UNIVERSITÄT
DES
SAARLANDES

Saarland University
Faculty of Mathematics and Computer Science
Department of Computer Science

# Biomedical Knowledge Base Construction from Text and its Applications in Knowledge-based Systems

**Patrick Ernst**

A dissertation submitted towards the degree
Doctor of Engineering
of the Faculty of Mathematics and Computer Science of
Saarland University

Saarbrücken, December 2017

| | |
|---|---|
| Dean: | Prof. Dr. Frank-Olaf Schreyer |
| Colloquium: | 6 March 2018 |

Examination Board

| | |
|---|---|
| Supervisor and First Reviewer: | Prof. Dr. Gerhard Weikum |
| Second Reviewer: | Prof. Karin Verspoor, Ph.D. |
| Third Reviewer: | Prof. Dr. Klaus Berberich |
| Chairman: | Prof. Dr. Christoph Weidenbach |
| Research Assistant: | Andrew Yates, Ph.D. |

# ABSTRACT

While general-purpose Knowledge Bases (KBs) have gone a long way in compiling comprehensive knowledge about people, events, places, etc., domain-specific KBs, such as on health, are equally important, but are less explored. Consequently, a comprehensive and expressive health KB that spans all aspects of biomedical knowledge is still missing. The main goal of this thesis is to develop principled methods for building such a KB and enabling knowledge-centric applications. We address several challenges and make the following contributions:

- To construct a health KB, we devise a largely automated and scalable pattern-based knowledge extraction method covering a spectrum of different text genres and distilling a wide variety of facts from different biomedical areas.

- To consider higher-arity relations, crucial for proper knowledge representation in advanced domain such as health, we generalize the fact-pattern duality paradigm of previous methods. A key novelty is the integration of facts with missing arguments by extending our framework to partial patterns and facts by reasoning over the composability of partial facts.

- To demonstrate the benefits of a health KB, we devise systems for entity-aware search and analytics and for entity-relationship-oriented exploration.

Extensive experiments and use-case studies demonstrate the viability of the proposed approaches.

# KURZFASSUNG

Universelle Wissensbanken, die durch automatische Wissensextraktion aus Internetquellen konstruiert wurden, beinhalten eine Fülle an Detailwissen über Personen, Orte, Ereignisse, etc. Wichtige domänenspezifische Anwendungsfälle, wie im Gesundheits- und biomedizinischen Bereich, haben ebenfalls viel Beachtung erhalten. Umfassende Wissensbanken, die alle Aspekte der Lebenswissenschaften widerspiegeln, fehlen allerdings. Das Hauptaugenmerk dieser Dissertation liegt daher auf der Methodenentwicklung zur Konstruktion solcher Wissensbanken und auf der Realisierung von darauf aufbauender, wissensbasierter Anwendungen. Die Arbeit adressiert drei Problembereiche und entwickelt folgende Lösungsvorschläge:

- Zur Konstruktion einer Wissensbank entwickeln wir einen weitgehend automatisierten und skalierbaren, musterbasierten Wissensextraktionsansatz. Ausgehend von einem Spektrum verschiedener Textgenres ist dieser in der Lage eine hohe Anzahl von Fakten zu extrahieren, die eine Vielzahl biomedizinischer Bereiche abdecken.

- In komplexen Domänen wie der Biomedizin ist es erforderlich, höherstellige Relationen zu betrachten. Um diesem Umstand gerecht zu werden, verallgemeinern wir das „Fact-pattern Duality“ Paradigma früherer Methoden. Ein Schwerpunkt liegt auf der Betrachtung von Fakten mit fehlenden Argumenten, die durch eine Erweiterung des Ansatzes auf partielle Muster und Fakten eingebunden werden. Die Vereinheitlichung partieller Fakten wird mittels logischer Deduktion realisiert.

- Um die Vorteile einer medizinischen Wissensbank zu demonstrieren, präsentieren wir Systeme zur entitätsbasierten Suche und Analyse sowie zur relationalen Faktenexploration.

Experimentelle Auswertungen und Anwedungsfallstudien zeigen die Tragfähigkeit der vorgeschlagenen Ansätze.

# SUMMARY

Large Knowledge Bases (KBs), such as DBPedia, YAGO, and Wikidata, contain millions of entities and billions of facts. They enable semantic search, analytics, question answering, and smart recommendations over Web contents or other kinds of Big Data. In the biomedical domain, KBs such as the Gene Ontology, the Disease Ontology, and the Foundational Model of Anatomy are prominent examples of the rich knowledge that is digitally available. However, each of these KBs is highly specialized and focuses only on a relatively narrow topic within the life sciences; comprehensive biomedical KBs linking all aspects of biomedical knowledge are missing. The goal of this thesis is the development of versatile methods that support bridging the gap between different topics within life sciences and to enable powerful and smart applications. To reach this goal this thesis makes the following contributions:

**KnowLife** is a versatile and scalable method for constructing a comprehensive KB, linking genes, anatomic parts, diseases, symptoms, treatments, as well as environmental and lifestyle risk factors for diseases. Following the fact-pattern duality paradigm and using a small number of seed facts for distant supervision of pattern-based extraction, we harvest a large number of fact candidates in an automated manner without requiring any explicit training. Our KB construction method combines statistics-based pattern matching for high recall of fact candidates with logics-based consistency reasoning for high precision of eventually accepted facts. We ran extensive experiments, creating different KBs based on different configurations. The best KB contains more than 500,000 facts at a precision of 93% for 13 relations.

**HighLife** generalizes the fact-pattern duality paradigm to higher-arity cases, going beyond harvesting binary facts, which have been the focus of most text-based knowledge extraction methods, including *KnowLife*. Higher-arity relations are crucial in advanced domains, such as health, where a more expressive knowledge representation is required. A key novelty lies in coping with the issue that higher-arity facts are often expressed only partially in texts. Our method is also able to integrate such partial facts, at both pattern-learning and constraint-reasoning stages by (1) extending our framework to partial higher-arity patterns and facts and (2) by extending reasoning over the consistency and composability of partial fact candidates into full facts. Experiments with health-related documents and with news articles demonstrate the viability of our method.

Leveraging versatile KBs, we devise two technical systems for exploring, analyzing and searching over large text collections on health.

**DeepLife** is an entity-aware search and exploration platform for the life sciences, which overcomes limitations of existing search engines which often perform poorly in retrieving informative contents for health-centric information needs. DeepLife integrates large KBs and harnesses entity linking methods to annotate articles in real-time in order to stay up-to-date on the latest health topics. The system is highly expressive in its capabilities for querying scientific literature, news feeds as well as social media, based on flexible combinations of keywords, phrases, biomedical entities, relational facts, and taxonomic categories. It also supports users by powerful entity-centric auto-completion suggestions, interactive query sessions, and entity-aware text analytics.

**KnowLife's One-stop Health Portal** is an exploration platform with the goal of comprising knowledge on all health-related aspects in an integrated manner. Based on KnowLife's versatile KB the portal offers powerful entity-relationship-oriented search and exploration features over large amounts of content. It automatically creates info boxes, which summarize important information about entities, link entites based on facts, and provide evidence for facts. By incorporating provencance information, such as textual patterns for relations, the system is able to automatically annotate any kind of input document, expert-level or layman style, with entities and relationships on the fly. The system aids users in "speed-reading" in order to efficiently consume textual medical knowledge. The value of the KnowLife portal is demonstrated by several use-case scenarios: laymen exploring health issues of personal interest, medical professionals searching for specific knowledge, and researchers "speed-reading" publications via entity-relationship synopses.

# CONTENTS

# Part I

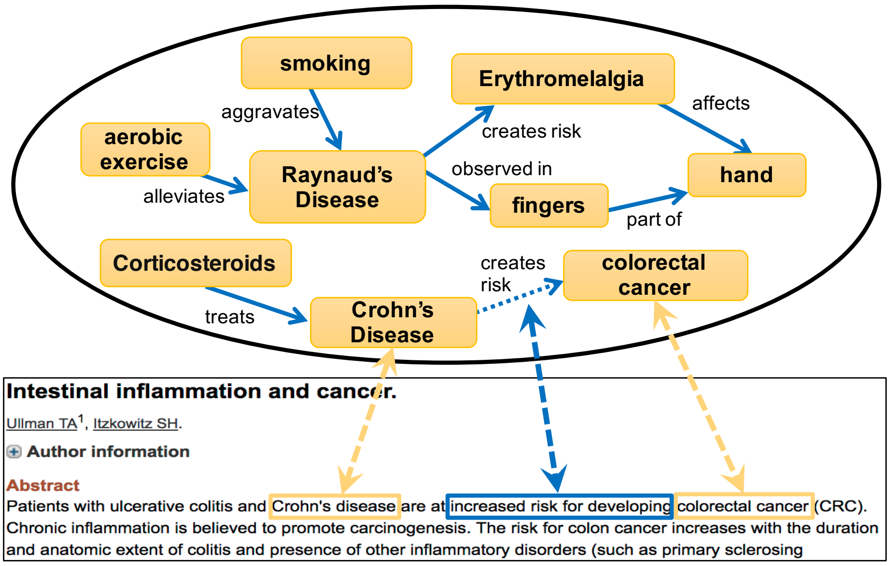
# BACKGROUND

# 1 INTRODUCTION

## 1.1 MOTIVATION

We have entered an era, where huge amounts of new data are generated continuously, for instance by laymen creating blog posts or social media entries on their smartphones or by professionals in publishing news reports or scientific results. This explosion of content poses a major challenge for knowledge dissemination, specifically when users seek specific amounts of relevant information in huge repositories, such as the Web. At the same time, this abundance of data also presents many opportunities for smart Information Extraction (IE) systems, able to harvest the buried information and discover knowledge. To make sense of this vast amount of textual content, these systems must be capable of capturing the semantics expressed in human language and transform unstructured and naturally expressed information within text into a structured logical representation suitable for computers.

Even though, approaches trying to overcome the obstacle date back to the early days of Artificial Intelligence (AI), these initial efforts failed due to limited scope and computing power. In contrast, contemporary methods are able to process and exploit the abundance of available data in a sophisticated manner. A common direction is to identify and harvest real-world entities (e. g. people, locations, diseases), their properties (e. g. height of people, size of organs), and relations (e. g. birth locations of people, symptoms of diseases) in order to construct extensive Knowledge Bases (KBs). KBs represent the knowledge in a structured and concise logical form, which computers can reason with to retrieve existing or deduce new knowledge. Figure 1 shows an example of such a KB representation. In recent years, notable research projects, like the Never-Ending Language Learning system (NELL) [107], YAGO [73] or WikiData [188] among many others, have produced comprehensive and highly interconnected KBs. Due to their versatility and extensive coverage, they have become key components for many applications, such as search, analytics, and smart recommendations over Web contents and other kinds of Big Data. Two prominent commercial use-cases are the Google Knowledge Graph, which since its inception in 2012 has become a key component for Google's search, and IBM's question answering system Watson [56], which has been released in 2010 and leverages KBs to analyze huge amounts of text content in order to answer natural language questions. Apart from general-purpose KBs, we see sim-



Intestinal inflammation and cancer.

Ullman TA[1], Itzkowitz SH.

⊕ Author information

Abstract

Patients with ulcerative colitis and Crohn's disease are at increased risk for developing colorectal cancer (CRC). Chronic inflammation is believed to promote carcinogenesis. The risk for colon cancer increases with the duration and anatomic extent of colitis and presence of other inflammatory disorders (such as primary sclerosing

Figure 1: Knowledge Base Construction Example

ilar problem settings in specialized domains, such as biomedicine and health, where there are constantly growing, huge repositories with millions of scientific reports like publications or clinical trials, as well as many encyclopedic health portals, and online communities where patients and physicians discuss health-related issues. Many structured resources, like the Gene Ontology (GO) [8], many Protein-Protein Interactions (PPI) databases, or the Unified Medical Language System (UMLS) [18], have been created in the past. However, they lack far behind in terms of versatility and comprehensiveness, compared to their general-domain counterparts, and this hinders the development of powerful knowledge-based applications for the health domain.

## 1.2 CHALLENGES

To construct a comprehensive KB, which spans all aspects of biomedical knowledge and can serve as foundation for powerful applications, the following shortcomings must be addressed:

DEPENDENCE ON MANUAL CURATION: Due to the high publication rate of health-related contents on the Web (over one million new PubMed scientific articles are published each year), we need to go beyond manually crafted KBs, that failed to keep pace with this fast publication rate.

NEGLECTION OF TEXTUAL KNOWLEDGE SOURCES: There is almost an exclusive focus on scientific publications to perform biomedical IE. However, there exists a vast universe of Web contents, which are still mostly neglected: encyclopedic portals, which aim to reach out to laymen, and online discussion forums are frequented by patients.

LIMITATION OF TARGET DOMAINS: Most prior work for extracting biomedical knowledge covers only special sub-domains. Protein-protein interactions, genetic mutations, and other relationships at the molecular level are prominent examples. However, an integrated view is still lacking, expanding the scope and connecting risk factors, symptoms, and drug side effects, etc.

LIMITED EXPRESSIVENESS: Many knowledge extraction methods for constructing Knowledge Bases only focus on binary relations between two entities, which is not sufficient to properly represent knowledge in advanced domains such as health.

LACK OF KNOWLEDGE-BASED SEARCH AND EXPLORATION: In life sciences, we are confronted with an overwhelming amount of information scattered over the Web. Thus, there is an urgent need for smart and powerful applications, which link information to KBs to guide users in their information seeking process.

## 1.3 PROBLEM SETTING

In this thesis, we consider the problem of Knowledge Base Construction (KBC) for the wider health domain from natural language text, i. e. the process of populating a Knowledge Base with entities, facts, or rules harvested from large amounts of data. In particular, we focus on the fact harvesting parts of the KBC problem, which aims to identify new facts, connecting multiple entities through logical relations.

In this thesis, we address four key aspects of the problem:

VERSATILITY: The proposed approaches should neither be restricted to a specific subdomain within biomedicine nor to specific text genres.

LIMITED TRAINING DATA: Since labelled data is costly to procure and due to the unavailability of a comprehensive pre-annotated corpus, only a limited amount of training data is available. The versatility aspect aggravates this further, since multiple kinds of texts and domains need to considered.

REUSE OF KNOWLEDGE: If appropriate, the solutions should take into account the already existing knowledge resources and augment them.

EXPRESSIVE KB: We want to generalize the notion of fact harvesting and go beyond binary to higher-arity relations to capture more expressive facts. For instance, we want to be able to capture which drug is used for which disease at which dosage (e.g. 2.5 mg/day) for which kinds of patients (e. g. , children vs. adults).

## 1.4 CONTRIBUTIONS

The main goal of this thesis lies in the development of versatile methods for the automatic construction of biomedical KBs, that span all sub-areas within the life sciences and thus enable smart applications for search, exploration and analysis. Specifically, we make the following contributions:

### 1.4.1 *Knowledge Base Construction*

*A Versatile Knowledge Base For Life Sciences*

We present a versatile, largely automated, and scalable approach for the comprehensive construction of a KB – covering a wide spectrum of different text genres as input and distilling a wide variety facts from different biomedical areas as output. Coupled with an entity recognition module that covers the entire range of biomedical entities, the resulting KB features a much wider spectrum of knowledge and use-cases. Our approach learns sequence patterns from text and combines statistics-based pattern matching for high recall of fact candidates with logics-based consistency reasoning, tailored to the different relations that connect diseases, symptoms, drugs, genes, risk factors, etc. This constraint checking eliminates many false positives that are produced by methods that solely rely on pattern-based extraction. Results of this research have been published as:

> Patrick Ernst, Amy Siu and Gerhard Weikum. "KnowLife: a versatile approach for constructing a large knowledge graph for biomedical sciences." In: *BMC Bioinformatics* 16.157 (2015).

*Beyond Binary Facts*

We introduce an approach to harvest higher-arity facts from textual sources. To achieve this, the method extends the fact-pattern duality paradigm to higher-arity cases. A major novelty lies in coping with the issue that higher-arity facts are often expressed only partially in texts and strewn across multiple sources. Our method is able to integrate partially observed facts, at both pattern-learning and constraint-reasoning stages by (1) extending our framework to partial patterns and partial facts and (2) by extending reasoning over the consistency and composability of partial fact candidates into full facts. This contribution has been submitted for publication and is currently under review:

> Patrick Ernst, Amy Siu and Gerhard Weikum. "HighLife: Higher-arity Fact Harvesting." *under submission*, 2017.

### 1.4.2 *Knowledge-based Applications*

*Exploratory Search*

We develop a health portal with the goal of providing an integrated and comprehensive view of available health-care contents, i. e. an one-stop portal where users can quickly digest any kind of health information. The portal demonstrates how a KB can be leveraged for efficient entity-relationship-oriented search and exploration of large amounts of health data. We also present on-the-fly IE capabilities, which annotate any kind of input documents, expert-level or layman style, with entities and relationships on the fly, as the user reads it. The value of the realized *One-stop Health Portal* is demonstrated by several use-case scenarios: laymen exploring health issues of personal interest, medical professionals searching for specific knowledge, and researchers "speed-reading" publications via entity-relationship synopses. The system is described in the following publication:

> Patrick Ernst, Cynthia Meng, Amy Siu and Gerhard Weikum. "KnowLife: A Knowledge Graph for Health and Life Sciences." In: *Proceedings of IEEE International Conference on Data Engineering*. ICDE '14. Chicago, IL, USA: IEEE Computer Society, 2014, pp. 1254–1257

*Entity-aware Search & Analytics*

We present an entity-aware search and analytics platforms for life sciences, which integrates large knowledge bases and harnesses entity linking methods to annotate documents in an ad-hoc manner from a continuous stream of scientific literature, newspaper feeds, and social media. Relying on this corpus, we describe a user interface for interactive search and query sessions based on flexible combinations of keywords, biomedical entities, facts, and taxonomic categories. Besides search, the system also provides functionality for entity-aware text analytics over health-centric contents. All salient new features are showcased by different use-case scenarios benefiting from the systems capabilities. The design and implementation of the search engine haven been published in:

> Patrick Ernst, Amy Siu, Dragan Milchevski, Johannes Hoffart, and Gerhard Weikum. "DeepLife: An Entity-aware Search, Analytics and Exploration Platform for Health and Life Sciences." In: *Proceedings of the Annual Meeting on Association for Computational Linguistics: System Demonstrations*. ACL '16. Berlin, Germany: Association for Computational Linguistics, 2016, pp. 19–24.

## 1.5 THESIS OUTLINE

This thesis is structured into four parts: Part I lays the foundation by explaining necessary preliminaries and the core concepts of Knowledge Base Construction (KBC) in Chapter 2 followed by a discussion of relevant related works (Chapter 3).

Part II and Part III present the core contributions of the thesis: Part II develops automatic KBC approaches for creating a versatile biomedical KB (Chapter 4), and describes our approach for harvesting higher-arity facts (Chapter 5). In Part III, we present two technical contributions, demonstrating how KBs can be used in knowledge-based applications. In Chapter 6 we describe an explorative one-stop health portal designed for efficient entity-relationship-oriented consumption of knowledge, and in Chapter 7 we present an entity-aware search and analytics platform for health-related content.

Part IV concludes the thesis with a summary and an outlook on future work.

# 2 PRELIMINARIES

## 2.1 KNOWLEDGE BASES

### 2.1.1 *Principles*

Knowledge Bases (KBs) are large networks about entities, their properties, and the relationships between entities. Before formally defining KBs we first introduce their core constituents, namely entities together with their types, relations with type signatures and facts.

The foundation of every KB is a set of entities $\mathcal{E}$ defining its universe of discourse.

**Definition 1 (Entity)** *An entity or named entity is a collection of all possible mentions (surface forms such as noun phrases, single or multi-word expressions, etc.) that refer to the identical real-world object or abstract concept.*

It should be noted that defining the real meaning of entity poses a philosophical challenge not tackled here, which is, for instance, discussed in [169]. Therefore, this is a rather pragmatic definition expressing the role of entities within the context of Knowledge Bases and knowledge extraction.

**Example 1 (Entity)** The mentions *Microsoft*, *NASDAQ:MSFT*, *the giant from Redmond*, and *The Microsoft Corporation* refer to the same entity of type company.
Similarly, *Heart Attack*, *Coronary Attack*, *Heart Infarction*, and *Myocardial Infarction* are all designated surface forms of the same disease entity.

Often several entities share the same surface forms, e. g. all people with the same first and last name, and it is necessary to disambiguate between them to identify the correct entity.

To categorize homogeneous entities into classes, types are introduced:

**Definition 2 (Type)** *An entity type is represented as a set $T \subseteq \mathcal{E}$ of entities sharing common characteristics. We define the set of all types $\mathcal{T} = \{T_1, \dots, T_k\}$.*

**Example 2 (Type)** The entities *Microsoft Corporation*, *Amazon, Inc.*, and *Facebook, Inc.* are members of the types *Internet Companies* and *NASDAQ Companies*, whereas *Myocardial infarction*, *Angina*, and *Coronary Artery Disease* are members of the type *Heart Diseases*.

Relations define templates how entities can be connected with each other:

**Definition 3 (Relation)** *An n-ary relation* $\mathsf{R}$ *over an entity set* $\mathcal{E}$ *together with a type signature* $\mathsf{T}_1 \times \ldots \times \mathsf{T}_n$ *constraining the set of* $\mathsf{R}$*'s possible arguments is defined as a set of tuples* $\mathsf{R} = \{(\mathsf{a}_1, \ldots, \mathsf{a}_n) \in \mathcal{E}^n | \mathsf{a}_1 \in \mathsf{T}_1 \wedge \ldots \wedge \mathsf{a}_n \in \mathsf{T}_n\}$*.*

In this thesis we call relations with two arguments binary relations and relations with more than two arguments higher-arity relations.

**Example 3 (Relation)** Consider the following binary relation and its type signature expressing symptom of diseases
$\mathsf{SymptomOf} : \mathsf{Symptoms} \times \mathsf{Disease}$
For the business domain, consider the higher-arity relation
$\mathsf{CompanyAcquired} : \mathsf{Organization} \times \mathsf{Organization} \times \mathsf{Date} \times \mathsf{Price}$

Based on templates declared by relations, facts connect multiple entities:

**Definition 4 (Fact)** *A fact is an instance of an n-ary relation. We write a fact in the form* $\mathsf{R}(\mathsf{a}_1, \ldots, \mathsf{a}_n)$ *where* $\mathsf{R}$ *is an n-ary relation and* $\mathsf{a}_1$ *through* $\mathsf{a}_n$ *are constants (i. e. , entities or literals including short phrases) of types that fit with the type signature of* $\mathsf{R}$*.*

As before, facts of binary relations are called binary facts and facts with more than two arguments are called higher-arity facts.

**Example 4 (Fact)** Based on the relation examples aforementioned and evidence from the following sentence:

1. "Chest pain is a classic sign of Myocardial Infarction."
2. "Google acquired Nest for a price of 3.2 billion U.S. dollars in January 2014."

could be expressed by the facts

1. $\mathsf{SymptomOf}(\mathsf{Chest\ Pain}, \mathsf{Myocardial\ Infarction})$
2. $\mathsf{CompanyAcquired}(\mathsf{Google}, \mathsf{Nest}, 201401, \$3.2\mathsf{billion})$

It is often the case that facts are incomplete, i. e. they miss certain arguments. This leads to partial facts:

**Definition 5 (Partial Fact)** *A partial fact is a fact where some arguments are unknown. Unknown arguments are specified as variables, for instance* $\mathsf{R}(\mathsf{a}_1, \mathsf{a}_2, \mathsf{X}_3, \mathsf{a}_4)$ *with variable* $\mathsf{X}_3$*. Logically, this denotes a formula* $\exists \mathsf{X}_3 : \mathsf{R}(\mathsf{a}_1, \mathsf{a}_2, \mathsf{X}_3, \mathsf{a}_4)$

**Example 5 (Partial Fact)** Considering the evidence from the following sentence:

- "Amazon bought Whole Foods in a deal valued at USD 13.7 billion."

we can only partially instantiate the $\mathsf{CompanyAcquired}$ relation

- $\mathsf{CompanyAcquired}(\mathsf{Amazon\ Inc}, \mathsf{Whole\ Foods}, X_3, \$13.7\mathsf{bln})$

We introduce facts of a special binary relation to state that an entity belongs to a particular type.

