un cancer de pulmon, es uno de los guionistas mas relevantes en la historia del cine.
  actualidad.terra.es/cultura/articulo/se_rafael_azcona_2372591.htm

5.4 Conclusions

In general terms, this chapter lays out two distinct approaches to multilingual definition QA systems operating on web snippets. In the first place, this deals with an answering technique based largely on redundancy. This method infers statistics from the set of answer candidates extracted from web snippets. These statistics are grounded on term frequencies, potential named entities, and content words. These content terms are determined from the semantic space provided by LSA, and they are particularly interesting as they are likely to be potential signals of essential characteristics of the *definiendum*. Above all, the multilinguality of this system lies in these statistics and its poor linguistic knowledge. This means this language portability is at the expense of avoiding the exploitation of linguistic tools. Some findings are the following:

1. Contrary to [Cui et al., 2004c], empirical results support the fact that web snippets can be a productive source of descriptive information. This challenge is supported by two evaluations: (a) web snippets yielded at least one nugget to a large fraction of the queries embraced by five testing question sets; and (b) The performance of this system is also competitive, even though its output is juxtaposed with the gold standard emanated from the TREC 2003 corpus. This was possible because web snippets embodied numerous nuggets within this ground truth.

2. In substance, results reveal that statistics collected from the candidate set cooperates on recognising the most statistically relevant descriptions. However, extra resources are indispensable for languages other than English, like Spanish. The figures for Spanish are, nonetheless, positively encouraging with relation to coverage and the CLEF ground truth. More importantly, these statistically relevant descriptions were identified without grabbing “annotated" descriptive knowledge produced by KBs.

3. On the whole, the outcomes show the significant impact of the redundancy of information across both the English and Spanish Web; in particular, the correlation of terms with the *definiendum* across answer candidates that match definition patterns.

On a different note, M-DefWebQA pioneers attempts by definitional QA systems to disambiguate descriptive utterances. One finding is that web snippets do not provide the necessary information for a complete disambiguation. To overcome this difficulty, external resources such as full documents, WordNet and/or extra queries could be explored as a source

for fetching extra information from the Internet. Nevertheless, this remains a very difficult problem as many *definienda* can bear slightly different senses (e.g., *George Bush*).

In the second place, the previous definition QA system was enriched with definition knowledge originated from Wikipedia abstracts, namely sentences matching definition patterns. More specifically, this enrichment encompassed templates and *word association norms* deduced from this array of training sentences. Contrary to the trend in TREC, this system benefits from these resources in an anonymous way, that is by learning common regularities across all descriptive sentences coming from all articles in place of accessing solely to articles connected with each particular *definiendum*.

The bottom line is that the positive advantage of these models is two-fold: (a) they can operate independently on the variations among *definienda* in terms of KB coverage; and (b) they were almost transparently employed to tackle definition questions in English and Spanish. Some conclusions in this regard are:

1. Overall, results demonstrate that the training corpus was instrumental for discerning nuggets that are low in frequency, and ergo undetected by the first system. Two ramifications are: (a) it is possible to automatically acquire a corpus of positive samples that can aid in building models capable of distinguishing these nuggets low in frequency; and (b) this procedure can port to Spanish, however, the figures suggest that the problem of data sparseness became graver.

2. Results also shed light on the pivotal role of syntactic information in the process of rating answer candidates. Additional syntactic knowledge can cause the proliferation of rewritings of the same descriptions in the output, put differently, an increment in the level of redundancy in the final response. Here, an extra challenge is recognising essential morpho-syntactical variations of descriptive sentences, which would help to decrease the redundancy of the output. At any rate, this redundancy can still be useful for discovering answers to definition queries in the context of the TREC/CLEF QA tracks, by projecting these redundant phrases to the corresponding corpus.

3. By the same token, *word association norms* showed to be a cost-efficient solution for inferring shallow pertinent lexico-syntactic relations, wherewith putative answers can be rated afterwards.

A final remark on multi-linguality regards the fact that only English and Spanish were the target of the experiments. Multi-linguality in definition QA is also an interesting research field, and the reader can dig deeper into this by referring to, for instance, [Sacaleanu et al., 2007, 2008] to study systems in other languages, including German.

Chapter 6

Using Language Models for Ranking Answers to Definitions Questions

> "But it must be recognized that the notion 'probability of a sentence' is an entirely useless one, under any known interpretation of this term." (Noam Chomsky)
>
> "Dependence is the norm rather than the contrary." (B. De Finetti, Theory of Probability)

6.1 Introduction

Chiefly, the discussion on chapter 4 was centred in Question Answering (QA) systems that profit from Knowledge Base (KB) articles about *definiendum* for finding answers to definition questions across the AQUAINT corpus. The prominent feature that these systems learn, consists of word frequencies (correlation) counts. These counts are projected into the set of answer candidates by means of purpose-built heuristics. Habitually, both resources and attributes play the vital role of modelling the positive evidence that heuristics exploit for scoring candidates.

In recent years, a lot of effort has gone into devising more efficient projection strategies. In this respect, incipient methodologies have been contemplating the use of Language Models (LM) for ranking answer candidates. Intrinsically, the reason to prefer LMs to other techniques lies in the fact that it is relatively easy to acquire positive samples (i.e., KB articles about the *definiendum*). In the opposite way, there is no dependable source of negative samples. Conventionally, their acquisition demands considerable efforts, normally in terms of manual annotations. Moreover, an additional reason that makes LMs a propitious tool is their efficiency in speech recognition and other related QA tasks [Zhai and Lafferty, 2004, Cui et al., 2007].

In other words, systems are moving from utilising heuristics to well-known statistical methods that can learn solely from one (positive) class. This shift is not only with the hope of enhancing the performance of QA systems, but also with the purpose of benefiting from widely known techniques, in such a way that they can deepen the understanding of the underlying linguistic phenomena behind definition QA.

Another focus of attention is broadening the coverage of KBs, since most approaches bank heavily on them, when constructing their models. In so doing, systems try to discover localised contexts carrying the *definiendum* across the Internet, and/or they also develop wrappers for a systematically augmenting number of KBs. However, latest strategies attempt to

capitalise on specific models of definitions, which are used in agreement with the type of *definiendum* in question. Typically, these models comprise persons, organisations and things, and they are gathered from KBs by setting apart articles targeting the respective classes. For the most part, this seems to be the most promising trend as it attenuates the sharp variations in or dependency on the coverage supplied by KBs for each *definiendum*.

From another angle, the trend of definition QA systems takes advantage of unigrams as attributes. In this regard, [Chen et al., 2006] studied the effect of unigrams and bigrams as well as biterms on LMs. In a statement, their results suggest that flexible word orderings are more salient features for rating answer candidates. The exploitation of *word association norms* shown in section 5.3 (page 128) also ratifies the importance of more informative properties.

This chapter addresses LMs for ranking putative answers to definitions questions, and it is organised as follows. The next section goes over the best Text REtrieval Conference (TREC) 2007 system. Singularly, this definition QA system merges the outcome of four different LMs for scoring candidate answers. Subsequently, section 6.3 contrasts the performance achieved by LMs combined with three distinct features (i.e., unigrams, bigrams, biterms), and posteriorly, section 6.4 touches on a strategy that integrates assorted LMs. More interestingly, this approach encompasses a definition model that synthesises: (1) one dependent on the type of *definiendum* (i.e., person, organisation, and thing); and (2) another relying on general definitions. Additionally, this framework is put together with a topic model homologous to the traditional models inferred from articles about the *definiendum* across KBs and/or the target collection of documents. Later, section 6.5 extends the idea of three specific to context LMs, which are automatically constructed on top of dependency tree representation of the training material. Ergo, they make allowances for richer syntactic information than shallow biterm and/or unigram attributes. A crucial aspect of *context models* is its over 40,000 distinct specific models, wherewith they reduce their reliance upon KBs. Lastly, section 6.6 draws some conclusions.

6.2 Language Models and Dependency Relations

In TREC 2007, the best run in the definition QA track belonged to the system implemented by [Qiu et al., 2007]. Their strategy ranked answer candidates by intermixing the influence of three distinct kinds of ingredients:

LMS

(a) Features learnt from LMs. This type branches into four specific properties corresponding to the probability of a sentence in conformity with LMs learnt from four training corpora: (1) AQUAINT+AQUAINT2, (2) the previous two collections, but in this configuration, they accounted for the selective substitutions of numbers and entities such as locations, people, and organisations as well (these replacements share the same spirit with the ones illustrated in table 4.12 on page 93), (3) articles related to the *definiendum* emanated from Wikipedia, and (4) one hundred snippets fetched from a commercial search engine. It is worth noting here that in these two configurations (3) and (4), [Qiu et al., 2007] made use of Dirichlet smoothing for the purpose of tackling the data sparseness of the LMs obtained from Wikipedia articles and web snippets.

FOUR DEPENDENCY RELATIONS

(b) Four types of syntactic dependency relations. This second sort of feature is premised on lexical relations outputted by Minipar[1]. Specifically, [Qiu et al., 2007] benefited from the chi-square test for selecting four class of dependency relations as binary attributes: "*punc*", "*appo*", "*pcomp-n*" and subject "*s*". Every time any of these four types of relations

[1]Minipar is available under: www.cs.ualberta.ca/∼lindek/minipar.htm

appears in a candidate answer, the respective feature turns to one if and only if: a term in the relation is not a stop-word, and this is contained in the *definiendum*, while at the same time, the other term must not belong to the *definiendum*. In any other case, the respective feature value is set by default to zero.

(c) The ranking value returned by the Information Retrieval (IR) engine. IR RANKING

In the TREC 2007 challenge, the best response of this system reaped an average $\mathcal{F}(3)$-Score of 0.329. The mean of all responses in this track was 0.118. Another aspect worth stressing is that their redundancy removal methodology checks the word overlap of the next answer candidate with all previously selected answers. The best run, at any rate, did not contemplate a redundancy removal step. PERFORMANCE

6.3 Unigrams, Bigrams or Bi-terms

In their work, [Chen et al., 2006] studied the influence of three types of ingredients on LMs designed particularly for ranking definitions. These three classes of features comprise: unigrams, bigrams, biterms as well. Generally speaking, unigram LMs assume that each word occurs within the answer candidate independently. This assumption is, however, barely truthful, and therein lies the attempt of [Chen et al., 2006] to test two extra models that also account for the local context when rating putative answers. In effect, the difference between bigram and biterms is that the former perceives pairs of consecutive tokens as attributes, while the latter loosens the order constraint. In practical terms, they took advantage of the *min-Adhoc* method, proposed by [Srikanth and Srihari, 2002], for estimating biterms probabilities: N-GRAMS BI-TERMS

$$P_{BT}(w_i|w_{i-1}) = \frac{Counts(w_{i-1}, w_i) + Counts(w_i, w_{i-1})}{min\left\{Counts(w_{i-1}), Counts(w_i)\right\}}$$

More exactly, [Chen et al., 2006] induced the values of P_{BT} from the ordered centroid representation of training sentences coming from web snippets. With regard to unigrams and bigrams, [Chen et al., 2006] utilised the maximum likelihood estimates (see section 4.9 on page 93). The web snippets used for training their models were fetched by expanding the query with task specific cues (see details in section 2.4 on page 35). Subsequently, they singled out the top-350 stemmed words that co-occur with the *definiendum* across the fetched surrogates in accordance with the centroid vector (see equation 4.8 on page 92). Training (web) sentences are posteriorly rewritten by making use of these 350 stemmed terms and preserving the original order. To illustrate this, [Chen et al., 2006] brought out the following example: ORDERED CENTROID VECTOR

Today's Highlight in History: On November 14, 1900, Aaron Copland, one of America's leading 20th century composers, was born in New York City.

The corresponding ordered centroid vectors become the words:

November 14 1900 Aaron Copland America composer born New York City.

Note that in this representation stop-words are ignored and terms are stemmed. LMs are accordingly built on top of this abstraction of the training sentences. These models are used thereafter for ranking an answer candidate A with n tokens t_i as follows:

	F(5)-Score with Query Expansion	F(5)-Score without Query Expansion
unigrams	0.472	0.459
bigrams	0.518	0.505
biterms	0.531	0.511

Table 6.1: Comparison of LMs coupled with unigrams, bigrams and biterms (source [Chen et al., 2006]).

$$P(A|\mu) = P(t_1|\mu) * \prod_{i=2}^{n} (\lambda * P(t_t|\mu) + (1-\lambda) * P(t_t|t_{t-1}, \mu))$$

BREVITY PENALTY

In this equation, μ denotes the LM inferred from the ordered centroid vector, and λ is the mixture weight, which is approximated by the Expectation Maximisation (EM) iterative procedure (see equation 4.9 on page 98). The final score of an answer candidate A is determined by weighing the logarithm of this factor with a brevity penalty:

$$Score(A) = log(P(A|\mu)) * exp(min(1 - \frac{L_{ref}}{L_A}, 1))$$

L_{ref} and L_A stand for the length of a reference and the candidate answer, respectively. Table 6.1 parallels the figures achieved by the three different features, when assessed by means of the TREC 2003 definition question set. This table also contrasts the performance of this system with and without the enrichment supplied by the query expansion method. For the most part, bi-gram LMs were observed to significantly improve the quality of the extracted answers. Furthermore, bi-term language models yield better results, showing that flexibility and relative position of lexical terms capture shallow information about their syntactic relation [Belkin and Goldsmith, 2002].

WORD ORDER

6.4 Topic and Definition Models

A different approach is due to [Han et al., 2006]. They modelled the definition QA task from two viewpoints: topic and definition. The difference between both models is illustrated by the following examples:

(a) John Hoyer Updike (March 18, 1932 - January 27, 2009) was an American novelist, poet, short story writer, art critic, and literary critic.

(b) **Danielle Fernande Dominique Schuelein-Steel** (born on August 14, 1947), better known as just Danielle Steel, is an American romantic novelist and author of mainstream dramas.

(c) In 1989, **Danielle Steel** was listed in the Guinness Book of World Records for having a book on the New York Times Bestseller List for the most consecutive weeks of any author-381 consecutive weeks at that time.

(d) **Danielle Steel** has a new book coming out! What a delight for her millions (and millions) of fans!.

(e) Never date a girl who reads **Danielle Steel**.

(f) Multiple editions of Windows are a horrible idea in the first place, but the anytime upgrade is a good idea.

The definition model is aimed specifically at rating answer candidates in congruence with their evidence of being definitions regardless of their relatedness to the *definiendum*. Inversely, the topic model is geared towards scoring candidate sentences in agreement with their evidence of relatedness to the *definiendum*, leaving aside their evidence measured in concert with the definition model. The following table clusters the working sentences in accordance with these two criteria:

DEFINITION AND TOPIC MODELS

	Definition (D)	Non-Definition ($\bar{D}$)
Topic (T)	b, c	d, e
Non-Topic ($\bar{T}$)	a	f

Table 6.2: Groups of sentences according to [Han et al., 2006].

In this working example, candidate sentences (b) and (c) put into words descriptions of the *definiendum* (i.e., "*Daniel Steel*"). Sentence (a) is a definition, but it is delineating another *definiendum*, whereas sentences (d) and (e) are related to the *definiendum*, but the information they verbalise is not descriptive. Lastly, sentence (f) is both non-definitional and out-of-topic. Desirable answers are therefore candidate sentences (b) and (c). Consequently, [Han et al., 2006] rendered the definition QA task as the maximisation of the joint probability $P(T, D|A)$, that is the likelihood that a sentence A is a definition and belongs to the topic of the *definiendum*. An interesting conjecture of this approach is that it assumes the likelihood of being a definition as independent of belonging to the topic. For instance, "*<definiendum> is a British politician*" is a definition regardless of the *definiendum*. This assumption helps to rewrite and simplify $P(T, D|A)$ as follows:

$$P(T, D|A) \approx \frac{P(T)P(A|T)}{P(A)} * \frac{P(D)P(A|D)}{P(A)}$$

Since $P(T)$ and $P(D)$ do not affect the ranking of a set of candidate sentences, this equation can be reduced to the following expression:

$$DS(A) = \frac{P(A|T)}{P(A)} * \frac{P(A|D)}{P(A)}$$

In light of this formula, [Han et al., 2006] specified $P(A|T)$ as the topic model, $P(A|D)$ as the definition model and $P(A)$ the general LM. To put it more precisely, the general LM $P(A)$ was approximated by working out the product of the unigram probability of each word in A. These unigram probabilities are estimated by the Maximum Likelihood Estimate (MLE) based on the entire target collection (see section 4.9 on page 93). The topic LM is predicated also on unigrams, and it linearly combines the evidence originated from three distinct kinds of sources:

UNIGRAMS

$$P(A|T) = \prod_{i=1}^{n}(\alpha * P(w_i|R) + \beta * P(w_i|E) + \gamma * P(w_i|W))$$

Where α, β and γ are the interpolation parameters, which are empirically set so that they add one. The probabilities $P(w_i|R)$, $P(w_i|E)$ and $P(w_i|W)$ are determined from the following three different sources of descriptive evidence:

INTERPOLATION

- $P(w_i|R)$ is the likelihood that the word w_i is generated from the five top-ranked documents R fetched from the target collection of documents by an Information Retrieval (IR) engine. This probability is estimated by means of the MLE and using Dirichlet smoothing [MacKay and Peto, 1994]:

$$P(w_i|R) = \frac{Counts(w_i) + \mu * P(w_i)}{\sum_{\forall w_j} Counts(w_j) + \mu}$$

  Here, $P(w_i)$ is the probability of w_i in consonance with the general model.

- In like manner, $P(w_i|E)$ and $P(w_i|W)$ are computed from a group of articles mined from a battery of authoritative KBs, and a group of ten web pages downloaded from the Internet using Google search engine. The aim of making allowances for web pages is alleviating the data sparseness that typifies the previous two models. The battery of KBs embraced:

KBS

  - Acronym Finder
  - biography.com
  - Columbia Encyclopedia
  - Wikipedia
  - FOLDOC
  - The American Heritage Dictionary of the English Language
  - Google glossary (see section 2.3 on page 30)
  - Online Medical Dictionary

  The reader is encouraged to consult table 2.4 (page 31) in order to compare this list of KBs with the resources utilised by other TREC definition QA systems.

The smoothing parameter μ is optimally set between 500 and 10,000, and the performance is robust when it is equal to 2,000 [Zhai and Lafferty, 2004]. For the purpose of building the definition LM, [Han et al., 2006] gathered definition articles for arbitrary *definienda* from KBs. With this corpus they created one general model and three domain-specific models: person, organisation and term. Candidate sentences are next ranked in agreement with the interpolation of the domain model with respect to the *definiendum* and the general model:

$$P(A|D) = \prod_{i=1}^{n} (\lambda * P(w_i|D_{<definiendum>}) + (1 - \lambda) * P(w_i|D_{general}))$$

RANKING

Where $P(w_i|D_{<definiendum>}$ and $P(w_i|D_{general})$ are the likelihood of finding the word w_i in the domain-specific and general corpus, respectively. These probabilities are estimated similarly to $P(w_i|R)$, and λ is an experimental interpolation parameter. For building these models, they constructed a definition corpus from their KBs (utilised for topic modeling) encompassing 14,904 people, 994 organisations and 3,639 terms. All things considered, [Han et al., 2006] ranked answer candidates gathered by means of the patterns utilised by [Han et al., 2005] in congruence with the next formula:

$$Score(A) = \sum_{i=1}^{n} log(\alpha * P(w_i|R) + \beta * P(w_i|E) + \gamma * P(w_i|W)) +$$
$$log(\lambda * P(w_i|D_{<definiendum>}) + (1 - \lambda) * P(w_i|D_{general})) - 2 * log(P(w_i))$$

The logarithm is used for coping with long sentences. Answers are chosen until a maximum length is satisfied and answers are discarded if they share many words with an already selected answer. In some cases, they also used WordNet for discarding additional redundant answers. In their evaluation, [Han et al., 2006] accounted for the 50 and 64 questions and gold standards supplied by the TREC 2003 and 2004 challenges, respectively. They extended the TREC 2004 ground truth by appending the answers to the factoid questions with relation to the context of the definition question. This makes sense because these factual pieces of information can be seen many times as parts of a response to a definition query.

ANSWER LENGTH

They carried out several experiments[2] so that they could obtain the best configuration of parameters that deals more efficiently with definition questions. First, they set λ and μ to 0.6 and 2,000, respectively. Second, they adjusted the values of α, β and γ, discovering that the best combination of values for the TREC 2003 set is 0, 1 and 0, respectively. This configuration reached an average $\mathcal{F}(3)$-Score of 0.3718 (recall = 0.4540; precision = 0.1990). Different settings obtained the best performance for the TREC 2004 data set: α=0.25, β=0.60 and γ=0.15. This setting finished with an average $\mathcal{F}(3)$-Score of 0.4211 (recall = 0.4511; precision = 0.2790). Homologously to [Zhang et al., 2005], the parameters, on the whole, reveal that their models depend heavily on the information supplied by KBs about each single *definiendum*. In deed, [Han et al., 2006] noticed that their pack of KBs yielded descriptions for 46 out of the 50 TREC 2003 definition questions, while solely for 55 out of the 64 definition TREC 2004 questions. The considerable change in the values of the parameter set goes hand in hand with the slight decrease (from 92% to 86%) in the coverage offered by the KBs, that is the less coverage they provide, the greater need of increasing the influence of other classes of resources. At any rate, [Han et al., 2006] conjectured that the help given by top ranked documents and web pages is insufficient.

PERFORMANCE

KBS COVERAGE

Incidentally, [Han et al., 2006] also tuned the value of λ. The best configuration set this value to one, signalling the relevance of the *definiendum* type when rating answer candidate. In the case of the TREC 2003 question set, this configuration achieved an average $\mathcal{F}(3)$-Score of 0.3691 (recall = 0.4506 and precision = 0.1983), while an average value of 0.4194 for the TREC 2004 dataset (recall = 0.4491; precision = 0.2784). One interesting finding is that the word distributions amongst their different domain-specific models markedly varies, which is deemed to be the reason for the relevance of the type of *definiendum* indicated by the empirical values of λ. Lastly, it is worth remarking that [Han et al., 2006] took advantage of the BBN Indetifinder [Bikel et al., 1999] for recognising the type of *definiendum*. The *definiendum* is labelled as a term, whenever it is not classified as a person or organisation by this tool.

definiendum-TYPE

6.5 Contextual Models and Descriptive Dependency Paths

A chief attribute that transpires most of definition QA systems that compete in the TREC challenge is their exploitation of the evidence derived from KBs. Most of the time, this evidence is the cornerstone of their answering strategy (a detailed list of systems and the KBs they exploit can be seen in table 2.4 on page 31). Essentially, in their analysis, [Zhang et al., 2005] revealed that the performance of this class of definition QA system depends heavily on the coverage supplied by external KBs. As [Zhang et al., 2005] also pointed out, this reliance on one particular resource can be mitigated by accounting for a larger number of KBs as they are very likely to express descriptions for different *definienda*. A natural and tangential consequence of augmenting the amount of KBs is that they can convey different paraphrases of the same descriptions, and it thus boosts the chances of matching actual definitions in the

KBS COVERAGE

KBS ADVANTAGES

[2] It is necessary to recall the fact that [Han et al., 2006] specified the $\mathcal{F}(3)$-Score in a different manner than the TREC 2003 and 2004 evaluations.

target (AQUAINT) corpus. A third advantage is that it also cooperates on discriminating the most pertinent facets on the grounds of redundancy across the respective articles in the KBs.

Following this trend, [Han et al., 2006] built their topic model on top of word distributions acquired from eight different KBs. Like other notable definition QA systems including [Qiu et al., 2007], they also tackled the data-sparseness of KBs by balancing this evidence with information emanated from web-pages and the target group of answer candidates as well as top-ranked documents retrieved from the target collection. Certainly, based on the findings of [H. Joho and M. Sanderson, 2000, 2001], making allowances for localised contexts of the *definiendum* makes perfect sense because descriptions can accompany instances of the *definiendum*, in particular when it is introduced in a piece of news for first time[3] (see section 4.2 on page 76). At any rate, the optimal combination of parameter values found by [Han et al., 2006] corroborates the findings of [Zhang et al., 2005], that is, the performance of their topic model relies largely on the evidence coming from KBs. On a different note, [Han et al., 2006] also underlined the fact that web pages and top-ranked documents fetched from the target collection are insufficient to tackle the data-sparseness that characterises KBs.

KBS DISADVANTAGES

Of course, using KBs facilitates the construction of systems grounded on heuristics that are capable of finding the top-relevant facts about the *definiendum*. But, in general, there are five factors that make this type of technique less attractive:

1. Studies indicate that the performance falls into a decline, when the supplied coverage is restricted or no evidence is found at all [Zhang et al., 2005, Han et al., 2006]. This is due basically to the fact that this class of system redefines this answering process as a "*more like this*" task. Therefore, when narrow coverage is found in KBs, this "*this*" is not well-defined, materialising a marked drop in the performance.

2. Even though this "*this*" can be well-defined by the KBs, the performance of this class of definition QA systems is limited because they fail to recognise descriptions dissimilar to the contexts supplied by the respective KBs articles.

3. The data-sparseness that typifies KBs inhibits the exploration of more efficient features that can be utilised for ranking answer candidates. That is, definition QA systems are restricted in the abstractions they can make in order to employ more complex structures (e.g., syntactic regularities, entities, trees and semantic classes) to recognise genuine answers from a target set of sentences. Although some systems do infer them from a large corpus of definitions [H. Joho and M. Sanderson, 2000, 2001, Blair-Goldensohn et al., 2003, Fernandes, 2004, Cui et al., 2007] (see sections 4.2 and 4.9 on pages 76 and 93, respectively), these inferences are normally in the form of constructs that are too general for recognising a diverse group of descriptions, or too specific to be highly frequently found in natural language texts. At the same time, these regularities do not yield a perfect accuracy, definition QA systems need to, therefore, combine them with extra evidence. All thing considered, this sort of system is fostered to rely chiefly on exploiting word overlaps.

4. Ultimately, it is the target collection of documents the one which rules the contexts and the senses of the *definiendum*, not a particular pack of KBs. This kind of assumption renders the problem the other way around.

5. The constrained use of attributes prevent shedding some light onto the understanding of the linguistic phenomena behind descriptions.

[3] The reader has to bear in mind that the target collection of these systems is the AQUAINT corpus, and this encompasses an array of news articles.

Contrary to [Han et al., 2006], and like several systems such as [Figueroa and Neumann, 2007] (see section 2.5 on page 36), [Chen et al., 2006] tried to allay this coverage problem by collecting descriptions coming from web snippets by means of a query expansion method that biases the search engine in favour of contexts that are probable to render descriptions of the *definiendum* (see details in section 2.4 on page 35). Their outcomes indicate that a powerful search technique, focussing on increasing the recall of localised contexts bearing descriptive information about the *definiendum* can cooperate on identifying more reliable distributions of words, ergo ameliorating the performance of their system.

In essence, the query expansion of [Chen et al., 2006] assisted in learning the top 350 stemmed terms ranked in conformity with the centroid vector (see section 4.8 on page 91), in this way they distinguished the array of words that are more probable to elucidate various facets of the *definiendum*. Later, they used these 350 terms to build an ordered centroid vector representation of each sentence (that embodies the *definiendum*) found across the retrieved web snippets. LMs were thereafter deduced on top of this representation.

BIGRAMS AND BITERMS

The attractive aspect of the models adopted by [Chen et al., 2006] is that they capture local shallow syntactic regularities. These clues are predicated on bigrams and biterms, which are possible to detect due to the rise in recall generated by their query expansion method. Results demonstrated that these regularities bring about an enhancement in performance. To reinforce how these regularities might help, consider the following sentence regarding "*American author Danielle Steel*"[4]:

According to **author** M. Kelleher, Danielle Steel's book about her son's illness has been a success in **America**.

A method grounded on simple stemmed word overlaps would find this sentence very likely to be an answer candidate, because the terms "*American*" and "*author*" are very likely to be found within localised and descriptive contexts about "*Danielle Steel*". Furthermore, words such as "*book*" and "*success*" could also be very delineative of "*Danielle Steel*". They can also cause this sentence to reach a very high score. The underlying idea of [Chen et al., 2006] is privileging sentences that also show syntactic regularities, for instance:

Danielle Steel is an **American author**.

As proven by [Chen et al., 2006], preferring the match of these shallow syntactic rules to simple term overlaps materialises a significant enhancement in performance. These regularities, however, are not the magic bullet that solve this problem. Consider the following case:

According to **American author** M. Kelleher, Danielle Steel's book about her son's illness is fascinating.

Another issue that makes this technique less attractive is the threshold of 350 stemmed terms. On the one hand, it assists in identifying the words that outline the most critical facets of the *definiendum*. On the other hand, in many cases, more terms can be necessary to express all pertinent facets. For instance, biographical people such as presidents, or writers with a long list of awards and works. Another aspect is that this threshold can cut-off several potential senses of the *definiendum* that are not prominent in the target corpus and/or the web snippets. Certainly, this problem is lessened by the size of the AQUAINT corpus, but when

DATA SPARSENESS

[4]For the sake of simplicity, only the string "*America author*" is deemed as the sole ordered centroid vector.

porting these strategies to a massive collection like the Web, the ambiguity of the *definiendum* can be considerably boosted. Another final thing is that some prominent paraphrases can "take over" the ordered centroid vector, causing a redundant output (e.g., "*American writer*" and "*author born in the U.S.*"). On the whole, this threshold can worsen the informativeness and diversity of the output presented to the user.

In juxtaposition to [Chen et al., 2006], [Han et al., 2006] profited solely from unigrams as features, but their approach mixes their topic with their definition model. This definition model puts together a general with a *definiendum*-type model. While the general model seeks to recognise expressions that are common across definitions of wide-ranging topics, the *definiendum*-type model comprises specific models for people, organisations and terms. There are two conclusions from this approach worth emphasising: (a) [Han et al., 2006] noticed that the distribution of words sharply differentiates from one type of *definiendum* to the other, and (b) the *definiendum*-type model greatly contributed to bettering the performance of their system. Still yet, there are two issues that make this technique less attractive:

TREC-TYPES MODELS

1. As a matter of fact, [Han et al., 2006] took advantage of an external tool for determining the type of *definiendum* prior to ranking. As aforementioned, a *definiendum* can bear various types corresponding to numerous senses, thus rating answer candidates would entail using several models accordingly. For instance, "*Calvin Klein*" can involve a "*person*", a "*company*" and/or a "*trademark*". Nonetheless, all potential senses can only be known after determining the set of answer candidates, and consequently the type identified by an external tool will not necessarily match the predominant sense in the target collection.

2. Grouping several types of *definienda* under the model of "*term*" somehow resembles the general model and does not take advantage of the finding that the distribution of words of distinct types of *definienda* widely vary. In effect, one can reasonably expect to find markedly divergent word distributions for types such as: "*disease*", "*song*" and "*company*", as well as "*language*".

One of the key motivations behind *context models* is the finding of [Chen et al., 2006]: the amalgamation of the semantic evidence (centroid words) with syntactic information (relative order of centroid words) causes an enhancement in performance. In light of this result, it can be concluded that syntactic information plays an essential role in ensuring that descriptive words are actually rendering a description within the candidate sentence, and it consequently helps to ameliorate the performance. Given this conclusion, [Figueroa and Atkinson, 2009] claimed that dependency paths offer a trade-off between lexical semantics and syntactic information required to typify definitions. To illustrate this, consider the following phrase[5]:

CONCEPT is a Entity novelist and author of Entity

Human readers would quickly notice that the sentence is a definition of a novelist, despite the missing concept and words. This is made possible due to the existence of two dependency paths ROOT→is→novelist, and novelist→author→of. The former acts as a *context indicator* showing the type of *definiendum* being described, whereas the latter yields content that is very probable to be found across descriptions of this particular *context indicator* (i.e., *novelist*). In substance, highly frequent directed dependency paths within a particular context are hypothesised to significantly characterise the meaning when describing an instance

[5]The placeholder *Entity* denotes a sequence of entities or adjectives that starts with a capital letter. This is explained in detail in the next section.

of the respective *context indicator*. The usage of this linguistic construct is a vital distinction between context models and the unigrams utilised by [Han et al., 2006], and the biterms and bigrams of [Chen et al., 2006]. In truth, this class of properties have aided in accomplishing a good performance with other kinds of QA systems [Cheng et al., 2009].

A key difference from the vast majority of TREC systems is that the inference is drawn by using contextual information conveyed by several descriptions of novelists, instead of using additional sources that provide information about a particular *definiendum* (e. g., "*Danielle Steel*"). These contextual models are inferred automatically from Wikipedia, and they are in disagreement with the tree *definiendum*-type models of [Han et al., 2006]. First, [Figueroa and Atkinson, 2009] do not capitalise on an external tool for distinguishing the type of *definiendum* prior ranking. In the opposite way, [Figueroa and Atkinson, 2009] identifies a *context indicator* for each answer candidate, brining about the opportunity of coping with several potential senses. Secondly, [Han et al., 2006] utilised three specific models: person, organisation and term, while [Figueroa and Atkinson, 2009] builds a model per *context indicator*. It is also worth recalling here that [Han et al., 2006] constructed also a general definition model, and they additionally took advantage of articles taken from eight distinct KBs for building their topic model, while [Chen et al., 2006] used task specific clues that biased the search engine in favour of articles from online dictionaries and encyclopedias. However, the approach of [Figueroa and Atkinson, 2009] sympathises and extends the idea of [Han et al., 2006] of utilising specific definition models, but their view of "specific" is at the sentence level, and broader in terms of the level of "specification".

CONTEXT MODELS

6.5.1 Building a Treebank of Contextual Definitions

Fundamentally, [Figueroa and Atkinson, 2009] benefit from training definition sentences for building a treebank of lexicalised dependency trees. These trees are thereafter automatically clustered in concert with their *context indicators*, and contextual n-gram LMs are constructed on top of these contextual dependency paths afterwards. The following is a breakdown of the steps taken for building the contextual treebank of definitions:

1. Abstracts are excerpted from the January 2008 snapshoot of Wikipedia, and all wiki annotations, such as brackets, are removed. With abstracts, it is meant the first section supplied by the document, which typically is a succinct summary of the key aspects, more important achievements and events of the corresponding *definiendum*. Accordingly, heuristics are used for removing undesirable abstracts corresponding to pages such as lists.

2. Then, the Stanford Named Entity Recogniser (NER)[6] is utilised for recognising named entities across all abstracts. The following classes are accordingly replaced with a placeholder (*Entity*): PERSON, DATE and LOCATION. This allows reducing the sparseness of the data and obtaining more reliable frequency counts later. In like manner, numbers and capitalised adjectives as well as sequences of words that start with a capital letter were mapped to the same placeholder. It is worth noting that any of these replacements is applied to an "entity" that has overlapping terms with the title of the article, and sequences of this placeholder are fused into a sole instance.

3. Sentences are identified by means of JavaRap.

4. Unlike other definition QA systems [Hildebrandt et al., 2004], definition patterns are applied at the surface level, similar to [Soubbotin, 2001, H. Joho and M. Sanderson,

[6] http://nlp.stanford.edu/software/CRF-NER.shtml

2000, 2001] (see section 4.2 on page 76). For this purpose, the lexico-syntactic constructs listed in table 3.1 (page 65) were used. Eventually, only matched sentences qualify for the following steps, and pronouns are acceptable occupiers of the slot corresponding to the *definiendum* in the array of definition patterns. At this step, soft patterns could also be exploited instead of or in conjunction with this group of rules (see section 4.9 on page 93).

5. Posteriorly, sentences are "anonymised", this means occurences of the topic/title of the article are replaced with a placeholder (CONCEPT). It is crystal clear that some sentences do not exactly match the pre-defined pack of clauses. Take, for instance, the next group of illustrative sentences:

   - In 1776 , he (John Edgar) was the commander of a British ship in the Great Lakes .
   - From 1990 to 1998 she (Monika Griefahn) was minister in Lower Saxony , since 1998 she is a member of the German Bundestag .
   - Currently , he (Joseph Pairin Kitingan) is the Deputy Chief Minister and Minister of Rural Development of Sabah and has held the post since March 2004 .
   - In 1936 he (Henrikas Radauskas) became an editor for the Lithuanian Commission of Book Publishing .

   An array of these underlined expressions, that precedes the topic/title of the article, was collected from the sentences that aligned the definition patterns. A set of templates was made out of these expressions by replacing numbers, possessive pronouns, and capitalised words with a placeholder. Here, the first personal pronoun was interpreted as the end of the template. The highest 4,259 frequent templates were kept, and every time any of these templates match a training sentence, the corresponding piece of text was removed. Most of these templates involve discourse markers, phrases that temporally anchors the sentence, and cataphoras. Notably, [Figueroa and Atkinson, 2009] assume that all pronouns refer to the concept dealt by the abstract of the article.

   In the previous step (4), only the existence of the lexico-syntactic clue was verified. Now, a full alignment is performed by means of the *Jaccard Meassure* (see section 3.4 on page 61) in conjunction with the experimental threshold presented in table 3.1 (page 65). Sentences satisfying these conditions, imposed by this group of cues, are kept and forced to start with the placeholder CONCEPT (some examples can be seen in table 6.4).

6. Preprocessed sentences are subsequently parsed by using a lexicalised dependency parser[7], in which extracted lexical trees are used for building a treebank of lexicalised definition sentences.

Overall, the source treebank encircles trees for 1,900,642 different sentences taken exclusively from anonymised Wikipedia abstracts. From the sentences in the treebank, the method of [Figueroa and Atkinson, 2009] identifies potential *context indicators*. This is in sharp contrast to [Han et al., 2006], who used three different models constructed at the *definiendum* level, and it is distinctly different from learning models on top of web snippets[Chen et al., 2006]. These involve words that signal what is being defined or what type of descriptive information is being expressed. *Context indicators* are recognised by walking through the dependency tree starting from the root node. Only sentences matching definition patterns and

CONTEXT INDICATORS

[7] http://nlp.stanford.edu/software/lex-parser.shtml

Indicator	$P(c_s)$	Indicator	$P(c_s)$	Indicator	$P(c_s)$	Indicator	$P(c_s)$
born	1.503	author	1.3160	novel	1.2398	title	1.1848
album	1.4604	term	1.3140	center	1.2390	used	1.1844
member	1.4505	series	1.3138	artist	1.234	officer	1.1837
player	1.3836	politician	1.3007	singer	1.2338	single	1.178
film	1.3738	group	1.2976	director	1.2297	coach	1.1741
town	1.3724	character	1.2946	community	1.2193	poet	1.1721
school	1.3521	actor	1.2880	program	1.2154	journalist	1.1708
village	1.3502	city	1.2856	known	1.2140	musician	1.1703
station	1.3446	writer	1.2738	site	1.2109	composer	1.168
son	1.3346	species	1.2492	professor	1.210	place	1.168
company	1.3281	footballer	1.2450	district	1.2058	painter	1.1666
game	1.3193	area	1.2443	leader	1.2056	daughter	1.1643
organization	1.3183	book	1.2435	team	1.199	producer	1.1596
band	1.3179	genus	1.2405	club	1.1907	language	1.159
song	1.316	actress	1.2398	episode	1.1901	home	1.1579

Table 6.3: Some interesting *context indicators* based on $log10$ of the frequencies (note:$P(c_s) * 10^4$).

that start with the placeholder CONCEPT are taken into account, so there are some clauses that are useful for the purpose of finding the corresponding *context indicators*. Since the root node itself is a *context indicator* whenever the node is a word contained in the surface patterns (e.g., *is*, *was* and *are*), the method walks down the hierarchy. In the case that the root has several children, the first child (different from the concept) is conceived as a *context indicator*. Note that the method must sometimes go down one more level in the tree depending on the expression holding the relationship between nodes (i.e., "*part/kind/sort/type/class/first of*"). Furthermore, the used lexical parser outputs trees that meet the projection constraint, and by the same token, the order of the sentence is preserved. Overall, 45,698 different *context indicators* were obtained during parsing. Table 6.3 shows the most frequent indicators acquired by this method, where $P(c_s)$ is the probability of finding a sentence triggered by the *context indicator* c_s within the treebank.