**Definition 6 (Type Facts)** *Instances of the binary relation* $\mathsf{Type}(\mathcal{E}, \mathcal{T})$ *assign entities from set* $\mathcal{E}$ *to types from set* $\mathcal{T}$.

**Example 6 (Type Facts)** The following facts express some of the type statements from Example 2:
$\mathsf{Type}(\mathsf{Amazon\ Inc}, \mathsf{Internet\ Companies})$,
$\mathsf{Type}(\mathsf{Amazon\ Inc}, \mathsf{NASDAQ\ Companies})$,
$\mathsf{Type}(\mathsf{Myocardial\ Infarction}, \mathsf{Heart\ Diseases})$

Finally, we can formally introduce Knowledge Bases as follows:

**Definition 7 (Knowledge Base)** *A Knowledge Base* $\mathcal{K}$ *is defined as a 4-tuple* $(\mathcal{E}, \mathcal{T}, \mathcal{R}, \mathcal{F})$*, where* $\mathcal{E}$ *is the* KB's entity repository, $\mathcal{T}$ is a set of possible types, $\mathcal{R}$ is a predefined set of relations with type signatures, and $\mathcal{F}$ is the collection of all its facts.

It should be mentioned, that $\mathcal{F}$ also covers entity types stated as type facts. A few prominent example KBs are described in Chapter 2.1.2.

### *Data Model*

As a machine-readable representation language proposed by the World Wide Web Consortium (W3C) the Resource Description Framework (RDF) framework is the de facto standard to encode data about any subject on the web and forms the foundation of the Semantic Web [47]. Thus, it is the data model of choice for most contemporary KBs. RDF represents an entity as an resource identified by a unique International Resource Identifiers (IRIs). Across the web Uniform Resource Locators (URLs) present a common and valid choice for IRIs to globally identify resources [3]. Instead of fully written out IRIs, namespaces can be introduced to capture prefixes enabling a shorter and more comprehensible notation as well as the definition of vocabularies.

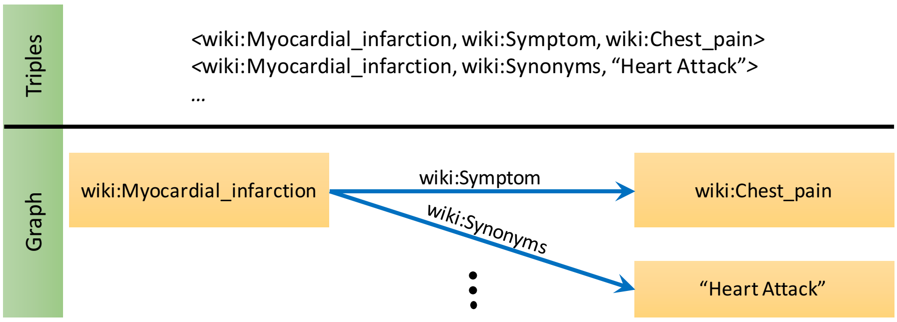

Triples
<wiki:Myocardial_infarction, wiki:Symptom, wiki:Chest_pain>
<wiki:Myocardial_infarction, wiki:Synonyms, "Heart Attack">
...
Graph
wiki:Myocardial_infarction
wiki:Symptom
wiki:Chest_pain
wiki:Synonyms
"Heart Attack"

Figure 2: Example RDF Triples

**Example 7 (RDF Resources)** The Wikipedia URL

- `http://wikipedia.org/wiki/Myocardial_infarction`

could be used to represent the disease *Myocardial Infarction*. By introducing the wiki prefix as

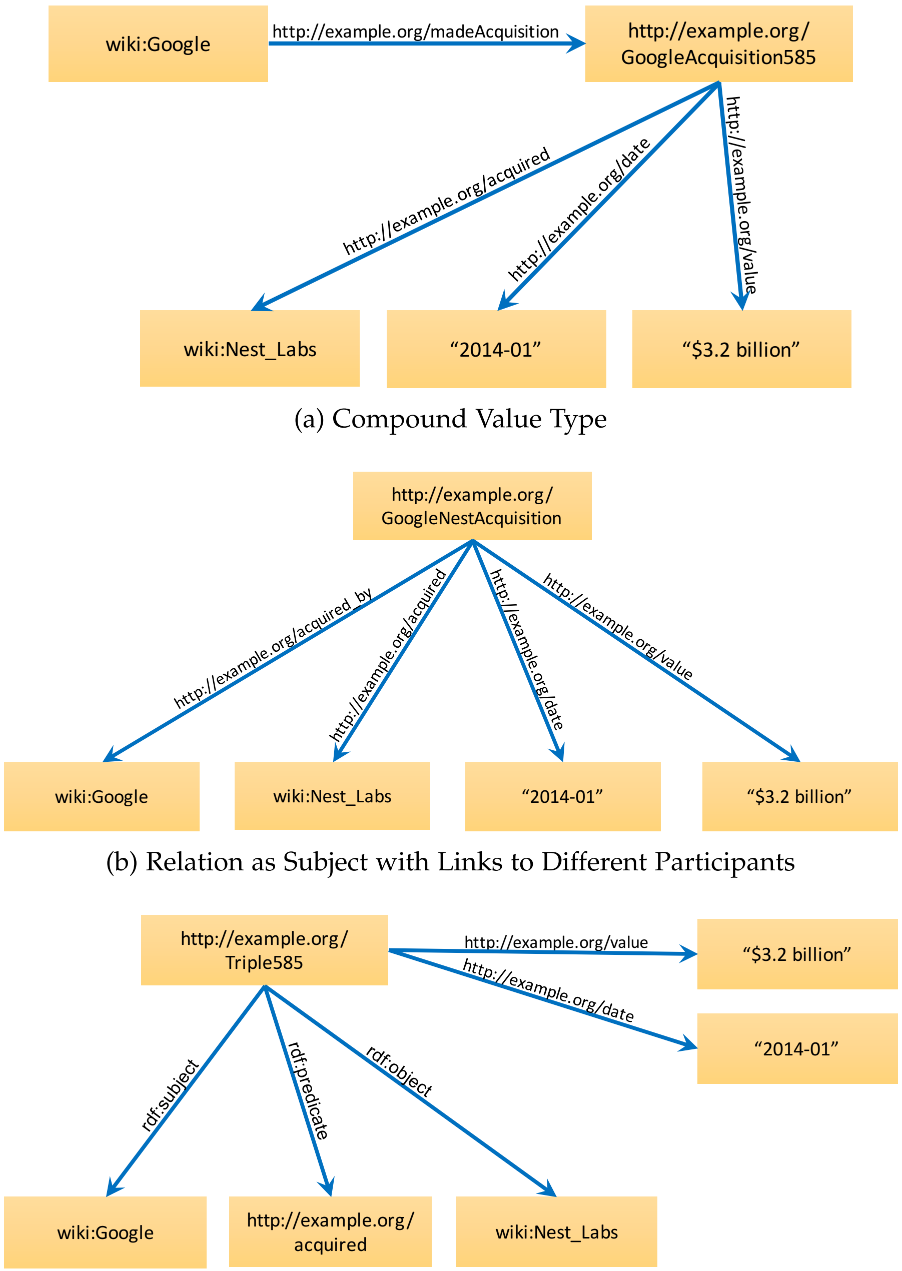
- `http://wikipedia.org/wiki/`

we can refer to same entity as *wiki:Myocardial_infarction*.

Blank nodes are unnamed resources without identifier and are commonly used to state unknown or anonymous values. Literals encode basic values, such as numbers, strings (e. g. "Heart Attack" in Figure 2), dates, etc. and can be constrained by datatypes. RDF links entities via relations with properties or other entities in the form of Subject-Predicate-Object (SPO) triple statements <*S*, *P*, *O*>, where

- the subject S can be any kind of resource,
- the predicate P represented as resource describes a relation or property between S and O,
- the object O is either any kind of resource or a literal.

Stating binary facts as triple statements is straightforward: the fact's first argument becomes the subject, the relation is used as predicate, and the second argument serves as object. For instance, besides plain text encoding, triples can also be viewed as single edges of a graph. Figure 2 shows an RDF encoding of the binary fact mentioned in Example 4 as plain text and graph representation. To state higher-arity facts in RDF usually four different techniques are applied, whereby three of them are also shown in Figure 3 and are representing the higher-arity fact from Example 4:



(a) Compound Value Type

(b) Relation as Subject with Links to Different Participants

(c) Reification

Figure 3: Possible Higher-arity Fact Representation in RDF

- In case we can identify one argument of a higer-arity fact as the "main actor" or "owner" of the relation, this main argument becomes the triple's subject, the relation becomes the predicate, and remaining arguments are encapsulated in a complex object capturing all different properties, aspects, or values. This strategy of encoding higher-arity facts is sometimes also called Compound Value Type (CVT). A corresponding example is depicted in Figure 3a.

- If such a main argument does not exist and all arguments participate in the relation, the relation becomes the main entity and serves as subject. Every argument of a fact serves as object of a new triple; each with a designated RDF predicates. Figure 3b shows an example graph of such a representation.

- Reification can be used if two arguments of a higher-arity fact can be identified as main arguments and all other arguments only state metadata or provenance information. Conceptually, we reduce the higer-arity to a binary fact between the main arguments, which is identified by an IRI resulting in a RDF resource reflecting an entire binary fact. This new resource is used to state further information about the fact, allowing us to create statements about statements, i. e. triples about triples. The resulting graph of such an encoding is depicted in Figure 3c.

- In the special case that arguments of a relation form an ordered list or sequence of arguments, RDF provides a special collection construct for the lists in such facts.

Even though some KBs such as YAGO represent higher-arity facts using reification, it is actually only intended for stating metadata or provenance information and thus the W3C recommends the other aforementioned approaches [125].

**Definition 8 (Knowledge Base Graph)** *A graph of a Knowledge Base $\mathcal{K}$ is the graph generated by all its facts encoded as* RDF triples forming the single arcs of the graph.

### 2.1.2 *Prominent Knowledge Bases*

Contemporary Knowledge Bases provide an extensive collection of facts linking a broad spectrum of knowledge. They have become important assets for numerous application domains, such as semantic search, analytics, and smart recommendations, etc. There are two big ecosystems which offer and maintain access to various comprehensive Knowledge Bases and RDF datasets of such flavor:

| Topic | Datasets | % |
|---|---|---|
| Government | 183 | 18.05% |
| Publications | 96 | 9.47% |
| Life Sciences | 83 | 8.19% |
| User-generated Content | 48 | 4.73% |
| Cross-domain | 41 | 4.04% |
| Media | 22 | 2.17% |
| Geographic | 21 | 2.07% |
| Social web | 520 | 51.28% |
| Total | 1014 | |

Table 1: LOD Datasets Overview[1]

- The Linked Open Data (LOD) Cloud contains thousands of interlinked KBs. The cloud already offers access to billions of triples, while its size is still continuously growing. The interconnected knowledge ranges over several domains, which are listed in Table 1.

- For the biomedical domain, the National Center for Biomedical Ontology (NCBO) maintains BioPortal, a web portal currently (2017, July) hosting 583 biomedical KBs with 8,130,260 entities [117]. Furthermore, a powerful user interface and various web services allow users to not only browse, search, and visualize the KBs in an abundance of ways, but also automatically annotate user provided text with entities [195].

Due to their large scope, versatility and generality a few central KBs form nuclei within these two extensive repositories, thus being especially suitable for bridging different domains and heavily used in a broad range of applications:

*YAGO*

Yet Another Great Ontology (YAGO) [73] was released in 2007 by the Max Planck Institute for Informatics and is now a joint project with Télécom ParisTech University. YAGO covers more than 10 million entities and more than 120 million facts about these entities. It is derived from an integration of Wikipedia, WordNet and GeoNames. Every Wikipedia article corresponds to a YAGO entity, whereas Wikipedia categories are

[1] http://lod-cloud.net/state/state_2014

leveraged as type information. A unique feature of YAGO is its clean type system achieved by linking Wikipedia type information to the WordNet taxonomy. Every entity and fact has a unique id, which can be used to formulate facts as well as facts about facts, so called meta-facts. Facts are harvested from Wikipedia infoboxes and Wikipedia category pages. Meta-facts are often used to embed facts into their textual context or on the spatio-temporal dimension. YAGO offers a 6-tuple representation to model such facts: its SPOTLX encoding, an RDF-like SPO representation extended by Time, Location, and conteXt. Reification is applied to encode the SPOTLX facts in an RDF compliant data model. YAGO covers more than 10 million entities, around half a million types, and more than 120 million facts about these entities. A manual evaluation of YAGO confirmed an overall precision of 95%.

*DBpedia*

Maintained by an open-source community project DBpedia [89] is also extracted from Wikipedia. Wikipedia articles represent entities and infoboxes are the major sources for facts. The main difference between YAGO and DBpedia lies in their strategies regarding coverage and quantity. DBpedia wants to stay very close to Wikipedia with the goal of providing an RDF version of Wikipedia, whereas YAGO's main goal is to create a consistent KB with high precision ensured by consistency constraints ruling out conflicting facts. Another difference is YAGO's inclusion of the WordNet taxonomy. Currently, DBpedia entails 6.6 million entities and around 167 million facts. RDF triples.

*Freebase*

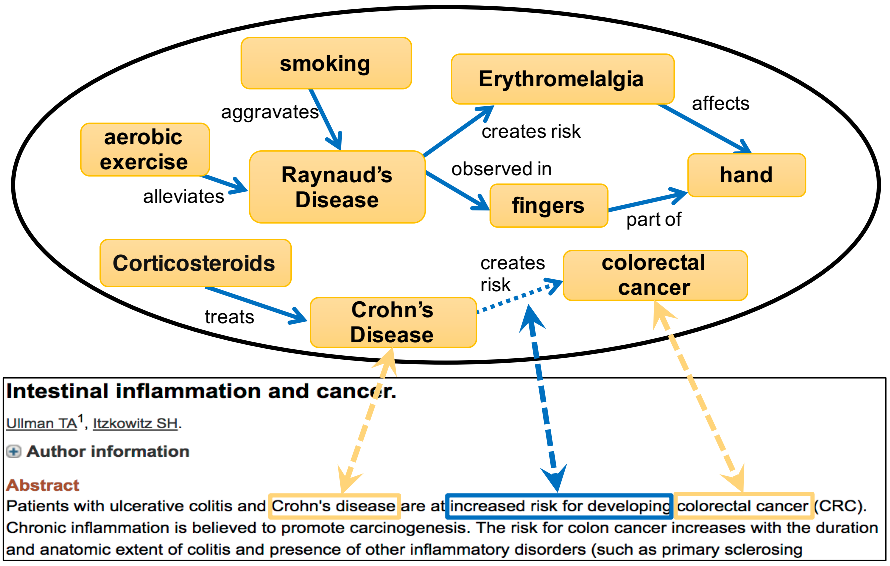
Originally released by Metaweb in 2007 and later acquired by Google in 2010, Freebase [19] was an extensive KB covering more than 3 billion facts about almost 50 million entities. Similar to YAGO and DBpedia, large parts of Freebase are derived from Wikipedia. However, Freebase also includes other sources, such as parts of MusicBrainz or direct input from a user base. If entities are added to Freebase, they are identified with special Machine Identifierss (MIDs), stable and consistent identifiers throughout an entity's lifetime. Binary facts are encoded as triples, whereas higher-arity facts are encoded using so called Compound Value Types (CVTs) (described in Section 2.1.1). In 2014 Google decided to focus on its proprietary Knowledge Base (KB) – the Google Knowledge Graph – and shut down Freebase.

### *Wikidata*

Wikidata [188] has been started in October 2012 by the Wikimedia Foundation as a collaborative KB. Entities in Wikidata are modeled as items, which are identified by special, unique identifiers, so called *qids*. Statements are represented as claims, which besides a fact also include the source of the information. By relying on this model it is possible to express contradictory facts in Wikidata without creating inconsistencies. For example, a border conflict can be expressed from different sources expressing different political points of view. Currently, Wikidata has 17,612 active contributors, and covers ca. 40 million items with 330 million statements.

### *Unified Medical Language System (UMLS)*

Created by the United States National Library of Medicine (NLM) the Unified Medical Language System (UMLS) [18] is an extensive, multi-lingual knowledge source of biomedical entities. It covers 3,221,702 entities with 12,842,558 entity names by integrating source vocabularies, i. e. the electronic versions of numerous thesauri, classifications, code sets, and lists of controlled terms used in patient care, health services billing, public health statistics, indexing biomedical literature, and/or basic, clinical and health services research, from different biomedical domains into a coherent metathesaurus. Notable source vocabularies are the Foundational Model of Anatomy (FMA), Medical Subject Heading (MeSH), Gene Ontology (GO), SNOMED Clinical Terms among others. Beyond its metathesaurus, UMLS entails a Semantic Network spanning over its entities and categorizing the entities into 127 different Semantic Types. On top of this network McCray et al. [103] created a set of Semantic Groups grouping the Semantic Types into consistent domains. All together, this results in a wide coverage and categorization of entities, i. e. entities about diseases, anatomy, genes, treatments, etc. However, the UMLS semantic type system is shallow: it only assigns 127 distinct types to more than 3 million entities. Even though the metathesaurus covers entities of all biomedical domains, it only provides a small amount of links between them.



**Intestinal inflammation and cancer.**

Ullman TA[1], Itzkowitz SH.

⊕ Author information

**Abstract**

Patients with ulcerative colitis and Crohn's disease are at increased risk for developing colorectal cancer (CRC). Chronic inflammation is believed to promote carcinogenesis. The risk for colon cancer increases with the duration and anatomic extent of colitis and presence of other inflammatory disorders (such as primary sclerosing

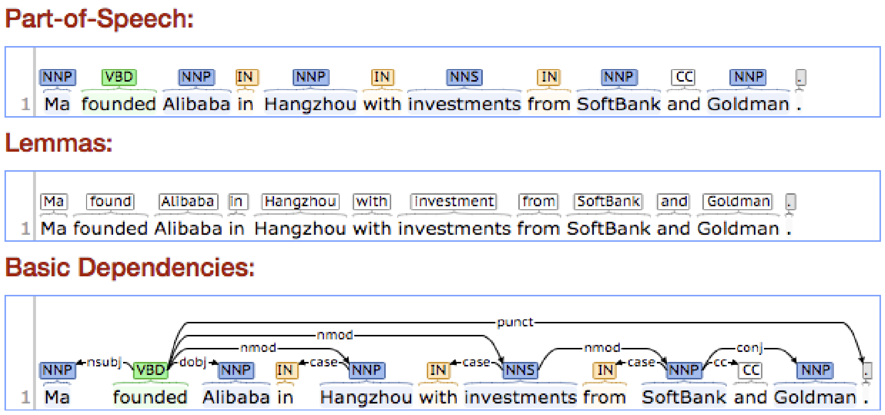
Figure 4: KBC Example

## 2.2 KNOWLEDGE EXTRACTION FROM TEXT

Understanding the knowledge and information expressed in natural language text is a hard task dating back to early days of Artificial Intelligence (AI). The goal is to capture the semantics expressed in human language in order to transfer unstructured and naturally expressed information within text to a structured or formal representation, which computers can process and reason with to deduce new information. The main challenge lies in coping with the ambiguity prevalent in natural language. Two common subproblems of this challenge are the recognition and disambiguation of entities and the extraction of concise logical facts between them from text documents. The gained knowledge is valuable for many tasks, such as Question Answering (QA), Information Retrieval (IR), etc., but especially for Knowledge Base Construction (KBC) which lies at the core of this thesis.

**Definition 9 (Knowledge Base Construction)** *Knowledge Base Construction* (KBC) is the process of populating a Knowledge Base with entities, facts or rules harvested from large amounts of input data. Popular input data includes text corpora, websites, databases, web tables, among others.

Figure 4 shows an example of the rich information, that can be found in text, and how it is useful to populate a KB. In the next sections three major research fields dealing with different types of ambiguity are introduced which are core components for most KBC pipelines: Natural Language Processing (NLP), Named Entity Recognition and Disambiguation (NERD), and fact harvesting.



Part-of-Speech:

1 Ma founded Alibaba in Hangzhou with investments from SoftBank and Goldman .

Lemmas:

1 Ma founded Alibaba in Hangzhou with investments from SoftBank and Goldman .

Basic Dependencies:

1 Ma founded Alibaba in Hangzhou with investments from SoftBank and Goldman .

NNP VBD NNP IN NNP IN NNS IN NNP CC NNP .
Ma found Alibaba in Hangzhou with investment from SoftBank and Goldman .
nsubj
dobj
nmod
nmod
punct
case
case
nmod
case
conj
cc

Figure 5: Stanford CoreNLP[98] analysis of "Ma founded Alibaba in Hangzhou with investments from SoftBank and Goldman."

### 2.2.1 *Natural Language Processing*

Natural Language Processing (NLP) is a subfield of AI concerned with computational methods for the linguistic analysis and understanding of human language. A wide spectrum of tasks, such as Machine Translation, Question Answering, Sentiment Analysis among others, are captured under the umbrella of NLP. Besides these tasks, resolving structural ambiguity, i. e. finding the correct grammatical interpretation of human text, is an important aspect of NLP, especially in the context of this thesis. The following core tasks involving the lexico-syntactic analysis of words and structure of natural language are most relevant:

TOKENIZATION: Segment a given text into a sequence of tokens (e. g. words, symbols, etc)

SENTENCE SEGMENTATION: Detect sentence boundaries within a sequence of tokens

PART OF SPEECH TAGGING: Categorize tokens into different linguistic groups, viz. Parts of Speech (POS) such as noun, verb, adjective, etc., sharing similar syntactic properties

MORPHOLOGICAL NORMALIZATION: Identify morphological base forms of tokens, e. g. remove inflection or determine lemmas for words

SYNTACTIC PARSING: Determine the grammatical structure of and dependencies between sequences of tokens in order to form a parse tree, e. g. aggregating groups of words to phrases or identifying subject and object of a verb

The Stanford CoreNLP software [98] is a popular and sophisticated suite of language technology tools which performs these tasks. Figure 5 shows an example analysis of the software. It often serves as the foundation for many text understanding applications as well as the approaches presented in this thesis.

### 2.2.2 *Named Entity Recognition and Disambiguation*

Named Entity Recognition and Disambiguation (NERD) methods face the challenges of detecting the mentions of entities in natural language text (Recognition) and of resolving the ambiguity of these mentions to canonical entities (Disambiguation). More specifically, Named Entity Recognition (NER) first extracts a set of boundaries for potential entity mentions from text. To disambiguate an ambiguous mention, Named Entity Disambiguation (NED) retrieves a list of candidate entities from a KB and ranks them using a combination of features [165], like

ENTITY POPULARITY: Based on prior probabilities or rankings this feature states how prominent an entity is, for instance with regard to a given candidate mention or globally in a corpus.

TEXTUAL CONTEXT: These features determine the similarity between the textual context of a particular mention and the associated textual information of an entity, derived from documents annotated with a particular entity or textual information extracted from a KB.

COHERENCE: Coherence measures the topical likeliness of a candidate entity with other entities occurring together in a chunk of text.

The top-ranked entity is usually picked as result.

**Example 8 (NERD)** Given the sentence

- "Ma founded Alibaba in Hangzhou with investments from SoftBank and Goldman."

Mature NERD systems would be able to infer that the terms Ma, Alibaba, and Hangzhou are mentions and mean the person Jack Ma, the Alibaba company, Hangzhou the capital of China's Zhejiang province, etc. which could be represented in YAGO.

For the general domain, there exist mature and versatile software which perform NERD with high accuracy, such as AIDA [72], Illinois Wikifier [148], and TAGME [55], to name a few.

Tapping in specific domains, such as Biomedicine, Named Entity Recognition and Disambiguation, here often coined BioNERD, is still a challenge. Mature and versatile approaches able to perform this task with high precision between genes, diseases, symptoms, anatomical concepts, etc. are still missing.

**Example 9 (BioNERD)** In the sentence

- "Potential effects of APRIL on HeV nuclear shuttling and gene expression regulation are discussed."

a BioNERD system needs to recognize that besides HeV nuclear shuttling, and gene expression regulation, APRIL is an entity mention for the month *April*, the gene *ANP32B* or the lab procedure *A Proliferation-Inducing Ligand Measurement* to then pick *ANP32B* as the correct disambiguation in this case.

The de facto standard tool suite for this task, Metamap [6], employs computationally expensive NLP methods, which results in low throughput and consequently creates a bottleneck within text mining pipelines processing large text corpora. Therefore, in this thesis we rely on customized methods inspired by Siu [167] to perform Biomedical Named Entity Recognition and Disambiguation.

### 2.2.3 *Fact Harvesting*

Fact harvesting methods aim to identify mentions of relation instances in text documents to harvest new facts. A relation mention is a sentence or piece of text expressing a relation between a tuple of entities, i. e. a fact as introduced in Definition 4. Besides domain-independent fact harvesting, there are also domain-specific approaches, which, for instance, harvest biomedical factual knowledge from text.

**Example 10 (Fact Harvesting)** Given the sentence

- "In 2017, Whole Foods was snapped up by Amazon in a deal valued at USD 13.7 billion."

as textual evidence, an advanced fact harvesting system detects that the passive voice form of "snapped up" expresses an acquisition between two companies, with a specific value indicated by "in a deal valued at", and on a specific date denoted by the preposition "in" to harvest the higher-arity fact

- $\mathsf{CompanyAcquired}(\mathsf{Amazon\ Inc}, \mathsf{Whole\ Foods}, 2017, \$13.7\mathsf{bln})$

By analyzing the sentence

- "Chest pain is a classic sign of Myocardial Infarction."

a biomedical fact harvesting system could extract the binary fact

- $\mathsf{SymptomOf}(\mathsf{Chest\ Pain}, \mathsf{Myocardial\ Infarction})$

In general, the two main types of features for fact harvesting are:

TEXTUAL CONTEXT: Typical textual features include some form of pattern extracted between recognized entities. They are based on the pattern-fact duality: principle [23]: given a set of patterns with high coverage and low error rate, we can construct a very good approximation to a set of target relations simply by finding all matches to all the patterns. Conversely, given a good set of relation instances, we can build a good set of patterns by finding all relation occurrences in a text corpus and discovering similarities in the occurrences. Patterns are often determined from a sequence of terms occurring between entities or from paths identified in dependency trees or document structures (e. g. HTML Document Object Model (DOM) trees). They can incorporate lexical information (e. g. words, POS tags, etc.), structural information (e. g. grammatical relations, document structure mark-up, etc.) as well as semantic information (e. g. entity ids or entity types occurring within a pattern).

SEMANTIC ASSETS: These features are usually formulated based on expert knowledge or inferred from a KB. They constrain the set of possible fact candidates by applying type signatures, mutual exclusion constraints, or other consistency rules.

One key feature of contemporary fact harvesting systems is exploiting redundancy in large text corpora. Instead of purely relying on the limited evidence of a single occurrence, they try to aggregate the evidence across all candidate occurrences witnessing the same fact in the input corpus. Often, the rationale is that every piece of text sharing the same entity tuple is highly likely to bear the same fact and thus expresses the same relation. Approaches are categorized based on their required prior knowledge:

### *Unsupervised Fact Harvesting*

Unsupervised approaches, often called Open Information Extraction (OIE) [10], just rely on a large input corpus to extract a large set of relation instances. Usually, these approaches still require some manual input in the form of heuristic rules or seeds to learn relation-independent extraction patterns to constrain the set of extractions. Clustering algorithms are often at the core of such approaches to aggregate fact candidates occurrences in text. However, OIE often suffers from uninformative fact extractions, since a canonicalization of the facts' relations and entities is not possible without prior knowledge about them, such as a unified entity dictionary or relation definitions specifying designated names, fixed arities, type signatures, etc. To distinguish such extractions from crisp facts, the result of OIE systems are often called relation tuples, substituting the logical relation predicates of facts with relational phrases. These phrases are potentially ambiguous and noisy textual representations of relation predicates.

### *Supervised Fact Harvesting*

Supervised methods formulate the task as classification problem and take advantage of a predefined set of relations together with large amount of training data supplied as text corpus with relation and fact annotations. Often the labeled text data is combined with semantic information inferred from a KB. Feature-based, kernel and DeepLearning methods are often employed as classifiers. They usually yield high accuracy, but incorporating new relations is cumbersome, since training data requires an exhaustive amount of manually labeled positive and negative examples for each relation.

### *Distantly Supervised Fact Harvesting*

Distantly supervised methods drop the requirement of huge amounts of labeled training data, instead they leverage a small amount of facts as seeds, which we define as follows:

**Definition 10 (Seed Fact)** *A seed fact for a relation is a relation instance presumed to be true based on expert statements or asserted by a* KB.

Based on the assumption, that every piece of text that mentions the entities of a seed fact, expresses the fact and thus the relation, it is possible to extract a very large number of potentially noisy training data given a large input corpus [23]. We call sentences, containing such mentions, seed sentences.

**Definition 11 (Seed Sentence)** *A seed sentence for a relation* $R$ *is a sentence that contains arguments of some seed fact of* $R$.

It is possible that a seed sentence might match a seed fact only partially, so that it contains a partial fact. Example 5 presents such a case. All seed sentences of a particular relation constitute the relation's training data. The challenge for distantly supervised methods resides in the noise of the training data, since it is not assured that every seed sentence really expresses a particular relation. To tackle this challenge, approaches usually perform statistical scoring to assign quality score to fact candidates followed by a pruning phrase, which constraints the set of plausible fact candidates (see Section 3.1.1).

# 3 RELATED WORK

In the following sections, we discuss relevant works, related to the contributions of this thesis. To allow for a better overview, we split this chapter into two parts: Section 3.1 presents general-domain KB construction methods, which do not target a specific use case. In Section 3.2, we discuss approaches, which target biomedical use cases and are applied to biomedical data.

## 3.1 GENERAL-DOMAIN KNOWLEDGE BASE CONSTRUCTION

### 3.1.1 *Fact Harvesting*

*Unsupervised Approaches*

The KnowItAll project [52] has been the origin for a number of open information extraction approaches. TextRunner [10] established a general framework for other works within the project: By applying heuristics on a small training corpus, relation-independent training data is gathered for building classifiers, constraining the set of accepted training instances. In a second step, using these classifiers for labelling candidates as trustworthy an extractor harvests relational tuple candidates from a large text corpus. Exploiting redundancy within the set of harvested candidates, an assessor assigns confidence scores to each candidate extraction and retains extractions with high confidence. The Wikipedia-based Open Extractor (WOE) [202] uses heuristic matches between Wikipedia Infoboxes and corresponding sentences as seeds to identify more meaningful training data than TextRunner. ReVerb [53] improves on TextRunner by introducing lexical and syntactic constraints to avoid incoherent, uninformative, and over-specific extractions, i. e. cases where extracted patterns have no meaningful interpretation. OLLIE [100] extends ReVerb by considering not only verb phrases and by taking the context of extracted tuples into account. High confidence extractions from ReVerb are used as seeds to learn more general syntactic and semantic patterns. Supervised logistic regression classifiers are trained on a curated corpus to detect, if extracted tuples are part of an attributional or conditional statement. Mausam [101] presents a summary of the KnowItAll project together with a discussion of further challenges and opportunities. The project also lead to the publicly available open information extraction system OpenIE 5[1].

[1] https://github.com/knowitall/openie

ReNoun [208] focuses on extracting nominal attributes using noun phrase patterns. It relies on a small set of relational pattern templates to identify seed facts in a text corpus, which are then used to derive extraction patterns.

ClausIE [43] first identifies clauses, minimal units of coherent information within sentence, by mapping grammatical relations from dependency parse trees to clause constituents. Based on the type of a clause, inferred by a decision tree algorithm and heuristics, ClausIE harvests multiple relational tuples from a clause, each exposing different pieces of information. MinIE [61] extends ClausIE by detecting polarity, modality, attribution, and quantities, enabling the removal of unnecessary contextual information and overly specific constituents from clauses to achieve more compact representations.

Stanford's OpenIE [4] relies on a classifier, which learns to extract similar self-contained clauses from dependency parse trees, which are the input for an inference framework to derive relational tuples.

EXAMPLAR [105] and PredPatt [197] employ deterministic, unlexicalized, syntactic rules over dependency parse trees to harvest relational tuples.

*Supervised Approaches*

As described in Section 2.2.3, supervised approaches require labeled data to train classifiers for harvesting facts. The two most used datasets are the SemEval-2010 Task 8 dataset [70] and the NIST Automatic Content Extraction (ACE) task corpus. The SemEval-2010 Task 8 dataset is based on nine relation types between nominals and a tenth type "Other" when there is none of these relations between two nouns. It contains 8,000 sentences for training, and 2,717 for testing. ACE defines a hierarchy of relations ranging from coarse-grained relations, such as employment/membership, geo-political entity affiliation, etc., to further refined relations.

Feature-based methods use a set of features manually selected by humans to build statistical models. Kambhatla [78] trained Maximum Entropy models using a combination of features derived from words, entity types and dependency parse trees. Zhou et al. [216] relies on a Support Vector Machine (SVM) classifier and extends the previous work by building feature vectors, including semantic information and adding chunking information as syntactic features. Incorporating these features, Li and Ji [93] perform a joint extraction of entity mentions and facts using structured perceptrons. Zhou et al. [217] extend the perceptron learning algorithm and SVMs to leverage relation hierarchies. Besides such hierarchies Chan et al. [33] leverages further background knowledge, such as type constraints, co-reference information, and Wikipedia derived information, to train regularized averaged perceptrons. Semantic Role Labeling (SRL) [62, 127, 140] aims to map single sentences onto structured frames with slots filled based on the verb-argument structure of a sentence, using supervised learning over fine-grained syntactic and lexical features. SRL methods strongly rely on labeled training data, and are focused on the frame repositories provided by PropBank [126] or FrameNet [9].

Supervised kernel methods are instance-based classifiers, which construct hypotheses directly from labeled data based on similarity measures between hypotheses and instances. Kernel functions define these similarities and are at the core of these approaches. Due to this reason, the contribution of fact harvesting kernel methods usually lie in the definition of more sophisticated kernel functions, and not in the design of new classifiers. In that sense, Zelenko et al. [210] introduce a kernel over shallow parse representations of text, Culotta et al.[42] rely on a kernel computing the similarity between two dependency parses and Bunescu et al. [26] exploit the similarity between two shortest paths occurring in dependency graphs. Mooney et al. [112] generalize this shortest path kernel to a kernel incorporating the similarity between multiple paths and subsequences within a dependency parse. Other approaches define kernels on different kinds of subtrees [144, 218], and differ in their spans depending on the context-sensitive information they include.

Lately, supervised DeepLearning approaches, which learn data representations automatically and hence do not require manually crafted features, have been introduced for fact harvesting. Such approaches train neural networks, composed of a cascade of multiple layers with the goal of learning a potentially non-convex function, e. g. a relation classifier. Different network architectures and different types of input data have been investigated to perform fact harvesting. We refer to [64] for an overview of DeepLearning networks, learning algorithms and their applications. Socher et al. [170] train a recursive neural network to harvest facts for the SemEval-2010 Task 8 relations. Besides surface words, they rely on WordNet hypernyms, POS, and NER tags of words as input vectors to classify if a sentence expresses a particular relation between two pre-annotated entities. Zeng et al. [211] drop the computationally expensive syntactical analysis of sentences and just rely on word, position and WordNet hypernym embeddings to train a Convolutional Neural Network (CNN). Also relying on CNNs, but dropping WordNet hypernyms to remove all features derived from lexical resources, Dos Santos et al. [160] learn class representations for every relation. Classification is performed by scoring and ranking the similarity of the vector representation of a new sentence to the class representations. Wang et al. [191] and Xiao et al. [203] introduce attention mechanisms to CNNs to determine which parts of the sentence are most influential for harvesting facts. Xu et al. [207] apply recurrent neural networks with Long Short Term Memory (LSTM) as network architecture to classify relations between two entities in a sentences. They embed information from the shortest dependency path between entities, such as words, POS tags, grammatical relations, as well as linguistic information, like WordNet hypernyms of words. Cai et al. [30] build a Convolutional Neural Networks and recurrent neural networks with LSTM with the goal of better capturing the information within shortest dependency paths.

*Distantly Supervised Approaches*

Dual Iterative Pattern Relation Expansion (DIPRE) [23] introduces the idea of harvesting facts from web documents using distant supervision and exploiting redundant appearances of textual patterns for finding good fact candidates. Snowball [1] extends DIPRE by including type constraints into extraction patterns and by introducing confidence measurements to score and find selective extraction patterns, i. e. patterns with high precision and recall. LEILA [175] is one of the first systems that uses a dependency parsing to determine more robust patterns, which are bound by counterexamples and generalized by machine learning.

Mintz et al. [106] use Freebase facts for distant supervision to detect seed occurrences in Wikipedia articles. Based on a number of predefined syntactical, lexical, and type features, they train a logistic regression classifier from these occurrences to harvest more facts. Based on the same features, Riedel et al. [152] train a probabilistic graphical model and cast classifying relations into a multi-instance classification problem. Their model is trained by constraint-driven semi-supervision, which learns parameters, which do not only optimize the prediction of target labels, but also satisfy user-defined constraints. By statistically modelling overlapping relations, MultiR [75] further enhances this graphical model.

Krause et al. [85] and Li et al. [92] apply a distantly supervised approach for learning extraction rules for relations from dependency graphs. Sar-graphs [86] aggregate this style of rules and incorporate lexical knowledge to construct a re-usable linguistic resource.

Combining pattern-based extractions with any kind of semantic constraint beyond just type constraints, SOFIE [176] presents a general logical framework constraining the space of fact candidates with consistency constraints formulated as Horn rules. PROSPERA [118] improves SOFIE by deriving richer patterns from ngram-itemsets and by computing more informative weights to judge the quality of patterns.

Starting from an initial definition of categories and target relations together with a seed examples the Never-Ending Language Learning system (NELL) project [107] has continuously crawled and read the Web since January 2010 with the goal of constructing a KB. NELL consists of many different learning tasks, represented as one or more functions performing category classification, entity resolution, fact harvesting, etc. To learn extraction patterns for relations the results of every learning function are coupled by constraints [31], e. g. type signatures are coupling constraints between category and relation classifiers.

Wikipedia completion methods, like Kylin [201] and LUCHS [74], match values and entities from Wikipedia Infoboxes with article sentences to heuristically produce training data. Hence they focus on predicting infoboxes for articles lacking such information, these systems are restricted to Wikipedia. Using a two step approach, an article classifier predicts a infobox schema to then pick and apply a small set of relation extractors.

DeepDive [166] is an end-to-end framework for building KBC systems and adapts Markov Logic to formulate constraints and inference rules. The CoType system [149] jointly harvests typed entities and facts by formulating a joint optimization problem, which constraints the learning of embeddings from text corpora and knowledge bases. Relation classification is performed by nearest neighbor search in the embedding space. Zeng et al. [212] shows the feasibility of Piecewise Convolutional Neural Networks for harvesting facts in a distantly supervised setting, taking the uncertainty of training data into account. They rely on word and position embeddings as input.

### 3.1.2 *Knowledge-based Applications*

The easy access to and extensive coverage of contemporary KBs (see Section 2.1.2) as well as the availability of sophisticated knowledge extraction approaches (see Section 2.2) lead to new opportunities for the development and advancement of text-centric applications. Especially, IR systems, such as search and exploration engines, benefit from the large amounts of structured knowledge and the improved semantic understanding of natural-language text [11, 44]. The semantic search system Broccoli [12] improves the retrieval of Freebase entities by performing NERD on Wikipedia to find descriptive text descriptions for entities, which help for query formulation. Based on YAGO, STICS [71] is an entity-aware search engine for news. User can formulate queries based on a combination of text phrases, KB entities and categories, which are automatically expanded to individual entities. Using the queries in combination with a NERD component (AIDA [72]) for annotating incoming news allows STICS to improve the retrieval and ranking of relevant documents. XKnowSearch! [213] exploits the multilingual contents of DBpedia and language agnostic NERD, and hence is able to perform multilingual and cross-lingual IR. `SemanticScholar.org` [185, 204] is an academic search engine based on a customized KB, i. e. a combination of Freebase with concepts inferred from corpus-extracted keyphrases, and a large corpus of scientific publications, annotated with entities. It matches query and

| Corpus | Size | Annotations |
|---|---|---|
| MEDLINE | 24 Million Journal Abstracts | MeSH Document Annotations |
| PubMed Central | 4.1 Million Full-text Articles | MeSH Document Annotations |
| GENIA [80, 81] | 1,999 MEDLINE abstracts | Entity and PPIs |
| BioInfer [141] | 1100 Sentences (MEDLINE) | Gene, Protein, RNA and PPIs |
| AImed [27] | 225 MEDLINE Abstracts | Protein and PPIs |
| LLL [121] | 80 Sentences (MEDLINE) | Gene, Protein, and Gene Interactions |
| IEPA [45] | 200 Sentences (MEDLINE) | Protein, PPIs |
| DrugDDI [162] | 579 DrugBank Documents | Drug and Drug Drug Interactions |
| PubMed45 [59] | 45 MEDLINE Abstracts | Abstract Meaning, Protein, PPIs |
| BioProp[2] | 500 MEDLINE Abstracts | Semantic Role Labels |

Table 2: Biomedical Text Corpora

documents on their entity embeddings incorporating semantics from its KB to improve ranking and retrieval of documents.

The exploration systems InstantEspresso [163], EVELIN [171], and Whoyouelect.com [58], leverage large entity graphs, which are automatically mined from extensive text corpora and are based on co-occurrence statistics. Relying on entities as query inputs all systems summarize the relations between input entities by determining informative subgraphs and retrieve relevant documents or sentences. Users can interact with the subgraphs, which allow for further exploration and analysis.

## 3.2 BIOMEDICAL KNOWLEDGE BASE CONSTRUCTION

Contrary to the general-domain approaches presented in the previous section, the main body of Information Extraction (IE) research in the biomedical domain targets specific use cases, with most approaches focusing on molecular entities and chemo-genomics. For instance, the discovery of gene associations or Protein-Protein Interactionss (PPIs) from scientific literature are among the first and most prevalent use-cases [17, 41]. These efforts have been driven by competitions, like the BioNLP Shared Task (BioNLP-ST) [142] or BioCreative [5].

Due to these tasks and the extensive amounts of biomedical scientific publications various text corpora are available which are usually annotated subsets of two openly accessible repositories of biomedical publications: MEDLINE and PubMed Central database. Table 2 lists the corpora relevant for the works discussed in the following sections. We refer to the survey by Neves [122] for an exhaustive and detailed discussion of biomedical corpora. Since we focus more on the research aspects in the next sections, Table 3 summarizes use-cases and underlying data of the discussed approaches.

[2] https://catalog.ldc.upenn.edu/LDC2009T04

| Supervision | Method | Target Use-case | Corpus | Reference |
|---|---|---|---|---|
| Unsupervised | Clustering | PPI, Gene Suicide Associations | AImed, PubMed Central | Quan et al. [145] |
| | Clustering | None specific | CALBC Corpus | Nebot et al. [120] |
| | Semantic Parsing | PPI | GENIA | USP [135, 136] |
| Supervised | Kernel-based | PPI | AImed, BioInfer, HPRD50, IEPA, LLL | Mooney et al. [112]<br>Airola et al [2]<br>Qian et al. [143, 144] |
| | | Gene Disease Associations<br>Drug Disease Associations<br>Drug Target Associations | AImed, BioInfer, HPRD50, IEPA, LLL | BeFree [22] |
| | Kernel-based, Semantic Parsing | PPI | PubMed45, AIMed | Garg et al. [59] |
| | Feature-based | Treatment Disease Relations | manually annotated PubMed abstracts | Rosario et al. [157] |
| | | Treatment Disease Relations<br>Gene Disease Relations | manually annotated PubMed abstracts | Bundschus et al. [25] |
| | | PPI | GENIA | Riedel et al. [151] |
| | | PPI, Protein Residue Relations | GENIA | Liu et al. [94, 95] |
| | | PPI | GENIA | TEES [15, 16] |
| | | PPI | GENIA | EventMine [108, 109] |
| | | SRL | BioProp | Zhang et al. [214] |
| | | Disease Symptom Relations | Electronic Heath Records | Rotmensch et al. [159] |
| | DeepLearning | PPI | AImed, BioInfer, HPRD50, IEPA, LLL | Zhao et al. [215] |
| | | PPI | AImed, BioInfer | McDepCNN [133] |
| Distantly Supervised | Inductive Logic Porgramming | PPI | Pubmed abstracts | Craven et al. [41] |
| | Markov Logic | Gene Interactions | Full-text PLOS Articles | DeepDive [97] |
| | Probabilistic Graphical Model | PPI, Cancer Pathways | GENIA, PubMed abstracts | Poon et al. [137] |
| | Semantic Parsing | PPI, Cancer Pathways | GENIA, PubMed abstracts | GUSPEE [131] |
| | Predicate-Argument Structure Patterns<br>Type Signatures | PPI, Drug Drug Interactions | AIMed, BioInfer, LLL, GENIA, DrugBank | PASMED [123] |
| | DeepLearning | Drug, Gene, Mutation Relations<br>Gene Regulation Relations | PubMed Central | Peng et al. [132] |

Table 3: Related Work: Fact Harvesting in the Biomedical Domain

### 3.2.1 *Fact Harvesting*

*Unsupervised Approaches*

Early approaches, which have been only based on entity cooccurrence statistics, have been pursued for the discovery of associations between genes and diseases [39, 91]. Wright et al. [200] performs association rule and itemset mining on structured electronic health records to discover links between medications, laboratory results and patient problems.

In order to find PPIs and Gene–suicide associations, Quan et al. [145] adapt polynomial kernels for computing similarities between extracted patterns, which are used for clustering candidate extractions. The versatile approach of Nebot et al. [120] also employs clustering, but additionally constrains possible extractions by lexico-syntatic patterns.

Unsupervised semantic parsing [135, 136] aims to transform dependency trees into logical formulas and clusters them to abstract away syntactic variations. Markov Logic is used to find plausible cluster and argument assignments in order to derive relations and facts.

*Supervised Approaches*

Due to being in the center of many tasks and due to the availability of annotated corpora, biomedical event extraction attracted substantial attention. Event extraction methods aim to identify occurrences of events from a predefined set of event types within a text corpus, including the identification of event triggers, arguments and interrelations. In the biomedical domain, such extraction systems usually target molecular events, like PPIs or gene associations. Often, IE methods in this line of research rely on kernel methods in combination with SVMs. Mooney et al.[112] also test their subsequence kernel, introduced in Section 3.1.1, for extracting PPIs. Airola et al. [2] first construct special graph structures out of the dependency parses and the linear structure of sentences. These graphs are then used as basis for a kernel defining a similarity between sets covering all possible paths between two vertices. Qian et al. [143, 144] use a kernel to extract PPIs, which computes the number of common subtrees of two parse trees. The BeFree system [22] combines two kernels, which incorporate shallow linguistic as well as deep syntactic information for the identification of associations between genes, diseases and drugs. Applying a hybrid method, Garg et al. [59] initially perform semantic parsing to deduce abstract meaning representations for sentences, which are used in combination with a customized contiguous subtree kernel [210] for harvesting biomolecular interactions. Furthermore, their approach aggregates evidences from multiple sentences and does not only rely on observations from single occurrences to harvest facts.