Candidate sentences are later grouped in congruence with the obtained *context indicators* (see table 6.4). One last remark on the *context indicators* is due to the definition pattern "<CONCEPT> [which | that | who] <description>". Occasionally, the indicator is a verb, but in praxis, it works in the same way as the nouns shown in table 6.3. The following sentences illustrate this similarity:

- CONCEPT which is **located** in Entity .
- CONCEPT that was **published** in Entity by the Entity .
- CONCEPT who **won** the Entity in Entity with a portrait of Entity .

Highly frequent directed dependency paths within a particular context are hypothesised to significantly characterise the meaning when describing an instance of the corresponding *context indicator*. This is predicated strongly on the extended distributional hypothesis [Lin and Pantel, 2001]: if two paths tend to occur in similar contexts, their meanings tend to be similar. In addition, the relationship between two entities in a sentence is almost exclusively concentrated in the shortest path between the two entities of the undirected version of the dependency graph [Bunescu and Mooney, 2005]. Ergo, one entity can be interpreted

EXTENDED DISTRIBUTIONAL HYPOTHESIS

Context Indicator	Terms
Author	CONCEPT was a Entity author of children's books . CONCEPT is the author of two novels : Entity and Entity . CONCEPT is an accomplished author . CONCEPT is an Entity science fiction author and fantasy author . CONCEPT is a contemporary Entity author of detective fiction . CONCEPT became an author after a career as an entrepreneur . CONCEPT , a Entity children's author . CONCEPT has been the author of several religious publications .
Player	CONCEPT is a Entity football player , who plays as a midfielder for Entity . CONCEPT is a Entity former ice hockey player . CONCEPT is a Entity jazz trumpet player . CONCEPT , a former Entity player for the Entity .
Disease	CONCEPT is a fungal disease that affects a wide range of plants . CONCEPT is a disease of plants , mostly trees , caused by fungi . CONCEPT is a chronic progressive disease for which there is no cure . CONCEPT , a disease in chickens and other birds , affects only hens .
Song	CONCEPT is a Entity song by Entity taken from the Entity album Entity . CONCEPT is a Entity song performed by the Entity band Entity . CONCEPT is a pop song written by Entity and Entity , produced by Entity for Entity's first album Entity . CONCEPT , the title of a song by Entity . CONCEPT , the theme song for the Entity film . CONCEPT , the theme song to the film performed by Entity . CONCEPT has been the official state song of Entity since Entity .

Table 6.4: Some examples of sentences grouped in concert with their *context indicators*.

as the *definiendum*, and the other can be any entity within the sentence. Paths linking a particular type of *definiendum* with a class of entity relevant to its type will therefore be highly frequent in the context (e. g., novelist→author→of→Entity). Note that using paths cushions the effect of knowing the exact category of the entity. For instance, the entity in the previous path will be a work because the linked sequence of words undoubtedly signals this. Some paths, nonetheless, can still be ambiguous: born→Entity→in.

TREEBANK ACCURACY

On a final note, a small random number (1,162 out of 1,900,642) of sentences in the treebank were manually checked for the purpose of estimating the amount of wrongly automatically annotated samples (false positives). In short, solely 4,73% of these selected sentences were judged as spurious descriptions.

6.5.2 Learning Contextual Language Models

For each context, all directed paths encompassing two to five nodes are collected. Longer paths are not taken into consideration as they are likely to indicate weaker syntactic/semantic relations. Directions are mainly perceived as pertinent syntactical information regarding word order is missed when going up the dependency tree. Otherwise, undirected graphs would lead to a significant increment in the number of paths as it might go from any node to any other node. Some illustrative directed paths obtained from the treebank for the *context indicator*: *author*, are shown below:

author→awarded→for→Entity
author→based→character→of→Entity

author→began→in→mid-90s
author→chairman→former
author→co-author→of→Entity→bestseller
author→contributed→to→Entity→journal
author→contributed→to→Entity→magazine
author→founder→of→Entity→movement
characterized→for→period→the
chief→of→editions
co-producer→of→film→entitled→Entity
columnist→with→Entity
editor→at→Entity

From the obtained dependency paths, an n-gram statistical LM ($n = 5$) was built as a means of estimating the most relevant dependency path for each context. The likelihood of a dependency path $\vec{dp}$ in a context c_s is defined by the likely dependency links that compose the path in the context c_s, with each link probability conditional on the last $n-1$ connected terms:

$$P(\vec{dp} \mid c_s) \approx \prod_{i=1}^{l} P(w_i \mid c_s, w_{i-n+1}^{i-1}) \tag{6.1}$$

Where $P(w_i \mid c_s, w_{i-n+1}^{i-1})$ is the probability of term w_i being linked with the previous word w_{i-1} after seeing the dependency path $w_{i-n+1} \ldots w_{i-1}$. Simply put, the likelihood that w_i is a dependent node of w_{i-1}, and w_{i-2} is the head of w_{i-1}, and so forth. The probabilities $P(w_i \mid c_s, w_{i-n+1}^{i-1})$ are usually given by computing the MLE:

$$P(w_i \mid c_s, w_{i-n+1}^{i-1}) = \frac{count(c_s, w_{i-n+1}^{i})}{count(c_s, w_{i-n+1}^{i-1})}$$

However, when utilising dependency paths, the word count $c(c_s, w_{i-n+1}^{i})$ can frequently be greater than $c(c_s, w_{i-n+1}^{i-1})$. For example, in the following definition sentence: "*CONCEPT is a band formed in Entity in Entity .*" The term "*formed*" is the head of two "*in*", hence the denominator of $P(w_i \mid c_s, w_{i-n+1}^{i-1})$ is the number of times w_{i-1} is the head of a word (after seeing w_{i-n+1}^{i-1}). The obtained 5-gram LM is smoothed by interpolating with shorter dependency paths [Chen and Goodman, 1996, Goodman, 2001, Zhai and Lafferty, 2004] as follows:

$$P_{interp}(w_i \mid c_s, w_{i-n+1}^{i-1}) = \lambda_{c_s, w_{i-n+1}^{i-1}} P(w_i \mid c_s, w_{i-n+1}^{i-1}) + (1 - \lambda_{c_s, w_{i-n+1}^{i-1}}) P_{interp}(w_i \mid c_s, w_{i-n+2}^{i-1})$$

The probability of a path is accordingly determined as shown in equation 6.1 by accounting for the recursive interpolated probabilities in place of raw Ps. Note also that $\lambda_{c_s, w_{i-n+1}^{i-1}}$ is calculated for each context c_s as proposed by [Chen and Goodman, 1996]. A candidate sentence A is ranked in consonance with its likelihood of being a definition as follows:

$$rank(A) = P(c_s) \sum_{\forall \vec{dp} \in A} P(\vec{dp} \mid c_s) \tag{6.2}$$

In order to avoid counting redundant dependency paths, only paths ending with a **leave node** are taken into account, whereas **duplicate** paths are discarded.

Why n=5?

Conventionally, the value of n normally oscillates between one and three. In this respect, [Figueroa and Atkinson, 2009] took into consideration longer n-grams as a mean of rewarding some prominent paths that can establish relations with other entities in candidate sentences. In order to verify this, the different contexts were scanned, and as a result, more than 1,200 paths of length four and five were found that are carried in more than fifty distinct contexts. For instance, some of these observed tetragrams and pentagrams include:

TETRAGRAMS

Tetragrams	Pentagrams
album→released→in→Entity	based→in→city→of→Entity
as→president→of→Entity	died→at→age→of→Entity
born→raised→in→Entity	known→for→work→with→Entity
built→between→Entity→Entity	located→in→heart→of→Entity
film→directed→by→Entity	named→in→honor→of→Entity
founded→based→in→Entity	one→of→founders→of→Entity
organization→based→in→Entity	served→as→president→of→Entity

Table 6.5: Predominant tetragrams and pentagrams in *context models* that can link the *definiendum* with an entity.

PENTAGRAMS

Furthermore, [Figueroa and Atkinson, 2009] also observed that longer paths tend to signal weaker relations. Nevertheless, some noticeable pentagrams can still imply a relationship between the *definiendum* and a pair of entities. Some noticeable paths of length five are as follows:

Pentagrams	
based→in→Entity→in→Entity	founded→by→Entity→in→Entity
born→in→Entity→in→Entity	founder→of→Entity→in→Entity
built→by→Entity→in→Entity	located→in→Entity→near→Entity
created→by→Entity→in→Entity	member→of→Entity→in→Entity
designed→by→Entity→in→Entity	professor→at→Entity→of→Entity
released→by→Entity→in→Entity	published→by→Entity→in→Entity
written→by→Entity→in→Entity	established→by→Entity→in→Entity

Table 6.6: Prominent pentagrams in *context models* that can connect the *definiendum* with two entities.

The reason why taking into consideration these longer paths is beneficial is two-fold:

(a) Definition QA systems are geared towards finding a group of sentences that express succinct and diverse information about the *definiendum*. Biasing the ranking in favour of sentences that can link the *definiendum* with several (two) entities is, therefore, naturally preferable than choosing two quite similar sentences that independently link both entities with the *definiendum*. This phenomena is chiefly due to the fact that human writers can use short form variants when the full meaning can be deduced from the context [Savary and Jacquemin, 2003]. These contexts (sentences) are thus more concise, and ergo more suitable to incorporate into the final output.

(b) Matching longer and higher frequent paths ensures grammaticality, that is, matching fuller ideas or descriptions. This might not be so important when dealing with doc-

uments that provide some structure like clearer sentence or paragraph delimitations. This is critical, however, when coping with a noisy and ungrammatical target corpus such as web snippets. It is worth recalling that search engines usually truncate web snippets by inserting intentional breaks (...), making them ungrammatical. In a nutshell, privileging longer paths assists in tackling truncations and biasing the ranking in favour of completer definitions. This also subsumes some paths that do not carry entities (e.g., species→of→plant→in→family).

6.5.3 Extracting Candidate Answers

The model of [Figueroa and Atkinson, 2009] sifts answers to definition questions from web snippets. Firstly, sentences are distinguished by means of truncations and JavaRap. Secondly, sentences matching the definition patterns at the surface level illustrated in table 3.1 (page 65) are singled out. Thirdly, matched sentences are "anonymised" and enforced on starting with the placeholder CONCEPT, similarly to the training sentences. Fourthly, these matched sentences are parsed in order to get the corresponding lexicalised dependency trees. Lastly, given an array of test sentences/dependency trees extracted from the surrogates, this approach discovers answers to definition questions by iteratively selecting sentences.

The general strategy for this iterative selection task can be seen in algorithm 1 whose input is the set of dependency paths (T). This first initialises a set ϕ which keeps the dependency paths belonging to previously chosen sentences (line 1). Later, *context indicators* for each candidate sentence are extracted so as to build an histogram $indHist$ (line 2). Since highly frequent *context indicators* show more reliable potential senses, the method favours candidate sentences based on their *context indicator* frequencies (line 3). Sentences matching the current *context indicator* are ranked in concert with equation 6.2 (lines 7 and 8). However, only paths $\vec{dp}$ in $t_i - \phi$ are taken into consideration, while computing equation 6.2. Sentences are thus ranked in conformity with their novel paths with respect to previously selected sentences, while at the same time, sentences carrying redundant information downgrade their rating value systematically. Highest ranked sentences are singled out after each iteration (line 9-11), and their corresponding dependency paths are added to ϕ (line 18). If the highest ranked sentence meets the halting conditions, the extraction task finishes. Halting conditions ensure that no more sentences to be chosen are left and no more candidate sentences embodying novel and reliable descriptive information are picked.

Unlike other techniques which control the overlap at the word level [Hildebrandt et al., 2004, Chen et al., 2006, Han et al., 2006], the basic unit is a dependency path, that is, a group of related words. Therefore, if a word comes from two distinct contexts, then this approach interprets them differently. To be more precise, [Chen et al., 2006] measured the cosine similarity of a new answer candidate to each of the previously selected answers. A threshold acted then as the referee for determining whether or not the new putative answer was similar to any of these previously selected answers, and by the same token, whether or not it must be incorporated into the final output. The approach of [Han et al., 2006] is very akin to the technique of [Chen et al., 2006], but they benefited from a relative word overlap measure instead of the cosine, and they additionally took advantage of WordNet for eliminating answer candidates that share the synset with any of the already chosen answers. A final aspect of both strategies is that they detach the redundancy factor from their ranking scores, this means a putative answer can still be rewarded for the same combination of words or features that can be found within an already selected answer.

BASIC UNIT

On the contrary, algorithm 1 integrates redundancy as an ingredient in the ranking score by nullifying the contribution of the redundant content. Accordingly, candidate sentences become less relevant as long as their overlap with all previously selected sentences be-

LOCAL VS. GLOBAL REDUNDANCY CHECK

```
1  φ = ∅;
2  indHis = getContextIndicatorsHistogram(T);
3  for highest to lowest frequent ι ∈ indHis do
4      while true do
5          nextSS = null;
6          forall t_t ∈ T do
7              if indHis(t_i)==ι then
8                  rank = rank(t_i,φ);
9                  if nextSS == null or rank > rank(nextSS,φ) then
10                     nextSS = t_i;
11                 end
12             end
13         end
14         if nextSS == null or rank(nextSS,φ) ≤ 0.005) then
15             break;
16         end
17         print nextSS;
18         addPaths(nextSS,φ);
19     end
20 end
```

Algorithm 1: ANSWER EXTRACTOR

comes larger. Thus, the method favours novel content, while at the same time, it makes a global verification of the redundant content. The idea behind this technique is in the same spirit as other scoring methodologies including the strategy of the best system in TREC 2006 [Kaisser et al., 2006], and the top-ten TREC 2007 system proposed by [Schlaefer et al., 2007] (for details, see section 4.12 on page 102).

The crucial difference stems from the decay factors and dependency paths, [Kaisser et al., 2006] systematically lowered the contribution of a term to the score of an answer candidate. This diminishment cohered with the number of times the word has already been subsumed in the set of previously selected answers, whereas algorithm 1 suppressed the contribution of the dependency paths already embraced in any of the previously chosen answers. The reason why [Kaisser et al., 2006, Schlaefer et al., 2007] needed the decay factors is that their procedure privileged terms that have a high correlation with the *definiendum*. These high-correlated words, for this reason, can be embodied in several distinct descriptions, and inclined to belong to the top-ranked answers. Therefore, suppressing the contribution of these highly correlated terms after their first selection, can result in the fact that the remaining words of some unselected descriptive contexts might not offer enough evidence to qualify as final answers, causing the definition QA system to miss these descriptions. In algorithm 1, dependency paths allay this issue by accounting for term combinations that convey fuller ideas, that is, it assesses the novelty of local contexts in place of isolated words.

DECAY FACTORS

6.5.4 Experiments and Results

In order to assess the context methods, the 189 definition questions obtained from TREC 2003-2004-2005 tracks were utilised. Since *context models* are aimed specifically at extracting answers from the Web, these TREC datasets were used solely as reference question sets. For empirical purposes, Wikipedia articles found to be related to these *definienda* were banned

from the training material, and three baselines were implemented. The four systems were accordingly provided with the same array of sentences.

With regard to the testing sentences, they were gathered from web snippets. First of all, the rewriting strategy presented in section 2.6.2 (page 47) was considered as a means to boost the retrieval of descriptive phrases within surrogates. The advantage of this technique is that it rewrites the original query in accordance with the syntactical structure of the surface rules shown in table 3.1 (page 65). Essentially, it generates ten search queries that bias the search engine in favour of web snippets that can be aligned with these clues. In the experiments, the MSN Search was used as an interface to the Internet. Each search query was aimed at fetching 30 web snippets, and hence for each question a maximum of 300 web snippets is retrieved. Second, all fetched text fragments were manually inspected, including those that mismatch the pre-defined battery of rules, in order to create a gold standard. All nuggets were equally weighted when computing the $\mathcal{F}(3)$-Score, and the evaluation adhered to the most recent standard [Lin and Demner-Fushman, 2006]. It is worth duly pointing out here that there was no descriptive information for eleven questions of the TREC 2005 dataset.

TESTING SENTENCES

The first baseline (BASELINE I) ranks candidate sentences in accordance with their likeness to the centroid vector [Yang et al., 2003, Cui et al., 2004b,c] (see section 4.8 on page 91). More specifically, final answers are singled out by using algorithm 1, and their respective words are added to ϕ; this way their contribution is nullified in the posterior iterations. Since the intention is studying strategies that are geared towards being independent of specific entries (e.g., "*Danielle Steel*" and "*John Updike*") in KBs, this centroid vector was inferred exclusively from all retrieved sentences bearing the *definiendum*. It is worth recalling here that:

BASELINE I: CENTROID VECTOR

(a) The impact of KBs (topic models) in the performance is well-known: the performance markedly improves or falls into a steep decline in agreement with the coverage yielded by KBs for each particular *definiendum* [Zhang et al., 2005, Han et al., 2006]. This finding makes study more robust methods, that ignore this sort of information, interesting.

(b) As pointed out by [Cui et al., 2004c], words highly correlated with the *definiendum* at the sentence level are likely to indicate some of its pertinent facets, and consequently they can be utilised for rating answer candidates to definition queries. For instance, the best system in the TREC 2006 took advantage of word frequency counts across web snippets related to the *definiendum* as a dominant feature for ranking putative answers taken from the AQUAINT corpus [Kaisser et al., 2006] (see section 4.12 on page 102).

In BASELINE I, all sentences are seen as candidates. It can, ergo, identify descriptions from sentences that misalign the pre-determined definition patterns, unlike the three other systems, which are targeted chiefly at increasing the accuracy of pattern matching. In short, the primary objective of this baseline is measuring how much recall (nuggets) is covered or can be inferred by means of redundancy. Here, redundancy is understood in terms of word correlation frequency counts.

In sharp contrast to this first baseline, the second one (BASELINE II) capitalises on the 1,900,642 preprocessed sentences harvested from Wikipedia abstracts. This baseline is predicated on the bi-term LMs adopted by [Chen et al., 2006] (see section 6.3). The difference between this baseline and the system presented by [Chen et al., 2006] lies in the fact that the LMs of the latter are inferred from the ordered centroid vector representation of sentences extracted from web snippets. This baseline, conversely, deduces the LMs from the array of 1,900,642 training sentences, where, like [Chen et al., 2006], training sentences are stemmed, and their stop-words removed. Still yet, there are two key aspects to bear in mind about this difference:

BASELINE II: BITERM LMS

(a) The ordered centroid representation maps a sentence to a sequence of its most describing (centroid) terms. This is very important when learning models from phrases emanated from non-authoritative sources, because this group of restricted words is very likely to elucidate some facets of the *definiendum*. Nevertheless, the centroid vector is computed for each particular *definiendum*, and having a threshold for the number of these centroid words is more suitable for technical or accurate/precise *definienda* (e.g., "*SchadenFreude*"), than for ambiguous or biographical *definienda* (e.g., "*Danielle Steel*") which need more words to describe many writings of their several facets. However, the training sentences are independent from a particular *definiendum*. They match surface definition patterns, while at the same time, they are distilled from an authoritative source, thereby ensuring to a great extent that they are actual descriptions.

(b) BASELINE I already ranks sentences in consonance with word correlation statistics deduced from web snippets. Hence, the goal is to liken three different models learnt from the same corpus and aimed at the same array of test sentences.

Like BASELINE I, this baseline then chooses sentences by means of algorithm 1, but candidate sentences are ranked in agreement with equation 6.3. That is, the unique difference between both baselines is the ranking function which is grounded on sharply distinct models. Correspondingly, bi-terms embraced in selected answers are added to ϕ, and analogously, their contribution to the next iterations is cancelled. Accordingly, the mixture weight λ was expirically set to 0.72 by using the EM algorithm [Dempster et al., 1977] (see equation 4.9 on page 98). By the same token, L_{ref} was experimentally set to fifteen words. In a statement, the intention behind this baseline is testing the performance of the LMs adopted by [Chen et al., 2006] against our test sentences and built on the training sentences.

BASELINE III: WORD ASSOCIATION NORMS

The third baseline (BASELINE III) is also incorporated into the framework provided by algorithm 1. This baseline is constructed on top of *word association norms* [Church and Hanks, 1990]. These norms were computed from the same set of 1,900,642 preprocessed sentences distilled from Wikipedia abstracts, and they comprise pairs I_2 and triplets I_3 (equations 5.1 and 5.2 on pages 131 and 131, respectively) of ordered words as seen in table 5.12 (page 131).

Sentences are subsequently ranked in agreement with the sum of the matching norms which are normalised by dividing them by the highest matching value. *Word association norms* compare the probability of observing w_2 followed by w_1 within a fixed window of ten words with the probabilities of observing w_1 and w_2 independently. They supply a methodology that is the basis for a statistical description of a variety of interesting linguistic phenomena, ranging from semantic relations of the professor/student type to lexico-syntactic co-occurrence constraints between verbs and prepositions (e.g., written/by) [Church and Hanks, 1990]. For this reason, BASELINE III offers a good starting point for measuring the contribution of dependency-based context LMs.

Since these three baselines do not account for *context indicators*, every sentence is assumed to have the same *context indicator*. All in all, these baselines supply distinct ways of deriving lexico-syntactic and semantic relations at different levels that typify descriptions of the *definiendum*, and exploit them for rating answers to definition questions afterwards.

OVERALL PERFORMANCE

Table 6.7 shows the figures achieved by the three baselines and the *context models* for the three test query sets. Broadly speaking, BASELINE III outperformed the other two baselines in all sets, and BASELINE II finished with better results than the first baseline. In terms of $\mathcal{F}(3)$-Score, *context models* surpassed BASELINE III in 5.22% and 11.90% for the TREC 2003 and 2004 datasets, respectively. To state it more clearly, the outcome was bettered for 81 (71.05%) out of 114 questions. These improvements are mainly due to *definienda* such as "*Allen Iverson*" and "*Heaven's Gate*" as well as "*Bashar Assad*". On the other hand, in 32 (28.07%) out of

	TREC 2003	TREC 2004	TREC 2005
Size	50	64	(64)/75
Baseline I			
Recall	0.27±0.23	0.27±0.16	0.24±0.17
Precision	0.20±0.19	0.20±0.19	0.18±0.23
$\mathcal{F}(3)$-Score	0.24±0.18	0.25±0.15	0.22±0.16
Baseline II			
Recall	0.45±0.18	0.40±0.17	0.38±0.19
Precision	0.28±0.22	0.19±0.11	0.21±0.17
$\mathcal{F}(3)$-Score	0.40±0.15	0.34±0.14	0.33±0.15
Baseline III			
Recall	0.52±0.18	0.47±0.13	**0.49**±0.20
Precision	0.27±0.14	0.26±0.11	**0.29**±0.24
$\mathcal{F}(3)$-Score	0.46±0.14	0.42±0.11	**0.43**±0.17
Context Models			
Recall	**0.57**±0.17	**0.50**±0.18	0.42±0.22
Precision	**0.39**±0.21	**0.40**±0.19	**0.29**±0.21
$\mathcal{F}(3)$-Score	**0.53**±0.15	**0.47**±0.17	0.38±0.19

Table 6.7: Results for TREC question sets.

these 114 questions, the performance deteriorated. For instance, *definienda* such as "*Rhodes Scholars*" and "*Albert Ghiorso*".

In terms of recall, the average raised from 0.52 to 0.57 (9.6%) for the TREC 2003 dataset, whereas 6.4% for the TREC 2004 dataset. Particularly, *definienda* such as "*Jennifer Capriatti*" and "*Heaven's Gate*" resulted in significant recall improvements, whereas "*Abercrombie and Fitch*" and "*Chester Nimitz*" dropped suddenly. A crucial factor behind the betterment of the performance eventuates from the privilege given to sentences belonging to prominent contexts. To exemplify, the clusters in relation to the contexts "*cult*" and "*religion*" contain twelve and nine sentences, respectively, where four and two of them were selected on the top of the ranking. All of these six selected sentences were actual definitions. It was observed that prioritising putative answers within prominent contexts cooperates on singling out some novel answers that were not selected by BASELINE III, because of the preference of some misleading answer candidates that obtained a higher score than these genuine definitions. These spurious sentences are usually, but not exclusively, connected with lowly frequent contexts. Interestingly, this improvement is obtained by enriching the selection algorithm with inferences drawn from the global context (all answer candidates) instead of solely using the attributes of each sentences for rating. Accordingly, a good example of answers chosen in concert with predominant contexts can be seen in table 6.8.

RECALL

As previously noted, the performance was diminished for 32 questions. In 26 out of these 32 cases, it was observed that the recall lessened in more than 10%, effectuating a significant drop in the $\mathcal{F}(3)$-Score. In order to qualify for the final output, an answer candidate must obtain a relatively high ranking value. The following are the three determining aspects integrated by the ranking strategy utilised by *context models*: (a) the probability $p(c_s)$ of its *context indicator*, (b) the frequency of its *context indicator* across answer candidates, and (c) the evidence brought forth by the paths that constitute its description, meaning the sum of its respective $P(\vec{dp} \mid c_s)$. Certainly, a harmony exists between $p(c_s)$ and the coverage of the corresponding context: if $p(c_s)$ is high, then the respective context will supply ampler coverage,

Tale of Genji
◇ The Tale of Genji is the story of a man, the son of the Emperor by his favorite consort ...
◇ The Tale of Genji is a fifty-four chapter epic novel written by Murasaki Shikubu.
◇ Written by Murasaki Shikibu The Tale of Genji is a Japanese story written in the beginning of the eleventh century by Murasaki Shikibu.
◇ The Tale of Genji is an ancient and grand novel which has themes, traditions, and prose that still sparkle in today's limelight.
◇ The Genji monogatari is a long work of prose fiction supposedly written in the early eleventh century by Murasaki Shikibu (978)...
◇ Waley's Tale of Genji is an English novel in its own right, a romantic escape, in prose, from the aftershock of war into an aestheticized realm of sensitive.
◇ Murasaki Shikibu's eleventh-century Tale of Genji is the most revered work of fiction in Japan.
◇ In what is perhaps the very earliest novel in the world, the Genji Monogatari (Tale of Genji), which dates back to around the eighth century CE, eroticism is treated as a central.
◇ The Tale of Genji' is the famous early eleventh-century novel by the court lady Murasaki Shikibu, relating the life and loves of the fictive Prince Genji ..
◇ The Tale of Genji is a full-length novel consisting of 54 individual chapters that was written at the beginning of the 11th century in ...
◇ The Tale of Genji is the first novel ever produced in the world.

Table 6.8: Sample output sentences regarding "*Tale of Genji*" (source [Figueroa and Atkinson, 2009]).

if it is low, the coverage becomes then narrower.

DATA SPARSENESS

Consequently, answer candidates matching a context that yields narrower coverage are more unlikely to be subsumed in the final output; in particular, whenever few putative answers are retrieved from the Web and the coverage supplied for their respective predominant contexts is limited. BASELINE III, oppositely, profits from statistics derived from the entire corpus, alleviating the data sparseness problem of some contexts with restricted coverage. This data sparseness can be, nonetheless, remedied by extending the treebank, and ergo the *context models*, with extra snapshots of Wikipedia and short definitions colleted from glossaries across documents on the Internet. These glossaries can automatically be extracted by identifying regularities in their lay-outs: tables, alphabetically sorted entries, etc. Additionally, both techniques can be put together, since the system knows which contexts are more trustworthy and prominent (across the fetched web snippets) than others. It is worth noting that this difference in coverage also explains the slight growth in dispersion.

TRUNCATIONS

Another reason for the worsening of the recall stems from the fact that some nuggets were missed due to the ungrammaticality of some web snippets. While the search procedure biases the output in favour of longer sentences [Figueroa and Neumann, 2007] (see table 5.6 on page 127), short and truncated sentences are still possible to be fetched. These truncations can eventually cause problems when obtaining the lexicalised dependency tree, missing some interesting nuggets. This is a great advantage when preferring surface statistics like word norms in place of methods, such as context LMs, that incorporate Natural Language Processing (NLP) processing. In summary, *context models* strengthened the recall for 50% of the TREC 2003-2004 *definienda*.

PRECISION

Incidentally, *context models* also achieved higher precision for two datasets. In the case of the TREC 2003, the increment was 44.44%, whereas it was 53.84% for the TREC 2004 question set. In other words, *context models* were capable of filtering out a larger amount of sentences that did not render descriptions, while at the same time, boosting the recall. Given these

IMPACT OF SYNTAX

outcomes, [Figueroa and Atkinson, 2009] concluded that these pieces of information were characterised by regularities in their contextual dependency paths, wherewith the accuracy of pattern matching was bettered. This finding is also ratified by the baselines, as long as more lexico-syntactic information was incorporated, the performance enhanced in terms of both precision and recall.

As a mean to understand some of the causes that still hurt the precision, the responses with relation to the different test questions was inspected. In fact, it was discovered that there are some misleading descriptions that pose a tough, but interesting, challenge to definition QA systems. Consider the following four sentences incorporated into the output of "*Jean Harlow*":

Jean Harlow is a red-headed secretary who hooks the company's married boss, while carrying on with chauffeur Charles Boyer.

Mona Leslie (**Jean Harlow**) is an up-and-coming Broadway actress, dancer, and singer, who leads a happy-go-lucky, freewheeling lifestyle; bailed out of jail by family friend Ned.

Crystal Wetherby (**Jean Harlow**) is an American widow left stranded in London with a stack of debts incurred by her late husband and barely a shilling to her name.

Jean Harlow is the secretary no wife wants her husband to have in Wife vs. Secretary.

The second and third cases elucidate the roles ("*Mona Leslie*" and "*Crystal Wetherby*") portrayed by "*Jean Harlow*" in two different movies. A definition QA system normally takes advantage of the parentheses pattern to align descriptions that include aliases of the *definiendum* (e.g., "*Abbreviation (organisation) is a/an/the* "). In these two cases, the roles are identified as they were "*Jean Harlow*" herself. In light of this observation, it can be deemed that the usage of a particular rule can be more appropriate in one context than in others. Homologously, the clause (e.g., "*<definiendum> became* ") captures good nuggets when dealing with contexts such as artists and sports, but they were inclined to be noisy when tackling contexts such as organisations and events. In the first and fourth descriptions, the *definiendum* replaces the name of the character in the movie, which is the actual concept being explained in the phrase. Of course, drawing this class of distinction or disambiguation would require deeper reasoning and understanding of the context. Although, there is uncertainty as to whether or not it is possible to resolve with a limited number of sentences and/or text snippets.

PATTERNS

Another fertile source of spurious answers is superlatives. The studies carried out by [Kaisser et al., 2006, Razmara and Kosseim, 2007, Scheible, 2007] (section 4.4 on page 83) unveiled that superlatives are useful for acquiring interesting descriptive nuggets from a collection of news documents. In juxtaposition, the Web is abundant in opinions and advertisements, which are highly likely to match superlatives, and consequently to convey the mistaken impression of actual descriptions: "*<definiendum> was/is the best man/player/group/band in the world/NBA*" and "*<definiendum> is the best alternative to ....*". It was also observed that the reason why these spurious sentences were singled out was two-fold: (a) these overmatched superlatives were normally included into the group of sentences belonging to the predominant sense (e.g., "*tenor*", "*singer*", "*band*" and "*actor*"), and (b) these misleading sentences obtain a relatively high score due partly to paths like band→best, band→the, band→is. It is worth remarking, nonetheless, that this class of path can be frequently found across definitions. To illustrate, take a band that won a prize for being "*the best band*" in a particular genre, country and year. Certainly, superlatives still play a pivotal role in definitions:

SUPERLATIVES

The **Nobel Prize** is the best known and most prestigious award in science, and California's universities and research institutes claim more....

The **Nobel Prize** is an annual international award for the best advances in science (among other disciplines).

On all sides, the evidence points to the need of a deeper context analysis for separating the wheat from the chaff. More precisely, one can be envision that a set of (anti-) *context models* encompassing negative examples would provide invaluable aid in recognising these misleading sentences.

With regard to BASELINE I, it is worth emphasising that the achieved results are comparable to the outcome obtained by [Zhang et al., 2005] in which they assessed a system that did not account for online specific resources. Unlike [Zhang et al., 2005] and the trends in TREC, *context models* do not make use of direct entities (e.g., "*Danielle Steel*") in order to discover relevant contextual information to be projected into the AQUAINT corpus. Instead, this learns contextual models which are used for discriminating answers directly from their context (i.e., with no projection into a target corpus). Additionally, the comparison between this baseline and BASELINE II entails the positive effect of the acquired corpus in the performance.

BASELINE I

As for TREC 2005, *context models* finished with a lower recall and $\mathcal{F}(3)$-Score. A closer look at the achieved results shows that *context models* enhanced the performance in 37 (57.81%) out of the 64 questions, while in 24 (37.5%) cases the performance was deteriorated. A key point is that in six of these 24 cases, *context models* obtained a recall of zero and so the $\mathcal{F}(3)$-Score values become zeros, and eventually, brought about a significant drop in the average score. Three of these six questions correspond to *definienda* such as "*Rose Crumb*" and "*1980 Mount St. Helens eruption*" as well as "*Crash of EgyptAir Flight 990*".

TREC 2005

Some common issues for these six scenarios were also observed. Firstly, few nuggets were found within the fetched surrogates. Secondly, these text fragments had a low frequency hence whenever *context models* missed any or all of them, the performance was detrimental. This situation becomes serious as the nuggets are uttered in contexts that are very unlikely to be in the *context models*. To measure the impact of these six cases, the average $\mathcal{F}(3)$-Score was compared by accounting solely for the other 58 questions: 0.43 for *context models*, and 0.41 for the third baseline.

Since $\mathcal{F}(3)$-Score does not assess the precision of the ranking order, the Mean Average Precision (MAP) of the top one and five ranked sentences was computed (see table 6.9) [Manning and Schütze, 1999]. MAP scores reveal that *context models* effectively contributes to improving the ranking of the sentences. The figures presented in table 6.9 did not only show that *context models* outperform the other three strategies in MAP terms, but they also finished with a higher precision in ranking, containing a valid definition at the top 80% of the cases. These achievements result from the bias of the ranking in favour of descriptive sentences that meet a combination of the following criteria:

RANKING ORDER

1. The top ranked sentences share more lexico-syntactic similarities with descriptive sentences in Wikipedia abstracts, and they have therefore access to more privileged positions in the ranking. As a logical consequence, this improvement in ranking signifies an increment in the accuracy of the matching surface patterns. For the purpose of investigating this, the ratio of selected to all fetched sentences that align the pre-determined battery of definition patterns was calculated. Table 6.10 highlights this improvement.

   NUMBER OF SELECTED SENTENCES

   It is worth recalling that `Baseline I` also chooses answer candidates that mismatch definition patterns. Therein lies their exclusion from this comparison. In a statement, the figures showed in this table along with their respective outcomes in terms of $\mathcal{F}(3)$-Score, precision and recall indicate an increase in the pattern matching accuracy.

	Baseline I	Baseline II	Baseline III	Context Models
TREC 2003				
MAP-1	0.16	0.56	0.64	**0.82**
MAP-5	0.21	0.57	0.64	**0.82**
TREC 2004				
MAP-1	0.27	0.67	0.66	**0.88**
MAP-5	0.25	0.59	0.62	**0.82**
TREC 2005				
MAP-1	0.18	0.58	0.77	**0.79**
MAP-5	0.24	0.53	0.70	**0.77**

Table 6.9: MEAN AVERAGE PRECISION.

	TREC 2003	TREC 2004	TREC 2005
Baseline II	0.23 ± 0.13	0.22 ± 0.06	0.27 ± 0.19
Baseline III	0.27 ± 0.12	0.26 ± 0.07	0.32 ± 0.18
Context Models	0.23 ± 0.07	0.21 ± 0.06	0.18 ± 0.1

Table 6.10: Average ratio of selected to fetched sentences (matching definition patterns only).

CONTEXT INDICATOR REDUNDANCY

2. More often than not, the highest ranked answers correspond to predominant and by the same token, more reliable potential senses, thus making the possibility of them verbalising a description of the *definiendum* more likely. To illustrate this, table 6.11 shows the most pertinent *context indicators* with relation to four different queries, including the example shown in table 6.8.

Tale of Genji		Teapot Dome Scandal		George Foreman		Chunnel	
Indicator	Frq.	Indicator	Frq.	Indicator	Frq.	Indicator	Frq.
book	2	example	2	boxer	3	film	5
novel	**12**	issue	6	**champion**	**17**	train	3
story	4	place	3	heavyweight	2	**tunnel**	**9**
work	4	**scandal**	**11**	hitter	2		
study	2	victory	8	medalist	2		
product	2			man	8		
popular	2			minister	2		
matter	2			symbol	5		

Table 6.11: Sample of relevant *context indicators* for some TREC definition queries.

Unlike TREC systems, the three baselines and the *context models* were evaluated by using sentences collected from the Internet. While the approach took advantage of sophisticated search engines, these are not optimised for QA tasks. In fact, this is the reason the model is required to capitalise on the purpose-built search strategy presented in section 2.6.2 (page 47). In addition, many TREC systems benefit from off-line processing on the AQUAINT corpus for the purpose of boosting the performance [Hildebrandt et al., 2004, Fernandes, 2004] (see sections 2.1 and 4.2 on pages 25 and 76), so that when scoring, extra features (i.e., entities) are used to recognise definitions. Instead, *context models* rank by accounting almost exclusively

for the lexical syntactic and semantic similarities to previously known definitions that describe other instances of the same kind of *definiendum*. The additional knowledge used when scoring is the frequency of the *context indicators*, which aids the model in ranking frequent potential senses, and more trustworthy sentences. Experiments thus showed that dependency paths supply key lexico-semantic and syntactic information that typifies definitions at the sentence level.