Feature-based methods have been considered as well. Rosario et al. [157] aims to discover seven different relations occurring between treatments and diseases, e. g. prevention, cures, etc., by employing different generative and discriminative graphical models. Bundschuss et al. [25] cast the problem of harvesting facts into a sequence labelling problem. They apply two cascaded conditional random fields defined over orthographic, lexical and dictionary features, i. e. one for NER and one for fact harvesting, to extract gene-disease associations and treatments. Riedel et al. [151] present statistical models for performing joint trigger, argument and correlation prediction of biomedical events, which ensure multiple consistency constraints, such as structural syntactic constraints or consistency between trigger and argument assignments of the same event. Liu et al. [94, 95] learn descriptive subgraphs from training data, which are used as extraction rules to label unseen sentences by applying their own approximate subgraph matching algorithm. The algorithm can cope with syntactical variations between learned graphs and unseen dependency graphs. Furthermore, the authors show that their approach is also applicable in a distantly-supervised setting. Turku Event Extraction System (TEES) [15, 16] and EventMine [108, 109] are two prominent and major developments for event extraction. They are based on intermediary graph representations from dependency parses and cast the relation classification into consecutive graph classification tasks. Rotmensch et al. [159] test three probabilistic models to derive a KB from structured electronic medical records gathered by the Beth Israel Deaconess Medical Center. Their best performing system, based on a importance measure derived from a Bayesian network modeling diseases and symptoms with noisy OR gates, has been used to harvest facts covering disease-symptom relationships.

Zhang et al. [214] adapt SRL for the biomedical domain. The system is trained on BioProp, a manually annotated biomedical SRL corpus.

Lately, DeepLearning systems have also been introduced for discovering PPIs in text. Zhao et al. [215] trains a deep neural network using input vectors encoding single words, POS tags and predicate-argument structures. The McDepCNN [133] system's input vectors represent sentences by concatenating word embeddings, POS tags, chunks, named entities, dependencies, and position features to feed them into a multichannel convolutional network architecture.

### *Distantly Supervised Approaches*

One of the earliest distantly supervised approaches [41] initiates logical predicates from dependency parses and exploits inductive logic programming to learn, when a particular parse leads to a logical fact. Leveraging known drug-disease pairs from clinical trials[3] as seeds,

[3] Clinicaltrials.gov

Xu [206] ranks the patterns most frequently appearing with seeds to manually select the most plausible ones for harvesting treatment facts from scientific publications. PASMED [123] uses predicate-argument structure patterns as constraints to find salient patterns, which are then passed as extraction rules to SemRep [153, 155]. SemRep is a rule-based fact extraction system, incorporating pre-defined textual extraction patterns as well as semantic features, such as type signatures. Poon et al. [137] customize the MultiR system and leverage prior knowledge from the Pathway Interaction Database to extract cancer pathways from PubMed abstracts. Fulfilling the same use-case, GUSPEE [131] learns semantic parses based on a expectation–maximization algorithm, which models spotted seeds as latent variables and incorporates a prior that favors semantic parses containing known events. The DeepDive system introduced in Section 3.1.1 has been adapted to extract gene interactions from over 100,000 full-text Public Library of Science (PLOS) articles [97]. Employing graph LSTM convolutional networks relying on custom document graphs, Peng et al. [132] is able to incorporate intra-sentential and inter-sentential dependencies to perform cross-sentence fact harvesting for discovering ternary relations covering drugs, genes and their mutations as well as binary gene regulation relations.

### 3.2.2 *Knowledge-based Applications*

Two types of applications, which heavily rely on KBC components in the biomedical domain, are disease surveillance systems as well as entity-aware search and exploration engines.

#### *Disease Surveillance Systems*

Disease surveillance systems track disease outbreaks or health hazards in online news and social media. Generally used for disease control and prevention by health organizations and medical institutes, such as the World Health Organization (WHO) or the Robert Koch Institute, their main purpose lies in the detection of emerging and re-emerging epidemics. Choi et al. [38] gives a systematic overview of eleven web-based infectious disease surveillance systems. HealthMap [57] and EpiSpider [79] rely on user-created ProMED reports and do not process documents automatically. Proteus-BIO [65] and the Medical Information System (MedISys) [156] in combination with PULS [174] leverages manually defined lexical patterns to detect diseases and their outbreaks on the Web. They produce a visualization, where outbreaks are linked to relevant source documents. Global Health Monitor [49] identifies disease outbreak events based on occurrence statistics from relevant medical news. A similar approach is used by the multilingual system DAnIEL [90], which extracts disease outbreaks by recognizing disease-location mentions in online news and matching them to a custom multilingual KB created from Wikipedia.

A different type of applications leverage social media for identifying disease outbreaks and spreading. Charles-Smith et al. [35] provide a survey of such systems, which usually rely on Social Media Analysis methods and thus are not further discussed in this thesis.

### *Entity-aware Search and Exploration Engines*

The majority of search systems focus on scientific publications stored in the PubMed repository. Kim et al. [80] use molecular entities for query expansion. The scopes of Textpresso [116], GoPubMed [48], FACTA+ [184], EVEX [186], BioTextQuest+ [130] and CRAB [66] are restricted to genes, proteins, or chemicals. Along these lines, Ferret [173] performs specialized NERD and fact harvesting on PubMed articles to retrieve and rank sentences mentioning gene-centric entities and facts. The system offers entity-aware query formulation and expansion and visualizes results as heat maps. Pang et al. [128] investigate exploratory search capabilities for health content and present the implementation of a slider-based UI for exploring and discovering information in health websites [129]. PolySearch2 [96] goes beyond scientific publications and offers 'Given *X*, find all associated *Y*s' queries, where *X* and *Y* are two types, e. g. diseases, toxins, etc, restraining the search together with optionally provided keywords. MEDIE [111] and GeneView [182] annotate PubMed articles with various kinds of biomedical entities and events, but both systems do not offer interactive real-time exploration and analytics. The exploration system ALIBABA [134] uses the PubMed search interface to graphically visualize information on associations between biological entities, such as proteins, cells, tissues, etc., extracted from retrieved search results. The interactive exploration system Life-iNet [150] system performs NERD using the entire UMLS dictionary and distantly supervised fact harvesting on scientific publications and articles from the Wikipedia health portal. The system provides relation-based exploration, where users can query the knowledge with triple patterns, explore the extracted factual knowledge using a network visualization, or digest entity summarizations for user-provided type queries. DELVE [68] is a modular faceted browser for exploring Pubmed search result, which provides different kinds of visualization, such as word clouds, phrase nets, and word trees, which can also leverage information from KB. Semedico [54] supports interactive entity- and relation-aware query formulation on PubMed articles with annotated entity and fact occurrences. The goal of the Literome [138] is to facilitate browsing, searching and reasoning over extracted genomic knowledge [137] from PubMed articles by allowing users to search for harvested facts using triple patterns or by traversing the extracted KB step by step.

Part II

# KNOWLEDGE-BASE CONSTRUCTION

# 4 KNOWLIFE: A LARGE AND VERSATILE KNOWLEDGE GRAPH FOR BIOMEDICAL SCIENCES

## 4.1 INTRODUCTION

As introduced in Section 2.1.2, large Knowledge Bases (KBs) about entities, their properties and interrelations, have become an important asset for semantic search, analytics, and smart recommendations over Web contents and other kinds of Big Data [177, 209]. The most notable projects along this line are BabelNet, DBpedia, Never-Ending Language Learning system (NELL), Yet Another Great Ontology (YAGO), Wikidata, and the Google Knowledge Graph with its public core Freebase.

In the biomedical domain, KBs such as the Foundational Model of Anatomy (FMA) [158], the Gene Ontology [8], the Disease Ontology [161], and the Unified Medical Language System (UMLS) (see Section 2.1.2) are prominent examples of the rich knowledge that is available in digital form. However, each of these KBs is highly specialized and covers only a focused domain within the life sciences (e. g. either focusing molecular or clinical aspects), resulting in very little inter-linkage between the KBs. This also holds for the Unified Medical Language System (UMLS). Even though its entity dictionary covers entities of all biomedical domains, it only provides a small amount of links between the different domains. Thus, in contrast to the general-domain KBs that power Web search and analytics, it is intrinsically hard to obtain an integrated view on all aspects of biomedical knowledge. The lack of a KB that spans biological, medical, and health knowledge, hinders the development of advanced search and analytic applications in this field, such as the ones presented in Part III of this thesis.

In order to build a comprehensive biomedical KB, the following three bottlenecks must be addressed.

**Dependence on manual curation.** Biomedical knowledge is advancing at rates far greater than any single human can absorb. Therefore, relying on manual curation of KBs is bound to be a bottleneck. To fully leverage all published knowledge, automated Information Extraction (IE) from texts is mandatory.

**Restriction to scientific literature.** Besides scientific publications found in PubMed Medline[1] and PubMed Central[1], there are substantial efforts on patient-oriented health portals such as Mayo Clinic[2],

[1] https://www.ncbi.nlm.nih.gov

[2] http://www.mayoclinic.org

Medline Plus[3], UpToDate[4], Wikipedia's Health Portal[5], and there are also popular online discussion forums such as Healthboards.com or Patient.co.uk. These resources constitute a rich universe of information, which is however scattered across many sources, mostly in textual, unstructured and sometimes noisy form. Prior work on biomedical IE has focused on scientific literature only, and completely disregards the opportunities that lie in tapping into health portals and communities for automated IE.

**Focus on molecular entities.** IE from biomedical texts has strongly focused on entities and relations at the molecular level; a typical IE task is to extract protein-protein interactions. Much less emphasis has been put on comprehensive approaches that link diverse entity types, spanning genes, diseases, symptoms, anatomic parts, drugs, drug effects, etc. In particular, no prior work on KB construction has addressed the aspects of environmental and lifestyle risk factors in the development of diseases and the effects of drugs and therapies.

### *Contributions*

In this chapter we present *KnowLife*, a large KB that captures a wide variety of biomedical knowledge, automatically extracted from different genres of input sources. *KnowLife*'s novel approach to KB construction overcomes and goes beyond the previously mentioned three limitations of prior work.

**Beyond manual curation.** Using distant supervision in the form of seed facts from existing expert-level knowledge collections, the *KnowLife* processing pipeline is able to automatically learn textual patterns and to use them to harvest a large number of relational facts. In contrast to prior work on IE for biomedical data, which relies on extraction patterns only, our method achieves high precision by specifying and enforcing logical consistency constraints that fact candidates have to satisfy. These constraints are customized for the relations of interest in *KnowLife*, and include constraints that couple different relations. The consistency constraints are available as supplementary material (see Appendix a.1). *KnowLife* is easily extensible, since new relations can be added with little manual effort and without requiring explicit training; only a small number of seed facts for each new relation is sufficient.

**Beyond scientific literature.** The *KnowLife* system scales to large text corpora – considering not only knowledge from scientific publications, but also tapping into previously neglected textual sources like Web portals on health issues and online communities with discussion boards. We present an extensive evaluation of 22,000 facts highlighting how these different genres of input texts affect the re-

[3] https://medlineplus.gov
[4] https://www.uptodate.com
[5] https://en.wikipedia.org/wiki/Portal:Health_and_fitness

sulting precision and recall of the KB. In Section 4.8 we present an error analysis that provides further insight on the quality and contribution of different text genres.

**Beyond molecular entities.** The entities and facts in *KnowLife* go way beyond the traditionally covered level of proteins and genes. Besides genetic factors of diseases, *KnowLife* also captures diseases, therapies, drugs, and risk factors like nutritional habits, life-style properties, and side effects of treatments.

In summary, the novelty of *KnowLife* lies in its versatile, largely automated, and scalable approach for the construction of a comprehensive KB – covering a spectrum of different text genres as input and distilling a wide variety facts from different biomedical areas as output. Coupled with an entity recognition module that covers the entire range of biomedical entities, the resulting KB features a much wider spectrum of knowledge and use-cases than previously built, highly specialized KBs. In terms of methodology, our extraction pipeline significantly extends previously proposed techniques, and is specifically customized to the life-science domain. Most notably, unlike prior work on biomedical IE, *KnowLife* employs logical reasoning for checking consistency constraints, tailored to the different relations that connect diseases, symptoms, drugs, genes, risk factors, etc. Constraint checking eliminates many false positives that are produced by methods that solely rely on pattern-based extraction.

In its best configuration, the *KnowLife* KB contains a total of 542,689 facts for 13 different relations, with an average precision of 93% (i. e. validity of the acquired facts) as determined by extensive sampling with manual assessment. The precision for individual realtions ranges from 71% (*CreatesRisk: Ecofactor* × *Disease*) to 97% (*SideEffect: (Symptom* ∪ *Disease)* × *Drug*). All facts in *KnowLife* carry provenance information, allowing to explore the evidence for a fact and filter by source.

## 4.2 RELATED WORK

The main body of IE research in biomedical informatics, relevant for the work presented this chapter, has focused on molecular entities and chemogenomics, like Protein-Protein Interactionss (PPIs) or gene-drug relations. These efforts have been driven by competitions such as BioNLP Shared Task (BioNLP-ST) [142] and BioCreative [5]. Each shared task offers pre-annotated corpora as gold standard, such as the GENIA corpus [80], the multi-level event extraction (MLEE) corpus [142], and various BioCreative corpora. Efforts such as the Pharmacogenetics Research Network and Knowledge Base (PharmGKB) [196], which curates and disseminates knowledge about the impact of human genetic variations on drug responses, or the Open PHACTS project [199], a pharmacological information platform for drug discovery, offer Knowledge Bases with annotated text corpora to facilitate approaches for these use cases.

Most IE work in this line of research relies on supervised learning, like Support Vector Machines [14, 29, 84, 109] or Probabilistic Graphical Models [25, 157]. The 2012 i2b2 challenge aimed at extracting temporal relations from clinical narratives [178]. Unsupervised approaches have been pursued by [21, 39, 91, 154], focusing on the discovery of associations between genes and diseases based on the co-occurrence of entities as cues for relations. To further improve the quality of discovered associations, crowdsourcing has also been applied [63, 146]. Burger et al. [28] uses Amazon Mechanical Turk to validate gene-mutation relations which are extracted from PubMed abstracts. [7] describes a crowdsourcing approach to generate gold standard annotations for medical relations, taking into account the disagreement between crowd workers.

Pattern-based approaches exploit text patterns that connect entities. [76, 116, 183, 193] manually define extraction patterns. [83] uses Hearst patterns [69] to identify terms that describe various properties of drugs. SemRep [153] manually specifies extraction rules obtained from dependency parse trees. Liu et al. [94, 95] learn descriptive subgraphs from training data, which are used as extraction rules to label unseen sentences by applying their own approximate subgraph matching algorithm. Furthermore, the authors show that their approach is also applicable in a distantly-supervised setting. Outside the biomedical domain, sentic patterns [139] leverage commonsense and syntactic dependencies to extract sentiments from movie reviews. However, while manually defined patterns yield high precision, they rely on expert guidance and do not scale to large and potentially noisy inputs and a broader scope of relations. Bootstrapping approaches such as [182, 205] use a limited number of seeds to learn extraction patterns; these techniques go back to [1, 23]. Our method follows this paradigm, but extends prior work with additional statistics to quantify the confidence of patterns and extracted facts.

A small number of projects, such as Sofie [176], PROSPERA [118], and NELL [32], have combined pattern-based extraction with logical consistency rules that constrain the space of fact candidates. Romero et al. [119] harness the IE methods of [118] for populating disease-centric relations. This approach uses logical consistency reasoning for high precision, but the small scale of this work leads to a very restricted KB. [114] used NELL to learn instances of biological classes, but did not extract binary relations and did not consider logical constraints either. Other work on constrained extraction tackles non-biological relations only (e. g. birthplaces of people or headquarters of companies). Our method builds on Sofie/PROSPERA, but additionally develops customized constraints for the biomedical relations targeted here.

Most prior work on biomedical Named Entity Recognition (NER) specializes in recognizing specific types of entities such as proteins and genes, chemicals, diseases, and organisms. MetaMap [6] is the most notable tool capable of recognizing a wide range of entities. For biomedical Named Entity Disambiguation (NED), there is relatively little prior work available. MetaMap offers limited NED functionality, while others focus on disambiguating between genes [67] or small sets of word senses [36]. A detailed discussion on Biomedical Named Entity Recognition and Disambiguation (BioNERD) can be found in [167].

Most prior IE work processes only abstracts of Pubmed articles; few projects have considered full-length articles from Pubmed Central, let alone Web portals and online communities. Vydiswaran et al. [189] addressed the issue of assessing the credibility of medical claims about diseases and their treatments in health portals. Mukherjee et al. [115] tapped into discussion forums to assess statements about side effects of drugs. White et al. [198] demonstrated how to derive insight on drug effects from search engine query logs. Building a comprehensive KB from such raw assets has been beyond the scope of these prior works. In 2015, Google updated its Knowledge Graph with common health knowledge representing real-life clinical knowledge, e. g. symptoms, treatments, etc. Facts are gathered from doctors as well as high-quality medical sources across the web and later curated by a collaborative effort led by a team of medical doctors at Google and the Mayo Clinic.

## 4.3 system overview

Our method for harvesting relational facts from text sources is designed as a pipeline of processing stages; Figure 6 gives a pictorial overview. As defined in Section 2.1.1, binary facts consist of two entities $e_1, e_2$, a relation $R$ between them and are denoted by $R(e_1, e_2)$. The components of our system, described in the following sections, rely on the following input sources:

**Dictionary** We use the Unified Medical Language System (UMLS) as the dictionary of biomedical entities. Being the largest collection of biomedical entities, the UMLS dictionary enables *KnowLife* to detect entities in text, going beyond genes and proteins by covering entities from the fields of anatomy, physiology, therapies, etc.

**Relations** *KnowLife* supports 13 binary relations between entities, each with a type signature constraining its domain and range (i. e. its left and right argument types). Table 4 shows that, for instance, the relation Affects only holds between diseases and organs, but not between diseases and drugs. Each type signature consists of multiple fine-grained semantic types defined by UMLS.

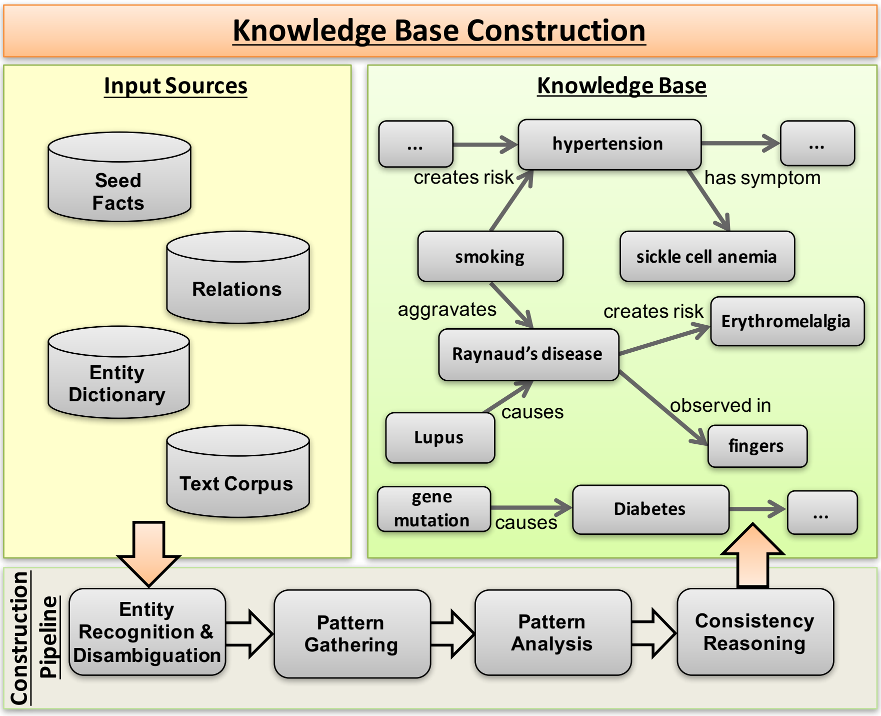


Figure 6: Overview of the *KnowLife* KB Construction Pipeline

| Relation | Domain | Range | Seed Facts |
|---|---|---|---|
| Affects | Disease | Organ | 23 |
| Aggravates | Ecofactor | Disease | 21 |
| Alleviates | Drug | Disease | 18 |
| Causes | Disease | Disease | 70 |
| ComplicationOf | Disease | Disease | 5 |
| Contraindicates | Drug | Disease | 26 |
| CreatesRisk | Ecofactor | Disease | 103 |
| Diagnoses | Device | Disease | 29 |
| Interacts | Drug | Drug | 9 |
| IsSymptom | Symptom or Disease | Disease | 69 |
| ReducesRisk | Drug or Behavior | Disease | 24 |
| SideEffect | Symptom or Disease | Drug | 12 |
| Treats | Drug | Disease | 58 |

Table 4: *KnowLife* Relations, their Type Signatures, and Number of Seeds

| Genre | Source | Documents | Sentences |
|---|---|---|---|
| Scientific Publications | PubMed Medline | 580,892 | 5,875,006 |
| | PubMed Central | 12,532 | 2,765,580 |
| Encyclopedic Articles | Drugs.com | 31,837 | 7,586,236 |
| | Mayo Clinic | 2,166 | 570,325 |
| | Medline Plus | 3,076 | 197,055 |
| | RxList | 2,515 | 1,102,791 |
| | Wikipedia Health | 20,893 | 787,148 |
| Social Sources | Healthboards.com | 752,778 | 37,270,371 |
| | Patient.co.uk | 44,610 | 1,081,420 |
| | **Total** | **1,451,299** | **57,235,932** |

Table 5: Overview of *KnowLife*'s Input Corpus

**Seed Facts** As introduced in Definition 10 a *seed fact* for a relation is a fact presumed to be true based on expert statements. We collected 467 binary seed facts (see Table 4) from the medical online portal Uptodate.com, a highly regarded clinical resource curated by physicians. These seed facts are further cross-checked in other sources to assert their veracity. The facts IsSymptom(Chest Pain, Myocardial Infarction) and CreatesRisk(Obesity, Diabetes) are two examples from our seed set.

**Text Corpus** A key asset of *KnowLife* is its ability to tap into different genres of text demonstrated in Table 5. PubMed documents are scientific texts with specialized jargon; they have been the de facto standard corpus for biomedical text mining. Starting with all PubMed documents published in 2011 that are indexed with disease-, drug-, and therapy-related Medical Subject Heading (MeSH) terms. We further prune out documents from inapplicable journals such as those not in the English language, or those about medical ethics. Web portals and encyclopedic articles are collaboratively or professionally edited, providing credible information in layman-oriented language. Examples include Uptodate.com, Mayoclinic.com, and the relevant parts of Wikipedia.org. In contrast, discussion forums of online communities, where patients and physicians engage in discussions (often anonymously), have a colloquial language style, and occasionally even use slang. We tap into all three genres of text to demonstrate not only the applicability of our system, but also the amount of information buried in all of them. We use Stanford CoreNLP to preprocess all texts, including Tokenization, Sentence Segmentation, Parts of Speech (POS) Tagging, Lemmatization, and Syntactic Parsing (see Section 2.2.1).

## 4.4 ENTITY RECOGNITION AND DISAMBIGUATION

The first stage in the *KnowLife* pipeline identifies sentences that potentially express a relational fact. We apply entity recognition to every sentence: a sentence with one or more entities is relevant for further processing. To efficiently handle the large dictionary and process large input corpora, we employ the method presented in [168]. The method uses string-similarity matching against the names in UMLS and is two orders of magnitude faster than MetaMap [6], the most popular biomedical entity recognition tool, while maintaining comparable accuracy. To quickly find matching candidates our system applies Locality Sensitive Hashing (LSH) [34] with min-wise independent permutations (MinHash) [24]. In particular, LSH probabilistically reduces the high-dimensional space of all character-level 3-grams, while MinHash quickly estimates the similarity between two sets of 3-grams. A successful match provides us also with the semantic type of the entity. If multiple entities are matched to the same string in the input text, we do not apply explicit disambiguation to determine the correct entity. Instead, using the semantic type hierarchy of UMLS, we select the most specifically typed entities. Later in the consistency reasoning stage, we leverage the type signatures to further prune out mismatched entities. As result of this processing stage, we obtain marked-up sentences such as

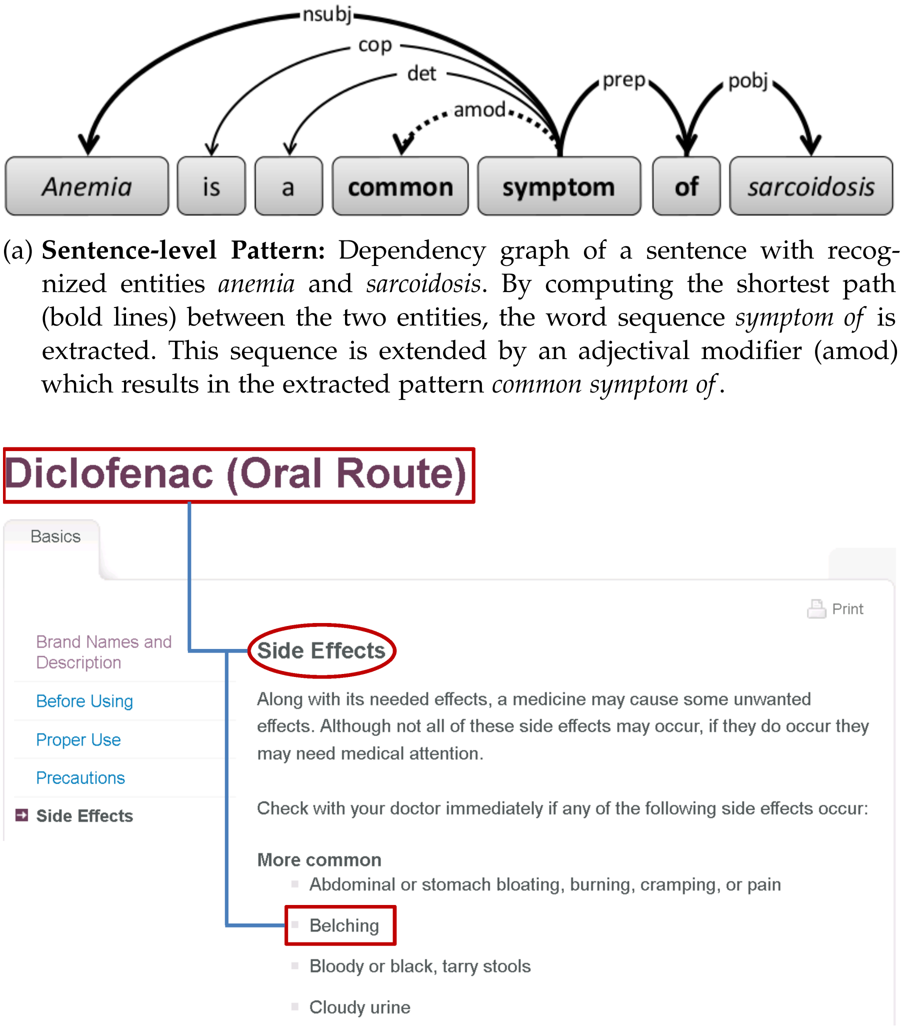
- "Anemia is a common symptom of Sarcoidosis."
- "Eventually, a heart attack leads to arrythmias."
- "Ironically, a Myocardial Infarction can also lead to Pericarditis."

where *myocardial infarction* and *heart attack* are synonyms representing the same canonical entity.

## 4.5 PATTERN GATHERING

The method extracts textual patterns that connect two recognized entities, either by the syntactic structure of a sentence or by a path in the Document Object Model (DOM) tree of a Web page. We extract two types of patterns:

**Sentence-level Patterns:** For each pair of entities in a sentence, we extract a sequence of text tokens connecting the entities in the syntactic structure of the sentence. Specifically, this is the shortest path between the entities in the dependency graph obtained from parsing the sentence. However, this path does not necessarily contain the full information to deduce a relation; for instance, negations are not captured, or essential adjectives are left out. Therefore, for every captured word, the following grammatical dependencies are added: negation,



(a) **Sentence-level Pattern:** Dependency graph of a sentence with recognized entities *anemia* and *sarcoidosis*. By computing the shortest path (bold lines) between the two entities, the word sequence *symptom of* is extracted. This sequence is extended by an adjectival modifier (amod) which results in the extracted pattern *common symptom of*.

Diclofenac (Oral Route)

Side Effects

Along with its needed effects, a medicine may cause some unwanted effects. Although not all of these side effects may occur, if they do occur they may need medical attention.

Check with your doctor immediately if any of the following side effects occur:

More common

- Abdominal or stomach bloating, burning, cramping, or pain
- Belching
- Bloody or black, tarry stools
- Cloudy urine

(b) **Document-structure Pattern:** The entity *Diclofenac* is found within the document title and *Belching* within an <li> element. Take *Diclofenac* as the left-hand entity. By traversing the DOM tree downwards and coming across the heading *Side Effects*, we extract the heading's text as a pattern. Further traversal leads us to *Belching*, which yields the right-hand entity for the pattern.

Figure 7: Pattern Gathering in *KnowLife*

adjectival modifiers, and adverbial modifiers. The resulting word sequence constitutes a sentence-level pattern. An example is shown in Figure 7a.

**Document-structure Patterns:** In Web portals like Mayo Clinic or Wikipedia, it is common that authors state medical facts by using specific document structures, like titles, sections, and listings. Such structures are encoded in the DOM tree of the underlying HTML markup. First, we detect if the document title, that is, the text within the <h1> tag in HTML markup, is a single entity. Next, we detect if an entity appears in an HTML listing, that is, within an <li> tag. Starting from the <h1> tag, our method traverses the DOM tree downwards and determines all intermediate headings, i. e. <h2> to <h6> tags, until we reach the aforementioned <li> tag. The document title serves as left-hand entity, the intermediate headings as patterns, and the <li> text as right-hand entity. These are candidates for a relation or an entity argument in a relational fact. Figure 7b shows an example.

## 4.6 PATTERN ANALYSIS

The goal of the pattern analysis stage is the identification of the most useful *seed patterns* out of all the pattern candidates gathered thus far. A seed pattern should generalize the over-specific phrases encountered in the input texts, by containing only the crucial words that express a relation and masking out (by a wildcard or POS tag) inessential words. This way we obtain high-confidence patterns.

We harness the techniques developed in the PROSPERA tool [118]. First, an itemset mining algorithm is applied to find frequent sub-sequences in the patterns. The sub-sequences are weighed by statistical analysis, in terms of confidence and support. We use the seed facts and their co-occurrences with certain patterns as a basis to compute confidence, such that the confidence for a pattern $q$ in a set of sentences $S$ is defined as

$$\mathrm{confidence}(q) = \frac{|\{s \in S \mid \exists (e_1,e_2) \in SX(R_i) \;\; q,e_1,e_2 \text{ occur in } s\}|}{|\{s \in S \mid \exists (e_1,e_2) \in SX(R_i) \cup CX(R_i) \;\; q,e_1,e_2 \text{ occur in } s\}|} \quad (1)$$

where $SX(R_i)$ is the set of all entity tuples $(e_1, e_2)$ appearing in any seed fact with relation $R_i$ and $CX(R_i)$ is the set of all entity tuples $(e_1, e_2)$ appearing in any seed fact without relation $R_i$. The rationale is that the more strongly a pattern correlates with the seed-fact entities of a particular relation, the more confident we are that the pattern expresses the relation. The patterns with confidence greater than a threshold (set to 0.3 in our experiments) are selected as seed patterns.

| Seed Facts | Seed Pattern | Relations | Confidences | Patterns | Harvested Facts |
|---|---|---|---|---|---|
| Causes(Tuberculosis, Pericarditis) | progress | CreatesRisk | 0.5 | progresses to | Causes(Pericarditis, Tamponade) |
| CreatesRisk(Obesity, Diabetes) | | Causes | 0.5 | still progressing | CreatesRisk(Wart, Skin cancer) |
| CreatesRisk(Obesity, Asthma) | risk factor | CreatesRisk | 1.0 | child risk factor | CreatesRisk(Wood Dust, Asthma) |
| CreatesRisk(Malaria, Stillbirth) | | | | have risk factors | CreatesRisk(Golf, Tendinitis) |
| | | | | known risk for | CreatesRisk(GBV − C, Hepatitis) |
| IsSymptom(Pain, Crohn's) | occur | Affects | 0.67 | occurs anywhere | Affects(Hashimoto's, Thyroids) |
| Affects(Pericarditis, Heart) | | IsSymptom | 0.33 | occurs patients | IsSymptom(Anemia, Sarcoidosis) |

Table 6: Examples of Seed Facts, Mined Seed Patterns, Automatically Acquired Patterns and Facts.

Each non-seed pattern $p$ is then matched against the seed pattern set $Q$ using Jaccard similarity to compute a weight $w$ associating $p$ with a relation.

$$w = \mathrm{max}\{\mathrm{Jaccard}(p, q) \times \mathrm{confidence}(q) \mid q \in Q\} \tag{2}$$

The pattern occurrences together with their weights and relations serve as *fact candidates*. Table 6 shows sample seed patterns computed from seed facts. The table also gives examples for automatically acquired patterns and facts.

## 4.7 CONSISTENCY REASONING

The pattern analysis stage provides us with a large set of fact candidates and their supporting patterns. However, the candidate set still contains many false positives. To prune these out and improve precision, the final stage of *KnowLife* applies logical consistency constraints to the fact candidates and accepts only a consistent subset of them.

We leverage two kinds of manually defined semantic constraints: i) the type signatures of relations (see Table 4) for type checking of fact candidates, and ii) mutual exclusion constraints between certain pairs of relations. For example, if a drug has a certain symptom as a side effect, it cannot treat this symptom at the same time. These rules allow us to handle conflicting candidate facts. The reasoning uses probabilistic weights derived from the statistics computed in the candidate gathering phase.

To reason with consistency constraints, we follow the framework of [176], by encoding all facts, patterns, and grounded (i. e. instantiated) constraints into weighted logical clauses. We extend this prior work by computing informative weights from the confidence statistics obtained in the pattern-based stage of our IE pipeline. We then use a weighted Max-Sat solver to reason on the hypotheses space of fact candidates, eventually computing a consistent subset of clauses with the largest total weight. Due to the NP-hardness of the weighted Max-Sat problem, we resort to an approximation algorithm that combines the dominating-unit-clause technique [124] with Johnson's heuristic algorithm [77]. Suchanek et al. [176] have shown that this combination empirically enables very good approximation ratios. The complete set of consistency constraints is available in Appendix a.1.

## 4.8 EXPERIMENTAL EVALUATION

We have conducted extensive experiments using the input corpora listed in Table 5, and created different KBs based on different configurations. We assess the size and quality of each KB, in terms of their numbers of facts and their precision evaluated by random sampling of facts. Tables 7 and 8 show the results, for different choices

| Relation | Precision | | | |
|---|---|---|---|---|
| | Encyclopedic sources | Scientific sources | **Encyclopedic + scientific sources** | Encyclopedic + scientific + social sources |
| Affects | 0.855±0.047 | 0.762±0.049 | **0.825±0.047** | 0.767±0.048 |
| Aggravates | 0.810±0.041 | 0.459±0.044 | **0.829±0.049** | 0.785±0.049 |
| Alleviates | 0.953±0.039 | 0.735±0.048 | **0.786±0.046** | 0.736±0.048 |
| Causes | 0.904±0.039 | 0.674±0.049 | **0.801±0.049** | 0.792±0.049 |
| Complication | 0.917±0.039 | 0.397±0.049 | **0.897±0.041** | 0.869±0.046 |
| Contraindicates | 0.874±0.048 | 0.710±0.000 | **0.961±0.030** | 0.908±0.048 |
| CreatesRisk | 0.878±0.047 | 0.569±0.049 | **0.720±0.040** | 0.620±0.049 |
| Diagnoses | 0.964±0.035 | 0.839±0.049 | **0.860±0.048** | 0.840±0.047 |
| Interacts | 0.964±0.035 | 0.709±0.000 | **0.965±0.034** | 0.957±0.034 |
| IsSymptom | 0.891±0.042 | 0.482±0.050 | **0.858±0.048** | 0.694±0.048 |
| ReducesRisk | 0.797±0.045 | 0.637±0.046 | **0.762±0.048** | 0.751±0.049 |
| SideEffect | 0.956±0.038 | 0.826±0.000 | **0.964±0.035** | 0.971±0.026 |
| Treats | 0.850±0.048 | 0.581±0.045 | **0.898±0.041** | 0.566±0.048 |
| Micro Average | 0.951 | 0.630 | **0.933** | 0.892 |

Table 7: Precision of Different Text Genres.

of input corpora, and Tables 10 and 11 the results for different configurations of the *KnowLife* pipeline. Recall is not evaluated, as there is no gold standard for fully comprehensive facts. To ensure that our findings are significant, for each relation, we computed the Wilson confidence interval at $\alpha = 5\%$, and kept evaluating facts until the interval width fell below 5%. An interval width of 0% means that all the facts were evaluated. Four different annotators evaluated the facts, judging them as true or false based on provenance information. As for inter-annotator agreement, 22,002 facts were evaluated; the value of Fleiss' Kappa was 0.505, indicating moderate agreement among all annotators.

### 4.8.1 *Impact of Different Text Genres*

We first discuss the results obtained from the different text genres:

i. scientific (PubMed publications),

ii. encyclopedic (Web portals like Mayo Clinic or Wikipedia),

iii. social (discussion forums).

Table 7 and 8 list the precision and number of facts for four different combinations of genres, respectively.

Generally, combining genres yielded more facts at a lower precision, as texts of lower quality like social sources introduced noise.

| Relation | Harvested Facts | | | |
|---|---|---|---|---|
| | Encyclopedic sources | Scientific sources | **Encyclopedic + scientific sources** | Encyclopedic + scientific + social sources |
| Affects | 1,278 | 450 | **2,388** | 5,053 |
| Aggravates | 130 | 371 | **432** | 708 |
| Alleviates | 903 | 4,433 | **4,530** | 6,790 |
| Causes | 28,119 | 19,203 | **47,463** | 62,407 |
| Complication | 1,011 | 1,475 | **1,524** | 1,566 |
| Contraindicates | 512 | 49 | **1,808** | 1,831 |
| CreatesRisk | 4,407 | 24,695 | **18,508** | 32,211 |
| Diagnoses | 813 | 5,920 | **4,832** | 9,743 |
| Interacts | 164,912 | 103 | **164,912** | 164,912 |
| IsSymptom | 4,878 | 2,320 | **6,395** | 11,017 |
| ReducesRisk | 1,712 | 4,684 | **4,489** | 5,865 |
| SideEffect | 270,600 | 139 | **270,709** | 271,416 |
| Treats | 11,915 | 9,318 | **14,699** | 35,803 |
| Total | 491,190 | 73,160 | **542,689** | 609,322 |

Table 8: Number of Harvested Facts for Different Text Genres.

The combination that gave the best balance of precision and total yield was scientific with encyclopedic sources, with a micro-averaged precision of 0.933 for a total of 542,689 facts. We consider this the best of the KBs that the system generated.

The best overall precision was achieved when using encyclopedic texts only. This confirmed our hypothesis that a pattern-based approach works best when the language is simple and grammatically correct. Contrast this with scientific publications which often exhibit convoluted language, and online discussions with a notable fraction of grammatically incorrect language. In these cases, the quality of patterns degraded and precision dropped. Incorrect facts stemming from errors in the entity recognition step were especially rampant in online discussions, where colloquial language (for example, *meds*, or short for *medicines*) led to incorrect entities (acronym for *Microcephaly, Epilepsy, and Diabetes Syndrome*).

The results vary highly across the 13 relations in our experiments. The number of facts depends on the extent to which the text sources express a relation, while precision reflects how decisively patterns point to that relation. Interacts and SideEffect are prime examples: the Drugs.com portal lists many side effects and drug-drug interactions by the DOM structure, which boosted the extraction accuracy of *KnowLife*, leading to many facts at precisions of 95.6% and 96.4%, respectively. Facts for the relations Alleviates, CreatesRisk,

| Genre | Source | Fact Occurrences |
|---|---|---|
| Scientific Publications | PubMed Medline | 39,266 |
| | PubMed Central | 6,979 |
| Encyclopedic Articles | Drugs.com | 461,130 |
| | Mayo Clinic | 35,300 |
| | Medline Plus | 6,559 |
| | RxList | 5,818 |
| | Wikipedia Health | 17,588 |

Table 9: Number of Fact Occurrences in Text Sources

and ReducesRisk, on the other hand, mostly came from scientific publications, which resulted in fewer facts and lower precision.

A few relations, however, defied these general trends. Patterns of Contraindicates were too sparse and ambiguous within encyclopedic texts alone and also within scientific publications alone. However, when the two genres were combined, the good patterns reached a critical mass to break through the confidence threshold, giving rise to a sudden increase in harvested facts. For the CreatesRisk and ReducesRisk relations, combining encyclopedic and scientific sources increased the number of facts compared to using only encyclopedic texts, and increased the precision compared to using only scientific publications.

A comparison of Table 7 and 8 shows, incorporating social sources brought a significant gain in the number of harvested facts, at a tradeoff of lowered precision. As [115] pointed out, there are facts that come only from social sources and, depending on the use case, it is still worthwhile to incorporate them; for example, to facilitate certain search and discovery applications where recall may be more important. Moreover, the patterns extracted from encyclopedic and scientific sources could be reused to annotate text in social sources, so as to identify existing information.

Taking a closer look at the best experimental setting, it is evident that scientific and encyclopedic sources in *KnowLife* contribute to a different extent to the number of harvested facts. Table 9 shows the number of fact occurrences in our input sources. Recall that a fact can occur in multiple sentences from multiple text sources. Our experiments show that encyclopedic articles are more amenable for harvesting facts than scientific publications.

| Relation | Precision | | | |
|---|---|---|---|---|
| | Full pipeline encyclopedic + scientific sources | Without document structure | Without statistical analysis | Without consistency reasoning |
| Affects | 0.825±0.047 | 0.882±0.044 | 0.821±0.048 | 0.171±0.051 |
| Aggravates | 0.829±0.049 | 0.833±0.036 | 0.598±0.049 | 0.592±0.053 |
| Alleviates | 0.786±0.046 | 0.778±0.050 | 0.320±0.049 | 0.289±0.062 |
| Causes | 0.801±0.049 | 0.800±0.046 | 0.631±0.048 | 0.490±0.069 |
| Complication | 0.897±0.041 | 0.781±0.048 | 0.376±0.050 | 0.739±0.050 |
| Contraindicates | 0.961±0.030 | 0.914±0.043 | 0.122±0.049 | 0.630±0.059 |
| CreatesRisk | 0.720±0.040 | 0.750±0.044 | 0.386±0.047 | 0.406±0.067 |
| Diagnoses | 0.860±0.048 | 0.887±0.044 | 0.802±0.049 | 0.303±0.063 |
| Interacts | 0.965±0.034 | 0.858±0.046 | 0.953±0.047 | 0.941±0.049 |
| IsSymptom | 0.858±0.048 | 0.691±0.050 | 0.625±0.049 | 0.328±0.064 |
| ReducesRisk | 0.762±0.048 | 0.729±0.050 | 0.228±0.046 | 0.406±0.067 |
| SideEffect | 0.964±0.035 | 0.938±0.048 | 0.941±0.046 | 0.879±0.050 |
| Treats | 0.898±0.041 | 0.784±0.050 | 0.549±0.050 | 0.402±0.067 |
| Micro Average | 0.933 | 0.784 | 0.777 | 0.707 |

Table 10: Precision Impact of Different Components.

| Relation | Harvested Facts | | | |
|---|---|---|---|---|
| | Full pipeline encyclopedic + scientific sources | Without document structure | Without statistical analysis | Without consistency reasoning |
| Affects | 2,388 | 2,350 | 4,088 | 29,477 |
| Aggravates | 432 | 431 | 592 | 1,730 |
| Alleviates | 4,530 | 4,387 | 18,142 | 16,943 |
| Causes | 47,463 | 30,563 | 66,833 | 91,784 |
| Complication | 1,524 | 700 | 4,812 | 2,955 |
| Contraindicates | 1,808 | 365 | 26,298 | 15,279 |
| CreatesRisk | 18,508 | 17,282 | 77,158 | 48,159 |
| Diagnoses | 4,832 | 4,002 | 7,467 | 35,326 |
| Interacts | 164,912 | 392 | 200,935 | 187,201 |
| IsSymptom | 6,395 | 2,920 | 9,543 | 29,776 |
| ReducesRisk | 4,489 | 4,043 | 11,023 | 14,729 |
| SideEffect | 270,709 | 924 | 270,427 | 338,645 |
| Treats | 14,699 | 14,057 | 23,473 | 45,439 |
| Total | 542,689 | 82,416 | 720,791 | 857,443 |

Table 11: Impact of Different Components on the Number of Harvested Facts

### 4.8.2 *Impact of Different Components*

In each setting, only one component was disabled, and the processing pipeline ran with all other components enabled. We used the *KnowLife* setting with scientific and encyclopedic sources, which, by and large, performed best, as the basis for investigating the impact of different components in the *KnowLife* pipeline. To this end, we disabled individual components: DOM tree patterns, statistical analysis of patterns, consistency reasoning – each disabled separately while retaining the others. This way we obtained insight into how strongly *KnowLife* depends on each component. Table 10 and 11 show the results of this ablation study.

**No DOM tree patterns:** When disregarding patterns on the document structure and solely focusing on textual patterns, *KnowLife* degrades in precision (from 93% to 78%) and sharply drops in the number of acquired facts (from ca. 540,000 to 80,000). The extent of these general effects varies across the different relations. Relations whose patterns are predominantly encoded in document structures – once again Interacts and SideEffect – exhibit the largest loss. On the other hand, relations like Affects, Aggravates, Alleviates, and Treats, are affected only to a minor extent, as their patterns are mostly found in free text.

**No statistical pattern analysis:** Here we disabled the statistical analysis of pattern confidence and the frequent itemset mining for generalizing patterns. This way, without confidence values, *KnowLife* kept all patterns, including many noisy ones. Patterns that would be pruned in the full configuration led to poor seed patterns; for example, the single word *causes* was taken as a seed pattern for both relations SymptomOf and Contraindicates. Without frequent itemset mining, long and overly specific patterns also contributed to poor seed patterns. The combined effect greatly increased the number of false positives, thus dropping in precision (from 94% to 77%). In terms of acquired facts, not scrutinizing the patterns increased the yield (from ca. 540,000 to 720,000 facts).

Relations, such as Interacts and SideEffect, mainly extracted from DOM tree patterns were not much affected. Also, relations like Affects and Diagnoses exhibited only small losses in precision; for these relations, the co-occurrence of two types of entities is often already sufficient to express a relation. The presence of consistency constraints on type signatures also helped to keep the output quality high.

**No consistency reasoning:** In this setting, neither the type signatures nor other consistency constraints were checked. Thus, conflicting facts could be accepted, leading to a large fraction of false positives. This effect was unequivocally witnessed by an increase in the number of facts (from ca. 540,000 to 850,000) accompanied by a sharp decrease in precision (from 93% to 70%).

| | | Percentage based on Text Genre | | |
|---|---|---|---|---|
| Percentage | Cause of Error | Encyclopaedic sources | Scientific sources | Social sources |
| 8.16% (62) | Preprocessing | 38.71% (24) | 3.23% (2) | 58.06% (36) |
| 27.24% (207) | Entity Recognition | 13.04% (27) | 45.41% (94) | 41.55% (86) |
| 32.11% (244) | Entity Disambiguation | 12.30% (30) | 26.23% (64) | 61.48% (150) |
| 1.97% (15) | Coreferencing | 13.33% (2) | 13.33% (2) | 73.33% (11) |
| 13.68% (104) | Nonexistent Relation | 23.08% (24) | 29.81% (31) | 47.12% (49) |
| 9.21% (70) | Pattern Relation Duality | 24.29% (17) | 27.14% (19) | 48.57% (34) |
| 3.29% (25) | Swapped Left/Right Entity | 28.00% (7) | 24.00% (6) | 48.00% (12) |
| 3.03% (23) | Negation | 17.39% (4) | 21.74% (5) | 60.87% (14) |
| 1.32% (10) | Factually Wrong | 40.00% (4) | 10.00% (1) | 50.00% (5) |

Table 12: Error Analysis (Number of Facts in Brackets)

The relations `Interacts` and `SideEffect` were least affected by this degradation, as they are mostly expressed in the via document structure of encyclopedic texts where entity types are implicitly encoded in the DOM tree tags (see Figure 7). Here, consistency reasoning was not vital.