REDUNDANCY

The use of relations between a group of words in place of isolated terms for ranking sentences also ensures a certain level of grammaticality in the candidate answers. Since web snippets are often truncated by search engines, relations allow singling out truncated sentences that are more plausible to convey a complete idea than others. This also leads to missing some relevant nuggets. On the other hand, two different dependency paths can yield the same descriptive information, materialising an increment in redundancy. A clear case of this is provided in table 6.12, which is an excerpt from the output concerning the *definiendum* "*Teapot Dome Scandal*". Basically, this output verbalises the next four ideas repeatedly:

under Harding presidency
in 1920s
involved government oil fields
major issue in 1924 election

In this example, the following three paths put into words the same fact about this *definiendum*:
took→during→presidency→of→Entity
took→during→administration→of→Entity
under→administration→of→Entity

Indeed, only three of the eight sentences would be enough to cover all these aspects. Other techniques to detect redundancy can be developed by recognising analogous dependency paths [Chiu et al., 2007]. This brings a key advantage of using dependency paths for answering definition questions. A TREC system, nevertheless, can find this redundancy very profitable when projecting the output into the AQUAINT corpus.

6.5.5 Expanding the Treebank of Descriptive Sentences

In [Figueroa and Atkinson, 2010], the *context models* were extended by taking into consideration two extra snapshots from Wikipedia: one corresponding to January 2007, and the other to October 2008. The intention here is to investigate whether or not the coverage can be widened by accounting for these extra snapshots of Wikipedia.

Following the same previously outlined procedure, two additional treebanks of dependency trees were built, and hence two extra sets of contextual n-gram LMs were generated. The ranking of a candidate sentence S (equation 6.2) was computed by making allowances for the average values of $p(c_s)$ and $p(\vec{dp} \mid c_s)$.

Accordingly, table 6.13 highlights the obtained figures for the two extensions accounting for two and three treebanks, respectively. By and large, the performance was weakened in terms of recall and precision. The gradual decrease in recall may be due to the averaging of two or three treebanks which diminishes the value of low frequent paths as they are not (significantly) present in all the treebanks. Ergo, whenever they match a sentence, the sentence is less likely to score high enough to surpass the experimental threshold (line 14 in algorithm 1). Here, some approaches could be used to exploit inter-treebank smoothing as

Teapot Dome Scandal
◇ NOTES: Presents an examination of the Teapot Dome scandal that took place during the presidency of Warren G. Harding in the 1920s.
◇ Teapot Dome Scandal was a scandal that occurred during the Harding Administration.
◇ This article focuses on the Teapot Dome scandal, which took place during the administration of U. S. President Warren G. Harding.
◇ The Teapot Dome Scandal was a scandal under the administration of President Warren Harding which involved critical government oil fields.
◇ Teapot Dome Scandal cartoon The Teapot Dome Scandal was an oil reserve scandal during the 1920s.
◇ The Teapot Dome scandal became a parlor issue in the presidential election of 1924 but, as the investigation had only just started earlier that year, neither party could claim full.
◇ The Teapot Dome scandal was a victory for neither political party in the 1920's, it did become a major issue in the presidential election of 1924, but neither party could claim full.
◇ The Teapot Dome Scandal was the first huge Federal Government corruption scandal in the 20th century if not in all US history.

Table 6.12: Sample containing issues regarding performance (source [Figueroa and Atkinson, 2009]).

a means of taking away probability mass of the high frequent paths (across treebanks) and distribute it across paths low in frequency in one of the treebanks, but absent in one of the others [Chen and Goodman, 1998]. This steady reduction in precision might stem from the following reasons:

- A diminishment in recall brings about a diminution in the length allowance [Voorhees, 2003].
- Algorithm 1 selected misleading or redundant definitions in place of the definitions matched by the original system, but missed by the two extensions.

On the other hand, highly frequent paths produce more robust estimates as they are very likely to be in all treebanks, having a positive effect in the ranking, as seen in table 6.14. In all question sets, these two extensions outperformed the systems in table 6.9. The growth in MAP values suggests that integrating approximations from various snapshots of Wikipedia cooperates on determining more pertinent and genuine paths. These estimates, along with the preference given by algorithm 1 to these paths, brings about the improvement in the final ranking. As a consequence, more genuine descriptive pieces of descriptive information tend to be conveyed in the highest position of the rank.

RANKING ORDER

6.5.6 Adding Part-of-Speech Tags Knowledge

Tables 6.15 and 6.16 present the outcomes obtained by enriching the *context models* with Part-of-Speech (POS) tags. These context models were constructed on top of the original models, but they account for a treebank in which words labelled with the tags below are mapped into a placeholder:

SELECTIVE SUBSTITUTIONS

DT, CC, PRP, PRP$,CD, RB, FW, MD, PDT, PRP, RBR, RBS, SYM

Additionally, the following verbs, which are normally used for discovering definitions, are mapped into a placeholder: *is*, *are*, *was*, *were*, *become*, *becomes*, *became*, *had*, *has* and *have*.

	TREC 2003	TREC 2004	TREC 2005
Context Models			
Recall	**0.57**±0.17	**0.50**±0.18	**0.42**±0.22
Precision	**0.39**±0.21	**0.40**±0.19	**0.29**±0.21
$\mathcal{F}(3)$-Score	**0.53**±0.15	**0.47**±0.17	**0.38**±0.19
Context Models II			
Recall	0.46 ± 0.17	0.46 ± 0.17	**0.42** ± 0.22
Precision	0.32 ± 0.19	0.38 ± 0.19	**0.29** ± 0.20
$\mathcal{F}(3)$-Score	0.43 ± 0.16	0.44 ± 0.15	**0.38** ± 0.19
Context Models III			
Recall	0.46 ± 0.17	0.44 ± 0.18	0.41 ± 0.21
Precision	0.31 ± 0.17	0.34 ± 0.17	0.28 ± 0.19
$\mathcal{F}(3)$-Score	0.43 ± 0.15	0.42 ± 0.16	0.37 ± 0.18

Table 6.13: Figures for TREC question sets (treebank expansion).

	Context Models	Context Models II	Context Models III
TREC 2003			
MAP-1	0.82	**0.88**	**0.88**
MAP-5	0.82	**0.88**	0.87
TREC 2004			
MAP-1	0.88	0.92	**0.94**
MAP-5	0.82	**0.88**	0.87
TREC 2005			
MAP-1	0.79	0.81	**0.82**
MAP-5	0.77	**0.78**	**0.78**

Table 6.14: MAP (treebank expansion).

The aim of these mappings is to consolidate the probability mass of similar paths, when computing context LMs. For example, the following paths:

was→author→a
is→author→the
is→author→an

are merged into: VERB→author→DT. A more specific example can be seen in figures 6.1 and 6.2, which parallel both models (with and without POS Tagging) for a small context. The idea behind this amalgamation is in the same spirit as the selected substitution introduced by [Cui et al., 2004a, 2007] (section 4.9 on page 93), and it is supported by the fact that some descriptive phrases, including "*Concept was an American author...*" and "*Concept is a British author...*", share some common structure that is very likely to render definitions, ergo consolidating their probability mass is reasonable. This certainly helps to tackle the data-sparseness of the *context models*. Particularly, figures 6.1 and 6.2 show the way the paths: co-lyricist→was and co-lyricist→is, are fused into co-lyricist→VERB. Naturally, the new likelihood is the sum of both original paths.

Table 6.15 highlights the figures achieved by this strategy when juxtaposed with the original model. In general, the three extensions outperformed the ranking with respect to the

	Context Models	Context Models + POS
	TREC 2003	
MAP-1	0.82	**0.88**
MAP-5	0.82	**0.88**
	TREC 2004	
MAP-1	0.88	**0.91**
MAP-5	0.82	**0.87**
	TREC 2005	
MAP-1	**0.79**	0.73
MAP-5	**0.77**	0.71

Table 6.15: MAP (POS Tagging).

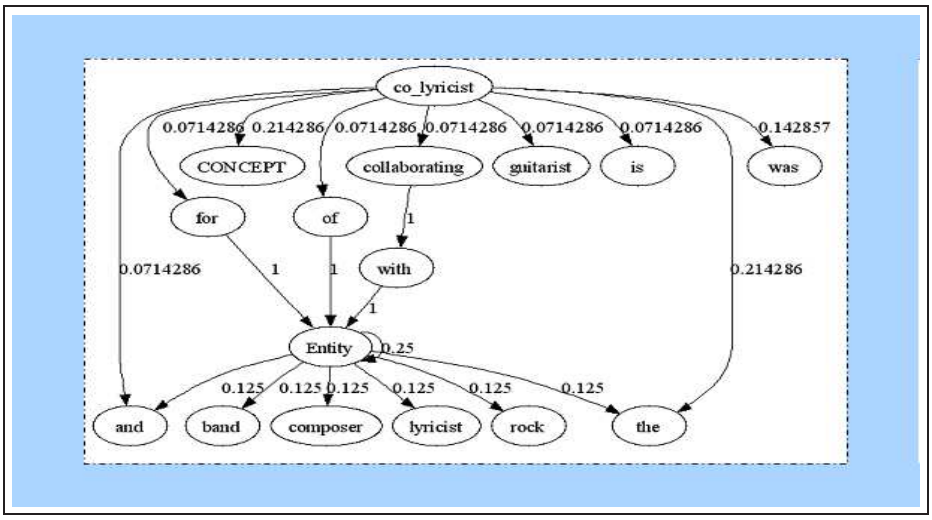

Figure 6.1: Bigram raw probabilities for c_s = "*co-lyricist*".

original *context models* (see table 6.14). Anyways, the experiments did not show a clear distinction on which is the best in this aspect. In the particular case of the POS-based method, results indicate an increase with respect to the original system for two datasets, but a lessening in the case of the TREC 2005 questions set. Unlike the two previous question sets, abstracting some syntactic categories leaded some spurious sentences to rank higher.

More interestingly, table 6.16 emphasises the marked decline in terms of $\mathcal{F}(3)$-Score for two datasets, while remarking a substantial improvement for the TREC 2005 question set, more exactly, when compared with the figures achieved by the original model. This improvement is due particularly to the growth in recall. This means the amalgamation of dependency paths cooperated on identifying a higher amount of genuine descriptive sentences. On the other hand, the addition of POS knowledge tagging assisted in matching more misleading and spurious sentences, and as a repercussion, it worsened the performance in terms of precision. This might also explain the decrease in the MAP value for this question set. From these observations, it is easy to see that the treebanks without POS tags cover less descriptive sentences embraced in this question set. In the TREC 2003-2004 question sets, the deprovement might result from the fact that some original paths are indispensable to recognise several sentences.

PERFORMANCE

	TREC 2003	TREC 2004	TREC 2005
Context Models			
Recall	**0.57**±0.17	**0.50**±0.18	0.42±0.22
Precision	**0.39**±0.21	**0.40**±0.19	**0.29**±0.21
$\mathcal{F}(3)$-Score	**0.53**±0.15	**0.47**±0.17	0.38±0.19
Context Models + POS			
Recall	0.56 ± 0.16	0.47 ± 0.15	**0.48** ± 0.21
Precision	0.24 ± 0.14	0.22 ± 0.09	0.24 ± 0.19
$\mathcal{F}(3)$-Score	0.48 ± 0.15	0.41 ± 0.12	**0.42** ± 0.18

Table 6.16: Figures for TREC question sets (POS Tagging)

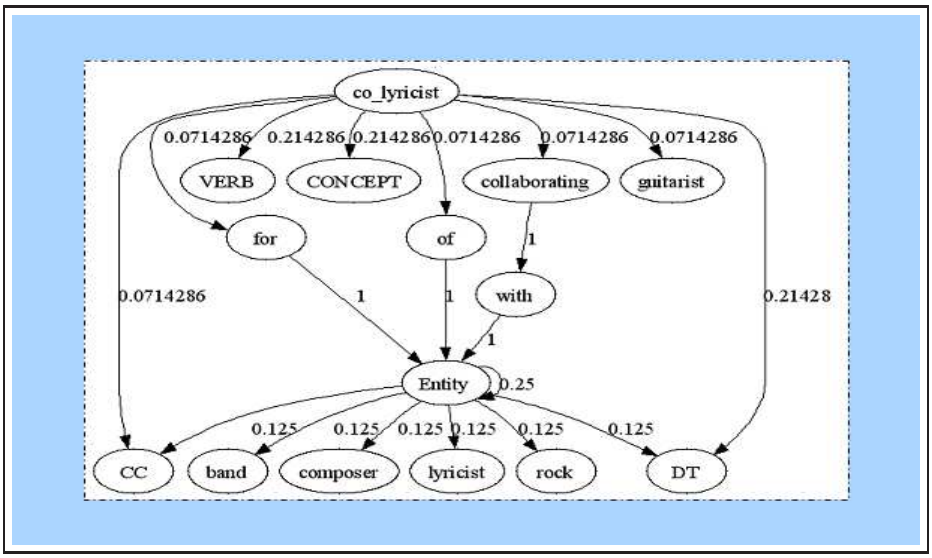

Figure 6.2: Bigram raw probabilities for c_s = "*co-lyricist*" accounting for POS taggings.

6.5.7 Improving Definition Extraction using Contextual Entities

The *context models* introduced so far have been built on the assumption that entities themselves do not provide meaningful information to recognise extra definitions. There are two legitimate reasons to assume this:

1. Replacing entities by a placeholder allows getting reliable dependency paths counts, and consequently, more trustworthy probability estimates.

2. Lexicalised dependency paths consolidate almost all the information about the type of relationship between the *definiendum* and some common type of relation in the respective context. For instance, the next paths with respect to the context "*novel*" underline that the implicit type of entity is a $writer$:

ROOT→novel→written→by→Entity
ROOT→novel→co-authored→by→Entity
ROOT→novel→authored→by→Entity

Hence, independently from what this entity really is, it is very likely to be a *writer*. This sort of method prevents *context models* from being reliant on any previously annotated or

automatically extracted list of *writers*. However, a major drawback is that some prominent entities within contexts could be useful in identifying extra descriptive content. To reinforce this point, consider the following descriptive sentences harvested from Wikipedia:

Wanderlust is a 1986 romantic novel, authored by Danielle Steel.
Kaleidoscope is a 1987 novel by Danielle Steel, published by Delacorte Press.
A Perfect Stranger is a Danielle Steele romance novel, published in 1981.

Hence, one would think that whenever a sentence has the *context indicator* "*novel*" along with the entity "*Danielle Steel*", this sentence will be very likely to verbalise descriptive information (i.e., a description expressed by paths not previously learnt).

In order to infer these relevant contextual entities, all pairs <*context indicator*, entity> were extracted and their respective mutual information H were posteriorly computed. It is worth pointing out that only pairs having a frequency higher than two were taken into account, and pairs with $H \leq 0$ were discarded. Table 6.17 stresses some relevant pairs in which capitalised adjectives were perceived as entities when constructing the treebanks.

MUTUAL INFORMATION

Entity	H	H + Alias Resolution
Agatha Christie	3.16423	4.62592
BBC Books	7.63855	8.99078
Baen Books	6.09107	8.10806
Ballantine Books	5.79151	8.6317
Bantam Books	5.41299	7.35776
Black Spring Press	7.05454	8.48454
British Nobel	7.79151	8.99537
DAW Books	6.62158	8.40758
Danielle Steel	7.69197	8.97289
Dell Publishing	7.05454	7.7426
Fyodor Dostoevsky	5.84397	7.35776
G. K. Chesterton	5.20654	7.85957
Hermann Hesse	5.79151	7.59002
Oxford University Press	3.1768	5.28791
Prize-winning	4.77358	6.89871
Pulitzer	3.60168	6.01041
Pulitzer Prize for Fiction	6.20654	7.52903
Science Fiction	4.80544	7.35314

Table 6.17: Some prominent entities within the context c_s="*novel*".

In order to assess the impact of leaving unconsidered entities when building *context models* and thus rating candidate sentences, a benchmark was carried out in which all sentences that were not selected by the algorithm 1 were checked as to whether they matched any pair <*context indicator*, entity>. If any match, the corresponding sentence is added to the end of the output. Ergo, the previous order of the top ranked sentences remains unmodified. This is premised on the observation that if entities were important to recognise novel information, the method would ameliorate the overall recall. This single strategy enlarged the output for 19 and 35 out of the 50 and 64 TREC 2003 and 2005 questions, respectively. No novel nuggets, however, were discovered.

EXPERIMENTS

PERFORMANCE

A slightly different view is seen in terms of the TREC 2004 question set in which the response was extended for 24 out of the 64 questions, but in four cases the recall was raised.

This means that this loose match helped to identify novel descriptive information with a better overall $\mathcal{F}(3)$-Score (see table 6.18).

Definiendum	Recall(R)	Precision(P)	$\mathcal{F}(3)$-Score
Frank Kakfa	0.47 → 0.5	0.51 → 0.54	0.47 → 0.5
Abercrombie and Fitch	0.14 → 0.32	0.24 → 0.53	0.14 → 0.33
Jack Welch	0.35 → 0.38	0.48 → 0.49	0.36 → 0.39
Chester Nimitz	0.18 → 0.23	0.18 → 0.22	0.18 → 0.23

Table 6.18: Performance improvement based on entities.

From these figures, entities were found to play no essential role in *context models*. Furthermore, entities are usually written in several ways. For instance, person names such as "*George Bush*" can be found as "*George W. Bush*", "*G.W. Bush*", "*President Bush*". These variations make it difficult to match entities in the models with entities in the target array of sentences, additionally they make it difficult to learn accurate distributions from the training data. For reasons already given, an alias resolution step is necessary when learning entities and rating new sentences. For this purpose, the alias repository presented in [Figueroa, 2008a] (section 2.6.1 on page 45) was utilised. Two entities in a context were deemed to be the same whenever there was an entry in this repository that maps one entity to the other. In every match, the entries in relation to both entities are removed, the frequencies of the matching entities consolidated, and one new entry is created embodying both entities and the new amalgamated frequency value. This algorithm is applied iteratively until no entry exists in the repository for any pair of entities in the context. The outcome of this iterative process is a set of entries, each containing a set of entities, and their consolidated frequencies. Accordingly, the values of H were now recomputed by considering entries and contexts.

ALIAS RESOLUTION

Table 6.17 outlines some illustrative variations in the values of H for some entities in the "*novel*" context. Merging aliases is geared towards augmenting the number of reliable entities in each context by clustering low and high frequent aliases. For instance, the amount of entities in the context "*novel*" boosted from 1364 to 2814 (e.g., "*Alternative Metal*", "*Elecktra Records*" and "*Ferret Records*"), while "*singer*" from 411 to 678 (e.g., "*BBC*", "*Latin Grammy Awards*" and "*Yngwie J. Malmsteem/Yngwie Malmsteem*"), "*language*" 443 to 579 (e.g., "*Haiti*", "*Iberia*" and "*Punjabi*"), and "*band*" from 974 to 1389 (e.g., "*Soho Press*", "*Lancer Books*" and "*Marina Lewycki*"). Despite this increment, a noticeable improvement was not observed.

6.5.8 Relatedness/Similarities Between Contexts

For the purpose of examining the relatedness and/or similarity between pairs of *context indicators*, a matrix M was built. Specifically, each cell M_{ij} in this matrix denotes the frequency of the path i in the context j. The dimensions of this matrix comprise 45,698 different *context indicators* and 26,490,042 different n-gram paths. As a means of strengthening the relationships between contexts j and their delineative paths, the value of each cell M_{ij} is rewritten by the mutual information value between the path i and the context j. This measure lowers the effect of paths that are highly frequent in many contexts, while it makes comparatively stronger paths that are highly frequent in few contexts, and thus more representative of those few contexts. This measure can also assign negative values to some paths in certain contexts. These paths can be interpreted as signal of weak relations. This matrix is subsequently utilised for computing the cosine similarity cosine(c_{s_1},c_{s_2}) between two contexts c_{s_1} and c_{s_2}. It is worth noting that determining the similarity of all pairs of contexts demanded ten days running on a four CPU server.

Figure 6.4 illustrates the strongest relations between the highest-frequent *context indicators* listed on table 6.3. This graph presents an isolated cluster of ten contexts (e.g., *area*, *city*, *community*, *district* and *town*) that can imply the description of physical places. Interestingly enough, the contexts *station* and *town* share some paths: founded→in→Entity and located→in→Entity, while at the same time, the context *station* encircles paths that are very unlikely to be found in descriptions of *towns*:

aired→from→Entity→Entity
broadcast→across→area→Entity

Certainly, the degree of similarity is naturally given by the overlap between both contexts, and in this scenario, transitivity must be carefully handled. For instance, the cosine drew a value of 0.28 for the pair *district* ↔ *town*, and 0.22 for the pair *town* ↔ *center*, but the pair *district* ↔ *center* obtained a lower value (0.159). This can be conceived as a result of likening two more specific contexts (*center* and *district*) with a more general or abstract context (*town*). Substantially, the semantic range of the more specific contexts is included within or is part of the more general context. Therefore, descriptions of instances of specific contexts are likely to include some aspects related to their respective more general context(s), causing a greater similarity between the specific and general contexts. However, each of the most specific contexts can emphasise radically different aspects of the general context. This along with the fact that the individualising facets of the specific contexts are pertinent when describing them, brings about a greater dissimilarity between the semantic extensions. Another good example happens when comparing *artist* ↔ *painter* (0.2385) and *artist* ↔ *singer* (0.2045) as well as *painter* ↔ *singer* (0.1061). TRANSITIVITY

On the other hand, the cosine gave the value of 0.208 for the tuple *town* ↔ *station*, while 0.17 and 0.47 for the tuples *station* ↔ *city* and *town* ↔ *city*, respectively. From a simple viewpoint, the contexts *town* and *city* are highly likely to be described by utilising the same types of nuggets (paths), the difference is due mainly to the value of some of their attributes (e.g., foundation date, location, size and number of inhabitants). In actuality, in many cases, due to this ambiguity, some individuals would raise an eyebrow when someone calls a *city* what they consider a *town*. However, the distinction is clearer between a *station* and a *city*, because the specific information that disambiguates a *station* from a *city* is indispensable in a definition, causing an increase in their dissimilarity. Another reason that boosts the divergence between these contexts is that *station* is ambiguous: an army base, a bus stop, and radio station. This ambiguity also stresses the need for incorporating discriminative aspects into the definition. AMBIGUITY

Analogously to the contexts *city* and *town*, the strong similarity between synonyms (e.g., *book* ↔ *novel* and *song* ↔ *single*) and genders (e.g., *actor* ↔ *actress*) can be explained. In like manner, some contexts, such as *son*, *daughter* and *born*, are closely related. In this particular case, the underlying reason is that the information about the birth and lineage/ancestry of a person is occasionally conveyed in the same descriptive sentence, causing them to appear as synonyms. Some illustrative examples: SYNONYMS

CONCEPT was a daughter born to a family of Entity artists five generations ago .
CONCEPT was the second son of Entity and was born in Entity , where his father was a banker .

Another conclusion regards some contexts corresponding to objects, which are closely related to a particular type of person. For instance, *album* ↔ *singer* and *author* ↔ *book*. Certainly, this is due to the fact that descriptions of albums usually include their singers, and descriptions of singers can include information about their albums. Further, this graph also displays some strong hyponymic relationships: *artist* ↔ [*painter*, *writer*, *singer*], *musician* ↔ [*composer*, PARTONOMIC RELATIONS

singer]. Furthermore, some part-of relations are also signalled as very close: *character* ↔ *series* and *band* ↔ *singer*.

In addition, it can be observed that contexts like *film*, *game*, *series*, and *novel* are largely related. These relationships can occur as a result of the fact that many films are based on novels, games on films, series on novels, making these types of definitions to have a clear overlap.

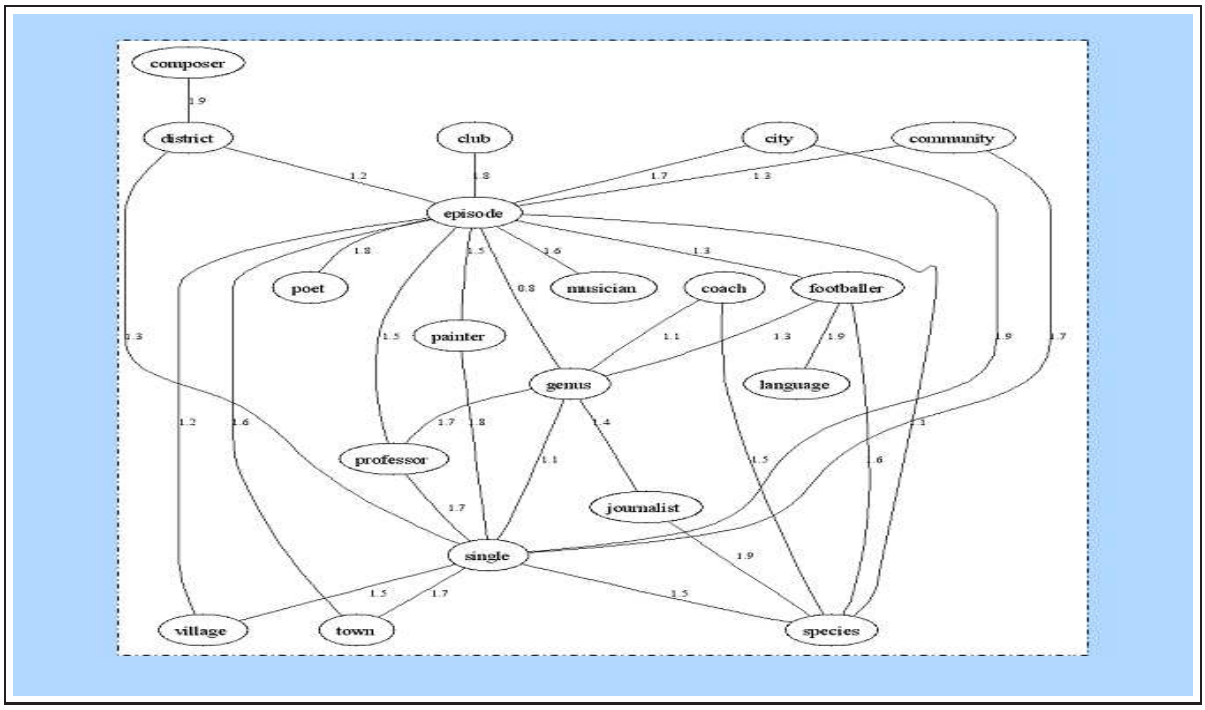

Figure 6.3: Highest-frequent dissimilar contexts (cosine(c_{s_1},c_{s_2})*100).

DISSIMILARITY

In the opposite way, figure 6.3 exhibits the most dissimilar pairs of contexts. Nodes with a high number of edges symbolise the most divergent contexts. Excellent examples are the nodes: *episode*, *single*, *species*, *genus*. Results demonstrate that descriptions of *episodes* are almost unrelated to definitions of type of people such as *professor* and *painter*, and also unrelated to descriptions of locations such as *city* and *club*. One interesting dissimilarity is due to the contexts *footballer* and *language*. In light of these results and concerning the three specific models utilised by [Han et al., 2006], it can be concluded that more specific types of *definiendum* must be considered when rating definitions. That is, the three types established by TREC are too general to model many divergent types of *definiendum*. The outcomes also reveal the need for a taxonomy that integrates the relevant types of nuggets for each class, and that specifies the association amongst different classes in the taxonomy.

In short, the similarities and dissimilarities amongst different highest-frequent *context indicators* support the initial hypothesis that discriminative *contexts models* can be inferred at the sentence level. More precisely, these models can be deduced from sentences matching wide-spread definition patterns, and used thereafter for scoring candidate sentences accordingly.

SENSE DISCRIMINATION

On the whole, *context indicators* provide hints about discriminating some potential senses. In a special manner, they can help to disambiguate people and places named after individuals, while they might not be that helpful in distinguishing between several versions of the same movie (e.g., "*Ben-Hur*"). In these cases, one can see that a more complex strategy that learns discriminative features by examining examples is necessary. For instance, dates can be

very important to recognise various versions of movies. Nevertheless, sense discrimination is a difficult problem [Chen et al., 2006]. All in all, the outcome emphasises the importance of the *context models*, and of the specific paths that characterise each particular context.

6.5.9 Some Semantic Relations Inside Contexts

As a matter of fact, there is always a semantic relationship between each pair of contexts. The intrinsic difference between these relations is their degree of strength. As mentioned earlier, there are contexts that bear a strong resemblance, whereas other pairs are clearly dissimilar. At any rate, it is not only inter-context patterns of similarity that matter, but also, interesting semantic relationships that can show up within each context. To be more specific, in some cases, the *context indicator* establishes a well-known semantic connection with some terms, which can be recurrently manifested across descriptions about the manifold *definienda* that can be derived from respective context.

A simple way of detecting the most important or relevant groups of semantic relations consists in checking synsets in WordNet that contain the *context indicator*, and their stipulated relations to other synsets thereof. More exactly, WordNet produces about twelve semantic relations including hypernym/hyponym, meronym/holonym. It also yields antonyms and pertanyms. Strickly speaking, when exploiting WordNet for distinguishing relationships across the whole treebank, 17,219 distinct pairs <*context indicator*, term> were labelled.

Context Indicator	Hyponyms/Hypernyms
place	abode, area, birthplace, center, coffin, contiguous, job, left, middle, park, point, rank, residence, right, sanctuary, square, status, stop, tomb, vicinity.
organization	activity, administration, alliance, body, brotherhood, coalition, company design, establishment, federation, government, party, regime, union.
area	acreage, arena, construction, environment, heart, land, middle, open, place, playground, resort, sanctuary, space, terrace, territory.
player	accompanist, bowler, flautist, footballer, guitarist, lutenist, oboist, performer, pianist, plant, scorer, seed, shooter, soloist, star, vocalist.
book	album, authority, booklet, catalogue, hardback, journal, notebook, novel, paperback, production, publication, record, register, text, ticket, tome.
musician	arranger, artist, composer, conductor, director, harpist, singer, soloist, virtuoso.
author	biographer, coiner, compiler, essayist, ghostwriter, lyricist, poet, scriptwriter.
film	create, docudrama, episode, footage, medium, negative, scene, sequence.
leader	boss, captain, chief, commander, father, head, imam, inspirer, politician.
language	expression, module, string, text, word.
genus	form, kind, plant genus, taxon, variety.

Table 6.19: Some samples of hypernyms/hyponyms related to a subset of the *context indicators* listed on table 6.3.

The first sort of relation recognised by WordNet is hypernym-hyponym pairs. Precisely, the amount of distinct pairs identified raised to 13,217, thus indicating the importance of this sort of relation. Overall, hypernymy encompassed 76.76% of the automatically labelled relationship pairs. To exemplify, table 6.19 underscores some hypernymy regarding the most prominent *context indicators*. As a rule, the number of times that a particular pair is found is relatively low with respect to the frequency of its context. By the same token, the number of global distinct pairs labelled by WordNet (17,219), is also comparatively low with regard to the total amount of context models (45,698). This can be attributed to data-sparseness, and/or due to the fact that the kinds of connections, discerned by WordNet, seldomly occur across definitions that match the rules illustrated in table 3.1 (page 65). Nonetheless, an

HYPERNYM/ HYPONYM

inspection of these matchings suggests that some relations might be fundamental and that they can be typified by some paths:

place→in→center→the
place→center→is
place→center→of

organization→body→for
organization→body→is
organization→body→member
organization→body→the

player→finished→scorer→of→Entity
player→scorer→leading

Context Indicator	Hyponyms/Hypernyms
place	...a major place for cement production in Entity, despite the fact that it is a populous place in the city *center*. ...the meeting place of the council for the municipality and is a service *center* for the surrounding farming community . ...the current resting place of the *heart* of Entity, commonly known as Entity. ...a popular place for weddings, as it has a historic *sanctuary* and is located in downtown Entity.
organization	...a standards organization and is the Entity member *body* for Entity. ...the official Entity non-profit organization charged with overseeing the international *coalition* of poetry slams.
area	...an area of *acreage* blocks and small farms. ...a small area of central Entity, named after the old market *square* at its heart. ...a historical area that lies in the *heart* of Entity.
player	...a major player in the bottled water business after Entity bought the small bottling *plant* in Entity. ...a former soccer player, the all-time leading *scorer* for the Entity national team.
book	...a book published by Entity, the fourth *tome* of the works of Entity. ...a book which compiles a *register* of numerous commercium songs .
language	...the Entity programming language plus a graphics *module* called Entity.
genus	...the one genus of the *taxon* Entity from the Entity site whereby the lower and upper jaws have been found united.

Table 6.20: Some definitions about the *context indicators* on table 6.19. These definitions express their connection with some of the listed hypernyms/hyponyms.

Table 6.20 highlights some descriptions bearing some of the pairs in table 6.19. Further, the amount of holonym/meronym discovered by WordNet reached 1,165 distinct matches, which is a number substantially lower than the previous type of relation. For example, the context *book* is linked with the words: *binding*, *cover*, and *text*, while the context *film* with: *credit*, *episode*, *scene*, *sequence* and *shot*. Other contexts such as *program* and *song* also signified interesting matches such as <*program*,*command*> and <*program*,*statement*> as well as <*song*,*chorus*>. Some illustrative sentences that underline the essentiality of this class of semantic relation are:

HOLONYM/ MERONYM

CONCEPT is a hard-rocking electronic song with retro synths and an infectious singalong chorus.

CONCEPT is a lightweight mail program which has a command syntax similar to ed.
CONCEPT is a children's book with text by Entity and illustrations by Entity.

These examples show that meronyms of the *context indicator* can be utilised with the goal of expressing specific characteristics of the *definiendum*, which are inherently connected with its meronym. To illustrate, take expressions such as: "*infectious singalong chorus*", "*command syntax similar to ed*", and "*text/illustrations by Entity*". This means the essential characteristic of the chorus (a part) of this song is the fact that it (the chorus) is "*infectious singalong*". Furthermore, WordNet labelled 448 different relations as *pertainym*. Some good examples are: <*film*,*cinematic*>, <*poet*,*poetic*> and <*title*,*titular*>:

PERTAINYM

CONCEPT is a title held by the Entity which signifies their titular leadership over the Entity of Entity.
CONCEPT is a Entity poet and the founder of poetic transrealism in contemporary poetry.
CONCEPT is a Entity action/comedy film, the second of the Entity cinematic releases.

What is more, WordNet discriminated 620 different antonyms of the *context indicator* that can potentially co-occur in a description. Take, for instance, the following sentence together with the pair <*leader*,*follower*>:

ANTONYM

CONCEPT was a leader of the Entity, a prominent follower of Entity and Entity revolutionary whose communist views of spreading wealth to the poorer classes earned him great popularity.

To sum up, an examination of descriptions belonging to the training material reveals that terms in a particular definition can signal specific semantic relations. This finding has the potential for aiding in the ranking of answer candidates. However, there were several relations undetected by WordNet, one can hence conjecture that some statistical techniques including Latent Semantic Analysis (LSA) can be helpful to: (a) broaden the coverage of the relations that WordNet can identify, and (b) tackle data sparseness head-on by inferring new semantic connections between context indicators and terms that do not directly appear in the respective context. Nevertheless, one can also envision that synsets in WordNet that match an existing relation can assist in combating data-sparseness. Put differently, one can assume that elements in each matching synset are interchangeable, analogously to the procedure used by [Han et al., 2006] for cushioning the effects of redundancy.

Context Indicator	Terms
album	country-rock, full-length, hard-glam, post-metal, two-CD
band	deathcore, four-piece, metal-trash, power-pop, sleaze
born	apprenticed, attended, baptised, graduated, gubernia, immigrated
character	fictional, legendarium, mutant, now-cancelled, portrayed, soap
company	biopharmaceutical, headquarted, manufactures, privately-held, publicly-traded
episode	first-season, one-hundred, reimagined, sit-com
film	action-comedy, black-and-white, crime-drama, directed, rockumentary
genus	algae, artiodactyl, catfishes, dromaeosaurid, meat-eating, ornithopod
member	co-chaired, constituents, entomologique, homegrown, senate
school	all-boys, boys-only, elementary, enrolls, fee-paying, secondary, tuition-free
town	atlas, census, ghost, notified, subregion

Table 6.21: Some terms highly statistically related to some of the *context indicators* listed in table 6.3.

In the previous section, the cosine, calculated in consonance with the matrix M_{ij}, threw some light into the semantic relatedness between distinct contexts. Although this matrix

STATISTICAL RELATIONS: WORDS

might not seem as attractive as a more powerful methods such as LSA, it can still cooperate on foreshadowing what might come to light when benefiting from more effective, but at the same time, more computationally demanding statistical techniques. Table 6.21 depicts some of the outcomes representing some of the terms that have a high M_{ij} value and occur more than ten times. In this table, for the sake of clarity, paths were omitted on purpose. Recurrently, specific adjectives seem to take over the most statistically significant relations. Still yet, one can notice other kinds of relations including <*born*,*baptised*>, <*company*,*manufactures*>, and <*film*,*directed*>. This connotes that both statistically and linguistically based methods can be integrated as a means of recognising a wider range of semantic relationships.

STATISTICAL RELATIONS: PATHS

What about paths? The matrix M_{ij} also signals the strong linkage between *context indicators* and some of its n-gram paths. Table 6.22 illustrates some representative relations that have a high M_{ij} value and which, at the same time, embody paths occurring more than ten times across the entire treebank of contextual descriptions. In this table, paths of all plausible lengths can be seen: short paths (e.g., winger→for) and longer paths (e.g., book→written→by→Entity→sociologist). Furthermore, this subset highlights the significance of longer paths carrying entities that typify their respective contexts, and therein lies the pertinence of this factor when scoring answer candidates. Note that these paths differentiate from the more "standard" relations shown in tables 6.5 and 6.6. Above all, this experimental observation supports the fundamental postulate of *context models* (see section 6.5): descriptions directed at a particular type of *definiendum* are mainly characterised by dependency paths connected with this type.

Context Indicator	N-gram paths
actor	actor→acted→in→films, actor→best→known→for, actor→starred→in→Entity, co-starred→in→Entity→Entity, remembered→for→role→in→Entity
album	album→by→Entity→songwriter, first→album→studio, group→of→name→the, recorded→during→tour→Entity, under→label→record, feature→Entity→drummer
book	book→published→originally, book→written→by→Entity→sociologist
character	appeared→in→Entity→created→by, character→played→by→Entity, portrayed→from→Entity→Entity, within→Entity→novel, character→antagonist
footballer	defender→for→Entity, midfield→for→Entity, without→club, winger→for, played→as→goalkeeper→for→Entity, striker→for, contract→for→Entity,
genus	genus→belonging→to→Entity→the, genus→extinct, species→of→trees, of→mammal→from→Entity, comprising→about→species, diversity→highest
member	board→of→directors→of→Entity, co-chaired→by→Entity, councils→tribal, constituents→in→Entity, of→faculty→at, alliance→led→by, dynasty→political
politician	politician→chairman→of→Entity, politician→father→of, served→as→to
term	coined→by, describe→types→of, meanings→two, applied→to→number→of

Table 6.22: Some n-gram paths highly statistically related to some of the *context indicators* listed on table 6.3.

6.5.10 Projecting Answers into the AQUAINT Corpus

As a means of finding documents related to the *definiendum* across the AQUAINT corpus, the collection of documents was indexed by Lucene, and the top one hundred documents were fetched by querying the *definiendum* to this Information Retrieval (IR) engine. The number of fetched documents and the IR Application Programming Interface (API) vary among different definition QA systems. For instance, [Han et al., 2006] retrieved the top 200 documents by means of OKAPI, and produced an output of a maximum of 2,000 bytes in length. Conversely, [Chen et al., 2006] fetched the top 500 documents by means of Lemur, and the best

TREC 2003 system retrieved a maximum number of 1,000 documents [Xu et al., 2003], outputting an answer of 4,000 bytes maximum in length. The best TREC 2004 system fetched a maximum of 800 documents per *definiendum* [Cui et al., 2004b].

In the first place, the one hundred documents retrieved from the AQUAINT corpus were split into paragraphs in agreement with the structure provided by the documents in the collection. In the second place, paragraphs containing a query term, excluding stop-words, are selected. In the third place, co-references are resolved within the chosen paragraphs by means of JavaRap. In the fourth place, each selected paragraph is divided into sentences in congruence with OpenNLP. Last, sentences carrying a query term, excluding stop-words, are returned as answer candidates. It is worth underlining here that a smaller group of documents was utilised as sources of putative answers in order to manually inspect them, hence making it possible to measure the performance independently from the quality of the retrieval system. Overall, more than 18,000 different sentences were manually annotated in consonance with the TREC ground truth, and larger amount of sentences would have made this manual task more demanding.

Even though many definitions match the pre-determined battery of rules across the ∼18,200 sentences, normally few matches occur per *definiendum*. Therefore, the direct application of rule-based *context models* would have few chances of achieving a significant recall, and ergo, a competitive $\mathcal{F}(3)$-Score. It is worth noting here that this score is biased towards recall. All things considered, *context models* were adapted or extended as follows in order to deal with the AQUAINT corpus:

CONTEXT MODELS FOR TREC

1. Top definition QA systems traditionally take advantage of KBs for learning a topic model or extracting nuggets that are projected into the set of candidate sentences afterwards. The dependence of these systems on the coverage of these KBs makes them less attractive [Zhang et al., 2005, Han et al., 2006]. For instance, [Han et al., 2006] capitalised on eight different KBs, while the best TREC 2003 and 2004 systems of six [Xu et al., 2003, Cui et al., 2004b] (see sections 2.4 and 4.8 on page 35 and 91). On the contrary, [Chen et al., 2006] profited from task specific clues in order to enhance the retrieval of web snippets from biography web-sites and on-line dictionaries (section 2.4 on page 35).

   The first adaptation consists in replacing these KBs with the output produced by *context models* applied to sentences that align definition patterns originating from web-snippets, that is, the output corresponding to the performance detailed in table 6.7. Certainly, this limited group of sentences supplies fewer redundancy and diversity of nuggets per *definiendum* than full pages acquired from six and/or eight KBs, and it is not as authoritative as these KB articles as well, but this set of sentences cushions the dependence on the coverage of KBs.

2. A topic model is deduced from this output obtained from the Internet. This topic model is built on top of the n-gram LMs and dependency paths presented in section 6.5.2, but it disregards contexts, that is, all sentences are seen as belonging to the same context. This restriction is slackened due to the fact that they are likely to be related to the predominant senses in the Web, and this model is aimed specifically at rating any answer candidate from the AQUAINT corpus, including those that mismatch the pre-determined array of definition patterns.

3. Answer candidates are scored in accordance with these topic models multiplied by a brevity penalty. As for the brevity penalty, the factor in equation 6.3 was utilised with the parameter L_{ref} empirically set to eight. Thus, algorithm 1 was utilised for singling out sentences up to a maximum length of 3,500 characters. It is worth pointing out

here that if there was still allowance, the output was augmented with the remaining highest, but usually few, ranked sentences that match definition patterns.

TREC-2003 System	$\mathcal{F}(5)$-Score	TREC-2004 System	$\mathcal{F}(3)$-Score
BBN	0.555	National Univ. of Singapore	0.460
[Chen et al., 2006] biterms+LMs	0.531	Fudan University	0.404
National Univ. of Singapore	0.473	National Security Agency	0.376
University of Southern California	0.461	University of Sheffield	0.321
Language Computer Corp.	0.442	University of North Texas	0.307
Context Models	**0.440**	***Context Models***	**0.299**
+ Projection	0.594	***+ Projection***	0.355
Univ. of Colorado/Columbia Univ.	0.338	IBM Research	0.285
ITC-irst	0.318	Korea University	0.247
University of Amsterdam	0.315	Language Computer Corp.	0.240
MIT	0.309	CL research	0.239
University of Sheffield	0.236	Saarland University	0.211
University of Iowa	0.231		

Table 6.23: Comparison with top-10 TREC Systems(sources [Voorhees, 2003, 2004, Chen et al., 2006]).

Table 6.23 depicts the performance of *context models* + the projection strategy with respect to top ten TREC 2003 and 2004 systems. For each dataset, two values are presented: the first and lower value is computed against the gold standard, and the second and higher value with respect to the nuggets in the ground truth which are also within the whole set of answer candidates. This last value signifies the performance with respect to the set of sentence inputted to the answer extraction module. That is, it filters out the effects of the retrieval and mistakes in the co-reference resolution step. This value was computed with respect to 46 and 62 sentences corresponding to the TREC 2003 and 2004 datasets, respectively. This means there were no vital nuggets across the fetched answer candidates for four and two queries pertaining to the TREC 2003 and 2004 question sets, respectively.

PERFORMANCE

From TREC 2003 to 2004, the $\mathcal{F}(3)$-Score, with respect solely to retrieved vital nuggets, diminished from 0.48 to 0.355. A reason for this worsening stems from the context questions incorporated into the 2004 track. Context questions comprise a sequence of queries about a specific target concept (*definiendum*). These queries encompass factoid and list questions, whose answers, as [Han et al., 2006] pointed out, could perfectly be part of the gold standard of its respective definition query. The crucial issue here is that systems are forced to remove these answers from the output of the definition question, and they are, for this reason, left unconsidered from the ground truth of the corresponding definition question. Since it is unfair to account for the right answers in the gold standard, because TREC systems in praxis do not know them, the answers to these previous queries were not taken into consideration when evaluating *context models* + the projection strategy. At any rate, the impact of these answers is certainly a decrease of $\mathcal{F}(3)$-Score, as many of these nuggets are captured, hence enlarging the response, but they do not contribute to the recall nor to the precision.

IMPACT OF TREC GROUND TRUTH

Another source of errors is anaphora resolution. In some cases, nuggets were missed, due to wrong inferences drawn by JavaRap. To reinforce this empirical observation, consider the next sentence:

The deal "will extend Rohm and Haas's technology platform beyond The deal's premier position in acrylic chemistry and electronic materials," Wilson said.

In this phrase, the resolved "*The deal*" should be replaced by the *definiendum* "*Rohm and Haas*", making it possible to recognise, and thus also annotate, one of the two vital nuggets for this *definiendum*: "*had a premier position in acrylic and electronic materials*".

Like [Han et al., 2006], *context models* + the projection strategy (CM+PS) also ranks among the vanguard definition QA systems when coping with these two question sets. One vital aspect that must be kept in mind when contrasting with the top TREC systems is that CM+PS makes allowance for a restricted group of sentences sifted automatically from the Web, whereas top TREC systems account for a battery of wrappers that take evidence from distinct KBs. This evidence is well-known to substantially boost the performance in these two sets [Zhang et al., 2005]. Additionally, CM+PS considers a small set of target documents, contrary to top TREC systems. This is also particularly relevant, because for some *definienda*, such as "*The Clash*", only unrelated documents were included in the top hundred hits returned by Lucene. Further, CM+PS does not learn local statistics or regularities from the respective target set of sentences. In light of these remarks and the performance achieved by CM+PS, it can be concluded that LMs and dependency paths can also offer a competitive alternative to inferring topic models from sentences regardless of the fact whether or not they match a definition pattern, and project them into the target corpus afterwards. Lastly, figures in table 6.23 additionally corroborate the performance of the web system premised on *context models* and definition patterns, as it seems to cover many of the TREC nuggets.

6.6 Conclusions and Further Work

This chapter mainly dissects the use of LMs for rating candidate answers to definition questions in English. In the first place, this chapter highlights a definition QA that capitalises on four distinct LMs for answering definition queries in the context of TREC 2007. These four language models are induced from different corpora, and as an achievement, this system captured the first place in the definition QA subtask. Another attractive feature of this system is the exploitation of four distinct dependency relations when ranking putative answers.

In the second place, this chapter fleshes out a comparison amongst LMs operating with three distinct features: unigrams, bigrams and biterms. The best performance was crystallised with the incorporation of biterms, shedding light to the pertinence of syntactic information to distinguish genuine answers, in particular the relative order of pairs of co-occurring words.

In the third place, this chapter outlines a method for scoring candidate answers that intermixes assorted unigram LMs: (a) one inferred from an array of documents fetched from the target collection; (b) another derived from eight KBs; (c) one deduced from the top ten hits returned by Google; and (d) a definition model that has two branches: one regarding three types of *definienda* and one general definition model. More exactly, the first three models are mixed as a means of creating to a topic model that inherently resembles traditional projection strategies. Their findings concern the decisive effect of KBs-based models in the good performance, the insufficiency of documents retrieved from the collection and the Internet, and the positive impact of their three specific models.

In the fourth place, this chapter elaborates on *context models*, which have their roots in the findings of the last two methodologies:

1. They assemble a set of 45,698 specific models in relation to various types of *definienda*. This framework is constructed automatically and it produces a sharper separation of semantic units. Further, *context models* do not capitalise on articles about *definiendum* across KBs, nor profit from a general definition model. A keystone principle of these

models is that definitions are principally typified by dependency paths connected with their types.

In this respect, the examination of the *context models* undoubtedly indicates that a semantic relation exists between them. To be more specific, empirical observations unveil that some pairs of contexts are more dissimilar than others (e.g., *painter* ↔ *singer*), while other pairs share a greater similarity (e.g., *town* ↔ *city*). This leads to the conclusion that three specific models do not offer the optimal solution. For instance, amalgamating contexts (e.g., *singer*, *footballer* and *politician*) under the umbrella of one model for persons would not capture essential disparities. Furthermore, results suggest that a semantic hierarchy might be necessary for modelling different relationships across models. At any rate, the construction of this hierarchy poses an interesting challenge.

From another viewpoint, *context models* also indicate that they produce enough granularity to encapsulate some semantic relations within the context (e.g., hypernymy and meronymy). Other semantic connections, including antonyms and pertainym, were also discovered when inspecting the variety of contexts.

2. They exploit sequences of terms (n-grams) given by the lexicalised dependency tree representation of sample and testing sentences, in contrast to other techniques in this chapter, which chiefly utilise shallow unigrams, bigrams and biterms as features.

   This materialised an enhancement in precision, as it helped to detect the most dependable nuggets, while diminishing the ranking score of those that gave the misleading impression of answers. Overall, experiments using this technique showed that lexicalised dependency paths serve as salient indicators for the presence of definitions in natural language texts.

In brief, *context models* take advantage of descriptive knowledge mined from Wikipedia for deducing some regularities, which characterise descriptions of instances of the same kind of *definiendum*. Specifically, these regularities are obtained from anonymised sentences expressing descriptive information about a large array of *definienda*. These sentences match a set of definition patterns, and they are automatically harvested from almost every article in Wikipedia. Exceptionally, *context models* attempt to eliminate the reliance of definition QA systems upon the coverage given by KBs to each particular *definiendum*.

Three baselines were implemented as a means of assessing and comparing the *context models*. All of these baselines ignore articles on the *definiendum* across KBs, and they were designed in such a way that they systematically increase their lexico-syntactic knowledge. Since the impact of KBs in the performance is well-known [Han et al., 2006, Zhang et al., 2005], robust methods were investigated to disregard this sort of information.

Generally speaking, the fact that *context models* outperform the three baselines indicates that they make a valuable contribution to enhancing the performance of definition QA systems. That is to say, they offer a solution to the problem of coverage exhibited by KBs, and in effect, there is nothing that prevents them from empowering models derived from KBs statistics.

More precisely, the figures signal a tangible betterment in terms of recall, revealing that *context models* aid in recognising more descriptive nuggets. This improvement eventuates from synergy between prioritising the selection of candidate sentences belonging to the most predominant *context indicators*, and basing the scoring on matching n-gram dependency paths. The former biases the ranking in favour of the more trustworthy sentences, while banning or leaving to posterior positions those misleading candidates which the baselines assigned a high ranked value and are in relation with less pertinent *context indicators*. The latter assists in detecting the most trustworthy candidates in each context as it singles

out those answer candidates bearing a closer similarity to the syntactic properties embodied in the LMs inferred from the automatically built treebank. Consequently, the final output encircles some correct answers seen as misleading by the baselines, but they are now preferred because of: (a) their membership to these prominent contexts; (b) the increment in their ranking; and (c) the reduction in the ranking score of those spurious answers picked by the baselines. These aspects really matter as the order of extraction is vital in controlling the redundancy.

In addition, the worsening of the score of misleading answer candidates and the boost of the ranking value of genuine answers were experimentally observed by the considerable enhancement in terms of precision. This betterment results from the fact that lexicalised dependency paths LMs do a better job when juxtaposing the syntactic properties of the putative answers to the regularities embraced by the models. By the same token, *context models* also help to ameliorate the ranking of answer candidates in terms of the top one and five sentences. In other words, the outcomes show that *context models* increases the precision of the definition patterns.

From another standpoint, and with regard to other features, results also showed that selective substitutions may enhance the recall of *context models* (i.e, TREC 2005 question set). In general, however, these attributes and the extensions based on extra snapshots of Wikipedia bring about an improvement in terms of raking order, but a diminution in terms of $\mathcal{F}(\beta)$-Score. This might be due to the fact that averaging models causes low frequency paths to lessen their weights, while high frequency paths receive more importance as they appear in more models with high weight. This promotes the research of smoothing techniques for *context models*, and their potential semantic hierarchy thereof.

Moreover, experiments attempting to incorporate information about contextual entities into *context models* revealed a deprovement. There were several issues on the exploitation of entities that must be scrutinised. Firstly, named entities are commonly denoted by means of several aliases, thus, it is easy to see that the variation within the answer candidate does not match any instance embodied in the models. Secondly, recurrently, dependency paths that link the *definiendum* to named entities seem to intrinsically contain enough information to rank the respective candidate sentence. Experiments also demonstrated that several mismatches can occur, despite the usage of an alias resolution strategy. By all means, an efficient -hopefully optimal- resolution of name aliases is a crucial component/subtask in the process of answering definition and list questions.

A more thorough analysis of the outcomes implies that the accuracy of definition patterns depends on the context(s)/type(s) of the *definiendum*, since they are more probable to hurt the performance when uttered in certain contexts than in others. Further, this analysis additionally points to the fact that negative samples are indispensable to enhance the recognition of descriptions expressed by superlatives.

An adaptation of *context models* to the TREC definition QA subtask highlights their competitiveness. To state it more precisely, this is manifested in the following two factors. First, instead of articles on the *definiendum* across KBs, the outcome of *context models* operating on web snippets was projected into the candidate sentences. This set is reduced in size, and not perfectly determined. Second, the achievements of the projection procedure ensure that the outcome returned by the web system resembles, to some extent, the TREC ground truth. Moreover, these figures corroborate that LMs built on top of dependency paths are a competitive projection strategy. All in all, the main objective of *context models* is to discern answers across web surrogates, not the AQUAINT corpus. Nonetheless, they accomplish a competitive performance without “annotated" resources when coping with this collection of news articles.

As future work, WordNet synset can be utilised for smoothing or grouping *context models*,

ergo enhancing the recall, and expectedly maintaining or improving their precision. Other strategies to detect redundancy can be developed by recognising similar dependency paths [Chiu et al., 2007]. This creates a key advantage of using dependency paths for answering definition questions.

On a final note, *context indicators* can aid in discriminating some of the potential senses of the *definiendum*. In a special manner, they can aid in disambiguating people and places named after individuals, while they might be fruitless in distinguishing strongly semantically closed senses, and probably also, not enough to separate different senses covered by the same *context indicator* (e.g., station). Nevertheless, some clustering approaches to automatically brach these contexts might be helpful. Lastly, the outcomes emphasise the importance of the *context models*, and of the specific paths that typify each particular context.

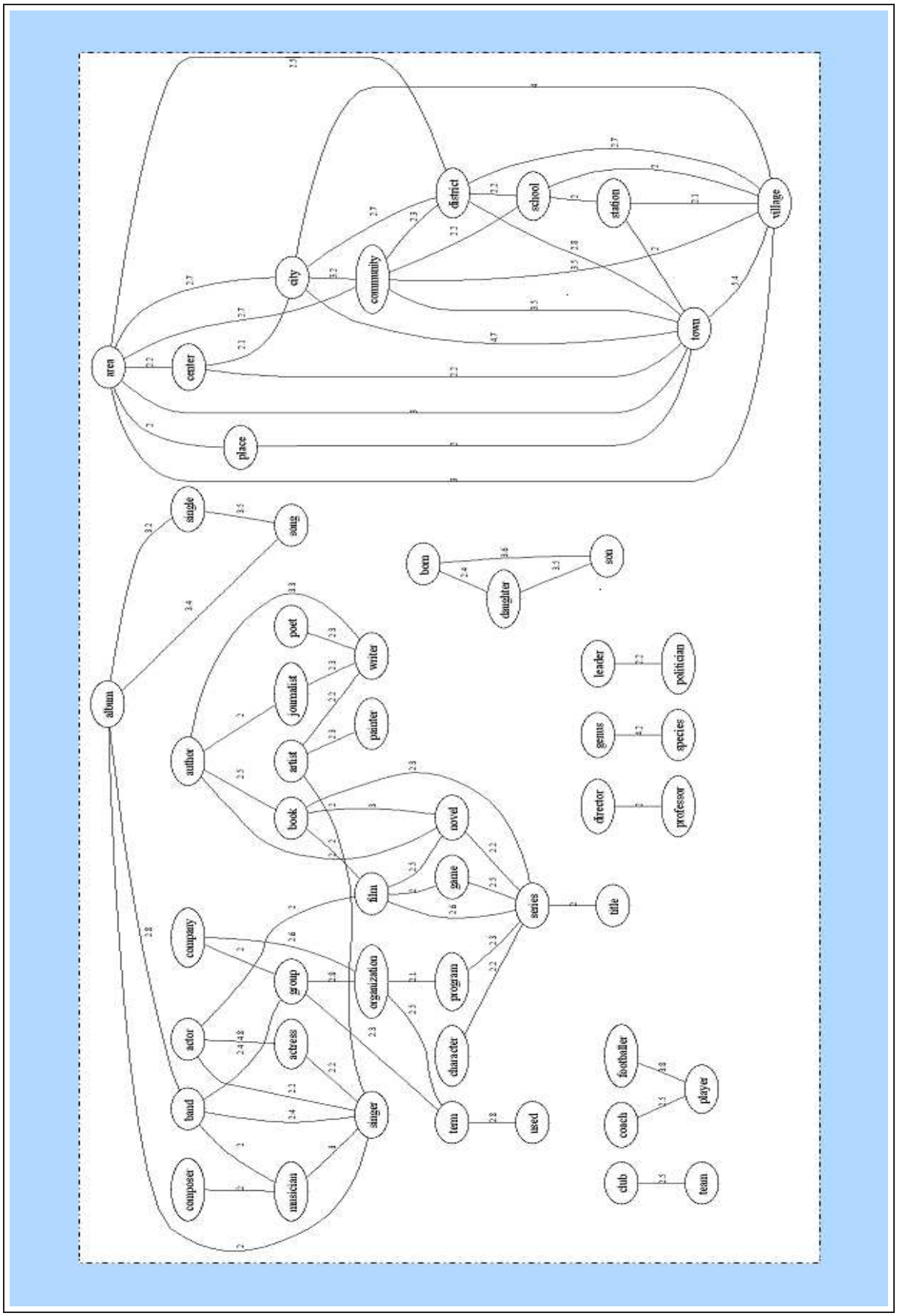

Figure 6.4: Highest-frequent similar contexts (cosine(c_{s_1},c_{s_2})*10).

Chapter 7

Discriminant Learning for Ranking Definitions

> "To know that we know what we know, and that we do not know what we do not know, that is true knowledge." (Henry David Thoreau)
>
> "The machine does not isolate man from the great problems of nature but plunges him more deeply into them." (Antoine de Saint-Exupery)
>
> "Science is the systematic classification of experience." (George Henry Lewes)

7.1 Introduction

As a rule, it is a standard practice of definition Question Answering (QA) systems to mine Knowledge Bases (KB) for reliable descriptive information about the *definiendum*. To a meaningful extent, this sort of practice makes intuitive sense, because the information embodied in KBs can be reasonably deemed as trustworthy. In other words, systems can perceive information units (e.g., sentences) harvested from KBs as positive examples, and ergo the task of answering definition questions can be rendered as identifying "*more like these*" positive examples across the set of answer candidates.

Strickly speaking, there are three ways of capitalising on positive training material: (a) models grounded solely on articles about the *definiendum*; (b) general definition models; and (c) models premised on the type(s) of the *definiendum*. In truth, there is nothing that prevents amalgamating from these models. However, the heart of the matter is that it is easy to gather "annotated" positive training data, but on the other hand, it is hard to automatically collect negative samples that can improve the performance. More accurately, these negative examples are contexts (e.g., sentences) bearing the *definiendum*, whereof no descriptive knowledge is conveyed. Take, for instance:

Posted by: CONCEPT[Chuck Moore] | April 19, 2007 8:35 AM
Starting in 1997 , the Union began to work with non-student instructional staff to join CONCEPT[CUPE 3902].

In effect, as matters stand today, the critical unresolved issue is the absence of authoritative sources of negative samples that can offer enough coverage to wide-ranging domains, and which, at the same time, can act as counterparts of the positive set. In practical terms, the annotation process is a complicated problem, because it is, in many cases, notoriously

difficult to clearly state whether a particular sample is positive or negative, even for human annotators.

In actuality, corpus annotation is not the only main issue. After specifying a dependable set of positive and negative examples, the next step implies another complex problem: determining the discriminative features; in other words, the identification of the characteristics that typify the elements of each category. These characteristics can be observed at different levels, such as words, chunks, sentences, shingles, and paragraphs, as well as, documents.

In this respect, the study of these characterising attributes consists of finding combinations of features that produce the best performance. Since this array of properties is typically dependent on the training and evaluation sets, this analysis can also entail the discovery of those attributes that port to other very different in nature target sets. Moreover, another key aspect of feature engineering is how much peformance is gained via the enrichment of Natural Language Processing (NLP)-based attributes. More often than not, extracting this class of property demands substantial computational resources. For this reason, it is always desirable to find out combinations of competitive features at the surface level that can aid in processing massive collections of documents.

This chapter is organised as follows: the next section deals at greater length with a strategy that categorises text snippets centred at the *definiendum* as putative answers. At heart, this classification methodology is predicated on a Support Vector Machine (SVM) with predominantly surface features. Furthermore, this strategy proposes a technique for automatically labelling training text fragments. Posteriorly, section 7.3 goes over the impact of a battery of surface attributes on the ranking of both sentences and paragraphs as answer candidates. In addition, this study examines the effect of considering a third group for those samples that are seen as ambivalent (positive and/or negative). Later, section 7.4 focuses on the influence of various discriminative learning models in the ranking of copula sentences in Dutch. This also explores the use of some NLP-oriented attributes emanated from named entities and dependency structures. Next, section 7.5 details the features exploited by a definition QA system in the context of the Text REtrieval Conference (TREC). This method has its roots in factoid QA, and it maps a *definiendum* type with an array of potential kinds of nuggets. Subsequently, section 7.6 presents an alternative technique for automatically annotating examples, and more important, it expands on the impact of assorted features on discriminative models for scoring open-domain pattern-independent sentences, and section 7.7 concludes this chapter.

7.2 Ranking Single-Snippet Answers

In their early work, [Miliaraki and Androutsopoulos, 2004] constructed a classifier that compartmentalises 250-characters text snippets into definitions or general texts. These text surrogates were harvested from an array of fifty documents downloaded from the Internet. To be more accurate, this collection of fifty documents conforms to the top fifty hits returned by a commercial search engine. Fundamentally, the core of their classifier is built on a SVM with a simple inner product kernel [Schölkopf and Smola, 2002]. In practical terms this means that the kernel learns an optimal linear classifier without moving to a higher-dimension space.

A special characteristic of these 250-character fragments is that they are centred at the *definiendum*, and accordingly, training instances were manually labelled as definitions or general texts. By and large, [Miliaraki and Androutsopoulos, 2004] observed that the positive samples (definitions) were much fewer than their counterparts, that is, the negative (non-definition) samples. In particular, they acquired 3,004 and 15,469 positive and negative examples respectively, as an outcome of this manual annotation process. These 18,472 exam-

IMBALANCE

ples were taken from TREC questions and documents. Due to the imbalance between both categories, [Miliaraki and Androutsopoulos, 2004] preferred to utilise the SVM as a ranker to as a classifier. In essence, they noticed that this imbalance causes their classifier to assign higher confidence scores to the negative class than the positive group. Since their SVM implementation returned confidence scores in congruence with the probability that a testing instance belongs to each category, the top five answer candidates are assumed to be correct answers, that is to say, the putative answers coinciding with the five highest confidence scores of the positive class.

As for features, [Miliaraki and Androutsopoulos, 2004] examined the performance of four different configurations of their definition ranker. The first configuration makes allowances for two numeric attributes: `SN` and `WC`. The first denotes the ordinal number of the sentence within the source document where it originally belongs to. The second embodies the percentage of the twenty highest frequent words across all answer candidates which are carried by the candidate being ranked. These twenty terms are stemmed and do not consider stop-words. Both features are, strictly speaking, aimed basically at capturing indicators of relevant global pieces of information that are likely to characterise the *definiendum*. In other words, these ingredients are predicated on the principle known as the Distributional Hypothesis [Harris, 1954, Firth, 1957], and they assist, for this reason, in ranking candidate answers in concert with the degree they typify the *definiendum*. By the same token, [Miliaraki and Androutsopoulos, 2004] inspected the utilisation of the ranking score returned by the Information Retrieval (IR) engine as another global feature. On the other hand, their local features encompassed thirteen binary numbers that cohere with the fulfilment of the nine rules of [H. Joho and M. Sanderson, 2000] (see 4.2 on page 76), and four extra definition patterns:

FEATURE: SENTENCE ORDER

FEATURE: TERM CORRELATION

FEATURE: IR ENGINE

1. <description> like <definiendum>
   ⇒ "*... diseases like swine flu*"

2. <definiendum> or <description>
   ⇒ "*... Germany or other countries*"

3. <definiendum> (can | refer | have) <description>
   ⇒ "*Leopards can hear five times more sounds than humans...*"

4. <description> (called | known as | defined) <definiendum>
   ⇒ "*... the giant wave known as tsunami*"

With regard to the second configuration, this extends the first one by adding an extra binary feature signalling whether or not the surrogate contains one of the best hypernyms of the *definiendum*. In addition, the third configuration enriches their ranker with features conforming to a group of automatically inferred patterns. These regularities, more specifically, encompass between 100 and 300 unigrams, bigrams and trigrams that occur immediately before or after the *definiendum*. In substance, [Miliaraki and Androutsopoulos, 2004] gathered these n-grams from the top documents returned by the search engine for the training queries. The highest ranked patterns in terms of precision that also have a frequency higher than ten were, subsequently, selected as acquired patterns. Here, the precision of a phrasal pattern is interpreted as the ratio of the amount of genuine definitions that it matches to the number of all its matching text snippets. The underlying idea behind this collecting process is to discover a set of new phrasal attributes that can imply potential descriptions. Some illustrative phrasal patterns embrace:

FEATURE: HYPERNYM

FEATURE: N-GRAMS PATTERNS

1. <definiendum> is one <description>
   ⇒ "*Israel is one of the most advanced economies in the Middle East.*"
   ⇒ "*Israel is one of 16 countries that make up the Middle East, a predominantly Arab region of the world at the crossroads of Africa, Europe, and Asia.*"

2. <description>, (a | an) <definiendum>
   ⇒ "*... A sudden health problem, a heart attack or*"

Moreover, [Miliaraki and Androutsopoulos, 2004] also observed that some of their discovered phrasal patterns are directed principally at a specific domain (e.g, "*people with*" and "*symptoms of*"). They suggested, therefore, that this automatic methodology for inducing phrasal patterns can be exploited by domain-specific QA systems to distinguish answers to definition questions. For instance, (definition) QA systems tackling medical documents. However, [Miliaraki and Androutsopoulos, 2004] additionally noticed that these regularities do not ensure perfect accuracy. Otherwise stated, they found out that these phrasal patterns do not rule out the presence of definitions reliably.

Eventually, [Miliaraki and Androutsopoulos, 2004] also attempted to discriminate dependable regularities on the grounds of metrics such as information gain [Yang and Pedersen, 1997] instead of using precision, but this led only to inferior results. Lastly, the fourth configuration ran with all their previous features, but it ignores the attribute that models the best hypernym for the *definiendum* incorporated by the second model.

Ranker	Correctly Answered (%)	
	TREC 2000 (160)	TREC 2001 (137)
`Baseline` [Prager et al., 2001, 2002]	50.00	58.39
`Configuration I`	61.88	72.26
`Configuration II`	63.13	73.72
`Configuration III`	72.50	84.67
`Configuration IV`	71.88	83.94

Table 7.1: Results achieved by the four configurations of attributes in comparison with a baseline system (adapted from [Miliaraki and Androutsopoulos, 2004]).

EXPERIMENTS

As to training, [Miliaraki and Androutsopoulos, 2004] conducted a 10-fold cross validation. In other words, they split the question set into ten parts, and at each iteration, the questions of a different part and their respective documents were utilised solely for testing, while the questions and documents of the remaining nine parts were used solely for training. In the case of configurations three and four, the deduction of phrasal patterns was accordingly repeated at each iteration. Consequently, table 7.1 emphasises their results in relation to the percentage of queries, for which one valid definition was found within the top-five ranked snippets. These figures underscore the performance reached by each of the four configurations. The marginal improvement and deprovement obtained by the second and fourth configurations respectively, largely appears to indicate that the hypernym is not as crucial as the automatically acquired phrasal attributes and patterns. This finding is particularly interesting because the best configurations do not account for deep linguistic information.

IMPACT OF HYPERNYM

IMPACT OF PHRASAL ATTRIBUTES

From a different angle, results in table 7.1 account solely for 200 n-gram features when employing configurations two and three, whereas table 7.2 depicts the outcomes accomplished by the same two configurations for both TREC question sets, but considering different amounts of n-grams. Particularly, [Miliaraki and Androutsopoulos, 2004] observed that,

when making allowances for 300 n-grams, both configurations finished with worse results because low reliable phrasal attributes begin to dominate the feature vector.

	Configuration III	Configuration IV
100	68.13/79.56	70.00/81.75
200	72.50/84.67	71.88/83.94
300	68.75/80.29	71.25/83.21

Table 7.2: Performance (%) varying the number of acquired phrasal attributes (adapted from [Miliaraki and Androutsopoulos, 2004]).

AUTOMATIC WINDOW TAGGING

Incidentally, [Androutsopoulos and Galanis, 2005] extended this method by replacing the training data acquisition process. They exploited definitions across on-line encyclopedia and dictionaries to obtain an arbitrary large amount of training windows. To be more precise, these KB definitions were then utilised for ranking training windows found across web pages. Accordingly, these training windows are consequently seen as positive or negative in congruence with their similarity or dissimilarity to the corresponding definitions across on-line KBs. The training windows, for which a clear distinction cannot be drawn, were discarded. Since definitions extracted from KBs are used solely for tagging training instances, which are taken exclusively from web pages. Allegedly, this is particularly important as they allow selecting lexical patterns that are indicative of definition within web pages, as opposed to patterns that signal descriptions in KBs. Presumably, this results also from the reliance of their strategy on the n-gram phrases attributes and both contexts left and right (their training windows are centred at the *definiendum*), and definitions emanated from online dictionaries do not observe this ordering, in general.

TAGGING SIMILARITY MEASURE

As for the similarity measure, [Androutsopoulos and Galanis, 2005] transformed the training instances into a new snippet by: (1) removing stop-words, any non-alphanumeric character and the *definiendum*, and (2) stemming the remaining terms. The similarity between this new form of training example W and the definitions C collected from the online KBs is measured in consonance with:

$$sim(W, C) = \frac{1}{\mid W \mid} * \sum_{i=1}^{|W|} sim(w_i, C)$$

In this formula, |W| denotes the number of distinct words in W, and $sim(w_i, C)$ is the similarity of the i-th distinct word of W to C in concert with:

$$sim(w_i, C) = fdef(w_i, C) * IDF(w_i)$$

Where $fdef(w_i, C)$ conforms to the percentage of definitions in C that contain w_i, and $IDF(w_i)$ is the Inverse Document Frequency (IDF) of w_i in the British National Corpus (BNC):

$$IDF(w_i) = 1 + log\frac{N}{df(w_i)}$$

In this equation, N coincides with the amount of documents in the BNC, and $df(w_i)$ corresponds to the number of BNC documents that contains w_i; every time a term is not found in the BNC, the lowest score of the BNC is assigned. Stated more concisely, $sim(w_i, C)$ privileges words with higher occurrence in all KBs and which are also rarely used in English. Accordingly, they took advantage of two different experimental thresholds for separating positive and negative as well as unlabelled examples.

THRESHOLDS Fundamentally, [Androutsopoulos and Galanis, 2005] conducted several experiments as a mean to set the values of these two thresholds. In so doing, they benefited from an array of 130 distinct training *definienda*, wherewith they fetched their top ten highest ranked hits in congruence with the positions of the results returned by Altavista. Since non-definition windows outnumber definition windows, they took into consideration only the first five from each web page. They manually tagged a group of 400 randomly chosen windows and tested various values of both thresholds in order to optimise the accuracy. The ratio of manual annotated examples was $\frac{0.37}{1}$. Of course, in both positive and negative examples, this similarity method can achieve a significant precision at the expenses of very low recall, that is very few training examples. Also, [Androutsopoulos and Galanis, 2005] ensured that the ratio of positive to negative examples that the method outputs was the same in order to avoid a bias in the training of their classifier. The empirical values obtained for the positive class was 0.5, which causes a precision above 0.72 and a recall of 0.49, while the value for the negative category was 0.34, bringing about a precision above 0.9. In their experiments, this last value oscillated from 0 to 0.34, so that it led to the ratio closest to the ratio observed in the 400 tagged windows.

7.3 Undecided/Indifferent Labels

Like [Miliaraki and Androutsopoulos, 2004, Androutsopoulos and Galanis, 2005], [Xu et al., 2005] also made use of SVM for ranking answer candidates to definition questions about technical terms. In their approach, however, they perceived paragraphs instead of fixed-length *definiendum*-centred windows as answer candidates. To be more precise, their technique gathers putative answers that embody the *definiendum* in the first base noun phrase of the first sentence of the paragraph. In addition, they interpreted two base noun phrases separated by "*of*" or "*for*" (e.g., "*Food for Peace Program*") as *definienda*. The next step in their process of detecting answer candidates consists in utilising definition patterns for filtering some candidate sentences. The pre-defined pack of rules used by [Xu et al., 2005] comprises:

PARAGRAPHS AS ANSWERS

DEFINITION PATTERNS

1. <definiendum> is (a | an | the) <description>
   ⇒ "*The Temple Mount is the site of the first and second Jewish Temples, destroyed in 586 BCE and 70 CE, respectively-a historic fact accepted even by Muslim authorities.*"

2. <definiendum>, * , (a | an | the) <description>
   ⇒ "*Judaism, founded in Israel around 4000 years ago by the Hebrew leader Abraham, the religion of morality.*"

3. <definiendum> is one of <description>
   ⇒ "*The New Israeli Shekel is one of the strongest currencies in world, the New Israeli Shekel is now traded on the international currency market and is freely interchangeable.*"

INDIFFERENT CLASS In this group of three patterns, the star stands for one or more words. As to ranking putative answers, [Xu et al., 2005] capitalised on SVM and Ranking SVM [Joachims, 2002]. The difference between both learning machineries is that the former compartmentalises test examples, while the latter is premised on ordinal regression. This regression is seen as assigning three distinct labels to the answer candidates: "*good*" and "*indifferent*", as well as "*bad*" in accordance with the scoring value returned by Ranking SVM. The underlying idea is that the risk of misclassifying an example as "*good*" or "*bad*" is considerably higher than labelling it as "*indifferent*". In the opposite way, the SVM clusters putative answers into only two groups "*good*" or "*bad*". Both learners were trained with the same set of features. These attributes encompass:

ATTRIBUTES

(a) a feature symbolising whether or not the *definiendum* appears at the beginning of the paragraph,

(b) if a determiner precedes the *definiendum*,

(c) all terms in the *definiendum* start with a capitalised letter,

(d) if the paragraph includes words such as "he", "she" or "said",

(e) if the *definiendum* contains pronouns,

(f) if the *definiendum* bears "for", "and", "," or "or",

(g) if the *definiendum* re-occurs in the paragraph,

(h) if the *definiendum* is followed by "is a", "is the" or "is an",

(i) number of sentences,

(j) amount of terms,

(k) number of adjectives,

(l) frequent words after the *definiendum* within a window.

PRONOUNS AS NEGATIVE EVIDENCE FEATURES: TERM CO-OCCURRENCE

According to [Xu et al., 2005], some members of this array of twelve properties, such as (g), indicate definitions, whereas other ingredients like (d) signal non-definitions. Both SVM and Ranking SVM also account for "*bag-of-words*" features. These attributes are gathered from high frequency words appearing immediately after the *definiendum* in training data, somehow following the intuition behind the phrasal patterns of [Miliaraki and Androutsopoulos, 2004]. In like manner, these terms are conceived as potential signals of descriptions. Simply put, whenever a paragraph carries any of these keywords, the respective feature value turns to one, otherwise it remains zero. Contrary to [Miliaraki and Androutsopoulos, 2004, Androutsopoulos and Galanis, 2005], [Xu et al., 2005] discarded redundant answers by means of the Edit Distance. More exactly, every time two candidates were too similar, the one having the lower score was removed.