*Lessons Learned*

Overall, this ablation study clearly shows that all major components of the *KnowLife* pipeline are essential for high quality (precision) and high yield (number of facts) of the constructed KB. Each of the three configurations with one disabled component suffered substantial if not dramatic losses in either precision or acquired facts, and sometimes both. We conclude that the full pipeline is a well-designed architecture whose strong performance cannot be easily achieved by a simpler approach.

### 4.8.3 *Error Analysis*

We analyzed the causes of error for 760 facts annotated as incorrect from the experimental setting using the full information extraction pipeline and all three text genres. This setting allows us to compare the utility of the different components as well as the different genres. As can be seen in Table 12, we categorize the errors as follows:

**Preprocessing:** At the start of the pipeline, incorrect sentence segmentation divided a text passage into incomplete sentences, or left multiple sentences undivided. This in turn leads to incorrect parsing of syntactic dependency graphs. In addition, there were incorrectly parsed DOM trees in Web portal documents. Not surprisingly, almost all preprocessing errors stem from encyclopedic and social sources due to their DOM tree structure and poor language style, respectively.

**Entity Recognition:** Certain entities were not recognized correctly. Complex entities are composed of multiple simple entities; examples include *muscle protein breakdown* recognized as *muscle protein* and *breakdown*, or *arrest of cystic growth* recognized as *arrest* and *cystic growth*. Paraphrasing and misspelling entities cause their textual expressions to deviate from dictionary entries. Idiomatic expressions were incorrectly picked up as entities. For instance, there is no actual physical activity in the English idiom *in the long run*.

**Entity Disambiguation:** Selecting an incorrect entity out of multiple matching candidates caused this error, primarily due to two reasons. First, the type signatures of our relations were not sufficient to further prune out mismatching entities during fact extraction. Second, colloquial terms not curated in the UMLS were incorrectly resolved. For example, *meds* for medicines was disambiguated as the entity *Microcephaly, Epilepsy, and Diabetes Syndrome*.

**Coreferencing:** Due to the lack of coreference resolution, correctly recognized entities were obscured by phrases such as *this protein* or *the tunnel structure*.

**Nonexistent relation:** Two entities might co-occur within the same sentence without sharing a relation. When a pattern occurrence between such entities was nevertheless extracted, it resulted in an unsubstantiated relation.

**Pattern Relation Duality:** A pattern that can express two relations was harvested but assigned to express an incorrect relation. For example, the pattern *mimic* was incorrectly assigned to the relation IsSymptom.

**Swapped Left and Right-hand Entity:** The harvested fact was incorrect because the left- and right-hand entities were swapped. Consider the example fact IsSymptom(Anemia, Sarcoidosis), which can be expressed by either sentence:

1. "Anemia is a common symptom of Sarcoidosis."
2. "A common symptom of Sarcoidosis is Anemia."

In both cases, the same pattern *is a common symptom of* is extracted. In sentence 2, however, an incorrect fact would be extracted since the order in which the entities occur is reversed.

**Negation:** This error was caused by missing negation expressed in the text. The word expressing the negation may occur textually far away from the entities, as in *It is disputed whether early antibiotic treatment prevents reactive arthritis*, and thus escaped our pattern gathering method. In other cases, the negation phrase will require subtle semantic understanding to tease out, as in *Except for osteoarthritis, I think my symptoms are all from heart disease*.

**Factually Wrong:** Although our methods successfully harvested a fact, the underlying text evidence made a wrong statement.

| Biomedical Areas | | Connections |
|---|---|---|
| Disorders | Chemicals | 310482 |
| Chemicals | Chemicals | 190160 |
| Disorders | Disorders | 36677 |
| Disorders | Procedures | 14169 |
| Chemicals | Physiology | 5397 |
| Disorders | Genes | 3831 |
| Disorders | Living Beings | 2539 |
| Chemicals | Drugs | 2455 |
| Disorders | Anatomy | 2895 |
| Disorders | Devices | 792 |
| Disorders | Activities | 592 |
| Disorders | Drugs | 511 |
| Disorders | Objects | 505 |
| Chemicals | Procedures | 544 |
| Disorders | Physiology | 370 |
| Procedures | Physiology | 123 |
| Procedures | Living Beings | 99 |
| Disorders | Geographical Areas | 82 |
| Genes | Physiology | 51 |
| Disorders | Phenomena | 50 |

Table 13: Top-20 pairs of inter-connected biomedical areas within *KnowLife*

*Lessons Learned*

Overall, this error analysis confirms that scientific and encyclopedic sources contain well-written texts that are amenable to a text mining pipeline. Social sources, with their poorer quality of language style as well as information content, were the biggest contributor in almost all error categories. Errors in entity recognition and disambiguation accounted for close to 60% of all errors; overcoming them will require better methods that go beyond a dictionary, and incorporate deeper linguistic and semantic understanding.

### 4.8.4 *Coverage*

The overriding goal of *KnowLife* has been to create a versatile KB that spans many areas within the life sciences. To illustrate which areas are covered by *KnowLife*, we refer to the *Semantic Groups* defined by [103]. Table 13 shows the number of acquired facts for pairs of the thirteen different areas inter-connected in our KB. This can be seen as an indicator that we achieved our goal at least to some extent.

The predominant number of facts involves entities of the semantic group *Disorders*, for two reasons. First, with our choice of relations, disorders appear in almost all type signatures. Second, entities of type clinical finding are covered by the group *Disorders*, and these are frequent in all text genres. However, this type also includes diverse, non-disorder entities such as *pregnancy*, which is clearly not a disorder.

## 4.9 SUMMARY

In this chapter we presented the first contribution: *KnowLife*, a large Knowledge Base for health and life sciences automatically constructed from different Web sources. We introduced three limitations of prior work, viz. most biomedical KBs depend on manual curation, exclusively rely on scientific publication as text sources, and only focus on the molecular level. We have addressed these shortcomings by a versatile and scalable approach to automatic KB construction. Using a small number of seed facts for distant supervision of pattern-based extraction, a huge number of fact candidates were harvested in an automated manner without requiring any explicit training. We extended previous techniques for pattern-based IE with confidence statistics, and combined this recall-oriented stage with logical reasoning for consistency constraint checking to eliminate false positives and achieve high precision. Our extensive experiments did not only target scientific publications, but also included encyclopedic health portals and online communities, this way creating different KB's based on different configurations. The best configured KB, *KnowLife*, contains more than 500,000 facts at a precision of 93% for 13 relations covering genes, anatomic parts, diseases, symptoms, treatments, as well as environmental and lifestyle risk factors for diseases. To showcase the usefulness of an integrated KB, such as *KnowLife*, we developed two web portals, that satisfy use-cases from speed-reading to semantic search along with richly annotated literature, the details of which are described in Part III of this thesis.

# 5 HIGHLIFE: HIGHER-ARITY FACT HARVESTING

## 5.1 INTRODUCTION

Knowledge Bases, such as KnowLife described in the previous chapter and the KBs presented in Section 2.1.2, have proven their usefulness for many applications. They are key components for search engines and recommender systems as well as domain-specific use cases, such as health care (e. g. curation of biological databases [94], medical question answering [190], and guided search and exploration of biomedical literature [51]).

However, a major limitation is that the majority of their facts refer to *binary* relations only, in the form of Subject-Predicate-Object (SPO) triples following the RDF data model. For example, DBpedia knows that Marie Curie has won the Nobel Prize in Physics, but it does not have any knowledge on which contribution it was for. YAGO knows that Marie Curie has won a Nobel Prize in 1903 and another one in 1911, but it does not keep the fields (Physics and Chemistry) as explicit predicates. KnowLife includes facts about drug treatment of diseases, but no information about the appropriate dosages and target groups. Freebase represented such complex relationships by means of Compound Value Types (CVTs), thus deviating from a simple SPO Resource Description Framework (RDF) data model (see Section 2.1.1). Information Extraction (IE) methods that distill knowledge from textual documents hardly capture these situations at all; they almost exclusively focus on binary relations.

Note that it is not always possible to decompose ternary or higher-arity relations into binary facts without losing information. If we only knew that Curie won both physics and chemistry Nobel prizes and we knew that she won prizes in 1903 and 1911, we would have no way to infer which prize was won in which year (and for which contribution). Going beyond the binary case is often crucial to capture more complete and deeper knowledge about events or multi-entity relationships. The following examples demonstrate this by text snippets that contain ternary or quaternary facts on prizes, business acquisitions and health (with relevant arguments for relations underlined).

- In 1978, Carl Sagan won the Pulitzer Prize for The Dragons of Eden.
- Google acquired Nest for $3.2 billion in January 2014.
- 2.5 mg Albuterol may be used to treat acute exacerbations, particularly in children.

- Salmonella infection is a common cause of bacteremia in Africa.

The problem that we tackle in this chapter lies in the automated harvesting of higher-arity facts from sentences of this kind.

Prior work on this problem is scarce. Notable contributions are Krause et al. [85] and Li et al. [92], that learn extraction rules for higher-arity relations based on training facts and dependency-parsing patterns. However, their methods produce large number of rules with fine-grained parse trees as rule body – these rules do not generalize beyond specific patterns. Krause et al. [86] published a resource of syntactic-semantic graph patterns for IE. However, this pattern collection is small and relies on manual curation. In contrast, our work is automated (with minimal supervision), scales well and can robustly cope with inputs that contain some but not all arguments of a higher-arity fact.

A related mature line of research is *Semantic Role Labeling* (SRL) [127], SRL for short. SRL methods are based on constrained learning, using fine-grained syntactic and lexical features. They depend heavily on training sentences, and are typically geared for the fixed set of frames in PropBank [126] or FrameNet [9]. In our experiments, we use the state-of-the-art SRL system [140, 172] of the Illinois NLP Curator software [40] as a baseline.

Generally, distant supervision approaches such as [106] have been widely used for harvesting facts. They usually rely on patterns incorporating syntactical and lexical features extracted from dependency parse trees. However, most of the earlier approaches focus exclusively on binary facts and neglect higher-arity facts. Our approach overcomes this limitation and is more general. We are able to harvest higher-arity relations by utilizing more complex pattern representations, i. e. trees instead of pure sequence patterns, and by considering partial facts, i. e. facts with unknown arguments.

Our method is twofold. We use seed facts as distant supervision to learn patterns, apply these patterns to extract fact candidates, and iterate these steps. This extends the fact-pattern duality paradigm [23] to higher-arity cases. While achieving high recall, this approach is susceptible to noise and target drifts. Therefore, we use constraint reasoning to eliminate spurious fact candidates. To this end, we extend the MaxSat-based reasoner of [118] to the higher-arity case. For example, we can apply type constraints to identify when facts about winning the Pulitzer prize are for movies or songs (instead of books), and we can exploit value constraints when confusing the numbers for amount and year on a company acquisition.

A key difficulty in this approach lies in the observation that higher-arity facts are often expressed only partially: with some but not all of their arguments. For example, we could have inputs such as "Google acquired Nest in 2014" without stating the amount, or "Google bought Nest for 3.2 Billion" without giving a date. We address this issue by extending our framework to partial facts, partial patterns and reasoning over the consistency and composability of partial fact candidates into full facts.

Our method is general and applicable to any domain and a wide range of text genres. For experimental studies, we test our method on two kinds of text corpora:

1. health-related texts about diseases and therapies from PubMed and other online sources, and
2. news articles about business acquisitions and athletes winning medals.

For unbiased evaluation, we obtain gold-standard assessments via crowdsourcing, using the CrowdFlower service. The experiments include comparisons to a state-of-the-art SRL system as a baseline.

## 5.2 RELATED WORK

**Knowledge Bases:** Contrary to Knowledge Bases, such as YAGO [73], WikiData [188] and Freebase, which extract the majority of n-ary facts from pre-structured resources (e.g. Wikipedia Infoboxes) or rely on human input, we focus on harvesting n-ary facts from text.

**Open Information Extraction:** Open information extraction approaches, such as OLLIE [100], ClausIE [43], and EXAMPLAR [105] are constrained by syntactic patterns on parse trees for extracting n-ary facts and canonicalize neither relations nor entities to a knowledge base. Thus, they suffer from ambiguous extractions.

**Semantic Role Labeling:** Semantic Role Labeling (SRL) [62, 127] aims to map single sentences onto structured frames with slots filled based on the verb-argument structure of a sentence, using supervised learning over fine-grained syntactic and lexical features. SRL methods strongly rely on labeled training data, and are focused on the frame repositories provided by PropBank [126] or FrameNet [9]. Adapting these methods to new domains is expensive, since it entails the specification of new frame types along with a large amount of manually labeled training data. In contrast, our distantly supervised approach requires only a moderate amount of seed facts and no explicit labeling at all. Since SRL is nevertheless closest to our approach, the experiments presented in Section 5.8 compare our method to the state-of-the-art baseline [140, 172], which is part of the Illinois NLP Curator software [40].

**Temporal and Spatial Anchoring of Facts:** The scope of temporal and spatial anchoring approaches is limited to assigning location or time information to facts [50, 60, 73, 88, 181, 192]. The goal of the TAC Knowledge Base Population task on Temporal Slot Filling [179] is related to this line of work. The systems for this task typically train classifiers with additional constraints, like temporal ordering or spatial consistency, which are not applicable to a general setting.

**Event Extraction:** Event extraction methods identify occurrences of events from a predefined set of event types within a text corpus. For example, extraction of Movement, Transfer, Creation and Destruction events was a task within the Automatic Content Extraction (ACE) program [46]. Named Event Mining distills structured event representations from text [87]. Events consist of a topic and multiple entities as actors, but they do not include relations between the actors beyond participation in the same event. Story mining aims to extract structured representations for linking different events [164]. Here, events are just topics, i. e. potentially ambiguous phrases, and links merely connect events without any further semantics. This is different to our use case, where clear semantics and canonicalization of entities are crucial for populating a knowledge base. In the biomedical domain, event extraction mostly focuses on binary relations between molecular entities, like protein-protein interaction or gene-drug relations (e.g., [94, 102, 110, 187]). Approaches in this area are typically based on dependency parsing and supervised learning, using different kinds of graph similarity kernels [26, 113].

**N-ary Fact Harvesting:** The Xart system [13] applies association rule mining to find highly co-occurring entities in dependency parse trees. Since the extracted rules require manual validation, the system relies on input by domain experts to discover instances of predefined n-ary medical relations from text. McDonald et al. [104] first trains a classifier to identify pairs of related entities which they use as input to construct a graph of all related entities within a sentence. Higher-arity relations then correspond to maximal cliques in the graph. The works [85, 92] apply a distantly supervised approach for learning extraction rules for n-ary relations from dependency graphs. These rules are highly specific and do not generalize well. Consequently, the method needs a large number of seed facts: several thousands per relation even for simple relations such as marriage (with date and place as additional attributes), while achieving moderate precision of ca. 50%. Sar-graphs [86] aggregate this style of rules and incorporate lexical knowledge to construct an easily re-usable linguistic resource. However, this resource is manually constructed and small. None of these methods is applicable to our setting with large-scale input corpora and a limited amount of distant supervision.

Peng et al. [132] present a graph-oriented LSTM neural network for learning how to extract ternary relations when the arguments are scattered across multiple sentences. However, this method is geared for named entities as arguments and does not cover arguments that are phrases for quantities (e.g., medical dosages) or general concepts (e.g., denominations for awards such as physics, medicine, peace, best actor, etc.). Experiments exclusively focus on the ternary interaction

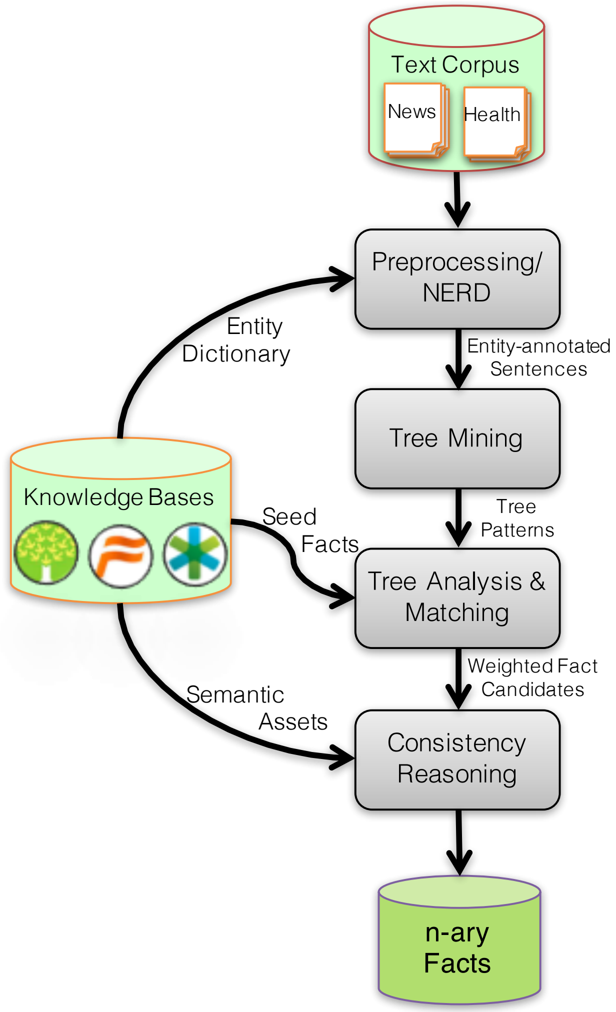


Figure 8: *HighLife* System Overview

of genes, drugs and gene mutations, and use extensive supervision from high-quality knowledge bases.

## 5.3 SYSTEM OVERVIEW

The goal of the *HighLife* system is to harvest n-ary facts from text corpora. One key feature is composing higher-arity facts from partial observations by joining arguments, e.g. one sentence referring to a drug, a disease and a target group and another one referring to the same drug, same target group, a dosage but not the disease are joined into a single fact containing all 4 pieces of information.

Figure 8 gives an overview of the HighLife system. To show the versatility of the approach, two different domains are considered in our experiments, namely health and news. A Named Entity Recognition and Disambiguation (NERD) component extracts a variety of entities from sentences. To identify fact candidates our system then constructs trees from parsed dependency graphs spanning over the entities. These trees either express a complete fact or have missing entities leading to unknown arguments and partial facts. Guided by distant supervision using seed facts, the extracted trees are analyzed and statistically weighted to determine good n-ary fact candidates. A logical consistency reasoner incorporates these weighted candidates together with specialized consistency rules as well as semantic information from knowledge bases to identify a consistent subset of true facts with a high total weight. Further, the reasoner composes complete facts out of partially expressed fact candidates as well as estimates an appropriate weight. The result is a set of n-ary facts, where each fact binds arguments that trace back to multiple, separate sources in the input texts.

## 5.4 ENTITY RECOGNITION AND DISAMBIGUATION

*HighLife* incorporates different entity recognition and disambiguation components that recognize entities from text and link them to knowledge bases. This allows us to incorporate a large variety of different kinds of entities into our fact extraction. As preprocessing, Stanford CoreNLP is applied on all texts.

**Biomedical Entities:** Similar to KnowLife, we rely on the UMLS (see Section 2.1.2) as biomedical entity dictionary, due to its extensive coverage of all kinds of biomedical entities, i. e. diseases, anatomy, genes, treatments, etc. To efficiently handle the large dictionary and process large input corpora, we again adapt the method presented in [168], using string-similarity matching against the names in UMLS. *HighLife* leverages entity type information and UMLS'es ranked list of entity preferences to disambiguate between multiple entity candidates matched to the same noun chunk in the input. In the first filtering step, we reduce the number of entity candidates by only retaining the most specifically typed entities according to the UMLS semantic type system. Taking into account that UMLS provides a ranked list of entities for every possible name, we further disambiguate between the remaining candidates by determining the highest ranked entities. In case two entities share the same rank, we determine their popularities by the number of occurrences in different UMLS source vocabularies and prefer the more popular entity.

**Quantities:** Numerical quantities are important quantifiers for many relations. Our system detects such information in text using powerful regular expressions incorporating entity types, POS tags, words and word classes. We developed a small set of expressions to detect quantities such as prices, percentages, and measurements among others. For instance, the expression

```
word:/USD|$/ [ word:IS_NUM | ner:MONEY ]+
```

denotes dollar prices such as *USD 1 billion*.

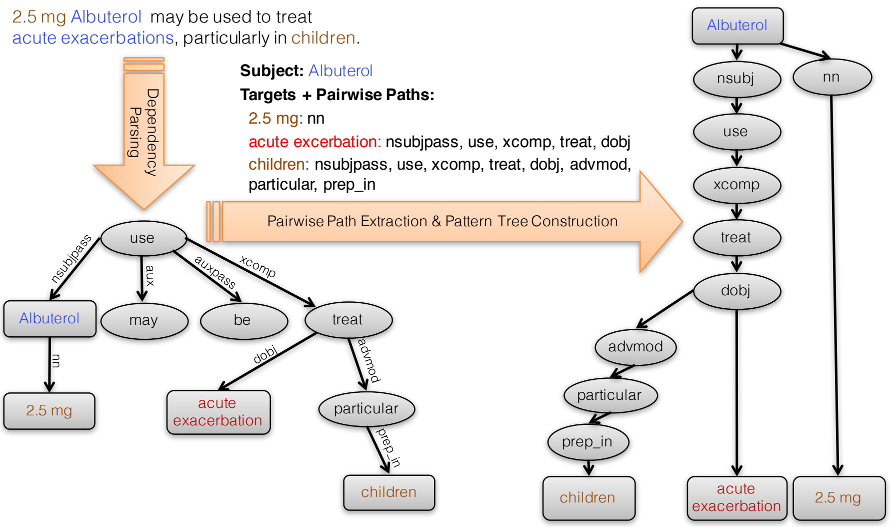

2.5 mg Albuterol may be used to treat
acute exacerbations, particularly in children.
Dependency Parsing
Subject: Albuterol
Targets + Pairwise Paths:
2.5 mg: nn
acute excerbation: nsubjpass, use, xcomp, treat, dobj
children: nsubjpass, use, xcomp, treat, dobj, advmod, particular, prep_in
Pairwise Path Extraction & Pattern Tree Construction
use
nsubjpass
aux
auxpass
xcomp
Albuterol
may
be
treat
nn
dobj
advmod
2.5 mg
acute exacerbation
particular
prep_in
children
Albuterol
nsubj
nn
use
xcomp
treat
dobj
advmod
particular
prep_in
children
acute exacerbation
2.5 mg

Figure 9: Pattern Tree Construction

**YAGO Entities:** To recognize and disambiguate entities in news we apply the AIDA system [72] which links entities to YAGO [73].

**WordNet Concepts:** We apply a most frequent sense disambiguation to map remaining noun chunks to concepts in WordNet.

**Temporal Expressions:** Using Stanford's CoreNLP sutime module we detect and normalize time expressions.

## 5.5 TREE MINING

Our method relies on constructing trees, called pattern trees, from typed dependency graphs to identify n-ary fact candidates in text. A fact candidate can be fully expressed by such a pattern tree or only partially. The goal is to construct pattern trees, which describe n-ary facts $R(e_1, \ldots, e_n)$ and reflect their complex structure.

**Definition 12 (Target)** *For a given sentence s with dependency-parse tree* $T(s)$*, the targets are the vertices of* $T(s)$ *denoting arguments of a (partial) fact (i. e. , entities, quantities, informative phrases).*

**Example 11 (Target)** The entities *Albuterol*, *acute exacerbations*, *children*, and the recognized quantity *2.5 mg* are the targets of the sentence given in Figure 9.

We assume that the targets in a sentence are already canonicalized whenever appropriate; for example, entity mentions are disambiguated into an entity of a KB, quantities are normalized and annotated with units, etc. Since targets may actually be multi-word phrases, we transform the dependency-parse tree to collapse all vertices that constitute a target phrase into a single vertex. This combined vertex is placed at the position of the phrase's head word in the original parse tree.

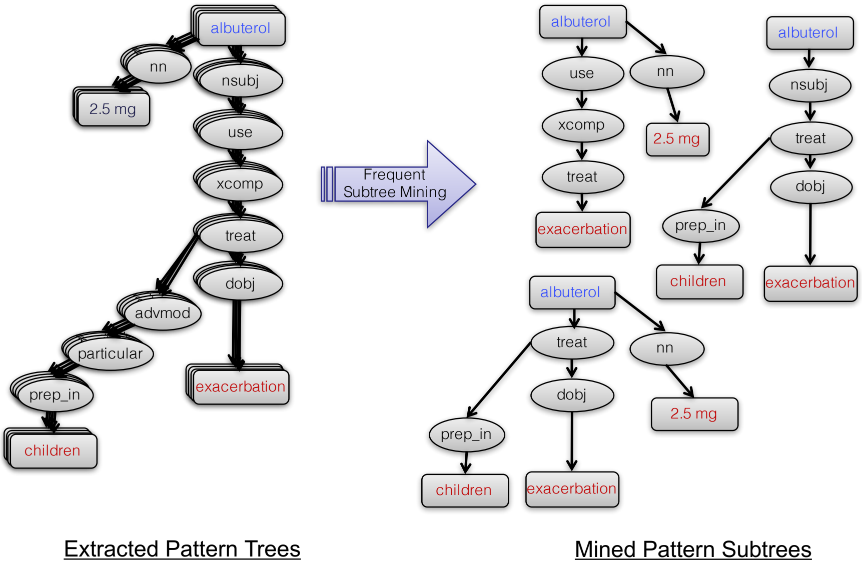


Figure 10: Subtree Mining Example

**Definition 13 (Pairwise Paths)** *For sentence s, the set of pairwise paths* $\mathsf{PP}(s)$ *contains all parse-tree paths linking a pair of targets.*

**Example 12 (Pairwise Paths)** Figure 9 depicts a few examples of pairwise paths starting from the entity *Albuteral*.

**Definition 14 (Matching Tree)** *For sentence* $s$, *the matching tree is the parse tree reduced to having only the sentence's targets as leaf vertices and all pairwise paths.*

**Definition 15 (Pattern Tree)** *Given a sentence* $s$, *the pattern tree* $P(s)$ *is the matching tree with the sentence's subject target as the root and all pairwise paths, with common subpaths that include the root represented only once.*

**Example 13 (Matching and Pattern Tree)** Figure 9 depicts the step-by-step construction of a pattern tree.

The goal is to construct pattern trees, which describe n-ary facts $R(e_1, \ldots, e_n)$ and reflect their complex structure.

## 5.6 TREE ANALYSIS

The harvested pattern trees can often be too large and over-specific, i. e. a sentence's pattern tree often contains more entities than allowed to represent a valid fact candidate. That is, a sentence may cover more entities than there are possible arguments for a relation. However, a subset of the entities and thus a subtree of the pattern tree could lead to a true fact. Also, not all internal vertices are often needed to express a relation and we only want to consider the important and necessary ones, e.g. as illustrated in Figure 10.

Our goals are to generalize the extracted pattern trees to mine subtrees by masking out inessential vertices, and to identify seed trees using our set of seed facts. The resulting trees, called salient subtrees, should syntactically and lexically express n-ary relations with high confidence. We use the seed trees to identify fact candidates, which are weighted subtrees where the weight describes the confidence that this tree expresses a particular relation.

### 5.6.1 *Salient Seed Tree Mining*

**Tree Generalization** We generalize the harvested trees to find salient trees by adapting the FreeTreeMiner algorithm [37] which mines all frequent subtrees satisfying a given support threshold. We extend this algorithm to incorporate lexical and semantic information into the tree mining. If vertices in trees do not occur often enough, our algorithm lifts them to either their part-of-speech tags, to a general wildcard, or to their semantic type.

**Seed Tree Mining** Guided by the seed facts (see Definition 10 in Section 2.2.3), which are either manually defined or harvested from a knowledge base, we can identify seed trees within the mined subtrees.

**Definition 16 (Pattern Subtree)** *Given a sentence* $s$ *from a set of sentences* $S$*, the pattern subtree* $PS(s)$ *is a mined subtree of the sentence's pattern tree, which only has the sentence's targets as leaf vertices and which occurs more often than a predefined threshold. A seed pattern subtree for a relation* $R$ *is a pattern subtree where the root and leaf vertices are the targets of a seed fact. Such a tree could represent a partial seed fact by matching the fact only partially.*

**Definition 17 (Pattern Subtree Confidence)** *Given a corpus of sentences* $S$ *and a set of entity tuples* $X$*, the support of a pattern subtree* $PS(s)$ *based on* $S$ *and* $X$ *is computed as:*

$$\begin{aligned} &\mathrm{support}(PS(s), X) = \\ &\qquad |\{s \in S \mid \exists (e_1, \ldots, e_n) \in X \wedge PS(s) \text{ contains } e_1, \ldots, e_n\}| \end{aligned}$$

*The confidence of a pattern subtree* $PS(s)$ *for a relation* $R$ *is:*

$$\mathrm{confidence}(PS(s)) = \frac{\mathrm{support}(PS(s), SX(R))}{\mathrm{support}(PS(s), SX(R) \cup CX(R))}$$

*where* $SX(R)$ *denotes the set covering all entity tuples of true facts of relation* $R$ *in our seed facts, and* $CX(R)$ *denotes the negative entity tuples, i.e. entity tuples which would be valid arguments of a relation, but do not lead to a true fact. A salient seed tree is a generalized seed tree having a confidence larger than a specific threshold.*

**Example 14 (Pattern Subtree Confidence)** A few example salient trees together with computed confidences can be seen in Figure 11.

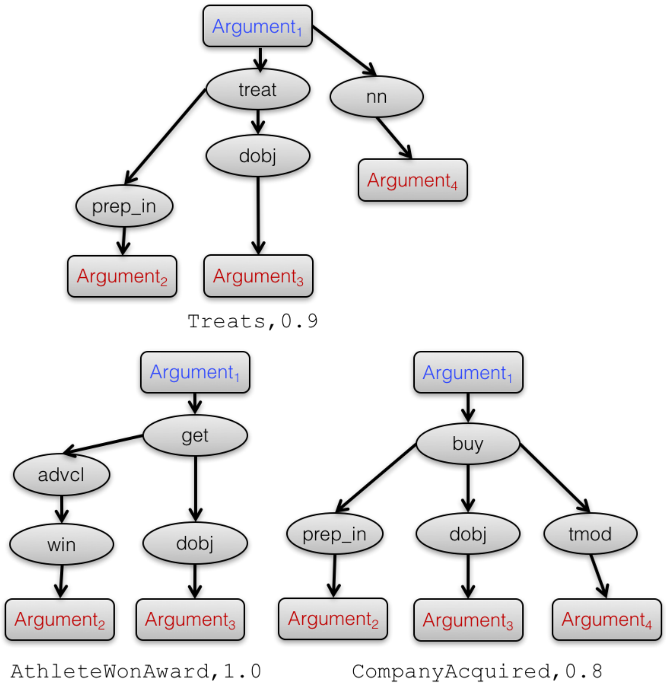


Figure 11: Salient Seed Tree Examples

### 5.6.2 *Partial N-ary Fact Candidates*

Every mined subtree is a potential candidate for an n-ary fact. Subtrees do not need to express facts completely, in which case they lead to partial fact candidates. To quantify the goodness of a subtree to be a fact candidate, the tree is matched against the salient seed trees, i. e. a weight is assigned describing the confidence that the tree expresses a particular relation.

**Tree Matching** Having the same number of leaf vertices is a necessary condition for two trees to be considered for matching. To define a similarity measure between trees, we first introduce a measure for computing the similarity between two sequences.

**Definition 18 (Vertex Similarity)** *Let* $1(i_1, i_2)$ *be the function that indicates, if two vertices are equal, i. e. if they represent the same word, grammatical relation, etc.*

$$1(v_1, v_2) = \begin{cases} 1, \textit{if } v_1 \textit{ and } v_2 \textit{ are equal} \\ 0, \textit{otherwise} \end{cases}$$

**Definition 19 (Similarity between Paths)** *The similarity* $sim_P(p_1, p_2)$ *between two pairwise paths* $p_1$ *and* $p_2$*, each of which connects two targets in pattern trees, is defined as*

$$sim_P(p_1, p_2) = \begin{cases} \frac{\sum_{i=0}^{|p_1|} 1(p_{1_i}, p_{2_i})}{|p_1|}, \textit{if } |p_1| = |p_2| \\ 0, \textit{otherwise} \end{cases}$$

| Source | Predicate | Description |
|---|---|---|
| Textual Evidence | $\mathrm{Express}(T, R)$<br>$\mathrm{Occur}(T, X_1, \ldots, X_n)$ | tree T expresses relation R<br>T occurs with $n$ entities in text |
| Relation Properties | $\mathrm{Type}(X, S)$<br>$\mathrm{Sig}(R, S_1, \ldots, S_n)$ | type S of an entity E<br>argument type signature of an n-ary relation R |
| Domain Knowledge | $\mathrm{OrganPartOf}(X, Y)$<br>$\mathrm{GroupInCountry}(X, Y)$ | organ X is part of organ Y<br>ethnic group X lives in country Y |
| Derived Output | $\mathrm{CompanyAcquired}(X_1, X_2, X_3, X_4, X_5)$<br>$\mathrm{Diagnoses}(X_1, X_2, X_3)$ | N-ary fact hypotheses |

Table 14: *HighLife* Predicate Examples

*where* $|\cdot|$ *denotes the length of a path.*

**Definition 20 (Similarity between Trees)** *Let* $P_1 = (p_{1_1}, \ldots, p_{1_n})$ *denote a possible arrangement of pairwise paths for pattern tree* $t_1$ *and* $P_2 = (p_{2_1}, \ldots, p_{2_n})$ *an arrangement for tree* $t_2$*, we can define a similarity measure between two pattern subtrees* $t_1, t_2$ *as follows:*

$$\mathrm{sim}_T(t_1, t_2) = \operatorname*{argmax}_{P_1, P_2} \prod_{i=1}^{n} \mathrm{sim}_P(P_{1_i}, P_{2_i})$$

Finally, we can formally describe tree fact candidates as:

**Definition 21 (Tree Fact Candidates)** *For a set of sentences* $S$ *and a seed tree set* $Q$*, the n-ary tree fact candidate multi-set* $C(S, Q)$ *is:*

$$\begin{aligned} C(S,Q) = &\{(PS(s), e_1, \ldots, e_n)[w] | \\ &\exists s \in S : PS(s)\ \mathrm{contains}\ e_1, \ldots, e_n \wedge \\ &w = \max\{\mathrm{sim}_T(PS(s), q) \times \mathrm{confidence}(q) | q \in Q\}\} \end{aligned}$$

## 5.7 CONSISTENCY REASONING

The tree fact candidate multi-set describes weighted pattern trees which potentially lead to full or partial facts and thus produce a set of weighted n-ary fact candidates. *HighLife* uses consistency rules to determine when such a tree becomes a true n-ary fact, i. e. the rules prune false positives out of the set of n-ary fact candidates and their supporting tree patterns provided by the tree analysis.

| Type | Rule |
|---|---|
| Tree pattern-fact duality<br>(Fact Hypotheses Generation) | $\mathsf{Express}(T,R) \wedge \mathsf{Occur}(T,X_1,\ldots,X_n) \wedge \mathsf{Sig}(R,S_1,\ldots,S_n) \wedge \mathsf{Type}(X_1,S_1) \wedge \ldots \wedge \mathsf{Type}(X_n,S_n)$<br>$\Rightarrow R(E_1,\ldots,E_n)$ |
| Mutual Exclusion | $\mathsf{Causes}(X_1,X_2,X_3,X_4) \Rightarrow \neg\mathsf{Treats}(X_1,X_2,X_3,X_4)$<br>$\mathsf{CompanyAcquired}(X_1,X_2,X_3,X_4,X_5) \Rightarrow \neg\mathsf{CompanyAcquired}(X_2,X_1,X_3,X_4,X_5)$ |
| Domain Constraints | $\mathsf{Diagnoses}(X_1,X_2,X_3) \wedge \mathsf{Diagnoses}(X_1,X_2,Y_3) \Rightarrow \mathsf{OrganPartOf}(X_3,Y_3)$<br>$\ldots$ |
| Equality Constraints | $\mathsf{AthleteWonAward}(X_1,X_2,X_3,X_4,X_5,X_6) \wedge \mathsf{AthleteWonAward}(X_1,X_2,Y_3,X_4,X_5,X_6) \Rightarrow E_3 = Y_3$<br>$\mathsf{CompanyAcquired}(X_1,X_2,X_3,X_4,X_5) \wedge \mathsf{CompanyAcquired}(X_1,X_2,X_3,Y_4,X_5) \Rightarrow E_4 = Y_4$ |

Table 15: Consistency Constraints

### 5.7.1 *Consistency Constraints*

Consistency constraints are encoded as rules that are composed of multiple different predicates. A predicate (see Table 14) can describe evidence extracted from text, logical relation properties, domain knowledge from a Knowledge Base, or it is derived as a result of executing a rule. The rules enforce consistency over the set of fact candidates and handle conflicting candidates. We rely on the different types of consistency constraints, shown in Table 15. Tree pattern-fact duality constraints describe when a tree pattern candidate becomes a fact candidate. Mutual exclusion constraints between relations rule out different fact candidates, which overlap in their arguments but conflict in their relations. Domain constraints restrict possible results by incorporating prior domain knowledge. Rules can also impose equality restrictions, specifying when arguments of two different facts are equal. Such constraints could express that facts making statements about the same athlete winning a medal on the same date must overlap in the athlete's type of sport. The consistency constraints are listed in Appendix a.2.

### 5.7.2 *Partial Fact Reasoning*

To reason with the aforementioned constraints, we ground the rules into weighted logical clauses. The clauses' weights are derived from the weights of the tree analysis phase. The goal is to compute a consistent subset of clauses with the largest total weight, i. e. to identify a subset of most plausible fact candidates. This task can be cast into a Weighted Max-Sat problem [176]. However, facts can have unknown arguments (partial facts), which cannot be handled by the weighted Max-Sat solver. The problem of determining constants for X and Y and groundings for unknown arguments in partial facts can be reduced to a unifica tion problem between logical literals.

**Example 15** The partially grounded fact candidates
$AthleteWonAward(Kerber, OlympicSilver, tennis, 2016, X)$
$AthleteWonAward(Kerber, OlympicSilver, Y, 2016, Rio)$ etc.
could be applied to the following formula:
$\exists X, Y AthleteWonAward(Kerber, OlympicSilver, tennis, 2016, X) \wedge$
$AthleteWonAward(Kerber, OlympicSilver, Y, 2016, Rio)$

We unify two partial n-ary fact candidates, if we can find a substitution between them, i. e. a mapping assigning constants to unknown arguments of partial facts. We use equality constraints defined as consistency rules to determine when arguments of two partial facts can be considered equal and thus can be substituted. For example, using a constraint which expresses that an athlete cannot win medals in more than one sports discipline on the same date, we can determine that $Y$ can only be substituted with $tennis$ in Example 15. Exploiting these constraints for defining equivalences, we implement the MGU algorithm described in [20, p. 71] to find most general unifiers between logical literals. This enables us to unify partially grounded fact candidates resulting in new fully grounded clauses. This unification combines information scattered in separate textual sources into a single, full-fledged n-ary fact, e.g. by substituting X and Y with constants (Rio for X, tennis for Y), we obtain the clause:

$$AthleteWonAward(Kerber, OlympicSilver, tennis, 2016, Rio)$$

However, we need to assign a weight to the clause to use it in the reasoning. The weights for the partial candidates correspond to observations of marginals over a 5-variate distribution. We need to estimate the hypothetical frequency for the full clause. In the absence of any other information, we can use a maximum-entropy estimator. This estimation problem is isomorphic to the cardinality estimation issue over multivariate datasets [99]. However, not all partial facts can be unified into fully grounded clauses. Therefore, we introduce special unknown arguments as placeholders to ground such facts. Due to the NP-hardness of the Weighted Max-Sat problem, we use the approximation algorithm SOFIE [176]) to reason over the created hypotheses space of grounded and weighted clauses which produces a set of n-ary facts we accept as plausible.

## 5.8 EXPERIMENTS

For empirical studies of the viability and comparative performance of our *HighLife* method, we designed various experiments using input texts and target relations from two domains: general news (on business, sports, etc.) and biomedical health. First, we compare *HighLife* to a state-of-the-art SRL baseline (Section 5.8.3). Second, we test the scalability of *HighLife* on two large corpora (Section 5.8.5). Third, we perform an ablation study with various components of *HighLife* enabled or disabled (Section 5.8.4). We start this section by discussing the general setup for these experiments.

### 5.8.1 *Setup*

**Datasets.** We run experiments on two different input corpora:

NEWS ARTICLES: a large collection of news articles, compiled from the STICS project [71] and the New York Times archive.

BIOMEDICAL TEXTS: a large and diverse collection of documents on biomedicine and health, consisting of i) PubMed articles with scientific content and specialized jargon, and ii) Web portals and encyclopedic articles (from MayoClinic, Wikipedia, etc.) with information geared for patients and doctors (see [51]).

| Domain | Relation | Arity | Signature |
|---|---|---|---|
| Biomedical | Treats | 5 | $Drug \times Disease \times Dosage \times DosageForm \times Targetgroup$ |
| | ReducesRisk | 4 | $(Drug \cup Behavior \cup Ecofactor) \times Disease \times Targetgroup \times Condition$ |
| | Causes | 4 | $Disease \times Disease \times Targetgroup \times Condition$ |
| | Diagnoses | 3 | $DiagnosticProcedure \times Disease \times (BodyPart \cup Organ)$ |
| News | AthleteWonAward | 6 | $Athlete \times Award \times TypeOfSport \times Event \times Location \times Time$ |
| | CompanyAcquired | 5 | $Organization \times Organization \times Date \times Price \times Organization$ |

Table 16: *HighLife*'s Harvested Relations with Type Signatures

| Domain | Genre | Source | Documents | Sentences |
|---|---|---|---|---|
| Biomedical | Encyclopedic Articles | Drugs.com | 31,837 | 7,586,236 |
| | | Mayo Clinic | 2,166 | 570,325 |
| | | Medline Plus | 3,076 | 197,055 |
| | | RxList | 2,515 | 1,102,791 |
| | | Wikipedia Health | 20,893 | 787,148 |
| | Scientific Publications | PubMed Medline | 580,892 | 5,875,006 |
| | | PubMed Central | 12,532 | 2,765,580 |
| News | | STICS Corpus | 1,462,294 | 30,252,627 |
| | | New York Times | 1,407,299 | 82,934,909 |
| | | **Total** | **3,523,504** | **132,071,677** |

Table 17: Text Corpora for Experiments

Table 17 shows the size and other properties of these corpora.

As for extraction targets, we focused on a small set of relations with different arities (i. e. , number of arguments), ranging from ternary to 6-ary. Table 16 gives an overview of these relations, and Table 23 shows sample facts extracted by *HighLife* for each of them.

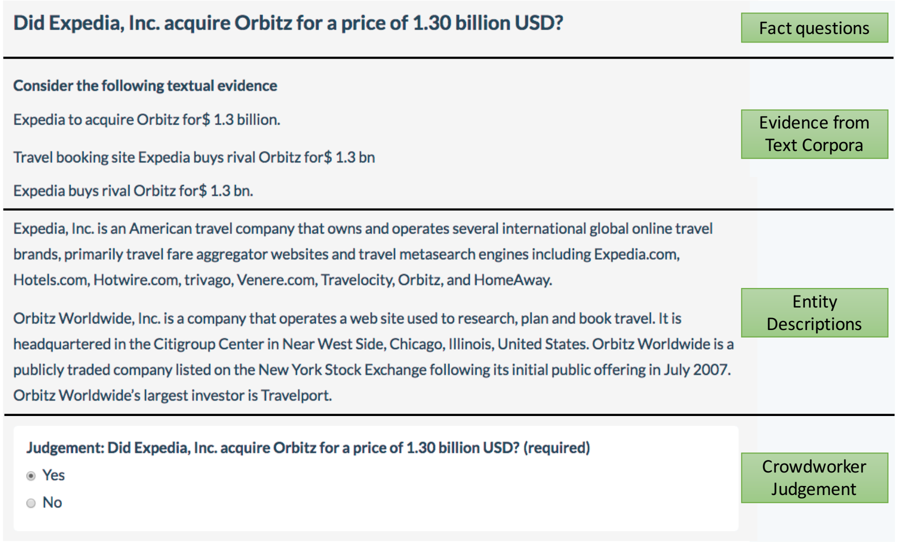
**Evaluation Metrics.** To assess the quality and coverage of the knowledge bases that *HighLife* can automatically build, we i) evaluate the correctness of randomly sampled facts and ii) report on the size of large-scale extractions (i. e. , the number of extracted facts per relation). The former is a precision measure, aggregated over all samples per relation. The latter can be seen as a proxy for recall. Note that the actual recall, in the sense of IR evaluations, cannot be estimated as it would require annotating a large number of entire documents with their maximally extractable facts. We also discuss the impact of the arity of facts (i. e. , the number of extracted arguments) on the resulting precision.

### 5.8.2 *CrowdFlower Setup*

To gather human judgments of extraction correctness and conduct unbiased experiments, we utilized crowdsourcing through the CrowdFlower platform.

To assess an extracted fact by judges of the crowdworkers pool we turn every fact into a short questionnaire, asking the judge if the fact is true or false. We provide two kinds of evidence to the judges:

- The textual sources from our input corpus where the fact was extracted from, and



**Did Expedia, Inc. acquire Orbitz for a price of 1.30 billion USD?**

**Consider the following textual evidence**

Expedia to acquire Orbitz for$ 1.3 billion.

Travel booking site Expedia buys rival Orbitz for$ 1.3 bn

Expedia buys rival Orbitz for$ 1.3 bn.

Expedia, Inc. is an American travel company that owns and operates several international global online travel brands, primarily travel fare aggregator websites and travel metasearch engines including Expedia.com, Hotels.com, Hotwire.com, trivago, Venere.com, Travelocity, Orbitz, and HomeAway.

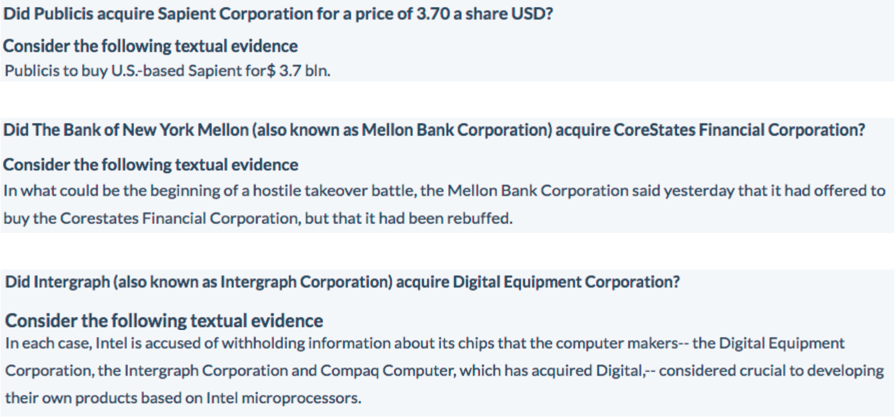
Orbitz Worldwide, Inc. is a company that operates a web site used to research, plan and book travel. It is headquartered in the Citigroup Center in Near West Side, Chicago, Illinois, United States. Orbitz Worldwide is a publicly traded company listed on the New York Stock Exchange following its initial public offering in July 2007. Orbitz Worldwide's largest investor is Travelport.