REDUNDANCY REMOVAL

In their experimental settings, [Xu et al., 2005] randomly singled out two hundred *definienda* about technical terms. Those having only one answer (paragraph) candidate were removed, and the remaining candidates were labelled as "*good*" and "*indifferent*" as well as "*bad*" by human annotators. As a result, they obtained a set consisting of 95 *definienda* and their respective 1,366 answer candidates: 225 good, 470 indifferent and 671 bad. All these putative answers came from an intranet collection.

	R-Precision	Precision at 1	Precision at 3
OKAPI	0.2886	0.2211	0.6421
SVM	0.4658	0.4324	0.8351
Ranking SVM	**0.5180**	**0.5502**	**0.8868**
Random Ranking	0.3224	0.3474	0.6316

Table 7.3: Outcomes obtained when considering paragraphs as answer candidates (adapted from [Xu et al., 2005]).

Table 7.3 contains the results accomplished by both learners and two baselines. These outcomes are in relation to the average performance achieved by conducting a 5-fold cross validation. As for baselines, they benefited from OKAPI and its ranking of paragraphs in agreement with their likeness to the query. As a second baseline, they randomly ranked answer candidates.

ADJECTIVE ATTRIBUTE

Their error analysis basically unveiled that the adjective feature caused some good answers to rank at the bottom. More precisely, they observed that 35% of the errors were due chiefly to this attribute. They additionally noticed that 30% of the errors stemmed from some indifferent or bad candidates ranked at the top. Of course, this has to do with the restricted potential of their features. Importantly enough, [Xu et al., 2005] detected that mislabelling samples brought about 5% of the errors. In addition, they noticed that, more often than not, the more adjectives a paragraph has, the less probable that the paragraph is a good definition. Nonetheless, they also took note of the fact that some good definitions can occasionally carry several adjectives. Furthermore, [Xu et al., 2005] observed that many paragraphs can verbalise a definition in the first sentence, while at the same time, the remaining sentences do not convey descriptive content at all.

PORTABILITY

On a different note, [Xu et al., 2005] carried out an extra experiment using the same trained classifiers. This time, however, they ranked putative answers originated from the TREC.gov data. In their experiments, they took into account 25 *definienda* and their corresponding 191 answer candidates: 67 good, 76 indifferent and 48 bad. Table 7.3 highlights their experimental outcomes with respect to this dataset. According to [Xu et al., 2005], the good performance ported to this new test set because their features are domain-independent. It is unclear, however, the extent of this conclusion, because the distribution of *definienda* in this test set is strongly biased towards abbreviations or organisations (e.g., NIST, MAP and IRS). These sorts of *definienda* could also be found across technical terms. From another standpoint, to probe that their attributes are domain-independent, it would be much more appropriate to test their approach against a wider variety of types of *definienda*, especially individuals, since their training material did not account for such sort of data. This kind of experimentation would draw stronger conclusions about the portability of their features to other domains. Another critical factor is which are the portable attributes. On the other hand, although they applied their rankers to testing instances coming from a corpus different in nature to their training data, their portability could be due to the fact that they utilise fixed syntactic structures (pre-determined definition patterns), and these regularities exist in both corpora.

	R-Precision	Precision at 1	Precision at 3
OKAPI	0.4267	0.4000	0.8000
Ranking SVM	0.5747	**0.6400**	0.8400
SVM	**0.5780**	**0.6400**	**0.9600**
Random Ranking	0.3307	0.3200	0.7600

Table 7.4: Results obtained when dealing with TREC.gov data (adapted from [Xu et al., 2005]).

SENTENCES AS ANSWERS

In addition, [Xu et al., 2005] conducted an extra experiment targeting sentences instead of paragraphs as putative answers. For this purpose, they made use of 78 *definienda* and their respective 670 candidate answers, which encompassed 157 good, 186 indifferent and 327 bad definitions. This new experiment required the inclusion of new features such as the position of the sentence in the paragraph.

	R-Precision	Precision at 1	Precision at 3
OKAPI	0.2783	0.2564	0.5128
SVM	0.6097	0.5972	0.8710
Ranking SVM	**0.6769**	**0.7303**	**0.9365**
Random Ranking	0.3693	0.3590	0.6795

Table 7.5: Results achieved when considering sentences as answer candidates (adapted from [Xu et al., 2005]).

Table 7.5 emphasises the portability of their method to the sentence level, in particular in relation to the top ranked answer. The performance of their strategy sharply increases when the second and third answers are taken into consideration.

7.4 Ranking Copular Definitions in Dutch

Another approach, outlined by [Fahmi and Bouma, 2006], focusses its attention on discovering responses to definition questions in the medical domain. Different to the techniques of [Miliaraki and Androutsopoulos, 2004, Androutsopoulos and Galanis, 2005, Xu et al., 2005], they targeted definition queries in Dutch. Their strategy, stated more precisely, extracted documents from the healthcare index of the version in Dutch of Wikipedia. These articles were parsed in order to obtain their dependency graphs, and additionally, their named entities were labelled. These entities included person, organization, and geographical entities. A vital distinction between this approach and the three discussed previously lies in the properties of the potential answer candidates. This strategy, to put it more clearly, considered only sentences that: (a) bears the verb "*zijn*" (*to be*) with a subject and a nominal predicative phrase as sisters; and (b) the predicative phrase precedes the subject, as in "*Onderdeel van de testis is de Leydig-cel*" (*the Leydig cel is part of the testis*). Furthermore, [Fahmi and Bouma, 2006] benefited from lexical filters to expunge some potential spurious and misleading candidates. Some of these lexical filters embrace the own translation into Dutch of: *example*, *result* and *symptom*.

FEATURES: NAMED ENTITIES

COPULA AS ANSWER CANDIDATE

With regard to the training and evaluation sets, [Fahmi and Bouma, 2006] manually labelled 2,500 sentences in accordance with three labels, homologous to [Xu et al., 2005]: *definition*, *non-definition*, and *undecided*. Conversely, in this annotation strategy, undecided sentences consisted chiefly of incomplete definitions such as "*Benzeen is carcinogeen*" (*Benzene is a carcinogen*). Table 7.6 remarks the distribution of instances within categories in the final labelled set.

UNDECIDED CLASS

	Definition	Non-definition	Undecided
first	831	18	31
other	535	915	170
total	**1,366**	**933**	**201**

Table 7.6: Amount of sentences annotated as definition, non-definition, and undecided versus the position within the document (source [Fahmi and Bouma, 2006]).

After discarding undecided sentences, a total of 2,299 samples were left, from which 1,366 are genuine definitions. More interestingly, [Fahmi and Bouma, 2006] realised that when the

FEATURE: SENTENCE POSITION

position of the sentence is taken into consideration as the unique feature for rating answer candidates, all first sentences are classified as definitions whereas all other sentences as non-definitions, a baseline accuracy of 75,9% is obtained. Incidentally, the list of ingredients that [Fahmi and Bouma, 2006] combined includes:

PROPERTIES

1. **Bag-of-words, bigrams, and root forms**. These attributes consider punctuations and stop-words. The reason for making allowances for stop-words was the fact that their experiments indicated a consistent decrease in accuracy when they are absent. In contrast to [Androutsopoulos and Galanis, 2005], [Fahmi and Bouma, 2006] accounted for all n-grams within a sentence as elements of the feature vector.

FEATURE: N-GRAMS AND ROOT FORMS

2. **The position of each candidate sentence in the document**.

ATTRIBUTE: SENTENCE ORDER

3. **Syntactic properties**. These features encompass the position of each subject in the sentence (initial, e.g. "*definiendum is ....*"; or noninitial, e.g. "*.... is definiendum*"). Their experiments unveiled that this attribute seems to be critical, because of the fact that sentence-initial subjects appeared in 92% of their definition sentences, while in 76% of their non-definition sentences.

   In addition, these properties incorporate two extra ingredients that encode the type of determiner (definite, indefinite, other) of the subject and predicative complement. Notably, [Fahmi and Bouma, 2006] observed that 62% of subjects in (Dutch) definition sentences have no determiner. For instance, they produced the following descriptive phrase: "*Paracetamol is een pijnstillend en koortsverlagend middel*" (*Paracetamol is a pain alleviating and a fever reducing medicine*), whereas in 50% of non-definition (Dutch) sentences subject determiners have the tendency to be definite, e.g. "*De werkzame stof is acetylsalicylzuur*" (*The active ingredient is acetylsalicylacid*). On the other hand, 64% of predicative complements tended to embody indefinite determiners in definition sentences like "*in een pijnstillend . . . medicijn*" (*a pain alleviating. . . medicine*), whereas in 33% of non-definitions, the determiner has a leaning to be definite. As in the next illustrative phrase: "*Een fenomeen is de Landsgemeinde*" (*A phenomenon is the Landsgemeinde*).

DETERMINER FEATURE

4. **Named Entities**. This attribute encapsulates the entity class of subjects. [Fahmi and Bouma, 2006] noticed that, contrary to non-definitions, definitions are more likely to have a named entity in their subjects (40.63% compared with 11.58%).

FEATURE: NAMED ENTITIES

For experimental purposes, [Fahmi and Bouma, 2006] selected the 200 highest ranked features of each class, which were scored in agreement with the information gain measure. Their training set was comprised of 1,336 definitions and 963 non-definitions sentences, respectively. Broadly speaking, all of their tried configurations outperformed their baseline (75.9%).

One of their interesting findings regards the best and relatively high accuracy (89.82%) achieved by Naïve Bayes when basic attributes, such as bigrams+bag-of-words, are the sole components of the feature vector. Further, their outcomes also unveiled that the addition of syntactic properties or position of sentences in documents results in some improvements. Their figures, on the other hand, suggest that an enrichment with root forms does not significantly enhance the performance.

INFLUENCE OF SENTENCE POSITION

INFLUENCE OF ROOT FORMS

In general, their experiments demonstrated that syntactic properties and the position of sentences within documents occupy a pivotal role in boosting the accuracy of their classifiers. The latter, in particular, cooperated on reaping the best performance of Naïve Bayes (90.26%), and gave better accuracy in all classifiers than syntactic properties. Needless to

	Naïve Bayes NB	Maximum Entropy ME	SVM Linear	SVM Polynomial	SVM Gaussian
bag-of-words	85.75 ± 0.57	85.35 ± 0.77	77.65 ± 0.87	78.39 ± 0.67	81.95 ± 0.82
bigrams	87.77 ± 0.51	88.65 ± 0.54	84.02 ± 0.47	84.26 ± 0.52	85.38 ± 0.77
bigrams+bag-of-words	89.82 ± 0.53	88.82 ± 0.66	83.93 ± 0.57	84.24 ± 0.54	87.04 ± 0.95
syntactic properties+ bigrams+bag-of-words	85.22 ± 0.35	89.08 ± 0.50	84.93 ± 0.57	85.57 ± 0.53	87.77 ± 0.89
syntactic properties+ entity classes+ bigrams+bag-of-words	85.44 ± 0.45	91.38 ± 0.42	86.90 ± 0.48	86.90 ± 0.53	87.60 ± 0.87
sentence position+ bigrams+bag-of-words	**90.26 ± 0.71**	90.70 ± 0.48	85.26 ± 0.56	86.05 ± 0.64	88.52 ± 0.92
root forms+ bigrams+bag-of-words	88.60 ± 0.81	88.99 ± 0.51	83.38 ± 0.38	84.69 ± 0.43	87.08 ± 0.87
syntactic properties+ sentence position+ bigrams+bag-of-words	86.40 ± 0.51	**92.21 ± 0.27**	86.57 ± 0.42	87.29 ± 0.47	88.77 ± 0.77
syntactic properties+ sentence position+ entity classes+ bigrams+bag-of-words	87.12 ± 0.52	90.83 ± 0.43	87.21 ± 0.42	**87.99 ± 0.53**	89.04 ± 0.67
root forms+ sentence position+ syntactic properties+ bigrams+bag-of-words	87.60 ± 0.38	91.16 ± 0.43	86.68 ± 0.40	86.97 ± 0.41	88.91 ± 0.68
all attributes	86.72 ± 0.46	91.16 ± 0.35	**87.47 ± 0.40**	87.05 ± 0.63	**89.47 ± 0.67**

Table 7.7: Achieved accuracies/standard errors versus feature and classifier configurations (adapted from [Fahmi and Bouma, 2006]).

say, when both ingredients were intermixed, the performance still improved, ergo implying a supplementary nature.

Another key finding refers to the fact that the addition of named entity classes was tended to increase accuracy. On the other hand, adding root forms did not enhance the performance. It is worth emphasising, however, that the best accuracies of Naïve Bayes (90.26%) and Maximum Entropy (92.21%) were accomplished without named entities and root forms. FEATURE: NAMED ENTITIES

On the whole, in nine out of eleven configurations, Maximum Entropy (ME) achieved the best performance. Interestingly enough, SVM-LINEAR and SVM-POLYNOMIAL did not finish with better accuracies than naïve Bayes when integrating basic features (bag-of-words and bigrams), while SVM-GAUSSIAN marginally outperforms Naïve Bayes in six out of the eleven configurations. CLASSIFIER

7.5 Answering Definitions in TREC

In TREC 2003, [Burger, 2003] extracted responses to definition questions from the top 25 documents fetched by Lucene. Sentences in these documents were pre-rated by summing the IDF scores of each word that overlaps with the query. Low scored sentences were thus eliminated. The remaining candidate sentences are then rated in congruence with conditional log-linear models, which were trained by means of the 24 TREC 1999/2000 definition questions plus the 25 definition evaluation questions given in 2001. The elements fused into their models combine: ATTRIBUTES

1. The IDF overlap score.
2. Raw count of overlapping terms with the query. These counts disregard stop-words.
3. The count of word bigrams in common with the prompted question.
4. This definition QA system focuses predominantly on named entities as answers. Accordingly, it first determines an open-domain phrase from the query that describes the entity being sought (e.g., "*first man on the moon*"). This attribute encodes the raw term count in this phrase.
5. Raw count of words interpreted as salient by the question analysis phase.
6. The number of words that could be synonyms of query terms. These features are computed with respect to WordNet.
7. Raw count of words that could be antonyms of question terms. These attributes are estimated in relation to WordNet.
8. The count of words in common between the candidate itself and the question.
9. Number of characters between the candidate and a term from the previously determined open-domain phrase.
10. Amount of characters between the candidate and the closest question word in the context.
11. The score assigned by WordNet.
12. A merge count implying the number of answers with the same text realisation.
13. A boolean attribute symbolising whether or not the answer matches the expected answer type.
14. A boolean feature signalling whether or not the answer is similar to the expected answer type.
15. Boolean attributes encoding twenty arbitrarily selected pairs of mismatching expected answer types and answer types.

DEFINITIONS AS FACTOID QUERIES

In fact, these features are aimed at factoid questions, but their definition QA module also exploited the same ranker to cope with definition queries. However, as a means of adapting their system to deal efficiently with this kind of question, they allow it to return windows of 90-characters around the answer candidate. In TREC 2005, [Burger and Bayer, 2005] benefited from features: (1-4), (6), (8-11), and (13-15). These attributes acted coupled with the next new two ingredients:

1. The logarithm of the consolidation of frequency counts pertaining to identical answer candidates.
2. Average character-level similarity between one putative answer and all the others. This attribute helps textually similar candidates to support each other, which might be particularly useful for dates and other kinds of answers that have multiple formats and representations.

Definiendum type	Candidate type
PERSON	DATE
PERSON	YEAR
PERSON	PERSON
PERSON	LOCATION
PERSON	COUNTRY
PERSON	fragment
ORGANISATION	LOCATION
ORGANISATION	COUNTRY
ORGANISATION	PERSON
ORGANISATION	fragment
unknown	fragment

Table 7.8: List of mismatching combinations of *definiendum*/answer types utilised for definition questions (source [Burger and Bayer, 2005]).

In this track, [Burger and Bayer, 2005] made a smarter use of their boolean attributes (15) by modelling potential combinations that can serve as answers to definition questions. Table 7.8 lists these pairs.

ANSWER TYPES MAPPINGS

Additionally, [Burger and Bayer, 2005] enriched their definition QA system with crude heuristics for identifying short fragments of descriptions occurring in appositional contexts, and consequently aided their system in recognising some non-entity candidates. In brief, the best run of this system reaped an average $\mathcal{F}(3)$-Score of 0.217 in the TREC 2005 track (median=0.156). That year, they trained their ME models using the question sets supplied from TREC 1999 through 2003, including the 25 TREC 2001 definition evaluation questions.

In the context of the TREC 2006 challenge, [Burger, 2006] made allowances for an array of features similar to [Burger and Bayer, 2005]. One difference is, for instance, the exclusion of attribute (6). In this track, they ran a definition QA system akin to the one used in TREC 2006, which typically retrieved the top fifty documents. Here, they trained ME models using the question sets from TREC 1999 through 2004, including the 25 TREC 2001 definition evaluation questions. The best average $\mathcal{F}(3)$-Score reached a value of 0.156, while the median of this track across all participants was 0.125 and the best run reached a value of 0.250.

7.6 Ranking Pattern-Independent Descriptive Sentences

Fundamentally, [Miliaraki and Androutsopoulos, 2004, Androutsopoulos and Galanis, 2005] adopted a technique to rate fixed 250-character length text snippets centred at the *definiendum*. These text windows were scored in agreement with models learnt from an array of previously annotated examples. Simply put, these samples consist of descriptions of *definienda* and general texts, both carrying the *definiendum* at their centres (see details in section 7.2). Contrary to this method, [Xu et al., 2005] ranked paragraphs and sentences that matched a group of pre-defined definition patterns (section 7.3), and [Fahmi and Bouma, 2006] focused solely on the copular structure in Dutch (section 7.4).

The following approach considers slackening the constraint that forces answer candidates to observe a set of pre-defined rules. This way the performance of discriminant learning operating on unconstrained or pattern-independent open-domain descriptive sentences can be studied. From another angle, it is worth recalling here that [Xu et al., 2005] aimed specifically at technical terms, whereas [Fahmi and Bouma, 2006] at the medical domain.

Most of all, these definition QA systems differentiate from each other in the attributes they exploit for rating putative answers. Table 7.9 juxtaposes some distinctive characteristics of these four approaches.

	section 7.2	section 7.3	section 7.4	section 7.6
answer granularity	snippets	paragraphs/sentences	sentences	sentences
type of answer	all	match patterns	copula Dutch	all
ranker	SVM	SVMs	SVMs,ME,NB	ME
manual annotations	yes/no	yes	yes	no
domain	open	technical	medical	open
size	∼20,000	∼2,000	∼3,000	$\sim 1.5*10^6$
surface features	yes	yes	yes	yes
named entities	no	no	yes	yes
chunking	no	no	no	yes
dependency trees	no	no	subject position	yes
categories	2	2/3	2	2

Table 7.9: Comparison amongst different methodologies.

SENTENCES, PARAGRAPHS, OR TEXT SNIPPETS AS ANSWERS

To begin with, it is fundamental to touch on the aspect of categorising sentences or larger spans of texts like paragraphs or fixed-length windows. The reason for singling out sentences as answer candidates instead of pre-specified length windows is two-fold: (a) by and large, sentences with resolved co-references embody the necessary context to understand their meaning; and (b) truncated *definiendum*-centred windows can trim essential descriptive content. With regard to paragraphs, one practical problem arises when dealing with target collections of documents from wide-ranging topics and from markedly different sources. On the one hand, in some cases, full paragraphs can be entire definitions, especially when they are sifted from biographical web pages. Conversely, this is much less probable when taking paragraphs from other sorts of web documents such as news and forums. Therefore, the requirement of trimming answer paragraphs will still exist, and hence probably entailing a strategy that can discriminate descriptions at the sentence level, so that answers can keep their integrity after trimming. Of course, it is always feasible to present to the user the answer in conjunction with the non-descriptive content, at the expense of quality.

TWO OR THREE CATEGORIES

Secondly, [Fahmi and Bouma, 2006] conducted an empirical study likening diverse classifiers integrated with some fixed groups of properties, and the reported figures suggest that Maximum Entropy (ME) models are a good choice as a classifier of definitions. Thirdly, strategies vary from grouping answer candidates into two ("*good*" or "*bad*") or three (adding "*undecided*" or "*indifferent*") classes. Three classes are definitively more suitable when the training corpus is manually annotated; this way annotators can assign this third label whenever they disagree on the annotation of an example. Discrepancies between annotators can arise often when labelling, this third category hence offers a workable solution to this problem as this uncertainty could be omitted or weighed in the models. In contrary, two groups seem to be more suitable when automatically labelling training data, which is naturally preferable, because it allows the creation of massive training corpora. Typically, automatic annotation procedures operate at the *definiendum* level and are predicated on lexical overlaps with its related articles across KBs. The underlying idea here is that those remarkably similar examples can be rendered positive and those extremely dissimilar samples can be allocated in the negative group. This method certainly does not bring out a perfect accuracy, but the obtained corpus can, nevertheless, aid in improving the performance [Androutsopoulos and Galanis, 2005]. Thus, those examples that do not fall into both automatically generated categories can

POSITIVE AND NEGATIVE FROM THE WEB

be expunged. Allegedly, the automatic elimination of these examples is a much better option than creating a third group, since they can still convey very good definitions that partly overlap with descriptions in KBs about their respective *definienda*.

Alternatively, large-scale training corpora can be constructed by assuming that descriptions supplied by KBs about a *definiendum* are positive examples, while those very dissimilar to them are negative. Then, the quality of the set of positive training sentences banks primarily on the structure of the KBs, and on the performance of the heuristic that harvests the descriptions therein. Of course, both factors also play a vital role in sifting positive examples from web pages, but in this case, the structure markedly varies from one web page to the other, and analogously, an efficient heuristic can become more complex. This difference is critical because of the natural imbalance noted by [Androutsopoulos and Galanis, 2005]: positive examples are harder to find across web pages than negative sentences, hence accounting for a fixed structure and authoritative sources for getting positive examples is a decisive advantage.

POSITIVE FROM KBS

Another prime consideration that must be taken into account is the diversity of the positive training set. On the one hand, ensuring relatively high levels of similarities between positive training sentences taken from the Internet and from articles about the a *definiendum* in KBs help to build a dependable positive set. This boost in reliability is, on the other hand, at the expense of diversity in training sentences, because many web sentences with more novel descriptive content, in relation to their related articles across KBs, might be discarded, and many facets elucidated in KBs articles are not necessarily covered by web sentences. Another aspect that makes the extraction of positive samples from web pages less attractive is that it requires experimentally tuning two different thresholds, while using positive sentences from KBs only one for the negative set. Incidentally, one also should account for the fact that both experimental thresholds can considerably vary from one training *definiendum* to the other. Thus, it is uncertain how to automatically set their optimal value, compelling the compromise between simplicity and accuracy. All things considered, experimentally setting the standard threshold(s) is the most attractive and workable solution.

Detractors to this approach can argue that these positive sentences would have a strong bias in favour of some specific types of wordings. However, there is always an inherent portability problem when exploiting data-driven methods, and admittedly, this heavy bias is present in the first sentences of the articles, but wordings of posterior sentences become less predictable, and they are therefore more plausible to offer an ampler diversity of nugget types.

7.6.1 Corpus Acquisition

Extracting Reliable Positive Examples From Wikipedia

As a matter of fact, Wikipedia pages are made up of several sections such as infoboxes, categories and the main body of the article. The body is usually the richer part in terms of descriptive information verbalised in natural language, and it is normally split into several thematic sections, which differ amongst *definienda*. The first section is called the abstract, because it typically put into words some sort of summary of the most pertinent aspects of the *definiendum*, whereas other sections tend to supply noiser or more irrelevant pieces of information. Equally important, this section is commonly found across all articles. Since Wikipedia articles are semi-structured, simple heuristics can extract this section from each page. In many cases, Wikipedia also produces official abstracts, which are shingles consisting of one to three sentences laying out the essence of the respective *definiendum*. However, these official abstracts are insufficient and not heterogeneous enough to build efficient clas-

WIKIPEDIA ABSTRACTS

sifiers capable of identifying a broad diversity of distinct descriptions. In addition, abstracts embodied in old snapshots of Wikipedia can be exploited, wherewith the coverage and the diversity of the positive set for some *definienda* can be widened, when needed.

Certainly, taking into account several Wikipedia resources alleviates the problem of finding few sentences in any of them, and more important, it mitigates potential obstacles when identifying the introductory section. As well as that, it lessens the impact of official abstracts bearing only structural and noisy information. Seemingly, these official abstracts are automatically extracted. In order to use these sentences for training, they are tokenised using OpenNLP, and in the case of duplicates, only one instance is left. To be more accurate, duplicate sentences are detected by simple string matching, and all sentences that do not overlap with any non-stop word belonging to the *definiendum* (title of the page) are expunged. Prior to this last task, co-references are needed to be resolved. For this purpose, the replacements adopted by [Keselj and Cox, 2004, Abou-Assaleh et al., 2005] are utilised (see details in table 3.5 on page 70). All instances of the *definiendum* are substituted with a placeholder (`CONCEPT`), this way the learnt models avoid overfiting any strong dependance between some lexical properties of the training *definienda* and their definitions across the training set. Some positive pairs <*definiendum*, positive training sentence> generated by this process are listed in table 7.10.

CORPUS PREPROCESSING

Definiendum	Training Sentences
A Handful of Dust	• `CONCEPT` is a novel by Evelyn Waugh published in 1934 .
A. G. Lafley	• Afterwards , `CONCEPT` studied at Harvard Business School , receiving `CONCEPT` 's M.B.A. in 1977 .
Zvi Elpeleg	• `CONCEPT` later entered academia , becoming an Arabist at the Dayan Institute . • `CONCEPT` was the military governor of Gaza , and was Israel 's first military governor of the West Bank .

Table 7.10: Samples of positive examples harvested from Wikipedia abstracts.

Obtaining Unlabelled Sentences

At this point, a group of pairs <*definiendum*, positive training sentence> has been distilled from Wikipedia abstracts. The next step is then finding their counterparts <*definiendum*, unlabelled sentence> across the Web. More precisely, unlabelled sentences carrying the *definiendum* are acquired by processing full-documents fetched from the Internet by means of Yahoo! Search. Since the amount of *definienda* determined in the previous step is huge (about 2,000,000), only *definienda* satisfying the next characteristics are submitted:

1. *definienda* consisting solely of numbers, letters and hyphens as well as periods.
2. *definienda* with more than two positive examples extracted in the previous step.
3. *definienda* disagreeing with purpose-built pages like lists (e.g., "*List of economists*") and categories (e.g., "*Category of magmas*").

In practice, a maximum of 100 hits were fetched per *definiendum*. Since there is a trade-off between the number of documents and the total download time, only hits embodying the exact wording of the *definiendum* within their web snippets were taken into consideration. This provided 3,810,512 documents with relation to 292,185 different *definienda*. Subsequently,

two kinds of hits were removed: (a) documents from Wikipedia, and (b) documents with a Multipurpose Internet Mail Extensions (MIME) type different from "text/html". The underlying idea here is to prevent the set of unlabelled sentences from polluting with binary snippets and/or positive examples. In the case of Wikipedia, rules designed to detect Wikipedia articles were implemented. The remaining hits were retrieved, tokenised and split into sentences via OpenNLP. Only sentences embracing the exact match of the respective *definiendum* were chosen, and accordingly, this matching *definiendum* was replaced with the same placeholder utilised for the positive examples. All in all, this procedure supplied about 1,600,000 unlabelled sentences pertaining to about 150,000 different *definienda*.

DOCUMENT PRE-PROCESSING

Labelling Unlabelled Training Sentences

There are several techniques intended for learning with positive and unlabelled data [Liu, 2006]. Most of these approaches are geared towards determining a set of reliable negative examples from the unlabelled samples, which are subsequently interpreted as the negative training set. Commonly, as aforementioned, the reliability of each negative example is measured in conformity to its resemblance to the positive set. In the case of definition QA, there are two possible ways of quantifying this similarity: (a) at the *definiendum* level, and (b) with respect to the entire positive set.

POSITIVE REFERENCE SET

Intuitively, automatically annotating unlabelled samples at the *definiendum* level seems to be more advantageous to accomplish a larger degree of lexical ambiguity in the training material. A chief obstacle for judging the likeness with respect to the whole array of positive examples stems from describing words like *actor* and *singer*. This kind of term is highly likely to be found across many descriptions in Wikipedia or any other encyclopedia, making many unlabelled examples that carry these words to create the misleading impression about being positive while they are actual negative examples. Since this kind of feature is not inherently discriminative as they rely heavily on the context, their complete removal induces a bias in the classifier which diminishes the performance. Simply stated, the classifier will be likely to learn that all sentences bearing these terms must be tagged as definitions, which is definitively false. On the other hand, operating at the *definiendum* level ensures that this sort of lexical bias is attenuated across *definienda*, that is, the word *organisation* can occur in the negative examples of singers, and vice verse. However, the efficiency of this technique resides largely in the coverage supplied by Wikipedia for each *definiendum* in the positive set, and for this reason, unlabelled definitions with few lexical overlap or corresponding to diverge senses are hard to detect. A way of allaying this data sparseness is to account for several KBs or several snapshots of Wikipedia. All things considered, there is not magic bullet to this labelling issue, it is expected, however, that the lexical features contained in the mislabelled negative examples are much more prominent in the positive category.

RANKING UNLABELLED SAMPLES

As a means to automatically annotate unlabelled examples, a centroid vector is constructed for each *definiendum*. This vector is formed of the terms in the positive examples excluding: (a) stop-words, (b) punctuations, (c) the concept placeholder, and (d) one character length tokens. Term frequencies in the positive examples were used as weights. Each unlabelled sentence was thereafter scored in accordance with the cosine similarity to this centroid vector. This similarity was computed for fragments of twenty consecutive words, and the highest similarity remained as the similarity of the whole sentence. Fixed-length fragments are utilised for dealing with long sentences carrying only few descriptive information. Some samples of ranked unlabelled sentences are shown in table 7.11.

THRESHOLD

Sentences rating lower than an empirical threshold (0.2) were labeled as negative, and the remaining were kept unlabelled. This experimental threshold also assures a slightly lexical overlap with the positive set. Higher values bring about a larger amount of mislabelled

Ranking	Unlabelled Sentence
0.000048	Maybe you 've confused it with CONCEPT , but it is nowhere where you put it
0.000052	Inside front cover has a map of Vinyalonde above a map of CONCEPT .
0.060302	In my new diorama , I show a deep-draught cargo vessel being built in the great Numenorean ship-yard of CONCEPT .
0.120623	Númenóreans are limited to coastal areas : they held CONCEPT on the Gwathló as a trading post , and Pelargir and Umbar were already founded .
0.235969	In the fiction of J.R.R. Tolkien , CONCEPT (also spelt Ened) was a great harbour in Eriador founded by the Númenóreans .
0.337174	It was also called CONCEPT meaning " Great Middle Haven " because it was located between the Grey Havens and Pelargir (though Pelargir was not established until 2350 S.A.) .
0.404519	Lond Daer , or CONCEPT is a great harbour at the mouth of the river Gwathló in Eriador.

Table 7.11: Some ranked unlabelled sentences for "*Lond Daer Enedh*".

negative examples. In substance, it was empirically observed for a group of fifty *definendums* that good values range from 0.2 to 0.3. However, the probability of polluting the negative set with definitions raises as the threshold turns to be more aggressive. For this reason, a conservative value was selected. Additionally, sentences within negative examples aligning the next eight definition patterns were re-tagged as "*unlabelled*":

1. <definiendum> (is | are | has been | have been | was | were) (a | the | an) <description>
   ⇒ "*AA Roadwatch is a service offered by the New Zealand Automobile Association, which provides a guide to travel and motoring in New Zealand, with traffic alerts, car parking.*"

2. <definiendum>, (a | an | the) <description> (, | .)
   ⇒ "*IMSA, the Insurance Marketplace Standards Association, is a voluntary, non-profit organization founded in 1996 to strengthen consumer trust and confidence in the life insurance, long ...*"

3. <definiendum>, or <description>
   ⇒ "*Instituto de los Mexicanos en el Exterior, or IME ...*"

4. <definiendum> (| ,) (| also | is | are) (called | named | nicknamed | known as) <description>
   ⇒ "*Norma Jean, a Country music singer, nicknamed 'Pretty Miss Norma Jean' (or was it, as she sometimes wrote it, Norma Jeane) Baker.*"

5. <definiendum> (<description>)
   ⇒ "*Indian Institute of Management (IIM) - Kolkata.*"

6. <definiendum> (become | became | becomes) <description>
   ⇒ "*AFC Bournemouth became the first community run club in 1997 after fans stepped in to prevent the club going into liquidation.*"

7. <definiendum> (| ,) (which | that | who) <description>
   ⇒ "*Motlatsi Molapisi, who heads the Botswana People's Party, urged his government to erect an electric fence along the common border, while the Botswana Congress Party urged the...*"

8. <definiendum> (was born) <description>
   ⇒ "*Attac-France was born in 1999 and other Attac groups emerged soon afterwards in many other European countries.*"

In this scheme, a strict rule matching of the *definiendum* is taken into consideration, this means the sentence must start with the placeholder of the *definiendum*. In order to remove additional mislabelled negative examples, a list of common descriptive phrases across all Wikipedia abstracts is constructed. These phrases are derived from sentences aligning the first pattern. From these matching sentences, the description is extracted. Next, this definition is trimmed at the first verb or punctuation sign by means of simple heuristics. Phrases consisting of more than one word are kept. For the purpose of getting reliable hits, these phrases must occur at least three times. As a result, about 46,000 distinct descriptive phrases were discovered including:

Phrases	Phrases	Phrases
1920 drama film	Croatian novelist	Nigerian author
2nd album	Czech chemist	anti-war activist
6th century manuscript	Debut album	charity organisation
8-bit character encoding	Democratic Senator	domestic airport
African American film	Evangelical Christian	herbaceous perennial plant
Albanian football club	General Secretary	independent film
Australian journalist	Gothic castle	leading mathematical physicist
Brazilian supermodel	High Priest	membrane protein
British ecologist	Inuit artist	newspaper journalist
Cameroonian football club	Labor candidate	prolific striker
Chinese actor	Liberal MP	second head coach

Table 7.12: Descriptive phrases used for reverting labels.

If any of these phrases appears within a window of five words to the left or to the right of the placeholder of the *definiendum*, then the label of respective negative sentence is reverted to "*unlabelled*".

Building a Balanced Training Set

Unfortunately, the training sets built in the previous steps are imbalanced, that is, the number of positives and negatives markedly differ. Balanced training sets are necessary to avoid biases, prevent overfiting, and learn discriminative feature distributions. In literature, several methods have been conceived to tackle this problem head-on. For instance, a survey can be found in [He and Garcia, 2009]. Alternatively, some ad-hoc purpose-built procedures can still be utilised in order to exploit some unique characteristics of a particular classification problem. Since these heuristics are easier to implement and the real benefit that can be reaped from more complex strategies is hitherto unknown, heuristics (or manual annotations) have been preferred so far (see section 7.2 and table 7.9).

Accordingly, a balanced training set is constructed by singling out approximately the same amount of positive and negative examples per *definiendum*. First, a base number of sentences k is stipulated as the smaller size of the two sets, and it is computed per *definiendum*. Second, the first k negative and positive sentences per *definiendum* are singled out for training. The order in the positive set is given by the article in Wikipedia, while in the negative set, by the similarity to the centroid vector. In the positive set, sentences are chosen from

the abstract of the article first, then the official abstract, and when needed, from abstracts in an extra snapshot of Wikipedia, whereas in the negative set, sentences with higher cosine similarity are preferred.

Third, since some *definienda* provide more negative training data, while others yield more positive examples, a set was allowed to exceed the other by a maximum of 10% of k. This way, more training material is obtained and a balance is kept across *definienda*.

CORPUS On the whole, this procedure supplied about 707,000 positive sentences, whereas about 736,000 negative examples. In order to check the degree of noise in the negative set, 1,000 randomly-selected sentences were manually inspected. Out of these, 13.5% were actual definitions, i.e., positive examples incorrectly labeled as negative. Typical examples in this group are sentences expressing a descriptive phrase not captured by the descriptive phrase acquisition algorithm, e.g., "*Congresswoman* CONCEPT". Conversely, 12,2% of 1,000 randomly chosen positive examples were non-definitions. In this array, some domains, such as music and movies, yielded more misleading sentences than other domains, typically due to Uniform Resource Locator (URL) names such as "*Official site*" and "*IMDB*". Another cause of mislabelling is wrong inferences during co-reference resolutions, and more fundamental, some sentences were definitions, but they described another concept related to *definiendum*.

7.6.2 Sentence Level Features

Since the intention is designing strategies that only work at the sentence level, attributes that can be taken exclusively from each isolated sentence were counted for scoring candidate sentences. As discussed earlier, this kind of constraint is relevant to definition QA systems edging towards discovering answers within any kind of text shingle, document surrogate or full-document. The list of ingredients distilled from the acquired training material is as follows:

Uppercased Words include terms that start with a capital letter or a number as attributes. Each word is associated with its frequency in the answer candidate when building its feature vector. During training, only elements with a frequency higher than two (across the training sets) are taken into consideration. Stop-words and one-character-length tokens are also removed.

Lowercased Words perceive terms that start with a lowercase letter as features. Similarly to the previous attribute, each word is associated with its frequency within the putative answer when forming its respective feature vector. Likewise, only elements with a frequency higher than two were considered when training, and stop-words and one-character-length tokens are also removed. The motivation behind separating words that start with a capital letter or a digit is that such words commonly refer to the structure of the page rather than its content (e.g., "*About*", "*Home*", "*Search*", and "*More*").

Semantic Classes benefit from SuperSense Tagger of [Ciaramita and Altun, 2006][1] for replacing sequences of terms with their corresponding semantic class. This tagger was configured with classes from the BBN Wall Street Journal (WSJ) Entity Corpus[2]. This comprises 105 distinct categories for entities, nominal concepts and numerical types. Some classes include:

DATE:AGE, DATE:DATE, DATE:DURATION, DATE:OTHER, FAC:AIRPORT, FAC:BRIDGE, FAC:ATTRACTION, FAC:BUILDING, FAC:HIGHWAY:STREET, FAC:OTHER, GPE:CITY,

[1] http://web.net/projects/supersensetag

[2] LDC catalog number LDC2005T33

Models	I	II	III	IV	V	VI	VI	VII	IX	X	Xa	X	XIa	XII	XIIa	XII	XI\	XV
Uppercased words	x		x	x	x	x	x	x	x									
Lowercased words	x	x	x	x	x	x	x	x	x									
Semantic classes			x	x			x		x	x	x	›	x	x	x	x	x	x
Concept positions				x	x	x	x	x	x									
Definition patterns						x	x	x	x									
Number of tokens+ determiner								x	x									
Syntactic chunks										x	x	›	x	x	x	x	x	x
Selected substitution										x		›		x	x	x	x	x
Concept positions (chunk)												›	x	x	x	x	x	x
Words in concept (chunks)												›	x	x	x	x	x	x
Shallow-syntax LMs														x	*	x	x	x
Verb position (chunk)																x	x	x
Number of chunks																	x	x
Biterms in concept (chunks)																		x

Table 7.13: Summary of models.

GPE:COUNTRY, GPE:OTHER, GPE:STATE:PROVINCE, LANGUAGE, LOCATION:BORDER, LOCATION:CONTINENT, LOCATION:LAKE:SEA:OCEAN, LOCATION:OTHER, LOCATION:REGION, LOCATION:RIVER, NORP:POLITICAL, NORP:RELIGION, ORGANIZATION:CITY, ORGANIZATION:CORPORATION, ORGANIZATION:EDUCATIONAL, ORGANIZATION:GOVERNMENT, ORGANIZATION:HOSPITAL, ORGANIZATION:HOTEL, ORGANIZATION:MUSEUM, ORGANIZATION:OTHER, ORGANIZATION:POLITICAL, ORGANIZATION:RELIGIOUS, PERSON, WORK_OF_ART:BOOK, WORK_OF_ART:OTHER, WORK_OF_ART:PAINTING, WORK_OF_ART:PLAY, WORK_OF_ART:SONG

The motivation behind profiting from a semantic tagger is that it alleviates data sparseness by replacing uncommon names/words with their semantic category. An example sentence labeled by this tagger is as follows:

CONCEPT is a GPE:COUNTRY military advanced technology demonstration project that is part of the ORGANIZATION:CORPORATION.

Equally to the other two ingredients, only classes with a frequency higher than two across the training set are considered.

Concept positions enrich the feature vector with information about the position of the *definiendum* in the candidate sentence. An element of the form "DefiniendumPosition=i" was added every time token i carried the placeholder "CONCEPT". For instance, the following sentence will generate the components "DefiniendumPosition=1" and "DefiniendumPosition=19":

CONCEPT lived in Paris for a year , studying philosophies of evolution on a Fulbright scholarship before completing CONCEPT's Ph.D. from the Biology Department of Yale in 1972 .

The addition of these positional attributes can eventually draw a clearer distinction of some positional regularities across negative and positive examples.

Definition Patterns incorporate features that indicate whether or not wide-spread definition patterns align candidate sentences. Analogously to [Miliaraki and Androutsopoulos, 2004, Androutsopoulos and Galanis, 2005], eleven boolean attributes were added in consonance with eleven definition patterns, which were derived from the unification of [Xu et al., 2005, Cui et al., 2007] (see sections 7.3 and 4.2). To control spurious matches, an alignment is considered valid if and only if a maximum of three words exists between the beginning of the sentence and the placeholder of the *definiendum*.

Determiner+Number Of Tokens add two of the ingredients mixed by [Xu et al., 2005] (see sections 7.3). The first element symbolises whether or not a determiner precedes the placeholder of the *definiendum* (feature (b) in section 7.3). The second attribute indicates the number of tokens in the sentence (attribute (j) in section 7.3).

Syntactic Chunks are recognised by means of MontyLingua[3]. Each chunk is modelled as a concatenation of its tokens, and is associated with its frequency within the sentence. Chunks consisting solely of punctuation and stop-words as well as chunks with a frequency lower than three across the training set were left unconsidered.

(CONCEPT) (was co-developed) (by) (Vinod Dahm , designer) (of) (the Intel processor) (.)

Selected Substitutions (POS) replaces words observing the following Part-of-Speech (POS) categories with their corresponding actual POS tags: DT, CC, PRP, RB, MD, PDT, RBR, RBS, PRP$ and CD. In addition, a placeholder (VERB) was utilised for verbs typically seen in definitions: is/VBP, is/VBZ, are/VBP, were/VDB, was/VDB, become/VB, becomes/VBZ, became/VBD, have/VB, had/VBD, have/VBP and has/VBZ. The following sentence corresponds to the previous illustrative example with these replacements:

(CONCEPT) (VERB co-developed) (by) (Vinod Dahm , designer) (of) (DT Pentium processor) (.)

These replacements share the same spirit with the selective substitutions used by [Cui et al., 2004a] (see table 4.12 on page 93). These inductions cooperate on coping with data-sparseness by making some generalisations on the training data.

Concept Position (chunks) equips the feature vectors with positional information at the chunk level. This is done by adding "DefiniendumChunkPosition=i" attributes indicating the chunk positions of *definiendum* placeholders. To illustrate, "DefiniendumChunkPosition=1" and "DefiniendumChunkPosition=6" are extracted from the next sentence:

(CONCEPT) (is grown) (by) (gardeners) (for) (CONCEPT 's striking appearance) (when) (in) (flower) (.)

Words in Concept Chunks add words within *definiendum* chunks (e.g., *"striking"* and *"appearance"* in the previous working example) as features. These also banned punctuation and terms co-occurring with the CONCEPT less than three times across the training set. Every time these ingredients are taken into account, words in other chunks are left unconsidered.

[3]For chunking and Part-of-Speech (POS) tagging, MontyLingua was used:
http://web.media.mit.edu/~hugo/montylingua/

Shallow-syntax LMs include information from shallow-syntax language models (LMs) acquired from the training corpus. LM statistics are gathered as follows: first, sequences of seven chunks in length are identified across the training set. Only sequences having a middle chunk containing the placeholder of the *definiendum* are preserved. Subsequently, the three left and right chunks across the training set is counted. Some illustrative chunk pairs discovered are:

<(published), (..CONCEPT..)>; <(is officially), (..CONCEPT..)>;
<(..CONCEPT..), (was elected)>; <(..CONCEPT..), (officially played)>

Only chunks patterns seen more than three times in training were considered. The difference between MODELS XII and XIIA (signalled with a star in Table 7.13) is that MODEL XIIA reduces each chunk to its last word. For instance, (was elected) is seen as (elected).

Verb position (chunk) incorporates positional attributes for definitional verbs. More precisely, the "VerbPosition=i" feature is added whenever the chunk i carries the placeholder VERB. For example, the following sentence produces "VerbPosition=2":

(CONCEPT) (VERB) (DT CD best player) (in) (DATE:DATE) (.)

It is worth noting that this attribute is put together with selective substitutions, which give the placeholder VERB.

Number of Chunks adds a feature storing the amount of chunks within the sentence. For instance, "NumberOfChunks=6" in the previous example.

Biterms in Concept Chunk enriches the feature vector with bi-terms co-occurring with CONCEPT in the same chunk. Here, only bi-terms seen more than three times across the training set are taken into account.

A final remark on attributes is due to the order. If the model accounts for semantic classes and/or selective substitutions, these are computed first, then the other features are extracted from the modified sentences.

7.6.3 Testing Sets and Baseline

Since the experiments carried out by [Fahmi and Bouma, 2006] showed that ME Models are inclined to provide a better performance in a similar task, the implementation of these models supplied by OpenNLP was used for the next experiments. Accordingly, two test sets and a baseline were taken into account for assessing the models:

Set A (in-domain) consists of 5,064 sentences. This group was constructed akin to the training set, but for a different array of *definienda*. Like the training set, these sentences were automatically annotated with the same strategy (i.e., positive examples from Wikipedia and negative examples from the Web): 2,360 (46.60%) positive and 2,704 (53.39%) negative examples. This corpus is expected to have the same lexico-syntactic properties as the training dataset.

Set	Baseline I	I	II	III	IV	V	VI	VII
A	44.77/44.23/44.17	71.64	64.42	71.15	73.10	74.22	74.58	73.40*
B	56.26/57.77/56.73	60.87	51.66	**62.13**	57.68	58.59*	58.41*	57.00

Set	VIII	IX	X	Xa	XI	XIa	XII	XIIa	XIII	XIV	XV
A	74.86	73.87	73.40	72.17	76.83	75.31	76.85*	76.36	77.30*	**77.46**	77.30*
B	58.49*	57.42*	58.56*	58.44*	58.38*	58.15	58.99	57.40	59.10	59.31	59.41

Table 7.14: Accuracy (%) accomplished by the different models and the baseline. For BASELINE I, three scores are shown pertaining to three distinct threshold values: 0.1, 0.2, and 0.3 (in order). Note that ME models were utilised as the ranker.

Set B (out-of-domain) comprises 5,165 sentences taken from the unlabelled set, that is, sentences harvested from the Internet, for which there was uncertainty as to whether or not they were negative examples. Accordingly, they were manually tagged: 2,345 (45.4%) and 2,820 (54.60%) positive and negative examples, respectively. Because of the fact that these sentences come from different resources than the training corpus (all Web instead of Wikipedia and the Internet) and the acquisition process was different, the characteristics of this corpus (both lexical and syntactic) are expected to be significantly dissimilar from the training dataset.

BASELINE I

Baseline I (centroid vector) discriminates candidate sentences on the grounds of their similarity to the centroid vector of the respective *definiendum* (see section 4.8 on page 91). More precisely, since implementations slightly different from one another, the blueprint presented in [Chen et al., 2006] was utilised for its construction. This centroid vector was built for each *definiendum* from a maximum of 330 web snippets, which were fetched by means of Yahoo! Search. As a search strategy, multiple queries per *definiendum* were submitted: (a) one query containing the *definiendum* and three task specific cues: "*biography*", "*dictionary*" and "*encyclopedia*"; (b) ten additional queries conforming to the structures listed in section 2.5 (page 36). Accordingly, hits from Wikipedia were removed as they can be included in the test set A.

Subsequently, co-references are resolved in congruence with the replacement method of [Keselj and Cox, 2004, Abou-Assaleh et al., 2005]. The centroid vector is then built from the snippet sentences that embody the *definiendum*. At evaluation time, sentences whose similarity to the centroid vector is lower than a threshold are labeled as negative. All others are marked as positive. Experimentally, it was observed that values of this threshold in the interval $[0.1, 0.3]$ performed the best.

7.6.4 Results and Discussion

Table 7.14 underlines the outcomes achieved by the baseline and the models presented in this section. In this table, each result is significantly better statistically than the next ranked score, unless it is noted with a star. This significance was computed on 20 samples generated using bootstrap resampling along with *ttest* considering $p\text{-}value < 0.05$. In light of the figures in table 7.14, the following observations can be pointed out:

ttest

TRAINING DATA

1. Results show that lexicalized models are crucial for definition QA systems. MODEL I finished with significantly higher scores than the baseline, for both testing corpora. This signifies that when enough training material is available, lexicalized features bring

Model	1	2	3	4	5	6	7	8	9	10
Baseline I(.1)	0.3925	0.4022	0.4153	0.4276	0.4425	0.4398	0.4388	0.4371	0.4389	0.4447
Baseline I(.2)	0.4032	0.394	0.4208	0.4347	0.4425	0.4418	0.4439	0.4425	0.4416	0.4454
Baseline I(.3)	0.3871	0.4049	0.4372	0.4418	0.4448	0.4458	0.4481	0.4449	0.4493	0.4454
Model I	0.9022	0.8729	0.858	0.8521	0.8329	0.8193	0.8151	0.7911	0.777	0.7603
Model II	0.7742	0.7556	0.7337	0.7274	0.7182	0.7097	0.7002	0.6966	0.6905	0.6905
Model III	0.8865	0.8583	0.8286	0.8323	0.8191	0.7953	0.7912	0.7806	0.7671	0.7581
Model IV	0.9027	0.9028*	0.8721	0.8598	0.83	0.81	0.7933	0.7722	0.7663	0.7607
Model V	0.9297*	**0.9171**	0.8928*	0.8735	0.842	0.8322	0.8125	0.7971	0.7869	0.7762
Model VI	0.9351	0.9116	0.8895	0.8735	0.841	0.83	0.8169	0.7984	0.7887	0.778
Model VII	0.9081	0.9028	0.869	0.8598	0.8329	0.8087	0.7943	0.7754	0.7634	0.7641
Model VIII	0.9297*	0.9144	0.8915*	0.8765	0.8475	0.8355	0.8232	0.8016	0.7896	0.7832
Model IX	0.8973	0.9	0.8728	0.8674	0.8367	0.8132	0.8014	0.7763	0.7707	0.7641
Model X	0.8602	0.8361	0.8363	0.8182	0.795	0.7958	0.7882	0.7779	0.7878	0.7754
Model Xa	0.8441	0.8169	0.8084	0.8076	0.8038	0.795	0.7867	0.7815	0.7769	0.7805
Model XI	0.9247	0.8852	0.8798	0.8772	0.8472	0.8377	0.8299	0.8205	0.8137	0.8056
Model XIa	0.9086	0.884	0.8508	0.8599	0.846	0.8366	0.8264	0.8152	0.8053	0.7935
Model XII	0.9086	0.888	0.8902	0.8869	0.8725*	0.8593*	0.8474*	0.8388	0.8376*	0.8273*
Model XIIa	**0.9462***	0.8962	0.885	0.8668	0.8435	0.8366	0.8236	0.8103	0.7968	0.7935
Model XIII	0.914	0.888	0.8889	0.8931	0.872	0.8554*	0.8445	0.8388*	**0.8393***	0.8234
Model XIV	0.914	0.8901	0.8927*	0.8931	0.8733*	0.8539	**0.8494***	**0.8406***	0.8372*	**0.8293***
Model XV	0.9086	0.9011	**0.8971***	**0.8967***	**0.8745***	**0.8615***	**0.8494***	0.8393*	0.8342	0.8254

Table 7.15: Average Precision at $k = 1, \ldots, 10$ for each system and model (set A). For BASELINE I, three scores are shown corresponding to three different threshold values 0.1, 0.2, and 0.3. The best results for every k are shown in bold face (ME models).

out good discriminative models. The performance suffers a steep drop for MODEL II, which is proof that words that begin with a capital letter contain essential information, regardless of the potential noise.

2. The outcomes also reveal that selective substitutions cooperate on accomplishing better results. In a special manner, POS tagging substitutions (MODELS X and XI) performed better than their counterparts without POS tagging information (MODELS XA and XIA) in both corpora. This is proof that POS tags assist in tackling data sparseness, and since they boosted the performance in both datasets, they aid in raising the portability of discriminative models. POS CATEGORIES

3. With regard to Set A, MODEL XIV is the best model taking advantage of chunking, and it outperforms the best model without syntactic information (MODEL VIII) by 2.60%. Contrarily, the best performance for Set B is achieved by MODEL III, which makes use of no syntactic information at all. Given this difference, it can be concluded that syntactic information provides strong hints in text that is syntactically clean, such as the content of KBs, but it is unreliable in (out-of-domain) web content. SYNTACTIC FEATURES

4. Observation 3 can be extended to definition patterns: MODELS VI and VII have a better performance in Set A than their counterparts without definitional patterns (MODELS V and IV respectively). The opposite is true in Set B, which is additional proof that (out-of-domain) web content is structured differently than texts extracted from KBs.

5. As for Set B, MODEL III finished with the highest accuracy. This model generalises word sequences to their semantic class. As expected, semantic analysis cushions data sparsity out-of-domain by replacing infrequent words with their semantic class and correctly learning that some semantic classes, e.g., DATE:DATE, are correlated with definitions. For example, 168 sentences bearing DATE:DATE entities turned from wrongly PORTABILITY OF SEMANTIC CLASSES

	1	2	3	4	5	6	7	8	9	10
Baseline I(.1)	0.5066	0.5385	0.5664	0.5629	0.5597	0.5497	0.5362	0.5221	0.5267	0.5306
Baseline I(.2)	0.5066	0.5405	0.5707	0.5644	0.5682	0.5607	0.5479	0.5443	0.5478	0.5503
Baseline I(.3)	0.5081	0.5445	0.5724	0.5727	0.5816	0.5724	0.5584	0.5527	0.5581	0.5624
Model I	0.6392	0.6526	0.6453	0.6304	0.6333	0.6596*	0.6349*	0.6463*	0.608*	0.6226*
Model II	0.6481	0.6398	0.6457*	0.6164	0.6354*	0.6117	0.6133	0.6083	0.6063	**0.6359***
Model III	0.6418	0.6617	0.6357	**0.6511***	**0.65***	**0.6691***	**0.6626**	**0.6587***	**0.6593***	0.6179*
Model IV	0.6477	0.6342	0.5833	0.5509	0.5938	0.5379	0.5625	0.5694	0.5802	0.46
Model V	**0.6748***	0.6656	0.6106	0.5833	0.6244	0.6369	0.5786	0.625	0.619*	0.5875
Model VI	0.674*	0.6667*	0.6087	0.5755	0.627	0.6319*	0.5882	0.5769	0.5397	0.5429
Model VII	0.6341	0.6213	0.5733	0.5469	0.5481	0.4907	0.4805	0.4643	0.3889	0.2333
Model VIII	0.6563	0.6526	0.6392	0.6154	0.6333	0.6304	0.6111	0.6083	0.5432	0.5556
Model IX	0.6464	0.6076	0.594	0.587	0.5931	0.553	0.6154*	0.5417	0.463	0.4
Model X	0.6009	0.6189	0.6333	0.628	0.6232	0.6059	0.5956	0.5521	0.5613	0.5438
Model Xa	0.6207	0.6369	0.6182	0.6184	0.5929	0.5931	0.6022	0.6281	0.6493*	0.6286*
Model XI	0.6214	0.6425	0.6439	0.6263	0.6086	0.5774	0.5649	0.5368	0.5897	0.5429
Model XIa	0.6105	0.652	0.6038	0.5955	0.597	0.6064	0.5897	0.5847	0.5822	0.4929
Model XII	0.6639	**0.6895***	0.6378	0.6224	0.6328	0.5761	0.5878	0.5968	0.5509	0.5947
Model XIIa	0.5926	0.6181	0.6366	0.6083	0.5941	0.5767	0.5615	0.5857	0.5299	0.5111
Model XIII	0.6734*	0.6694*	**0.6527***	0.6432*	0.6349*	0.5833	0.6134	0.63*	0.5789	0.5647
Model XIV	0.6742*	0.6772*	0.6477*	0.6436*	0.6317	0.5814	0.5952	0.6343	0.5444	0.5765
Model XV	0.6618	0.6622	0.6478*	0.6374*	0.6486*	0.617	0.6163*	0.6078*	0.5397	0.5632

Table 7.16: Average Precision at $k = 1, \ldots, 10$ for each system and model (set B). For BASELINE I, three scores are shown corresponding to three different threshold values 0.1, 0.2, and 0.3. The best results for every k are shown in bold face (ME models).

to correctly classified in MODEL III, while only 113 did the opposite. The same improvements were not seen on the in-domain corpus, where it can be conjectured that lexico-syntactic properties capture mostly the same signal as the semantic features due to the higher syntactic and lexical overlap with the training set. In other words, the generalisation of dates assists in alleviating the ambiguity of numbers and months, when they are considered isolated as in MODEL I and Set B. This ambiguity is, oppositely, mitigated by syntactic information in the case of Set A.

6. The best model in Set B scores over 15% lower than the best model in set A. The examination of the corresponding confusion matrix for the results in Set B indicates that about 78% of errors are false negatives (definitions tagged as non definitions). The reason for this is two-fold: (a) lack of coverage of the models, and (b) the structure of the text harvested from web content. For example, for MODEL III, most errors in Set B are generated for very short and very large sentences with little descriptive content. Specifically, in the case of positive sentences shorter than 200 characters, MODEL III labels 51.90% as negative, while this systematically increased to 62.86%, 65.67% and 73.41% when classifying sentences of 201–300, 301–400 and longer than 400 characters, respectively. In short, the training material does not supply enough samples with this syntactic property: sentences with only a small descriptive portion. This makes sense, because sentences across Wikipedia abstracts (positive samples) are likely to be entirely descriptive. Consequently, the large portion of non-descriptive content makes this sort of testing instance to resemble more the negative than the positive examples, therein lies the negative label misassigned to this type of concise descriptions.

SENTENCE LENGTH

7. Surprisingly, the inclusion of the eleven definition patterns had a more significant repercussion on enhancing the recognition of answers that misalign than match these regularities. In the case of Set A, eight candidates that observe definition patterns

DEFINITION PATTERNS

shifted from being wrongly to correctly labelled, while seven candidates did the opposite (MODELS IV and VII). This means only one (8-7) out of the sixteen overall improvements comes from answer candidates that align the definition patterns, whereas the remaining fifteen are originated by other classes of sentences.

The same picture is seen if MODELS V and VI are juxtaposed, nine answer candidates that observe definition patterns turned from wrongly to correctly labelled, while only four did the contrary. That is, solely five (9-4) out of the nineteen overall betterments emanate from putative answers that match definition clauses, while the other fourteen from other kinds of candidates.

In light of these figures, it can be concluded that definition cues were more likely to ameliorate the detection of in-domain answers that mismatch definition patterns.

In the case of Set B, definition patterns tended to deteriorate the performance: eleven candidates that observe definition patterns went from correctly to wrongly labelled, whereas only one did the contrary (MODEL V and VI). When MODEL IV and VII are contrasted: twelve candidates turned from correctly to wrongly tagged, while only five did the opposite. This certainly showed that definition clauses negatively affect the performance for out-of-domain test cases.

It is important to underscore that this counting also made allowances for those few sentences with more than three words between the placeholder of the *definiendum* and their beginning, which is a restriction when aligning definition cues.

8. The best configuration for the Set A capitalises on semantic classes and selected substitutions as well as several syntactic features: syntactic chunks, verb and concept positions, words correlated with the *definiendum* in the same chunk, shallow syntax LMs, and number of chunks. These elements assisted in boosting the accuracy from 71.64% to 77.46% in relation to the basic configuration encompassing only lexical features (MODEL I). This is evidence pointing out to the positive contribution of the proposed attributes to deal with in-domain test cases. PROPERTIES

In addition, tables 7.15 and 7.16 yield another view of the achieved results. Both tables present the outcomes in terms of precision at k. In these tables, significance tests were performed utilising two-tailed paired *t-test* at 95% confidence interval on twenty samples of the corresponding test corpus. Here, each sample is obtained using bootstrap resampling, and has the same size as the original test corpus. That is to say, statistical significances were computed at twenty times the size of the questions in the test corpus. In the particular case of Set A, on $20 \times 5{,}064 = 101{,}280$ sentences. To neatly illustrate, the statistical significance implies that the precision at 1 of MODEL XIIA is not statistically different from the next ranked model MODEL VI (Set A). Some remarks concerning the outcomes in tables 7.15 and 7.16:

1. At various ranking levels k, MODEL I significantly outperforms BASELINE I in terms of precision for both data-sets. This outcome reaffirms that training discriminative models on the acquired corpus is beneficial for ranking answer candidates to definition questions, in general. In particular, given the fact that simple attributes were used by MODEL I. TRAINING CORPUS

2. In both testing corpora, models intermixing more complex attributes outperformed MODEL I, corroborating that the feature study for definition QA systems is promising.

3. At most ranking levels k, MODELS XV and XIV give the best peformance for Set A. Both models are equipped with most of the presented attributes, hence confirming

their usefulness though they operate on automatically acquired corpora. Conversely, MODELS III reaped the best results for $k = 4 \ldots 9$ for Set B, whereas models that account for more syntactic information were inclined to accomplish better results solely for the top three ranking positions. This is due to the bias in favour of the few answer candidates in Set B that bear syntactic similarities with the training corpus.

SHALLOW VS. DEEP SYNTACTIC FEATURES

4. The best outcomes for the top-two ranked answers are MODEL V and XIIA (Set A). The fact that both models are heavily lexicalised and profit from very limited amount of syntax, lead to the conclusion that top-ranked positions can be assigned by means of shallower features, while improving the ranking at lower positions (greater values of k) requires deeper syntactic information. This is in direct contradiction to the results obtained for out-of-domain testing cases.

   Distinctly, MODEL V takes advantage of all terms as attributes in conjunction with the position of the *definiendum* within the sentence. This position plays the role of a very simple detector of syntactic role for the *definiendum*. On the other hand, MODEL XIIA eliminates some sparsity caused by words through the substitution with semantic classes or POS tags. This model also incorporates extra targeted lexicalisation through lexical patterns detected by reduced shallow syntactic LMs. Here, shallower is understood with respect to MODEL XII, which profits from full chunks. In a nutshell, in order to enhance the performance at lower ranking levels, deeper syntactic similarities with the training corpus are necessary.

COVERAGE

5. Results achieved by MODEL I suggest that learning term distributions from large training corpora brings about better results than from localised and contextualised web snippets (as BASELINE I does), because a significant amount of redundancy is necessary to discover sentences that delineate the numerous facets of the *definiendum*. MODEL I, for instance, amalgamates the evidence yielded by all training *definienda*, this way the chances of distinguishing descriptions that are verbalised with words not found or barely found within web snippets. Consequently, this combination of evidence assists in reducing the lack of coverage, which usually markedly diminishes the performance of lexicalized definition QA systems [Zhang et al., 2005, Han et al., 2006].

In order to hold a general view of the outcomes in terms of precision at k, the average precision was calculated as the average value of the k levels ($k = 1, \ldots, 10$). Accordingly, figure 7.1 highlights the accomplished average precisions. Broadly speaking, these results substantiate previous observations:

(a) In domain, the best models (MODELS XIII, XIV, and XV) take advantage of most of the presented features. These models almost double the performance reached by BASELINE I. On the other hand, a significant negative impact eventuates from the removal of words that start with a capital letter (MODEL II).

(b) Out of domain, MODEL III finishes with the best performance, suggesting the pertinence of semantic class abstractions as a sparsity reduction technique.

On the whole, the inferences drawn from the training material did not fully port to the corpus extracted from the Internet. This means there are additional regularities that express descriptions, which are not typically found across KBs.

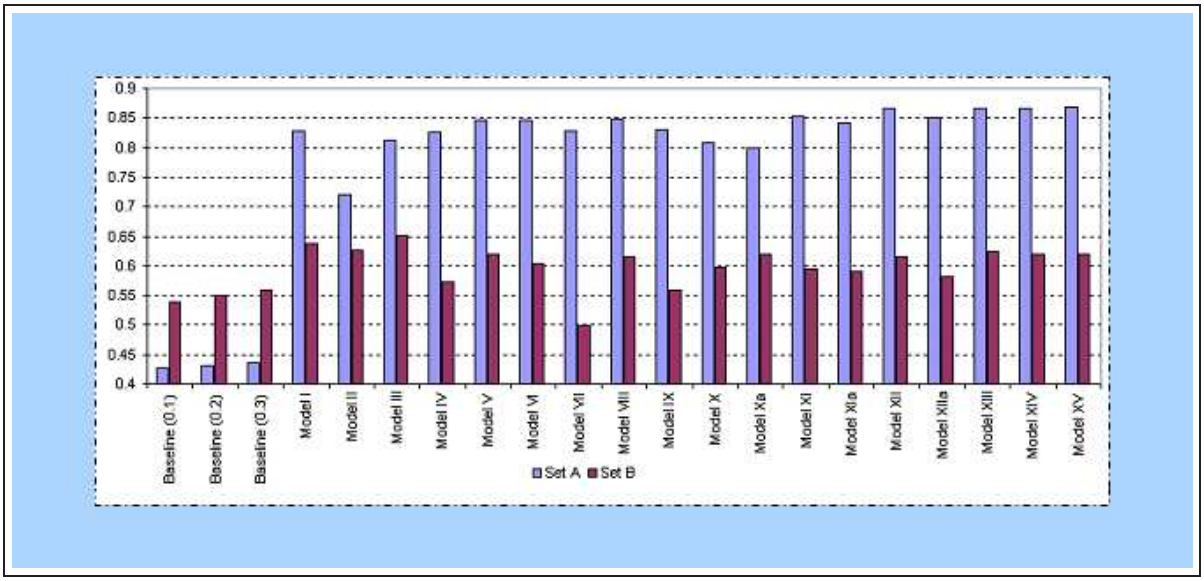

Figure 7.1: Overall comparison between the results obtained by the different models in terms of average precision.

7.6.5 Features Based on Dependency Trees

The previous study dissected the influence of assorted features on discriminant learning models. Exactly, these properties were inferred from chunks, named entities, POS taggings, and surface information. Contrarily, the next analysis extends the exploitation of dependency trees as a source for acquiring attributes. In detail, the following are the new ingredients distilled from the lexicalised dependency graph representation of the sentences contained in the training material:

Bigrams incorporate pairs of ordered connected words in lexicalised dependency trees into feature vectors. The order of these pairs is given by their hierarchy *gov→dep*. Lexicalised dependency graphs are obtained by means of Stanford Dependency Parser[4]. For instance, some illustrative attributes taken from the sentence "CONCEPT *is a science writer and lecturer on environmental matters.*" are as follows (see also figure 7.2):

ROOT→writer→science
ROOT→writer→lecturer
ROOT→writer→CONCEPT
START→on→matters
START→matters→environmental

Since nodes closer to the root usually convey the general gist of the sentence, a placeholder was introduced (ROOT or START) as a means to discriminate paths starting at the root from any other part of the tree. Each path was associated with its frequency in the sentence.

Trigrams enrich the feature vector with triplets of ordered connected words. Homologously to bigram attributes, these ingredients are extracted from lexicalised dependency paths, and correspondingly, they are represented in the same manner as bigrams. Some

[4]This parser is available at: http://nlp.stanford.edu/software/lex-parser.shtml

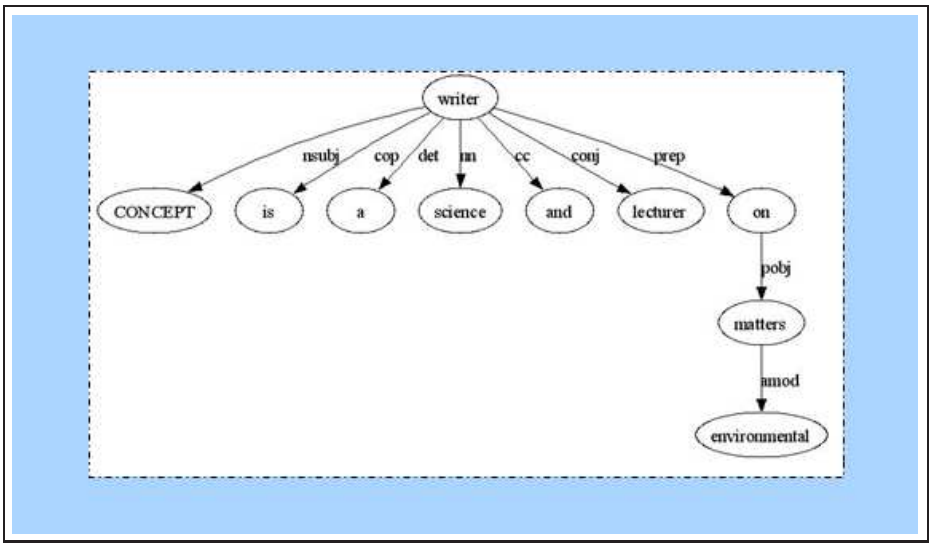

Figure 7.2: Lexicalised dependency tree for the description "CONCEPT *is a science writer and lecturer on environmental matters.*"

illustrative attributes belonging to the previous working sentence are as follows (see figure 7.2):

ROOT→writer→on→matters
START→on→matters→environmental

Tetragrams equip the feature vector with tuples of four ordered and connected words. Analogously to trigrams attributes, these ingredients are taken from lexicalised dependency paths, and accordingly, they are modelled akin to trigrams. An attribute with relation to the working sentence in figure 7.2 is ROOT→writer→on→matters→environmental.

Root to *definiendum* path adds to the feature vector all dependency paths that go from the root node to all nodes carrying the placeholder of the *definiendum*. For instance, figure 7.2 provides the path ROOT→writer→CONCEPT, whereas figure 7.3 produces ROOT→Signature→American→CONCEPT.

***Definiendum*-rooted subtrees** augment the feature vector with two distinct attributes extracted from subtrees rooted at the *definiendum*. The first property conforms to paths from the placeholder to each sub-node. To exemplify, the following paths in figure 7.4 would be included: CONCEPT→lecturer→writer and CONCEPT→lecturer→and. The second component regards extra bigrams in these subtrees such as lecturer→writer.

Root node inserts an attribute into the feature vector symbolising the word that occupies the root node role in the dependency tree (e.g., "RootNode=writer", "RootNode=Signature" and "RootNode=popularize").

Root-*definiendum* distance adds an element "ConceptLevelX=1" into the feature vector, where X denotes the number of nodes from the root to each instance of the *definiendum*.

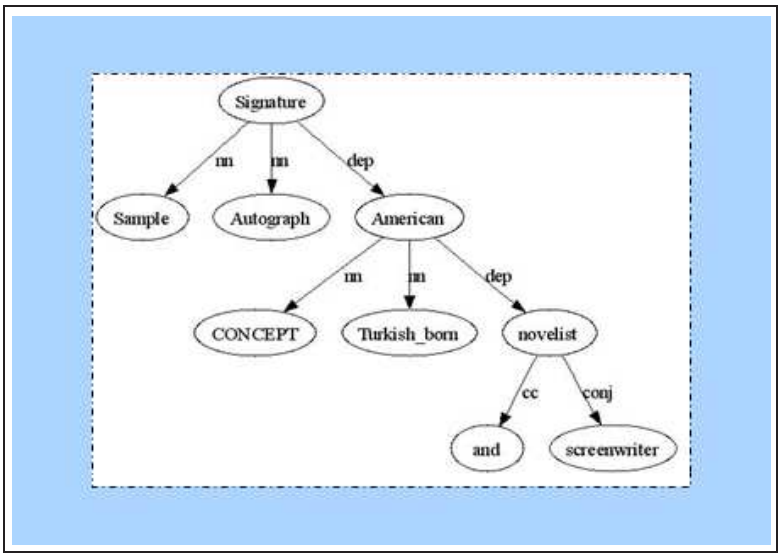

Figure 7.3: Lexicalised dependency tree for the sentence "*Sample Autograph Signature* CONCEPT *Turkish-born American novelist and screenwriter*".

Compactness acts as an indicator of the global properties of the graph. This is an attribute of the form "GraphCompactness=X", where X is worked out in concert with the following equation [Navigli and Lapata, 2007]:

$$Compactness(G) = \frac{Max - \sum_{u \in V} \sum_{v \in V} d(u,v)}{Max - Min}$$

In this formula, V stands for the set of nodes in the dependency graph G. Min stands for the minimum compactness, and it is given by $|V| * (|V| - 1)$, whereas Max denotes the maximum compactness $K*|V|(|V|-1)$. Here, $K = |V|$ and d(u,v) is the length of the shortest path between the nodes u and v.

Children of the root node paths put attributes into the feature vector signalling the path from the root to each of its children, that is the first level of the dependency tree. This level is deemed to convey most of the gist of the sentence. In figure 7.4, this attribute adds: ROOT→popularize→helped and ROOT→popularize→astronomy.

Shallow predicates are similar to the previous attribute, but this also considers the labels of the paths, and it also substitutes the root of the tree with a placeholder. This is aimed essentially at being a shallow predicate analysis of the general structure of the sentence and the idea behind the replacement of the root node is detecting regularities that transpire numerous roots. In figure 7.4, this feature adds: ROOT→**advmod**→astronomy and ROOT→**csubj**→helped.

Syntactic rewritings equip the feature vector with some syntactic inferences about definitions bearing certain characteristics. These inferences are premised on rewritings of copular structures that can also entail descriptions (see for example, figure 7.2). In order to check whether or not the definition observes a copular structure, the following two conditions are checked: (1) the existence of a node connected to the root that observes a **cop** dependency type (e.g., ROOT→writer→**cop**→is in figure 7.2); and (2) the *definiendum* is directly connected to the root node (e.g., ROOT→writer→**nsubj**→CONCEPT in figure 7.2). Note that the relation

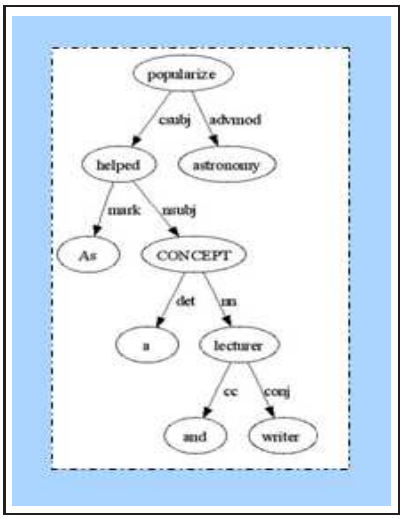

Figure 7.4: Lexicalised dependency tree for the sentence "*As a lecturer and writer* CONCEPT *helped popularize astronomy*".

type **nsubj** could be enforced, but it is not always the case that this is the label given by the parser (see [Marneffe et al., 2006] for an exhaustive list of the possible labels). Whenever both conditions are fulfilled, then the following elements are added:

1. In the event of an **amod** relation between the root node and any of its children, the next feature is added: CONCEPT→<children>.
2. If there is any **nn** relation between the root and one of its children, then the following attributes are added: CONCEPT→<children> and <children>→CONCEPT. For instance, figure 7.2 yields the path ROOT→writer→**nn**→science. In this case, the next two paths are inserted into the feature vector: CONCEPT→science and science→CONCEPT.
3. Invert the root node with the placeholder of the *definiendum*. To illustrate, the path in figure 7.2 writer→CONCEPT implies CONCEPT→writer. This, for example, would lead to augmenting the similarity to the following description: "*American writer* CONCEPT *published his first novel X in 1978*".

Root to semantic classes paths enrich the feature vector with the paths from the root to every placeholder pertaining to a semantic class. These semantic classes were annotated in concert with SuperSense Tagger [Ciaramita and Altun, 2006][5].

Root-definition verb distance is a property conforming to "VerbLevel_Y=X", where Y coheres to the position of a definition verb in the sentence, and X coheres with the distance of the respective node to the root of the dependency tree. It is worth recalling that definition verbs are those replaced with a placeholder (VERB), when performing selected substitutions. At any rate, this element can be partially used with these replacements together or independently.

Root position inserts an attribute of the form "RootPosition=X" into the feature vector, where X coincides with the position of the root node in the sentence. For example, in the sentence shown in figure 7.4, the corresponding attribute is "RootPosition=8".

[5] http://web.net/projects/supersensetag

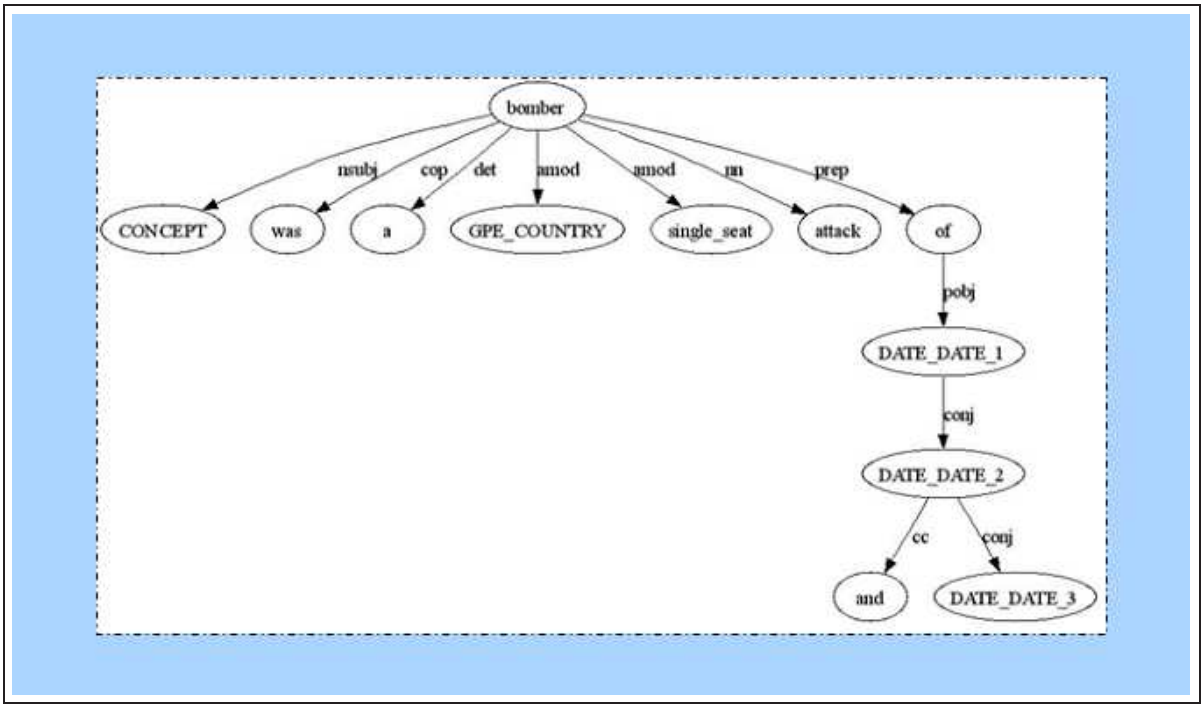

Figure 7.5: Lexicalised dependency tree for the sentence "CONCEPT *was a* GPE:COUNTRY *single-seat attack bomber of* DATE:DATE, DATE:DATE *and* DATE:DATE".

Betweenness is based on a simplification of the notion explicated by [Navigli and Lapata, 2007]. This enriches the feature vector with an attribute of the shape "B_Y=X", where Y and X stand for a word within the sentence and the amount of different directed paths between two distinct nodes that goes through it, respectively. To neatly illustrate, consider the word "*novelist*" within the sentence in figure 7.3. This example includes four different directed paths going through the term "*novelist*":

Signature→American→novelist→and
Signature→American→novelist→screenwriter
American→novelist→and
American→novelist→screenwriter

More exactly, the resulting property is as follows: "B_novelist=4". It is worth underlining that only paths that do not start or end with Y are taken into account when counting X. In short, more crucial nodes in the dependency graph are assumed to be involved in a higher amount of paths.

Shallow predicate II (only labels) is an attribute that concatenates the root node with the sorts of relations with all its children. In this concatenation, the order of the labels is preserved. For instance, the resulting features with relation to the sentences in figures 7.2 and 7.4 are: "writer_nsubj_cop_det_nn_cc_conj_prep" and "Signature_nn_nn_dep". For a list of the potential types, the reader can refer to [Marneffe et al., 2006].

Verb Positions II incorporates positional attributes for definition verbs. This element is of the form "VerbPosition2=X", where X is the position in the sentence. An NLP-free version of this ingredient can be obtained by means of a tokeniser based on simple regular expressions,

Features (without NLP)	Accuracy	Features (with NLP)	Accuracy
unigrams	78.36%	bigrams	78.67%
+definition patterns	78.87%	+unigrams	81.48%
+verb positions II	79.11%	+shallow predicates	82.25%
+number of tokens	79.52%	+number of tokens	82.58%
+determiner	79.56%	+shallow predicates II	82.86%
		+*definiendum* rooted subtrees	83.08%
		+root node	83.20%

Table 7.17: Best feature configurations for Set A (with/without NLP-based attributes).

namely, the occurrences of blank spaces within the sentence.

7.6.6 Results: Features Extracted from Lexicalised Dependency Graphs

By and large, the previous models integrate various lexical, semantic and syntactic features; in particular, syntactic attributes extracted from the structures produced by chunking. However, an alternative source of syntactic knowledge is dependency trees, which, in essence, produce a labelled graph representation of a sentence. Principally, the succeeding experiments explore the effects of features coming from this graph representation of the training and evaluation corpora. Note that the following assessments disregard attributes from chunking, because preliminary experiments showed that simple models equipped with bigram features from dependency graphs outperform the best configurations designed in the previous evaluation, along with the memory constraints imposed by the servers where the experiments were carried out.

Another vital difference between the preceding and the following assessments regards the way attributes are selected. In the previous study, the combinations of features were manually chosen, whereas in the following experiments, the configurations are automatically formed via the greedy selection algorithm proposed by [Surdeanu et al., 2008]. The reason for this change lies in the substantial growth in the number of properties, making a manual procedure more time consuming, whilst an automatic procedure would naturally be preferable.

SELECTION ALGORITHM

To be more specific, this procedure incrementally selects features without pre-assuming any of them as fixed. At each iteration, this algorithm tests all attributes individually, and singles out the one that crystallises or gives shape the higher increment in performance. Once a property is chosen, it is kept fixed, while each of the remaining features is tried in conjunction with the fixed attributes. When no addition brings about an enhancement, the selection procedure stops, and the set of fixed properties is returned as the best set.

As a matter of fact, there are two extra remarks here: (a) computationally speaking, this selection process demands enormous resources, thus accuracy was the only metric considered when seeking the optimal group of attributes (potentially, one could also optimise for *precision at* k); and (b) the figures in relation to the same configuration may slightly vary with respect to the previous study, because surface features were inferred from the tokanisation returned by the dependency parser instead of the previous output given by OpenNLP. In substance, the underlying idea is attempting to account solely for the linguistic information yielded by one tool, because processing each sentence with each NLP tool signifies a rise in the processing time and in terms of memory usage (due to the distinct views of the training material). This is also a technical reason why features deduced from chunking were not included in this evaluation.

Features (without NLP)	Accuracy	Features (with NLP)	Accuracy
unigrams	60.52%	unigrams	60.52%
+number of tokens	60.76%	+semantic classes	61.18%
+selective substitutions	60.94%	+selective substitutions	62.09%

Table 7.18: Best feature configurations for Set B (with/without NLP-based attributes).

In addition to Set A and B, this new assessment capitalises on three extra arrays of testing sentences: A', TREC 2003 and TREC 2004. The first in-domain group was constructed analogously to Set A, but for a different battery of *definienda*. This pack is comprised of 5,982 sentences: 2,888 positive and 3,094 negative. Contrarily, both out-of-domain TREC sets were acquired from the AQUAINT corpus as follows. In the first place, this collection was indexed by Lucene, and the top one hundred documents were fetched by querying the *definiendum* to this IR engine. In the second place, these retrieved documents were split into paragraphs in conformity with the structure provided by the documents in the collection. In the third place, paragraphs containing a query term, excluding stop-words, were singled out, and co-references were accordingly resolved within these chosen paragraphs by means of JavaRap. In the fourth place, each selected paragraph is divided into sentences in concert with OpenNLP. Last, sentences carrying a query term, excluding stop-words, are returned as answer candidates. At this point, duplicates were eliminated by means of string comparisons, and some samples were missed because the dependency parser did not produce an output. Subsequently, sentences were manually annotated in consonance with the TREC ground truth. In detail, the TREC 2003 testing set encompassed 7,968 sentences: 1,091 positive and 6,877 negative, whereas TREC 2004 consisted of 10,565 instances: 1,011 positive and 9,554 negative.

SETS A' AND TREC 2003/2004

Tables 7.17 to 7.21 underline the group of features discovered for each array of testing sentences. Each table is compounded of two columns: one listing the set of properties obtained when considering all features, and the other column when taking advantage solely of attributes at the surface level. For the sake of clarity, by surface features, it is meant those properties that do not exploit dependency graphs. Although, for reasons already stated, the respective tokenisation was used for extracting these attributes (a very lightweight tokeniser/heuristic could be a proper replacement).

SURFACE FEATURES

A bird's eye view of the figures considering NLP features points out three attributes that seem to be the most salient, hence discriminative, ones: unigrams, trigrams and the root node. To put it more accurately, these features were selected in three distinct configurations. Even though all three were not chosen together, the three potential pairs did appear in at least one set. Roughly speaking, one could conceive the triplet determined for Set A (i.e., unigrams, bigrams and root node) as an approximation of the combination of these three essential properties. The bottom line is that these are the most instrumental attributes, that is to say, they are the most powerful (of the list of tested features) for separating the wheat from the chaff. Further, granted that lexical features are still picked together with NLP-oriented attributes, it can be concluded that these properties remain suitable, presumably, for enhancing the recognition of some borderline examples, ergo for tackling data sparseness.

HIGH IMPACT ATTRIBUTES

In a special manner, an examination of the most frequent values for the attribute root note across the positive class reveals the following frequency distributions: `CONCEPT`(22,465), *born* (9,562), *known* (8,049), *played* (6,944), *one* (6,842), *is* (6,404), *has* (6,261), *made* (5,488), *was* (5,453), and *had* (5,305). In the opposite category, one can observe the next frequencies: `CONCEPT`(58,967), *is* (7,517), *said* (7,466), *are* (3,376), *was* (3,047), *found* (2,888), *have* (2,881), *had* (2,612), *made* (2,504), and *include* (2,478). Simply put, half of the top ten highest frequent val-

ROOT NODE

Features (without NLP)	Accuracy	Features (with NLP)	Accuracy
unigrams	79.12%	bigrams	82.68%
+concept positions	80.89%	+shallow predicates	84.72%
+number of tokens	81.13%	+unigrams	85.64%
+definition patterns	81.83%	+trigrams	85.87%
+verb positions II	81.85%	+definition patterns	85.99%
+determiner	81.86%	+number of tokens	86.14%
		+root position	86.38%
		+shallow predicates II	86.48%
		+syntactic rewritings	86.53%
		+compactness	86.64%
		+root to definiendum distance	86.71%
		+determiner	86.74%

Table 7.19: Best feature configurations for Set A′ (with/without NLP-based attributes).

ues for this attribute permeate both classes, but their frequency distributions substantially differ. In effect, since both sets are balanced in terms of the number of examples, this difference proves that the negative group is more diverse in nature. As a means of verifying this, the number of distinct values for this attribute was counted: 84,387 and 32,667 for the negative and positive class, respectively. On top of that, values such as "*said*" and "*include*" can be found in the negative category, ratifying the observation of [Schlaefer et al., 2006] (see section 4.5 on page 87).

TRIGRAMS

More important, the five values that show up in both categories can be utilised for examining trigrams that typify each group. Interestingly enough, some regularities emerge as a result of this inspection:

- Some prominent trigrams subsumed in the negative class that appear in tandem with a value of "*was*" for the root node element are listed below:

  ROOT→was→at/in/by/with→CONCEPT
  ROOT→was→CONCEPT→Between
  START→Link→on→page
  ROOT→was→said→CONCEPT
  START→accused→of→murdering

  These paths show prepositional phrases compounded by the placeholder of their respective training concepts. Accordingly, one can deem that this regularity symbolises sentences that do not convey a description, in general, especially when this is the sole instance of the placeholder within the sentence. To illustrate, consider the next examples:

  Tony Macmillan was at CONCEPT[RAF Bampton Castle] between 1979 and 1989.
  That post was by CONCEPT[Alex Tabarrok] . posted by : Eric Slusser on 04.21.05 at 08:39 AM

  In truth, these paths can also be utilised for expressing descriptive information, normally in the form of a brief description (one to three words) embodied in the same

Features (without NLP)	Accuracy	Features (with NLP)	Accuracy
lowercased terms	47.25%	betweenness	53.21%
+selective substitutions	56.06%	+trigrams	55.11%
+definition patterns	57.53%	+root node	56.23%
+number of tokens	60.26%	+tetragrams	56.73%
		+definition patterns	56.98%
		+compactness	57.38%
		+root position	57.87%

Table 7.20: Best feature configurations for TREC 2003 (with/without NLP-oriented attributes).

prepositional phrase (e.g., in newspapers). To exemplify, one can think about the case of descriptive adjectives anteceding the placeholder. When the placeholder is isolated in the prepositional phrase, it is more probable to materialise non-descriptive contexts or a sentence that talks about a different topic. Additionally, the second example suggests that blogs/forums are potential sources of negative examples as bloggers typically rename themselves after people they admire, or they just simply bear the same name (also check the introductory sample). Still yet, definitions can be found across this sort of web document. Other trigrams also signal some interesting findings, for instance, the expressions "*Link on page*" and "*accused of murdering*" are strong indicators of negative examples. Note that these paths do not start at the root of the lexicalised dependency tree. In the positive category, trigrams including: ROOT→was→in→charge, ROOT→was→role→CONCEPT, ROOT→was→known→for can be found.

PREPOSITIONAL PHRASE

FORUMS

- In juxtaposition to the value "*was*", it can be observed that many genuine descriptions automatically labelled as negative are embodied in the value "*made*". Take, for instance, the sentence below and the next group of trigrams that are prevalent in a setting with root node equal to this value: ROOT→made→use→of, ROOT→made→possible→by, ROOT→made→by/of/to/with→CONCEPT.

  Some of the most dramatic improvements in field methods were made by CONCEPT[John Bell Hatcher], who started working for Marsh in 1884 and would later gain fame collecting for Andrew Carnegie .

  These properties are exploited in conjunction with other attributes that might be more delineative of the positive category, therefore decreasing the impact of these misannotations.

A second tier of features is given by those showing up in two distinct sets: bigrams, shallow predicates, number of tokens, shallow predicates II, definitions patterns, root position, and betweenness. Interestingly enough, two groups of glued (correlating) attributes can be distinguished in this group. There is a first array of four properties that are conspicuous in in-domain testing sets: bigrams, shallow predicates, number of tokens and shallow predicates II. A second group of two features that are visible in both in-domain and out-of-domain: root position and definition patterns. This cohesive set of properties reveals that: (a) definition patterns and the root position are discriminate features in the event of in-domain, but more

DEFINITION PATTERNS

Features (without NLP)	Accuracy	Features (with NLP)	Accuracy
lowercased terms	44.26%	betweenness	50.62%
+selective substitutions	54.21%	+root node	53.21%
+definition patterns	55.47%	+trigrams	54.62%
+number of tokens	57.14%		
+determiner	57.25%		

Table 7.21: Best feature configurations for TREC 2004 (with/without NLP-based attributes).

fundamental, (b) they are pertinent for tackling sentences taken from the TREC corpus. This last observation coheres with the utilisation of these rules by the vast majority of systems taking part into the definition subtask of the TREC challenge (see section 4.2 on page 76).

BETWEENNESS Another interesting finding is the portability of the attribute: betweenness. Plainly speaking, the value of this property gets higher in conformity to the amount of paths that go through the corresponding node (word). That is to say, nodes with a high value can be perceived as the neural centre of the sentence as they are indispensable to establish a relation between numerous pairs of words contained therein. On a side note, it is worth recalling here the Shortest Path Principle [Bunescu and Mooney, 2005]. From another viewpoint, regularities in terms of pairs, encompassing both a word and its numeric value, trigger or encode a synthesisation of the shallow syntactic and semantic properties of the respective nodes. Hence, one can affirm that these shallow semantic and syntactic structures port to out-of-domain data. On a different note, two additional observations are: (a) the essentiality of the number of words can be deduced from the configurations obtained for both with and without NLP; and (b) in the event of Set B, both settings differiantiate in a sole property: semantic classes were replaced by the amount of tokens, bringing about a drop in performance from 62.09% to 60.94%. In brief, no syntax ported to Set B, whereas some syntactic knowledge did it to the testing sets emanated from the AQUAINT corpus.

To return to the point of the betweenness, there is a salient triplet of properties discovered for TREC testing sentences: trigrams, root node, and betweenness. The relation between root node and trigrams was dissected earlier. The connection of the root node to the betweenness of some sentence level terms also produces very interesting findings:

1. Broadly speaking, sharp differences were observed between pairs of values between both categories. Some conspicuous regularities across the positive category are:

   (a) In many cases, when the root note was equalised to sorts of persons, the word "*born*" gets the value of six for the betweenness: *actor* (793), *player* (751), *politician* (568), *actress* (439), *footballer* (299), *singer* (271), *author* (258), *writer* (234), *runner* (134), *journalist* (114) and *member* (99). Note that the number in parentheses indicates the frequency of the respective combination across the positive set. These co-occurrence patterns were not significantly observed in the negative class. Other noticeable values for betweenness that materialise this kind of distribution are: 10 and 12. To reinforce the idea behind this property, consider:

   [Betweenness:6] ⇒CONCEPT[John Boehner] (pronounced "Bay-ner"), born November 17 1949, is an American **politician** of the Republican Party who serves as House Minority Leader in the 110th Congress, and a U.S. Representative from, which includes parts of the city of Dayton as well as several of CONCEPT's suburbs.
   [Betweenness:6] ⇒CONCEPT[David Duke] (born July 1, 1950) is a former Republican **member** of the Louisiana House of Representatives, a candidate in presidential

primaries for both the Democratic and Republican parties, and former Grand Wizard of the Knights of the Ku Klux Klan.
[Betweenness:10] ⇒CONCEPT[Christian Prudhomme] (born November 11, 1960 in France) is a sports **journalist** and general director of the Tour de France since 2005.
[Betweenness:12] ⇒CONCEPT[David Gerrold], born Jerrold David Friedman (January 24, 1944), is an award-winning science fiction **author** who started CONCEPT's career in 1966 as a college student by submitting an unsolicited story outline for the television series Star Trek .

In short, these examples solidify how the betweenness encapsulates some essential syntactic regularities. A case in point, one of the factors that increases the betweenness is the amount of nodes in the sub-tree rooted at the word, for which the betweenness is calculated. Ergo, a very frequent number of nodes (words) in the sub-tree signifies a characteristic (shallow) syntactic structure. One can notice, consequently, that a value of six symbolise succinct dates (e.g., "*born November 17 1949*"), while a value of 10 more lengthy information (e.g., "*born November 11, 1960 in France*"), and a value of 12 implies a more wordy nugget (e.g., "*born Jerrold David Friedman (January 24, 1944)*").

(b) In the same way, the word "*based*" strongly correlates with values of three, when at the same time, the value of the root node property is directed at organisations or group of people. One can empirically observe the following frequency distributions: *team* (85), *club* (68), *company* (62), *band* (44), *airline* (44) and *newspaper* (23). Also, in this case, a value of four for the betweenness is very prominent. Good examples of this are the following:

[Betweenness:3] ⇒CONCEPT[The Christian Post] is a pan-denominational, Evangelical Christian **newspaper** based in Washington, D.C.
[Betweenness:4] ⇒CONCEPT[SC Paderborn 07] is a German football **club** based in Paderborn, North Rhine-Westphalia and currently playing in the Second Bundesliga.

To sum up, these examples signal that the shallow syntactic information encapsulated by the betweenness feature is also applicable to other semantic relationships.

(c) From another angle, with a fixed value for the root note, some contextual regularities occur with fixed values for the betweenness. Consider the case of "*species*" and a value of eight: *family* (1836), *bird* (554), *plant* (358), *fish* (104), *legume* (97), *frog* (93), *native* (85), *toad* (74), *rodent* (53) and *conifer* (50). Note that the numbers in parentheses represent the corresponding frequencies across the positive training material. Some illustrative samples are given below:

The CONCEPT (Pyrgulopsis trivialis) is a **species** of gastropod in the Hydrobiidae family.
The Hairy Acacia (CONCEPT) is a **species** of **legume** in the Fabaceae **family**.

(d) More relevant are the words "*single*" and "*song*" which share a relation to the word "*album*" put together with the following array of values: 6, 9, and 12. However, a more essential characteristic arises when examining the terms such as "*written*" and "*released*". The first one is connected with "*song*" and the other with "*single*". Both pairs bear the prominent values of three and six, showing their syntactic and semantic similarities across definitions:

[Betweenness:3] ⇒CONCEPT[You Enjoy Myself], known in short as YEM by Phisheads, is a Phish **song written** by Trey Anastasio.

[Betweenness:6] ⇒In 1962 CONCEPT[James Cecil Dickens] **released** " The Violet and the Rose," CONCEPT's first top ten **single** in twelve years.
[Betweenness:6] ⇒CONCEPT[The Perfect Kiss] was the first New Order **song** to be included on a studio **album** at the same time as CONCEPT's release as a single.
[Betweenness:9] ⇒CONCEPT[Rocky Mountain Way] is a 1973 **song** by rock guitarist Joe Walsh and also a 1985 compilation **album** by Walsh which features the song.
[Betweenness:12] ⇒CONCEPT from the **album** The Masterpiece Debut **single** from London rapper Nathan.

(e) Superlatives, including "*largest*", "*smallest*" and "*worst*", as well as "*best*", normally exhibit a betweenness of zero across both categories, but connected with different values of the root node, and markedly different frequencies. In the event of adjectives, like "*big*" and "*large*", the same observation holds. This signals the inherent relation between both properties.

2. Essentially, this alliance between the values of the root note and the betweenness of some crucial words is weaker or more disperse in the negative class. Some relatively predominant tuples include: "*found*" with *people* (3) and *helpful* (3), "*come*" with *work* (14) and *sources* (3), and "*feel*" with *article* (4). Note that numbers in parentheses refer to the betweenness and the respective frequencies of these tuples range from 600 to 1,132.

3. Evidently, some root nodes only present noisy relations, from which some can be attributed to data sparseness.

Type of Relations	Shallow Predicates
nn	CONCEPT (5,481), football (3,531), rock (1,640), film (1,640).
partmod	known (2,922), located (2,573), born (1,710), playing (1,365), used (1,255), written (1,122), released (1,029), directed (945), based (826), working (765), starring (737), published (729), serving (715), founded (678), formed (678), making (634).
dobj	CONCEPT (5,186), career (2,330), debut (1,247), number (1,043).
rcmod	played (1,571), plays (1,389), competed (545), served (503), known (463), won (425), born (351), worked (350), appeared (313).

Table 7.22: Most prominent shallow predicates with respect to some interesting dependency types (positive class).

SHALLOW PREDICATES AND BIGRAMS

In addition, another combination of properties that was found to be fruitful for recognising descriptions consists of bigrams and shallow predicates. The former encodes local lexico-syntactic relations between pairs of words, while the latter attempts to capture some regularities across second level semantic relationships that are found across distinct root nodes (first level). Tables 7.22 and 7.23 parallel some prevalent features extracted by shallow predicates in agreement with four types of dependency relations "*nn*", "*partmod*", "*dobj*" and "*rcmod*". From the point of view of definition QA systems, these four relations seem to be the most interesting as they carry some semantic meaning. Above all, the contrast shows that the distributions in both categories are clearly different. Still yet, one can observe the term "*playing*" in both classes and carrying the same relation type. Distinctly, this term is useful for understanding how bigram features interact with shallow predicates. To be more precise, the shallow predicate ROOT→partmod→playing simultaneously manifests with the group of bigrams determined by START→playing→?, where the question mark stands for

five prepositions "*for*", "*with*", "*in*" and "*at*" as well as "*on*". This last observation holds for both categories, but in the positive class, one can also find some prominent relationships with "*currently/professionally/mainly*", "*as/from/before*", "*guitar/piano/music*" and "*role*" as well as "*baseball/football*". In the opposite class, one cannot notice predominant substitutes for the question mark other than the prepositions listed earlier. In light of this experimental observation, one can understand the synergy between both features when they are synthesised as a mean to distinguish descriptions.

Type of Relations	Shallow Predicates
nn	`CONCEPT` (13,084), Search (1,250), Home (1,139), New (875).
partmod	blog (825), using (378), featuring (285), making (203), playing (168), working (144), related (131), showing (119).
dobj	`CONCEPT` (14,051), material (982), it (854), book (831).
rcmod	`CONCEPT` (503), feature (265), include (265), to (253), used (146), is (120), Buy (94), made (92), said (81).

Table 7.23: Most prominent shallow predicates with respect to some interesting dependency types (negative class).

As for NLP-free configurations, eight attributes proved to be useful, in particular the number of tokens showed to be the most instrumental. This feature is embodied in all NLP-free settings, and further, it was chosen in conjunction with NLP-based properties. From another standpoint, this attribute indicates versatility as it was incorporated into in and out-of-domain configurations. Figure 7.6 plots the distribution of the value of this property in both categories. Broadly speaking, this graph reveals that short sentences are more likely to be descriptive, while longer non-descriptive, 35 tokens being the amount for the turning point. Another vital attribute to be emphasised is the determiner, which was singled out for four out of the five testing cases. In actuality, [Xu et al., 2005] also benefited from two features homologous to the determiner and the amount of tokens (see section 7.3). Given this evidence, it can be concluded that both properties are instrumental for rating definitions.

NUMBER OF TOKENS

DETERMINER

Incidentally, when features in Set A, A′ and B are compared with attributes in both TREC sets, it can be noticed that the first group capitalised on all lexical items, while TREC sets solely on lowercased[6] items. On this account, one can affirm that these uppercased items are essential for identifying descriptions across web pages, but they do not port to news articles. One can consequently postulate that capitalised words collected from the Internet (e.g., "*Search*", "*Home*" and "*Link*") do not significantly occur across news articles, ergo being of little use when tackling the AQUAINT corpus.

NLP-FREE PROPERTIES

It can be further observed that definition patterns aided in the same four sets as the determiner, thus denoting that both features are complementary. Furthermore, the positions of the placeholder of the *definiendum* and describing verbs showed to be of less relevance, and selective substitutions were fundamental only in out-of-domain sets. Substantially, these replacements share the same spirit with the semantic classes, which were useful solely for Set B. Simply put, the best NLP-free features coincide with the attributes typically exploited by QA systems in the definition subtask of the TREC challenge.

SELECTIVE SUBSTITUTIONS

At any rate, while strong relations between properties, found across the distinct testing sets, indicate meaningful outcomes, it is also true that the final figures accomplished by those models are not. The underlying reason for this is that the greedy algorithm profits

[6]For the sake of simplicity, by uppercased/lowercased terms it is meant tokens that start with a capital/lowercase letter.

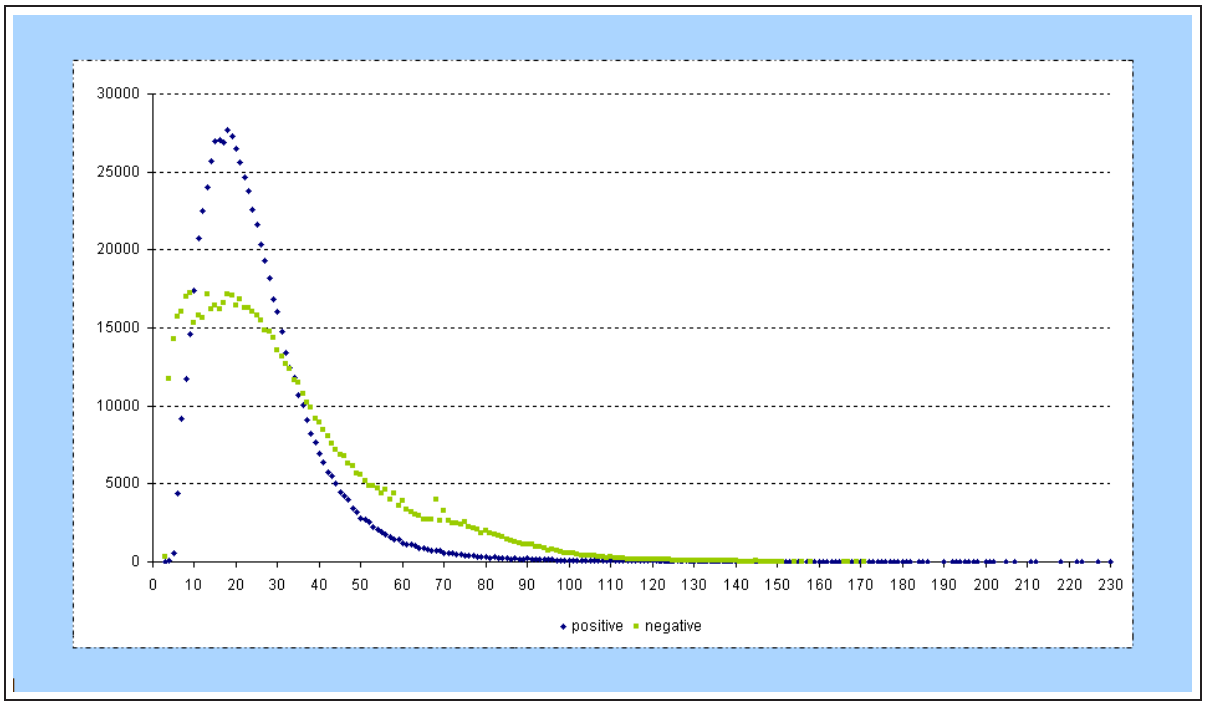

Figure 7.6: Number of samples vs. Number of tokens (reference: training material).

from the labels when automatically selecting the most propitious properties. In a realistic scenario, this does not happen. As a means of drawing meaningful conclusions in terms of quantifying performance, additional experiments must be carried out. For this purpose, the group of features automatically determined for one set was employed to rate testing instances embodied in the other four sets. This concerns both types of arrays of attributes, that is with and without NLP-based properties.

PERFORMANCE: SET A AND A'

In general terms, the figures are quite motivating (see table 7.24) both configurations acquired for Set A (i.e., with/without NLP) finished with the best outcome when tackling Set A'. The contrary also holds: the best performance for Set A was reached by means of the properties obtained for Set A'. Additionally, the features stipulated for the TREC 2003 (with NLP) accomplished a competitive performance for both Set A and A'. Interestingly enough, the accuracy declined about 5% with respect to the top-scores when profiting solely from NLP-free attributes. In light of this gap, one can quantify the trade-off between performance and computational speed.

PERFORMANCE: OUT-OF-DOMAIN

On the other hand, one can find the diametrical opposite. NLP-free models finished with the best results in all three out-of- domain cases (see table 7.25). In the case of Set B, there is a slight variation in accuracy between NLP-free and NLP-oriented configurations. Conversely, a more noticeable detriment was manifested in the case of TREC 2003, more precisely, almost 4%. As a rule, analogously to the in-domain arrays of sentences, the best outcomes are substantiated by the group of attributes specified for the counterpart group of instances (i.e., TREC 2003 $\leftrightarrow$ TREC 2004). This entails the dissonance amongst the lexico-syntactic properties of the testing sets, and it can also be corroborated by contrasting the different lists of features. All in all, the five models combined took advantage of about twenty of the proposed features, thus signifying the effectiveness of these properties. Of course, unselected attributes might aid in coping with testing sentences of another nature.

For the most part, an inspection of the outcomes with respect to the TREC sets unveils that errors chiefly emanated from negative instances tagged as positive. In this respect, the

Best Attributes Found for	Applied to Set A′ Accuracy	Set A Accuracy
Set A without NLP	81.16%	
Set A with NLP	**85.94%**	
Set A′ without NLP		78.28%
Set A′ with NLP		**83.04%**
Set B without NLP	76.86%	75.50%
Set B with NLP	63.19%	74.61%
TREC 2003 without NLP	71.63%	69.22%
TREC 2003 with NLP	79.89%	76.29%
TREC 2004 without NLP	71.80%	69.18%
TREC 2004 with NLP	78.25%	74.77%
Baseline I (0.3)	54.13%	44.17%
Baseline I (0.2)	54.63%	44.23%
Baseline I (0.1)	56.25%	44.77%

Table 7.24: Accuracy obtained when applying the arrays of best features to in-domain sentences: Set A and A′.

confusion matrix reveals that 21.88% of the instances were negative perceived as positive by the models, 6.33% were positives seen as negative, and 11.60% were unlabelled testing examples. These numbers regard the best outcome for the TREC 2003 set, and unlabelled samples refer to those that were attributed an equal probability for both categories. A sound and plausible reason for the large amount of false positives is given by [Hildebrandt et al., 2004]: many good descriptive sentences can still be outputted by the system, but since they are not embraced in the ground truth, they worsen the performance. By all means, the quantification of this diminishment relies on the criteria of the assessor(s) criteria for singling out the true positives.

Impact of TREC Ground Truth

As to TREC 2004, one can observe in table 7.25 that the figures consistently lowered when applying the same model to both TREC sets. A reason for this consistent worsening is given by the context questions introduced in the 2004 subtask. These queries consist of a sequence of enquiries about a *definiendum*, and their types embrace factoid and list. As [Han et al., 2006] mentioned, the truth of the matter is that the answers to context questions can also be perceived as part of the ground truth of the corresponding definition query.

In practice, the critical issue here is that definition QA systems are compelled to eliminate answers to the other context queries from the final response to the definition question. For this reason, they are left unconsidered from the pertaining ground truth. Since it is unfair to account for the right answers in the gold standard because TREC systems do not know them, the responses to these context queries were not taken into consideration in the assessment. Categorically, the influence of these answers is certainly detrimental as some of these nuggets are identified (tagged as positive by the models), but their corresponding annotations in the ground truth is negative. In this regard, the confusion matrix confirms this situation: 25.33% negative testing examples where labelled as positive (an increment of 3.45% in relation to the TREC 2003), while 4.62% positives labelled as negative (a diminution of 1.71%). Here, it is also worth highlighting that 12.89% sentences were left unlabelled (a growth of 1.29%), that is, both classes were assigned the same probability.

Baseline-I

With respect to BASELINE-I, this system slightly outperformed the best configurations for the TREC 2003 and TREC 2004 by 0.71% and 0.99%, respectively. On the surface, these figures

Best Attributes Found for	Applied to Set B Accuracy	TREC 2003 Accuracy	TREC 2004 Accuracy
Set A without NLP	59.86%	47.53%	44.26%
Set A with NLP	58.71%	49.57%	49.27%
Set A' without NLP	58.25%	35.79%	30.53%
Set A' with NLP	58.44%	51.64%	50.00%
Set B without NLP		47.50%	44.69%
Set B with NLP		45.68%	40.89%
TREC 2003 without NLP	**60.09%**		57.15%
TREC 2003 with NLP	59.11%		56.04%
TREC 2004 without NLP	60.05%	60.14%	
TREC 2004 with NLP	59.35%	56.24%	
Baseline I (0.3)	56.26%	**60.85%**	**58.14%**
Baseline I (0.2)	57.77%	51.69%	46.34%
Baseline I (0.1)	56.73%	42.20%	37.28%

Table 7.25: Accuracy obtained when applying the arrays of best features to out-of-domain sentences: Set B and TREC 2003 as well as TREC 2004.

appear to be modest outcomes, but one has to keep in mind that both testing sets are out-of-domain. More specifically, this implies that these results could substantially be improved by replacing the training material with some corpus acquired from the AQUAINT collection. Certainly, this will also involve the determination of the respective arrays of properties afterwards.

7.7 Conclusions and Further Work

Generally speaking, this chapter discusses discriminant models for scoring answers to definition questions. There are four major challenges posed by the construction of this class of ranker: answer granularity, corpus acquisition, feature selection and the learning machinery. In deed, each of these factors plays a pivotal role in the performance of answer extractors predicated on this sort of approach.

In the first place, this chapter stresses a strategy designed to tag text fragments of 250 characters in length centred at the *definiendum* as a definition or general text. This ranker is premised on a SVM with surface features, and a manually annotated training material coming from the Internet. Later, this technique was revisited, and the manual annotation was substituted with an automatic labelling of positive and negative samples. These annotations conform to the similarity and dissimilarity of text fragments to the battery of descriptions encircled by KBs about the pertaining *definiendum*. Their list of features include: term correlation, hypernym, n-gram phrasal regularities, definition patterns and the ranking returned by the IR engine.

In the second place, this chapter touches on a methodology that capitalises on two distinct SVM rankers for rating answer candidates to definition queries on the technical domain. Notably, this study analyses two distinct levels of answer granularity: paragraph and sentence. More accurately, their putative answers match a pre-specified array of definition patterns, and in the event of paragraphs, the first sentence must align these patterns. In a special manner, in some configurations, this approach benefits from three labels for the train-

ing instances: positive, negative and indifferent. As for the features of their classifiers, this method profits from determiners, pronouns, patterns, number of sentences, words and adjectives, and frequent words after the *definiendum* in a window of fixed length. Their figures showed three classes tend to enhance the quality of the final output, and their outcomes additionally suggest that rating sentences is a much more appropriate level of granularity than paragraphs, and that adjectives are inclined to hurt the performance.

In the third place, this chapter expands on the influence of various discriminative learning models on the ranking of copula sentences in Dutch. In this study, the primary goal is to discern definitions in the medical domain. For this reason, they harvested the medical index of Wikipedia, and manually annotated some samples. These models accounted for assorted properties including bag-of-words, bigrams, root forms, named entities in the subject, the position in the document, and some syntactic features derived from the dependency tree representation of the sentences in the corpus. Their figures show that Naïve Bayes and ME produced the best results. As to the ingredients of these classifiers, they found out that root forms did not materialise an improvement, while named entities and an indicator of the position of the sentence did.

Lastly, this chapter turns to an approach targeted at recognising open-domain pattern-independent answers (sentences) to definition questions. To put it more precisely, this strategy scores putative answers by means of ME models built on top of a host of automatically annotated examples. This corpus was obtained by balancing positive examples originated from Wikipedia and negative instances from the Internet. An exciting aspect of this study regards an exhaustive inspection of numerous features dealing with testing sets of various natures. This analysis is divided into two parts:

1. The first part examines the performance of a set of configurations formed by attributes derived from chunking, named entities, POS tags, and other surface properties. Many configurations were tried as a mean to rank answer candidates in two distinct groups: one in-domain and one out-of-domain. The former bears strong lexico-syntactic similarities to the training material, while the later is quite dissimilar. In terms of features, some findings are as follows:

   (a) Given the fact that the absence of words that start with a capital letter causes a steep decline in performance, one can conclude that they are pivotal for building efficient discriminant models.

   (b) By the same token, experiments reveal that selective substitutions work well in accompanying syntactic knowledge deduced from chunked sentences. These replacements help to ameliorate portability and to combat data sparseness by generalising chunks.

   (c) Syntactic information is fruitful for rating answer candidates that share the same lexico-syntactic properties with the training data.

   (d) Oppositely, semantic classes port to out-of-domain candidates, whereas syntactic information did not manifest an improvement when tackling out-of-domain candidate answers.

   (e) From the various kinds of named entities, the replacement of dates by a placeholder is the most beneficial. Presumably, this substitution allays the ambiguity of numbers.

   (f) When scoring out-of-domain putative answers, the accuracy systematically deteriorated in tandem with the length of the candidates. To state it more exactly, the models failed to correctly recognise out-of-domain sentences with only a small descriptive portion.

(g) Unexpectedly, definition patterns had a greater impact on the recognition of answers that mismatch these regularities than on those providing alignment.

(h) In terms of ranking order, experiments indicate that syntactic knowledge betters the precision of the top positions of the ranking. More crucial is the fact that this was observed in both kinds of sets, but for the out-of-domain test set, only in a very slight way.

2. The second part of this study investigates the effects in ME models of numerous features emanating from the surface and the dependency tree view of the corpus. Contrary to the first part, this analysis profits from a greedy algorithm for automatically configuring the models. In a nutshell, this automatic configuration consists in systematically singling out the more efficient attributes. This study also extends the amount of arrays of sentences to five, which are utilised for assessing the impact of these features (two in-domain and three out-of-domain). A summary of the main conclusions is as follows:

(a) Overall, the obtained models show that three attributes proven to be the most instrumental: unigrams, trigrams and the root note. Further, the models unveil that these features are supplementary, that is they can be members of the same configuration. Note that the repeated use of unigrams is apparently a product of data sparseness.

(b) The values exhibited by the root node indicate considerably different frequency distributions across both positive and negative sets; therein lies the importance of this attribute. Plainly speaking, the node root often yields a substantial part of the semantics of the sentence.

(c) Exceptionally, if the *definiendum* is contained isolated in a prepositional phrase and the semantic of the sentence is given by a root node equal to "*was*", then one can observe that it is improbable that this sentence delineates a pertinent facet of the *definiendum*.

(d) Results point to a combination of definition patterns and the position of the root node as part of a battery of attributes effective in ranking in-domain answer candidates, and tackling out-of-domain sentences taken from collections of news articles.

(e) By the same token, the obtained configurations show that (a mix of) the following attributes are also good for ranking in-domain answer candidates: bigrams, shallow predicates, number of tokens, and shallow predicates II.

(f) More interestingly, the root node and the betweenness were utilised for rating out-of-domain testing instances collected from news articles. This empirical fact signifies that the shallow semantic and syntactic structures encapsulated by both properties port to this type of collection.

(g) Further, the outcomes show that the integration of shallow predicates and bigrams helps to recognise descriptions. As a rule, the acquired models highlight the importance of n-gram properties. In this respect, they are very likely to be chosen, especially trigrams proved to be effective (discriminative) in ranking in-domain candidate sentences and putative answers originating from the collection of news documents. However, distinct configurations can take advantage of n-grams of different lengths.

(h) Above all, experiments confirm that the amount of tokens is a leading feature. Particularly, this attribute proves to be instrumental in both NLP-free and NLP-oriented models, and it also ratified its versatility by being selected in assorted arrays of testing sentences (in-domain and all three out-of-domain).

(i) In summary, NLP-free outperformed NLP-based configurations when coping with out-of-domain data, while the opposite holds for in-domain testing instances.

(j) Contrary to chunking, selective substitutions did not cooperate with properties taken from dependency trees.

(k) From a different viewpoint, the figures specify the gap (5%) between the performance of the best models with and without NLP-based features. This outcome is particularly relevant as it measures the trade-off between quality and speed. As a logical conclusion, it is plausible to build competitive systems capable of processing massive collections of documents by capitalising solely on surface features. In a real situation, the output of this class of system can be exploited for creating a purpose-built index, wherewith definition QA systems can fetch the top hits, and process them with more effective configurations afterwards.

As future work, one can envisage more powerful approaches to acquire a balanced training material. As a matter of fact, a larger corpus is necessary to build context discriminant models (e.g., persons, organisations, and diseases). One can hypothesise that this level of abstraction can bring about an enhancement in terms of accuracy, and allegedly, also in terms of ranking order.

Glossary

Altavista

A web search engine owned by Yahoo!, which provides web and newsgroup, as well as paid submission services.
Homepage: www.altavista.com, pp. 29, 193

answers.com

A web-site that delivers answers by combining two answering approaches: (a) aggregations from various sources including dictionaries, encyclopedias, and atlases; and (b) a QA platform powered by the collaborative efforts of a global knowledge community.
Homepage: wiki.answers.com, pp. 2, 30, 31

AQUAINT

This corpus consists of newswire text data in English, drawn from three sources: the Xinhua News Service (People's Republic of China), the New York Times News Service, and the Associated Press Worldstream News Service.
Homepage: www.ldc.upenn.edu/Catalog/docs/LDC2002T31/, pp. 5, 17, 18, 20, 28, 31, 32, 46, 55, 59, 75, 79, 81, 90, 91, 96, 101, 103–106, 108–110, 112, 115, 116, 122, 124, 125, 133, 143, 144, 149–151, 161, 166–168, 180, 181, 185, 222, 226, 229, 231

Britannica Encyclopedia

It is a general English-language encyclopaedia, and it is the oldest encyclopaedia still in print in this language.
Homepage: www.britannica.com, pp. 30, 32, 103, 244

Columbia Encyclopedia

It is an electronic version of the encyclopedia produced by Columbia University Press, and it is available and licensed by several different companies for use over the World Wide Web.
Homepage: www.bartleby.com, pp. 30, 31

EFE

The EFE corpus of Spanish contains about 450,000 news articles from the EFE agency corresponding to the year 1994. The corpus is made available by the research group TALP of the Universitat Politécnica de Catalunya,
Homepage: www.talp.upc.es, page 51

Excite

An Internet portal, which was once one of the most recognised brands on the Internet. Nowadays, it provides a variety of services, including search, news, email, personals, and portfolio tracking.
Homepage: www.excite.com, page 29

FreeLing

It is an open source language analysis tool that can deal with numerous languages. FreeLing is designed to be used as an external library from any application, and it currently supports: Spanish, Catalan, Galician, Italian, English, Welsh, Portuguese, and Asturian.
Homepage: www.lsi.upc.edu/∼nlp/freeling/, pp. 136, 139

gold standard

A group of ranked answers to an array of pre-specified definition questions. This is typically utilised for assessing different answering strategies, pp. 5, 15, 16, 18–22, 44, 52, 85, 124, 127, 128, 133, 134, 137, 141, 161, 182, 231

Google

A web search engine owned by Google Inc. Nowadays, it is the most-used search engine on the Web, and it provides more than 22 special features beyond the original word-search capability. These include synonyms, maps, earthquake data, and sports scores.
Homepage: www.google.com, pp. 26, 28–32, 35–37, 40, 41, 47, 78, 91, 103, 108, 148, 183, 244

Google n-grams

It is a collection of sequences of words supplied by Google. These sequences include strings comprising from one to five tokens that are commonly found across the Internet.
Homepage: googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html, pp. 40, 41, 44, 53, 55, 98, 129, 247

Google Timeline

A feature of Google search engine that produces a timeline of events related to a given input (e.g., concept or *definiendum*).
Homepage: www.google.com/views?q=Iran+view:timeline, pp. 29, 30, 32, 111, 244

ground truth

See "Gold Standard", pp. 5, 15, 16, 18–22, 51, 88, 124, 128, 133, 134, 137, 141, 148, 181, 182, 222, 230, 231

JavaRap

An implementation of the Resolution of Anaphora Procedure (RAP) in the Java programming language. It resolves lexical anaphors, third person and pleonastic pronouns.
Homepage: wing.comp.nus.edu.sg/∼qiu/NLPTools/JavaRAP.html, pp. 153, 159, 181, 182, 222

Lucene

An open source text search engine library written in Java, suitable for nearly any application that requires full-text search.
Homepage: lucene.apache.org, pp. 26, 31, 199

Merriam-Webster dictionary

An online version of the Merriam-Webster's Collegiate Dictionary (11th Edition). It includes a main A-Z listing of terms, abbreviations, and biographical names.
Homepage: www.merriam-webster.com, pp. 30–32, 109

MSN Search

The current web search engine from Microsoft (a.k.a. Bing). It is the third largest by query volume.
Homepage: www.live.com, pp. 26, 29, 37, 161

nugget

A relevant fact or piece of information that it is perceived as an answer to a definition question, pp. 5, 15–18, 20–22, 28, 38, 43, 44, 52, 82, 86, 88, 100, 101, 106, 107, 123, 127, 128, 141, 203, 227

OpenNLP

A battery of NLP tools implemented in Java, which perform task such as sentence detection, tokenisation, POS-tagging, chunking, parsing, named-entity recognition, and co-reference resolution.
Homepage: opennlp.sourceforge.net, pp. 181, 204, 211, 222

Oxford Dictionary

A dictionary of English created by A. S. Hornby.
Homepage: www.oup.com, pp. 1, 2, 11–13

Who2

A online database of brief biographies and vital statistics pertaining to celebrities, historical figures, and famous people.
Homepage: www.who2.com, page 30

Wikipedia

It is a free, web-based, collaborative, multilingual encyclopedia project supported by the non-profit Wikimedia Foundation.
Homepage: en.wikipedia.org, pp. VI, 2, 5, 10, 20, 21, 26, 28, 30–34, 40, 43–47, 50, 53, 55, 66–68, 70, 72, 84, 88, 91, 96, 103, 110, 111, 128, 129, 132, 133, 136, 138, 139, 142, 144, 153, 154, 160–162, 164, 166, 168, 169, 172, 184, 185, 197, 203–205, 207, 211, 212, 214, 233, 244, 247

WordNet

A free and public lexical database for English. It clusters words into sets of synonyms, provides brief definitions, and records assorted semantic relations between these synonym sets.
Homepage: wordnet.princeton.edu, pp. 28, 30–32, 46, 55, 59, 81, 91, 96, 106, 107, 141, 148, 159, 177–179, 185, 200, 244, 246, 247

Yahoo Search

A web search engine, owned by Yahoo! Inc. It is the 2nd largest search engine on the web by query volume.
Homepage: www.yahoo.com, pp., 26, 29, 34, 38, 42, 58, 204, 212

List of Acronyms

Application Programming Interface (API)

An interface provided by a software program; which allows other applications to interact with it.

British National Corpus (BNC)

A collection of 100 million words found in written and spoken language. These terms were harvested from a wide range of sources.
Homepage: www.natcorp.ox.ac.uk

Cross Language Evaluation Forum (CLEF)

A forum that promotes the research and development in multilingual information access. It develops an infrastructure for the testing of different information retrieval systems operating on European languages in both monolingual and cross-language contexts.
Homepage: www.clef-campaign.org

Expectation Maximisation (EM)

In statistics, this algorithm is used for finding maximum likelihood estimates of parameters in probabilistic models, where the model depends on unobserved latent variables. It is an iterative method which alternates between performing an expectation (E) step, which computes an expectation of the log likelihood with respect to the current estimate of the distribution for latent variables, and a maximisation step, which computes the parameters which maximise the expected log likelihood found on the E step. These parameters are then used to determine the distribution of the latent variables in the next E step. Source: Wikipedia article retrieved 16th Nov. 2009

HyperText Markup Language (HTML)

It is the predominant markup language for web pages.

Inverse Document Frequency (IDF)

It is a weight used in information retrieval and text mining, which measures how important a word is.

Information Retrieval (IR)

It is the field that studies techniques for searching information within documents.

Knowledge Base (KB)

It is an authoritative resource of descriptive information. Usually, it is specific, online and accessible by any system. Some examples are online encyclopedias and dictionaries.

Language Model (LM)

It is statistical model that conforms a sequence of words to a probability distribution.

Latent Semantic Analysis (LSA)

A technique in natural language processing of analysing relations between documents and their terms.

Mean Average Precision (MAP)

It is a measure for assessing the ranking order of a given output (see details in section 1.7).

Maximum Entropy (ME)

A categorisation framework for synthesising information from numerous heterogeneous sources.

Multipurpose Internet Mail Extensions (MIME)

An internet standard.

Maximum Likelihood Estimate (MLE)

It is a statistical method utilised for fitting a statistical model to data, and providing estimates for its parameters.

Named Entity Recogniser (NER)

A tool for identifying -normally sequences of- words that belong to some predetermined class.

National Institute of Standards and Technology (NIST)

It is a measurement standards laboratory.

Natural Language Processing (NLP)

The area, involving computer science and linguistics, directed at studying the relationship between computers and human languages.

Portable Document Format (PDF)

A file format created by Adobe Systems for document exchange.

Profile Hidden Markov Model (PHMM)

Statistical tools that can model the regularities across sequences. They allow position dependent insertion and deletion penalties.

Part-of-Speech (POS)

A linguistic category of words.

Question Answering (QA)

Abbreviation.

Singular Value Decomposition (SVD)

A factorization of a matrix.

Support Vector Machine (SVM)

A supervised learning method utilised predominately for classification and regression; consisting basically in the construction of a hyperplane in a high dimensional space.

Term Frequency-Inverse Document Frequency (TF-IDF)

A weight often used in information retrieval and text mining.

Text REtrieval Conference (TREC)

It is a conference that supports research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies. According to: trec.nist.gov

Uniform Resource Locator (URL)

It is a subset of the Uniform Resource Identifier. This specifies the location of an identified resource and the mechanism for retrieving it.

Wall Street Journal (WSJ)

Abbreviation.

Index

Index of Citations

Bibliography

T. Abou-Assaleh, N. Cercone, J. Doyle, V. Keselj, and C. Whidden. DalTREC 2005 QA System Jellyfish: Mark-and-Match Approach to Question Answering. In *Proceedings of TREC 2005*. NIST, 2005.

D. Ahn, V. Jijkoun, G. Mishne, K. Müller, M. de Rijke, and S. Schlobach. Using Wikipedia at the TREC QA Track. In *Proceedings of TREC 2004*. NIST, 2004.

K. Ahn, J. Bos, J. R. Curran, D. Kor, M. Nissim, and BonnieWebber. Question Answering with QED at TREC–2005. In *Proceedings of TREC 2005*. NIST, 2005.

I. Androutsopoulos and D. Galanis. A practically Unsupervised Learning Method to Identify Single-Snippet Answers to Definition Questions on the web. In *HLT/EMNLP*, pages 323–330, 2005.

M. Belkin and J. Goldsmith. Using eigenvectors of the bigram graph to infer gram- matical features and categories. In *Proceedings of the Morphology/Phonology Learning Workshop of ACL-02*, 2002.

D. M. Bikel, R. L. Schwartz, and R. M. Weischedel. An algorithm that learns what's in a name. *Machine Learning*, 34(1-3):211–231, 1999.

M. Bilenko, R. Mooney, W. Cohen, P. Ravikumar, and S. Fienberg. Adaptive name matching in information integration. *IEEE Intelligent Systems*, 18(5):16–23, September/October 2003.

S. Blair-Goldensohn, K. R. McKeown, and A. H. Schlaikjer. A Hybrid Approach for QA Track Definitional Questions. In *Proceedings of TREC 2003*, pages 185–192. NIST, 2003.

M. D. Boni and S. Manandhar. The use of Sentence Similarity as a Semantic Relevance Metric for Question Answering. In *Proceedings of the AAAI Symposium on New Directions in Question Answering*, 2003.

P. F. Brown, S. A. D. Pietra, V. J. D. Pietra, and R. L. Mercer. Word-sense disambiguation using statistical methods. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, pages 264–270, 1991.

R. Bunescu and R. J. Mooney. A Shortest Path Dependency Kernel for Relation Extraction. In *Proceedings of HLT/EMNLP*, 2005.

J. D. Burger. MITRE's Qanda at TREC-12. In *Proceedings of TREC 2003*, pages 436–440. NIST, 2003.

J. D. Burger. MITRE's Qanda at TREC-15. In *Proceedings of TREC 2006*. NIST, 2006.

J. D. Burger and S. Bayer. MITRE's Qanda at TREC-14. In *Proceedings of TREC 2005*. NIST, 2005.

J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st ACM-SIGIR International Conference on Research and Development in Information Retrieval*, pages 335–336, 1998.

Y. Chali and S. R. Joty. University of LethbridgeŠs Participation in TREC–2007 QA Track. In *Proceedings of TREC 2007*. NIST, 2007.

S. Chen and J. Goodman. An Empirical Study of Smoothing Techniques for Language Modeling. In *Proceedings of the 34th Annual Meeting of the ACL*, pages 310–318, 1996.

S. Chen and J. Goodman. An Empirical Study of Smoothing Techniques for Language Modeling. Technical report, Computer Science Group, Harvard University (TR-10-98), 1998.

Y. Chen, M. Zhon, and S. Wang. Reranking answers for definitional qa using language modeling. In *Coling/ACL-2006*, pages 1081–1088, 2006.

X. Cheng, P. Adolphs, F. Xu, H. Uszkoreit, and H. Li. Gossip galore: A self-learning agent for exchanging pop trivia. In *Proceedings of the Demonstrations Session at EACL 2009*. Association for Computational Linguistics, 3 2009.

A. Chiu, P. Poupart, and C. DiMarco. Generating Lexical Analogies Using Dependency Relations. In *Proceedings of the 2007 Joint Conference on EMNLP and Computational Natural Language Learning*, pages 561–570, 2007.

K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990.

M. Ciaramita and Y. Altun. Broad Coverage Sense Disambiguation and Information Extraction with a Supersense Sequence Tagger. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 594–602, 2006.

W. Cohen, P. Ravikumar, and S. Fienberg. A comparison of string metrics for matching names and records. In *Proceedings of the IJCAI-2003 Workshop on Information Integration on the Web*, pages 73–78, 2003.

H. Cui, M.-Y. Kan, and T.-S. Chua. Unsupervised learning of soft patterns for definitional question answering. In *Proceedings of the Thirteenth World Wide Web Conference (WWW 2004)*, pages 90–99, 2004a.

H. Cui, K. Li, R. Sun, T.-S. Chua, and M.-Y. Kan. National University of Singapore at the TREC 13 Question Answering Main Task. In *Proceedings of TREC 2004*. NIST, 2004b.

H. Cui, M. Kan, and T. Chua. Generic soft pattern models for definitional question answering. In *Proceedings of SIGIR 2005*, pages 384–391, 2005.

H. Cui, M.-Y. Kan, and T.-S. Chua. Soft pattern matching models for definitional question answering. *ACM Trans. Inf. Syst.*, 25(2), 2007.

T. Cui, M. Kan, and J. Xiao. A comparative study on sentence retrieval for definitional question answering. In *SIGIR Workshop on Information Retrieval for Question Answering (IR4QA)*, pages 383–390, 2004c.

I. Dagan, A. Itai, and U. Schwall. Two languages are more informative than one. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, pages 130–137, 1991.

H. T. Dang, J. Lin, and D. Kelly. Overview of the TREC 2006 Question Answering Track. In *Proceedings of TREC 2006*. NIST, 2006.

H. T. Dang, D. Kelly, and J. Lin. Overview of the TREC 2007 Question Answering Track. In *Proceedings of TREC 2007*. NIST, 2007.

C. de Pablo-Sánchez, A. González-Ledesma, A. Moreno-Sandoval, and M. T. Vicente-Díez. MIRACLE Experiments in QA@CLEF 2006 in Spanish: Main Task, Real-Time QA and Exploratory QA Using Wikipedia (WiQA). In *CLEF*, pages 463–472, 2006.

C. de Pablo-Sánchez, J. L. Martínez-Fernández, A. González-Ledesma, D. Samy, P. Martínez, A. Moreno-Sandoval, and H. T. Al-Jumaily. Combining Wikipedia and Newswire Texts for Question Answering in Spanish. In *CLEF*, pages 352–355, 2007.

S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing By Latent Semantic Analysis. *Journal of the American Society For Information Science*, 41:391–407, 1990.

A. Dempster, N. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39:1–38, 1977.

C. Denicia-Carral, M. M. y Gómez, L. V. Pineda, and R. Hernández. A Text Mining Approach for Definition Question Answering. In *FinTAL*, pages 76–86, 2006.

A. Echihabi, U. Hermjakob, E. Hovy, D. Marcu, E. Melz, and D. Ravichandran. Multiple-Engine Question Answering in TextMap. In *Proceedings of TREC 2003*, pages 772–781. NIST, 2003.

I. Fahmi and G. Bouma. Learning to Identify Definitions using Syntactic Features. In *Proceedings of the Workshop on Learning Structured Information in Natural Language Applications*, 2006.

A. Fernandes. Answering definitional questions before they are asked. Master's thesis, Massachusetts Institute of Technology, 2004.

A. Figueroa. Finding Answers to Definition Questions across the Spanish Web. In WWW'09:WWW in Iberoamerica an Alternate Track, 2009.

A. Figueroa. Mining Wikipedia Resources for Discovering Answers to List Questions in Web Snippets. In *4th International Conference on Semantics, Knowledge and Grid*, pages 133–140, 2008a.

A. Figueroa. Mining Wikipedia for Discovering Multilingual Definitions on the Web. In *4th International Conference on Semantics, Knowledge and Grid*, pages 125–132, 2008b.

A. Figueroa. Boosting the recall of descriptive phrases in web snippets. In *In LangTech 2008*, 2008c.

A. Figueroa and J. Atkinson. Answering Definition Questions: Dealing with Data Sparseness in Lexicalised Dependency Trees-Based Language Models. In *LNBIP 45*, pages 297–310, 2010.

A. Figueroa and J. Atkinson. Using Dependency Paths For Answering Definition Questions on The Web. In *5th International Conference on Web Information Systems and Technologies*, pages 643–650, 2009.

A. Figueroa and G. Neumann. A Multilingual Framework for Searching Definitions on Web Snippets. In *KI*, pages 144–159, 2007.

A. Figueroa and G. Neumann. Finding distinct answers in web snippets. In *WEBIST (2)*, pages 26–33, 2008.

A. Figueroa, J. Atkinson, and G. Neumann. Searching for definitional answers on the web using surface patterns. *IEEE Computer*, 42(4):68–76, April 2009.

J. R. Firth. A synopsis of linguistic theory 1930-1955. *In Studies in Linguistic Analysis*, pages 1–32, 1957.

R. Gaizauskas, M. A. Greenwood, M. Hepple, I. Roberts, H. Saggion, and M. Sargaison. The University of SheffieldŠs TREC 2003 Q&A Experiments. In *Proceedings of TREC 2003*, pages 782–790. NIST, 2003.

R. Gaizauskas, M. A. Greenwood, M. Hepple, I. Roberts, and H. Saggion. The Univeristy of Sheffield's TREC 2004 Q&A Experiments. In *Proceedings of TREC 2004*. NIST, 2004.

W. A. Gale, K. W. Church, and D. Yarowsky. One Sense Per Discourse. In *Proceedings of the DARPA Speech and Natural Language Workshop*, 1992.

J. Goldstein, V. Mittal, J. Carbonell, and M. Kantrowitz. Multi-document summarization by sentence extraction. In *NAACL-ANLP 2000 Workshop on Automatic summarization*, pages 40–48, 2000.

J. T. Goodman. A bit of progress in language modeling. *Computer Speech and Language*, 15: 403–434, 2001.

H. Joho and M. Sanderson. Retrieving Descriptive Phrases from Large Amounts of Free Text. In *9th ACM conference on Information and Knowledge Management*, pages 180–186, 2000.

H. Joho and M. Sanderson. Large Scale Testing of a Descriptive Phrase Finder. In *1st Human Language Technology Conference*, pages 219–221, 2001.

K. Han, Y. Song, and H. Rim. Probabilistic model for definitional question answering. In *Proceedings of SIGIR 2006*, pages 212–219, 2006.

K.-S. Han, H. Chung, S.-B. Kim, Y.-I. Song, J.-Y. Lee, and H.-C. Rim. Korea University Question Answering System at TREC 2004. In *Proceedings of TREC 2004*. NIST, 2004.

K.-S. Han, Y. Song, S. Kim, and H. Rim. Phrase-based definitional question answering using definition terminology. In *AIRS 2005, LNCS 3689*, pages 246–259, 2005.

S. M. Harabagiu, D. I. Moldovan, C. Clark, M. Bowden, J. Williams, and J. Bensley. Answer mining by combining extraction techniques with abductive reasoning. In *TREC*, pages 375–382. NIST, 2003.

D. K. Harman and E. Voorhees. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, 2005.

Z. Harris. Distributional structure. In *Distributional structure. Word, 10(23)*, pages 146–162, 1954.

H. He and E. A. Garcia. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, September 2009.

A. Hickl, J. Williams, J. Bensley, K. Roberts, Y. Shi, and B. Rink. Question Answering with LCC's CHAUCER at TREC 2006. In *Proceedings of TREC 2006*. NIST, 2006.

A. Hickl, K. Roberts, B. Rink, J. Bensley, T. Jungen, Y. Shi, and J. Williams. Question Answering with LCC's CHAUCER-2 at TREC 2007. In *Proceedings of TREC 2007*. NIST, 2007.

W. Hildebrandt, B. Katz, and J. Lin. Answering Definition Questions Using Multiple Knowledge Sources. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 49–56, 2004.

P. Jaccard. The distribution of the flora of the alpine zone. *New Phytologist*, 11:37–50, 1912.

V. Jijkoun, G. Mishne, C. Monz, M. de Rijke, S. Schlobach, and O. Tsur. The University of Amsterdam at the TREC 2003 Question Answering Track. In *Proceedings of TREC 2003*, pages 586–593. NIST, 2003.

T. Joachims. Optimizing Search Engines Using Click-through Data. In *8th ACM Conference on Knowledge Discovery and Data Mining*, 2002.

A. Juárez-Gonzalez, A. Téllez-Valero, C. Denicia-Carral, M. M. y Gómez, and L. Villaseñor-Pineda. INAOE at CLEF 2006: Experiments in Spanish Question Answering. In *Working Notes for the CLEF 2006 Workshop*, 2006.

M. Kaisser and T. Becker. Question Answering by Searching Large Corpora with Linguistic Methods. In *Proceedings of TREC 2004*. NIST, 2004.

M. Kaisser, S. Scheible, and B. Webber. Experiments at the University of Edinburgh for the TREC 2006 QA Track. In *Proceedings of TREC 2006*. NIST, 2006.

B. Katz, M. Bilotti, S. Felshin, A. Fernandes, W. Hildebrandt, R. Katzir, J. Lin, D. Loreto, G. Marton, F. Mora, and O. Uzuner. Answering multiple questions on a topic from heterogeneous resources. In *Proceedings of TREC 2004*. NIST, 2004.

B. Katz, G. Marton, G. C. Borchardt, A. Brownell, S. Felshin, D. Loreto, J. Louis-Rosenberg, B. Lu, F. Mora, S. Stiller, Ö. Uzuner, and A. Wilcox. External knowledge sources for question answering. In *Proceedings of TREC 2005*. NIST, 2005.

B. Katz, G. Marton, S. Felshin, D. Loreto, B. Lu, F. Mora, Ö. Uzuner, M. McGraw-Herdeg, N. Cheung, A. Radul, Y. K. Shen, Y. Luo, and G. Zaccak. Question answering experiments and resources. In *Proceedings of TREC 2006*. NIST, 2006.

B. Katz, S. Felshin, G. Marton, F. Mora, Y. K. Shen, G. Zaccak, A. Ammar, E. Eisner, A. Turgut, and L. B. Westrick. CSAIL at TREC 2007 Question Answering. In *Proceedings of TREC 2007*. NIST, 2007.

V. Keselj and A. Cox. DalTREC 2004: Question Answering Using Regular Expression Rewriting. In *Proceedings of TREC 2004*. NIST, 2004.

T. Kikuchi, S. Furui, and C. Hori. Automatic speech summarization based on sentence extraction and compaction. In *Proceedings of ICASSP*, 2003.

J. H. Kil, L. Lloyd, and S. Skiena. Question Answering with Lydia (TREC 2005 QA track). In *Proceedings of TREC 2005*. NIST, 2005.

W. Kintsch. Predication. *Cognitive Science*, 25:173Ű–202, 1998.

L. Kosseim, A. Beaudoin, A. Keighbadi, and M. Razmara. Concordia University at the TREC-15 QA track. In *Proceedings of TREC 2006*. NIST, 2006.

D. Lin. Principar – an efficient, broad-coverage, principle-based parser. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING-94)*, pages 484–488, 1994.

D. Lin and P. Pantel. Discovery of Inference Rules for Question Answering. In *Journal of Natural Language Engineering, Volume 7*, pages 343–360, 2001.

J. Lin and D. Demner-Fushman. Automatically evaluating answers to definition questions. In *Proceedings of the 2005 Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, pages 931–938, October 2005.

J. Lin and D. Demner-Fushman. Will pyramids built of nuggets topple over? In *Proceedings of the 2006 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL 2006)*, pages 383–390, June 2006.

B. Liu. *Web Data Mining*. Springer, December 2006.

D. A. Lupsa and D. Tatar. Some Remarks about Feature Selection in Word Sense Discrimination for Romanian Language. *Studia Univ. Babes-Bolyai, Informatica*, L(2), 2005.

D. J. C. MacKay and L. C. B. Peto. A hierarchical Dirichlet language model. *Natural Language Engineering*, 1(3):1–19, 1994.

B. Magnini, D. Giampiccolo, P. Forner, C. Ayache, P. Osenova, A. Peñas, V. Jijkoun, B. Sacaleanu, P. Rocha, and R. Sutcliffe. Overview of the CLEF 2006 Multilingual Question Answering Track. In *Working Notes for the CLEF 2006 Workshop*, 2006.

C. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999.

M. Marneffe, B. Maccartney, and C. D. Manning. Generating typed dependency parses from phrase structure parses. In *In LREC 2006*, 2006.

Á. Martínez-González, C. de Pablo-Sánchez, C. Polo-Bayo, M. T. Vicente-Díez, P. Martínez-Fernández, and J. L. Martínez-Fernández. The MIRACLE Team at the CLEF 2008 Multilingual Question Answering Track. In *CLEF*, 2008.

S. Miliaraki and I. Androutsopoulos. Learning to identify single-snippet answers to definition questions. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 1360–1366, 2004.

M. Montes-y-Gómez, L. Villaseñor-Pineda, M. Pérez-Coutiño, J. M. Gómez-Soriano, E. Sanchis-Arnal, and P. Rosso. INAOE-UPV Joint Participation in CLEF 2005: Experiments in Monolingual Question Answering. In *Working Notes for the CLEF 2005 Workshop*, 2005.

R. Navigli and M. Lapata. Graph connectivity measures for unsupervised word sense disambiguation. In *In IJCAI-07*, pages 1683–1688, 2007.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL2002)*, pages 311–318, 2002.

M. Paşca. *Answering Definition Questions via Temporally-Anchored Text Snippets*. Proceedings of IJCNLP, 2008.

J. Prager, D. Radev, and K. Czuba. Answering What-Is Questions by Virtual Annotation. In *1st Human Language Technology Conference*, pages 26–30, 2001.

J. Prager, J. Chu-Carroll, and K. Czuba. Use of WordNet Hypernyms for Answering What-Is Questions. In *TREC-2001*, 2002.

J. Prager, J. Chu-Carroll, K. Czuba, C. Welty, A. Ittycheriah, and R. Mahindru. IBM's PIQUANT in TREC2003. In *Proceedings of TREC 2003*, pages 283–292. NIST, 2003.

A. Purandare and T. Pedersen. Word Sense Discrimination by Clustering Contexts in Vector and Similarity Spaces. In *HLT-NAACL 2004 Workshop: Eighth Conference on Computational Natural Language Learning (CoNLL-2004)*, pages 41–48, 2004.

X. Qiu, B. Li, C. Shen, L. Wu, X. Huang, and Y. Zhou. FDUQA on TREC2007 QA Track. In *Proceedings of TREC 2007*. NIST, 2007.

D. Ravichandran and E. Hovy. Learning Surface Text Patterns for a Question Answering System. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 41–47, 2002.

M. Razmara and L. Kosseim. A little known fact is . . . answering other questions using interest-markers. In *Proceedings of the 8th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2007)*, pages 518–529, 2007.

M. Razmara, A. Fee, and L. Kosseim. Concordia University at the TREC 2007 QA track. In *Proceedings of TREC 2007*. NIST, 2007.

D. E. Rose and D. Levinson. Understanding user goals in web search. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 13–19, 2004.

D. Roussinov, Y. Ding, and J. A. Robles-Flores. Experiments with web qa system and trec2004 questions. In *Proceedings of the TREC 2004*. NIST, 2004.

D. Roussinov, E. Filatova, M. Chau, and J. Robles-Flores. Building on Redundancy: Factoid Question Answering, Robust Retrieval and the "Other". In *Proceedings of TREC 2005*. NIST, 2005.

B. Sacaleanu, G. Neumann, and C. Spurk. DFKI-LT at QA@CLEF 2007. In *In Working Notes for the CLEF 2007 Workshop*, 2007.

B. Sacaleanu, G. Neumann, and C. Spurk. DFKI-LT at QA@CLEF 2008. In *In Working Notes for the CLEF 2008 Workshop*, 2008.

G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.

G. Salton and M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

A. Savary and C. Jacquemin. Reducing Information Variation in Text. In *Text- and Speech-Triggered Information Access, LNCS 2705/2003*, pages 145–181, 2003.

S. Scheible. *A Computational Treatment of Superlatives*. PhD thesis, University of Edinburgh, 2007.

N. Schlaefer, P. Gieselmann, and G. Sautter. The Ephyra QA System at TREC 2006. In *Proceedings of TREC 2006*. NIST, 2006.

N. Schlaefer, J. Ko, J. Betteridge, G. Sautter, M. Pathak1, and E. Nyberg. Semantic Extensions of the Ephyra QA System for TREC 2007. In *Proceedings of TREC 2007*. NIST, 2007.

B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, 2002.

D. Shen, J. L. Leidner, A. Merkel, and D. Klakow. The Alyssa System at TREC 2006: A Statistically-Inspired Question Answering System. In *Proceedings of TREC 2006*. NIST, 2006.

D. Shen, M. Wiegand, A. Merkel, S. Kazalski, S. Hunsicker, J. L. Leidner, and D. Klakow. The Alyssa System at TREC QA 2007: Do We Need Blog06? In *Proceedings of TREC 2007*. NIST, 2007.

M. M. Soubbotin. Patterns of Potential Answer Expressions as Clues to the Right Answers. In *Proceedings of the TREC-10 Conference*, pages 293–302. NIST, 2001.

M. Srikanth and R. Srihari. Biterm language models for document retrieval. In *Proceedings of the 2002 ACM SIGIR Conference on Research and Development in Information Retrieval*, 2002.

R. Sun, J. Jiang, Y. Tan, H. Cui, T.-S. Chua, and M.-Y. Kan. Using Syntactic and Semantic Relation Analysis in Question Answering. In *Proceedings of TREC 2005*. NIST, 2005.

M. Surdeanu, M. Ciaramita, and H. Zaragoza. Learning to Rank Answers on Large Online QA Collections. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL 2008)*, pages 719–727, 2008.

J. Surowiecki. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. Little, Brown, 2004.

N. Swartz. Definitions, Dictionaries, and Meanings. Available online http://www.sfu.ca/philosophy/swartz/definitions.htm, 1997.

A. Vallin, D. Giampiccolo, L. Aunimo, C. Ayache, P. Osenova, A. P. nas, M. Rijke, B. Sacaleanu, D. Santos, and R. Sutcliffe. Overview of the CLEF 2005 Multilingual Question Answering Track. In *Working Notes for the CLEF 2005 Workshop*, 2005.

E. M. Voorhees. Overview of the TREC 2004 Question Answering Track. In *Proceedings of TREC 2004*. NIST, 2004.

E. M. Voorhees. Evaluating answers to definition questions. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 109–111, 2003.

E. M. Voorhees and H. Dang. Overview of the TREC 2005 Question Answering Track. In *Proceedings of TREC 2005*. NIST, 2005.

E. Whittaker, P. Chatain, S. Furui, and D. Klakow. TREC 2005 Question Answering Experiments at Tokyo Institute of Technology. In *Proceedings of TREC 2005*. NIST, 2005.

E. Whittaker, J. Novak, P. Chatain, and S. Furui. TREC 2006 Question Answering Experiments at Tokyo Institute of Technology. In *Proceedings of TREC 2006*. NIST, 2006.

E. W. D. Whittaker, M. H. Heie, J. R. Novak, and S. Furui. TREC 2007 Question Answering Experiments at Tokyo Institute of Technology. In *Proceedings of TREC 2007*. NIST, 2007.

P. Wiemer-Hastings and I. Zipitria. Rules for Syntax, Vectors for Semantics. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, 2001.

L. Wittgenstein. *Philosophical Investigations*. Blackwell, 1953.

L. Wu, X. Huang, L. You, Z. Zhang, X. Li, and Y. Zhou. FDUQA on TREC2004 QA Track. In *Proceedings of TREC 2004*. NIST, 2004.

L. Wu, X. Huang, Y. Zhou, Z. Zhang, and F. Lin. FDUQA on TREC2005 QA Track. In *Proceedings of TREC 2005*. NIST, 2005a.

M. Wu, M. Duan, S. Shaikh, S. Small, and T. Strzalkowski. ILQUA–An IE-Driven Question Answering System. In *Proceedings of TREC 2005*. NIST, 2005b.

M. Wu, C. Song, Y. Zhan, and T. Strzalkowski. UAlbany's ILQUA at TREC 2007. In *Proceedings of TREC 2007*. NIST, 2007.

J. Xu, A. Licuanan, and R. Weischedel. TREC2003 QA at BBN: Answering Definitional Questions. In *Proceedings of TREC 2003*, pages 98–106. NIST, 2003.

J. Xu, Y. Cao, H. Li, and M. Zhao. Ranking definitions with supervised learning methods. In *WWW2005*, pages 811–819, 2005.

J. Xu, R. Weischedel, and A. Licuanan. Evaluation of an Extraction-Based Approach to Answering Definitional Questions. In *SIGIR'04*, pages 418–424, 2004.

H. Yang, H. Cui, M. Maslennikov, L. Qiu, M.-Y. Kan, and T. Chua. QUALIFIER In TREC-12 QA Main Task. In *Proceedings of TREC 2003*, pages 480–488. NIST, 2003.

Y. Yang and J. O. Pedersen. A Comparative Study on Feature Selection in Text Categorization. In *14th International Conference on Machine Learning*, pages 412–420, 1997.

C. Zhai and J. Lafferty. A Study of Smoothing Methods for Language Models Applied to Information Retrieval. *ACM Transactions on Information Systems*, 22(2):179–214, 2004.

Z. Zhang, Y. Zhou, X. Huang, and L. Wu. Answering Definition Questions Using Web Knowledge Bases. In *Proceedings of IJCNLP 2005*, pages 498–506, 2005.

Y. Zhou, X. Yuan, J. Cao, X. Huang, and L. Wu. FDUQA on TREC2006 QA Track. In *Proceedings of TREC 2006*. NIST, 2006.