**Judgement: Did Expedia, Inc. acquire Orbitz for a price of 1.30 billion USD? (required)**

● Yes

○ No

Figure 12: CrowdFlower Task



**Did Publicis acquire Sapient Corporation for a price of 3.70 a share USD?**

**Consider the following textual evidence**

Publicis to buy U.S.-based Sapient for$ 3.7 bln.

**Did The Bank of New York Mellon (also known as Mellon Bank Corporation) acquire CoreStates Financial Corporation?**

**Consider the following textual evidence**

In what could be the beginning of a hostile takeover battle, the Mellon Bank Corporation said yesterday that it had offered to buy the Corestates Financial Corporation, but that it had been rebuffed.

**Did Intergraph (also known as Intergraph Corporation) acquire Digital Equipment Corporation?**

**Consider the following textual evidence**

In each case, Intel is accused of withholding information about its chips that the computer makers-- the Digital Equipment Corporation, the Intergraph Corporation and Compaq Computer, which has acquired Digital,-- considered crucial to developing their own products based on Intel microprocessors.

Figure 13: Different Types of False Test Questions

- additional descriptions of the entities appearing as fact arguments.

Figure 12 shows an example for the CrowdFlower task on a candidate fact for the relation CompanyAcquired.

We took several measure for quality assurance. First, we designed a set of test questions for every task, which are prejudged and cross-checked with external sources by ourselves. Second, we balanced the numbers of true and false candidate facts shown to judges, so that crowdworkers were not biased towards quickly guessing the assessment. To prevent judges from giving superficial results without carefully reading the question and context, we specifically included test questions with false components in the candidate facts: differences in the numerical quantities, textual statements that contain negations, and entities that spuriously co-occur without any real relationship. Figure 13 depicts some examples. We paid 0.5 cents for each judgement on business and sports news, and 0.83 cents for each judgement on biomedical health (the latter requiring more expertise and careful reading).The final ground truth for the samples to be assessed was determined by a weighted voting scheme among the judgements for each sample. The weights were proportional to the confidence of each judge, derived from the test questions with prejudged truth. On average, each sampled fact candidate was assessed by three crowdworkers.

### 5.8.3 *Evaluation of Extraction Quality*

In this subsection, we compare the extraction quality of *HighLife* with a state-of-the-art SRL system. We focus on the two relations from general news articles: *CompanyAcquired* and *AthleteWonAward*, for which the SRL system has high-quality frame types and has been intensively trained on. For the biomedical relations, it would be unfair to the baseline to compare *HighLife* against SRL without specific engineering and training. We evaluate the precision of the extracted facts, for varying arities (by ignoring some arguments of the relations), based on samples assessed by the CrowdFlower judges.

**Seed Facts.** For the relation CompanyAcquired, Freebase (see Section 2.1.2) provides us with ternary seed facts: the acquiring and the acquired company as well as the date of the acquisition. We manually extended these ternary facts to 5 arguments by incorporating acquisition prices and including the previous owner of an acquired company. For the relation AthleteWonAward, we gathered the seed facts from the WikiData KB WikiData stores the events (e.g. 2016 Summer Olympics) an athlete participates in together with medals won and the specific date. Combining this with other WikiData facts, such as the type of sport an athlete performs and the location of the event (e.g. Rio for the 2016 Olympics), we constructed instances of the 6-ary AthleteWonAward relation

Overall, we compiled 593 binary, 279 ternary, 45 quaternary, and 3 quintary seed facts, together with 42 binary and 28 ternary negative seed facts manually defined. Note that no 6-ary facts were spotted in any of the sentences of the corpus. However, *HighLife* can still extract 6-ary facts by combining lower-arity facts from different sentences in the reasoning stage.

**Competitors.** As discussed in Sections 5.1 and 5.2, SRL is the prior work most related to HighLife. Therefore, we selected the state-of-the-art SRL system of the UIUC Illinois Natural Language Processing (NLP) Curator software [40] as our baseline (for the software and an online demo, see (`cogcomp.org/page/software_view/Curator`). The system integrates NERD [147] and nominal relation modeling [172] into SRL [140].

| Relation by Extracted Arguments | System | Precision | | | |
|---|---|---|---|---|---|
| | | 2 | 3 | 4 | 5 |
| AthleteWonAward | HighLife-FULL | 0.80 | 0.80 | 0.81 | 0.81 |
| | HighLife-NT | 0.37 | 0.70 | 0.66 | 0.00 |
| | SRL-T | 0.68 | 0.86 | 0.86 | 0.67 |
| | SRL | 0.47 | 0.38 | 0.20 | 0.00 |
| CompanyAcquired | HighLife-FULL | 0.70 | 0.77 | 0.88 | 0.88 |
| | HighLife-NT | 0.23 | 0.53 | 0.78 | 0.83 |
| | SRL-T | 0.88 | 0.87 | 0.78 | - |
| | SRL | 0.39 | 0.20 | 0.08 | 0.00 |

Table 18: Comparison of *HighLife* against SRL Baselines

The target relations are mapped to frames in PropBank (which is the basis for SRL) as follows. CompanyAcquired is modeled by the Propbank roleset *acquisition.01*, with five argument slots corresponding to the arguments of the *HighLife* relation. For AthleteWonAward, by disregarding the third argument (*TypeOfSport*), we are able to map it to the roleset win.01 in PropBank. This makes our relations compatible with the SRL frames.

Since SRL methods and *HighLife* are still not fully comparable, we further added two pre-processing steps to the SRL system, giving it additional benefits. First, we increase the coverage of rolesets by considering all predicates that

- specify the same type of roles and
- fall into the same verb classes as defined by PropBank.

For example, for acquisitions, we manually incorporated also the predicates buy, purchase and get and their respective frame types. Second, we restrict the input in the experiment to sentences in the corpus where at least two possible arguments of the relation are mentioned. For example, sentences mentioning two companies are candidates for CompanyAcquired, and sentences mentioning mentioning an athlete and a medal are candidates for AthleteWonAward. We further implemented an extended version of the SRL system, by incorporating type constraints for the candidate extractions, thus giving SRL more power closer to what *HighLife* does.

In the following we present results for four competitors:

SRL: the native SRL system (with the pre-processing steps added as benefit),

SRL-T: the extended SRL with type constraints,

| Relation by | HighLife | Precision | | | | | Micro |
|---|---|---|---|---|---|---|---|
| Extracted Arguments | Config. | 2 | 3 | 4 | 5 | 6 | Avg. |
| AthleteWonAward | FULL | 0.80 | 0.81 | 0.82 | 0.82 | 1.0 | **0.80** |
| | NT | 0.37 | 0.70 | 0.66 | 0.0 | - | **0.39** |
| | NR | 0.73 | 0.76 | 0.76 | 0.76 | 1.0 | **0.74** |
| | NT-NR | 0.28 | 0.56 | 0.53 | 0.5 | 1.0 | **0.27** |
| | NU | 0.83 | 0.80 | 0.77 | - | - | **0.82** |
| CompanyAcquired | FULL | 0.70 | 0.77 | 0.88 | 0.88 | | **0.74** |
| | NT | 0.23 | 0.53 | 0.78 | 0.83 | | **0.30** |
| | NR | 0.70 | 0.76 | 0.87 | 0.88 | | **0.74** |
| | NT-NR | 0.33 | 0.35 | 0.63 | 0.57 | | **0.34** |
| | NU | 0.67 | 0.73 | 0.87 | - | | **0.70** |

Table 19: HighLife Ablation Study Precision.

highlife-full: the full-fledge *HighLife* extractor,

highlife-nt: the *HighLife* extractor without type constraints (making *HighLife* as type-agnostic as the native SRL).

**Results.** Table 18 shows the results of this comparison. For each of the two relations, 500 samples were evaluated by crowdsourcing.

The different columns for precision refer to different arities of the two target relations. We projected the extracted facts onto subsets of their arguments. Smaller arities focus on the main arguments (e.g., the acquiring and the acquired company and the date, but ignoring the price); so smaller arities are easier to extract correctly.

The results in Table 18 show that SRL in its type-extended variant SRL-T performs well for lower arities. For CompanyAcquired SRL-T is even the best system when focusing only on the 2 or 3 main arguments of the relation. For AthleteWonAward it is slightly better than *HighLife* for the cases of 3 and 4 arguments. SRL without type awareness is substantially inferior to all other competitors.

HighLife-Full consistently performs close to the best competitor, and is the clear winner when all arguments of the relations are to be extracted. In these full-arity cases, both of the SRL variants degrade. For example, for the 5-ary CompanyAcquired relation, the native SRL extracts only incorrect facts – hence precision 0.0; the type-enhanced SRL-T does not yield any output facts at all in this case. The type-agnostic HighLife-NT also drops significantly in output quality compared to HighLife-Full, but mostly stays at a reasonable level.

Overall, HighLife-Full shows its robustness and superiority over the SRL approach, although SRL is given the benefits of pre-filtered sentences and even when it is extended with type constraints.

| Relation by | HighLife | | | Precision | | | |
|---|---|---|---|---|---|---|---|
| Extracted Arguments | Config. | 2 | 3 | 4 | 5 | 6 | **Sum** |
| AthleteWonAward | FULL | 3,804 | 1,089 | 224 | 23 | 2 | **5,142** |
| | NT | 40,728 | 2,206 | 11 | 2 | 0 | **42,947** |
| | NR | 3,939 | 1,873 | 243 | 17 | 2 | **6,074** |
| | NT-NR | 40,728 | 2,246 | 265 | 23 | 2 | **43,264** |
| | NU | 3,804 | 1,078 | 44 | 0 | 0 | **4,926** |
| CompanyAcquired | FULL | 2,304 | 1,253 | 452 | 11 | | **4,020** |
| | NT | 19,027 | 4,090 | 787 | 17 | | **23,921** |
| | NR | 2,649 | 1,505 | 583 | 13 | | **4,750** |
| | NT-NR | 20,805 | 4,584 | 993 | 22 | | **26,404** |
| | NU | 2,306 | 1,263 | 165 | 0 | | **3,734** |

Table 20: *HighLife* Ablation Study Harvested Facts.

### 5.8.4 *Ablation Study*

We deactivated various component(s) of the full system in order to assess the contribution of the individual components. Table 19 shows the precision and Table 20 the number of harvested facts under different configurations and numbers of known arguments. FULL refers to the full system with all components enabled, which performs the best. When entity types are disregarded (NT), the relations' type signatures no longer apply and all entities in matching patterns lead to fact candidates. As shown in Table 19, the number of facts increased tremendously, but precision also drops tremendously under this setting. We see similar but less severe effects when consistency rules are not applied (NR) and conflicting fact candidates are no longer pruned out. The configuration NR-NT denotes disabled entity type and consistency constraint checking, leading to an increased number of higher-arity facts while binary and ternary facts remain largely unaffected. We observe the trends for all configurations that the higher the arity, the higher the precision. When deactivating unification (NU), partial facts are no longer combined to form more complete facts and the harvesting of higher-arity facts is negatively impacted; we observe that unification is essential to harvesting facts with 4 or more arguments.

### 5.8.5 *Large-Scale Experiments*

In order to demonstrate that our proposed method works well across different domains and to demonstrate that the method scales well, we performed large-scale experiments for news and biomedicine.

| Relation by<br>Extracted Arguments | Precision<br>2 | <br>3 | <br>4 | <br>5 | **Micro<br>Average** |
|---|---|---|---|---|---|
| Treats | 0.85 | 0.85 | 0.93 | 1.00 | **0.86** |
| ReducesRisk | 0.81 | 0.83 | 0.99 | | **0.82** |
| Causes | 0.80 | 0.80 | 0.85 | | **0.80** |
| Diagnoses | 0.88 | 0.96 | | | **0.89** |

Table 21: Precision in the Biomedical Domain

#### 5.8.5.1 *News Text*

Contrary to the extraction quality experiment we apply *HighLife* on the entire corpus by using the same seeds. In addition, AthleteWonAward incorporates one more possible argument increasing the arity from five to six, since we do not need to be compatible with Propbank frames.

**Results** The best performing system configuration (FULL) achieves an average precision of 0.77%. Table 21 and 22 show the precision and the number of harvested facts. In terms of sources of error, our results suffered most from unquestioningly accepting statements of speculation as facts. Speculation is prevalent in news, especially for acquisitions when some company is reported to consider acquiring another company prior to the actual transaction. We believe that speculation detection such as [180] is a complementary method that can be orthogonally applied in addition to our method for fact harvesting. Another source of error is the repetitive nature of CompanyAcquired's type signature. Coupling the triple appearance of organization in the signature with the numerous non-acquisition-related relationships (such as companies suing, competing with, etc. one another) between them, the signature is not distinctive enough to separate the arguments. Annotators were presented with 600 randomly selected AthleteWonAward facts and 500 CompanyAcquired facts. As for inter-annotator agreement, the value for Fleiss' Kappa was 0.568 for CompanyAcquired and 0.483 for AthleteWonAward, which indicates a moderate agreement among annotators.

#### 5.8.5.2 *Biomedical Text*

The biomedical relations have signatures with 3 to 5 types, some of which are applicable in multiple relations (see Table 16). The relation *Treats* describes not only drug treatments for diseases, but also critical information about dosage (e.g. 2.5 mg), dosage form (e.g. topical cream), and target groups (children or women). *ReducesRisk* facts describe a drug, a behavior (e.g. exercise), or an ecological factor (e.g. sunlight) that reduces the risk of a disease for a certain target group

| Relation by Extracted Arguments | #Facts 2 | 3 | 4 | 5 | **Sum** |
|---|---|---|---|---|---|
| Treats | 10,472 | 3004 | 198 | 5 | **13,769** |
| ReducesRisk | 5,339 | 1,541 | 72 | | **6,952** |
| Causes | 21,254 | 2,517 | 70 | | **23,841** |
| Diagnoses | 5,607 | 1,170 | | | **6,777** |

Table 22: Harvested Facts in the Biomedical Domain

carrying a condition (e.g. pregnancy). *Causes* describes one disease that causes another disease in the context of a target group and a certain condition. *Diagnoses* states which medical procedure diagnoses which disease manifesting in a certain body part or organ.

**Seed Facts** We manually collected 474 seed facts from medical online portals Uptodate.com, Drugs.com. 294 seed facts are binary, 165 ternary, 14 quaternary, and 2 quinary.

**Results** Our system achieved an average precision of 0.83%. Table 21 shows the precision and Table 22 the number of harvested facts under different numbers of known arguments. As for inter-annotator agreement, the values of Fleiss' Kappa were between 0.46 to 0.49 for relations Treats, Causes and Diagnoses, which indicates a moderate agreement among annotators; for ReducesRisk it was 0.37 which indicates fair agreement. Precision is promising, with the lowest at 0.80 and other settings above 0.90. Contrary to the intuition that the higher the arity, i. e. more known arguments, the more difficult it is to correctly capture all the arguments thus leading to a lower precision, our results instead show that precision increases with arity. When the arity is higher, the trees gathered are more comprehensive, which in turn contribute to more expressive patterns for capturing a relation. On the other hand, without unification the number of higher-arity facts drops significantly, effectively shutting down the possibility of harvesting facts with 5 or more known arguments. Errors made by our method can be attributed to two main sources. First, sentence structures are often complex in biomedical text, especially in scientific publications. This leads to errors in dependency parse trees, which further cascades into errors in the tree patterns. Second, entity typing in UMLS is not fine-grained enough to support clear-cut delineation in the relation property predicates during constraint reasoning.

## 5.9 CONCLUSIONS

In this chapter we presented *HighLife*, a versatile approach to harvest higher-arity facts from texts. The described method combines the mining of tree patterns based on dependency parses for high recall of fact

| Fact | Textual Evidence/Observations |
|---|---|
| AthletewonAward(DavidReid, GoldMedal, SummerGames, Atlanta, 1996, Boxing) | Reid won the United States' only gold medal in boxing, on Sunday. |
| | David Reid won a gold medal in the Summer Olympic Games in Atlanta in 1996. |
| CompanyAcquired(Hewlett, Compaq, 2002/05, USD 19 bln, Unknown) | Yesterday's report was the second filing of results since Hewlett-Packard acquired Compaq last May. |
| | Hewlett has not achieved the promised benefits from its $19 billion purchase of Compaq Computer. |
| Treats(ImmuneGlobulin, Immunodeficiencies, 10%, Intravenous, Humans) | Immune Globulin Intravenous (human) 10% is indicated for the treatment of immunodeficiency disorders. |
| Causes(Smoking, Miscarriage, Unknown, Pregnancy) | Smoking cigarettes during pregnancy can cause low birth weight, miscarriage, or stillbirth. |

Table 23: Harvested N-ary Fact Examples

candidates, and consistency reasoning to prune out false candidates for high precision of eventually accepted facts. A key feature is the use of unification during consistency reasoning, which merges multiple partial facts into full facts, thus enabling the harvesting of facts with 4 or more known arguments. The *HighLife* approach is also versatile: it can extract facts with any number of arity, and it is not limited to specific domains. We used diverse sets of relations from the Biomedical and News domains in our experiments to demonstrate that the approach harvests facts with higher-arity (up to 6 known arguments) with high precision. Table 23 shows some sample facts harvested for different relations.

# Part III

# KNOWLEDGE-BASED APPLICATIONS

# 6 KNOWLIFE'S ONE-STOP HEALTH PORTAL

## 6.1 INTRODUCTION

Knowledge Bases (KBs) have become great assets in interpreting and enriching Web contents for entity-relationship-oriented search and browsing, recommendations, and analytics [177, 209]. The Google Knowledge Graph and the IBM Watson technology are prominent examples. Projects of this kind have looked at entities and facts in a broad, general-purpose manner. However, interesting use-cases often require domain-specific knowledge at a depth and coverage that universal KBs do not provide. This is especially critical for the health and life sciences domain, which we target in this thesis.

Databases on the Web contain a wealth of information about proteins, genes, and molecular pathways, and there is also an enormous amount of health-oriented, textual information on diseases and drugs available in specialized portals and discussion forums. Besides scientific publications found in PubMed Medline, physicians as well as laymen also consult health portals on the Web. Examples are Mayo Clinic[1], Medline Plus[2], UpToDate[3], or Wikipedia's Health Portal[4].

Moreover, there are rapidly growing medical online communities, such as Healthboards.com or Patient.co.uk, sharing experience and knowledge about health issues, such as side effects of drugs and drug combinations, or symptoms of diseases.

All this constitutes a rich universe of health information, but the information is spread across many sources, mostly in textual form, and unorganized – far from being anywhere near a universal semantic health portal. Our research presented here aims to fill this gap by leveraging a powerful KB, serving as foundation for an one-stop portal that comprises knowledge on all health-related aspects in an integrated manner.

A suitable KB should take a holistic view of biomedicine: for example, relating genetic factors of diseases to other risk factors such as nutritional habits and life style, looking into side effects of combinations of drugs rather than single drugs alone, or analyzing the experience of patients on mass diseases such as asthma or diabetes.

The portal demonstrated in this chapter, called *KnowLife One-stop Health Portal*, builds on the work described in Chapter 4, but extends it by constructing a much richer one-stop KB targeted to provide an

[1] mayoclinic.org

[2] medlineplus.gov

[3] uptodate.com

[4] en.wikipedia.org/wiki/Portal:Health_and_fitness

integrated and holistic view of available health-care contents. In contrast to prior work, we tap into both life-science publications and health-related online forums, and integrate the extracted facts with biomedical backbone knowledge. The acquired knowledge, including textual patterns for relations, is used to annotate any kind of input document, expert-level or layman style, with entities and relationships on the fly, as the user reads it. The value of the realized *One-stop Health Portal* is demonstrated by several use-case scenarios: laymen exploring health issues of personal interest, medical professionals searching for specific knowledge, and researchers "speed-reading" publications via entity-relationship synopses.

The salient contributions of this chapter can be highlighted as follows:

- Constructing a large KB on a wide range of health-centric relations with entity linking to UMLS and integrated external sources in order to provide an integrated and holistic view of available health-care contents to end-users

- Automatically annotating newly seen documents from the scientific literature or from social media with relevant entities and relationships mentioned in natural-language form

## 6.2 RELATED WORK

The majority of related works try to enhance search capabilities on scientific publications and are restricted to molecular entities. In particular, Kim et al. [80] Textpresso [116], GoPubMed [48], FACTA+ [184], EVEX [186], BioTextQuest+ [130] and CRAB [66] focus on facilitating information retrieval by improving result rankings or search queries via query expansion or synonym search. Along these lines, Ferret [173] performs specialized NERD and fact harvesting on MEDLINE articles to retrieve and rank sentences mentioning gene-centric entities and facts. The system offers entity-aware query formulation and expansion and visualizes results as heat map. PolySearch2 [96] goes beyond scientific publications and offers 'Given *X*, find all associated *Y*s' queries, where *X* and *Y* are two types, e. g. diseases, toxins, etc, restraining the search together with optionally provided keywords. MEDIE [111] and GeneView [182] annotate PubMed articles with various kinds of biomedical entities and events. Semedico [54] supports interactive entity- and relation-aware query formulation on PubMed articles with annotated entity and fact occurrences. However, all the aforementioned systems do not allow interactive exploration and browsing of biomedical facts over a large corpus.

On the contrary, ALIBABA [134] uses the PubMed search interface to graphically visualize information on associations between molecular entities, which can be interactively browsed and explored. However, the system is still restricted to one domain and is not able to annotate user provided content. The interactive exploration system Life-iNet [150] provides a relation-based exploration, where user can query the knowledge with triple patterns, explore the extracted factual knowledge using a network visualization, or digest distinctive entity summarizations for user-provided type queries. The goal of the Literome [138] is to facilitate browsing, searching and reasoning over extracted genomic knowledge [137] from PubMed articles by allowing users to search for harvested facts using triple patterns or by traversing the extracted KB step by step. DELVE [68] is a modular faceted browser for exploring Pubmed search result, which provides different kinds of visualization, such as word clouds, phrase nets, and word trees, which can also leverage information from KB. In contrast to the *Knowlife One-stop Health Portal*, these systems do not support interactive and on-the-fly annotation of user-provided content.

Pang et al. [129] build a slider-based UI for exploring and discovering information in health websites, but do not leverage a KB nor BioNERD to analyze documents.

## 6.3 AN ONE-STOP HEALTH KNOWLEDGE BASE

The core of the *KnowLife One-stop Health Portal* is a Knowledge Base (KB), which covers a large entity dictionary, cross-domain facts connecting large parts of this dictionary, together with textual provenance information as well as linked external sources. Figure 14 gives an overview of the portal's architecture.

**UMLS:** As entity dictionary, we rely on UMLS because of its broad coverage (see Section 2.1.2). Furthermore, we use its semantic type system of 127 universal types as upper hierarchy spanning over 3 million entities. The types are further subdivided and grouped into 15 *Semantic Groups* defined by McCray et al. [103]. A group corresponds to a particular biomedical sub-area, e. g. *Physiology*, *Disorders*, *Genes & Molecular Sequences*, etc.

**KnowLife:** We connect UMLS entities with cross-domain facts from KnowLife, i. e. relationships connecting different biomedical domains. As described in Chapter 4, KnowLife contains more than 500,000 of such facts at a precision of 96% connecting different biomedical areas such as genes, diseases, anatomic parts, symptoms, treatments, as well as environmental and lifestyle risk factors for diseases.

**Text Corpus:** The portal is based on the text corpus described in Chapter 4, Table 5. We encode provenance information of the harvested logical facts in KnowLife and UMLS entities, i. e. their mentions and occurrences within the text corpus, as well as the relational patterns of each relation in the KB, so that users can explore the evidence for relational facts and entities.

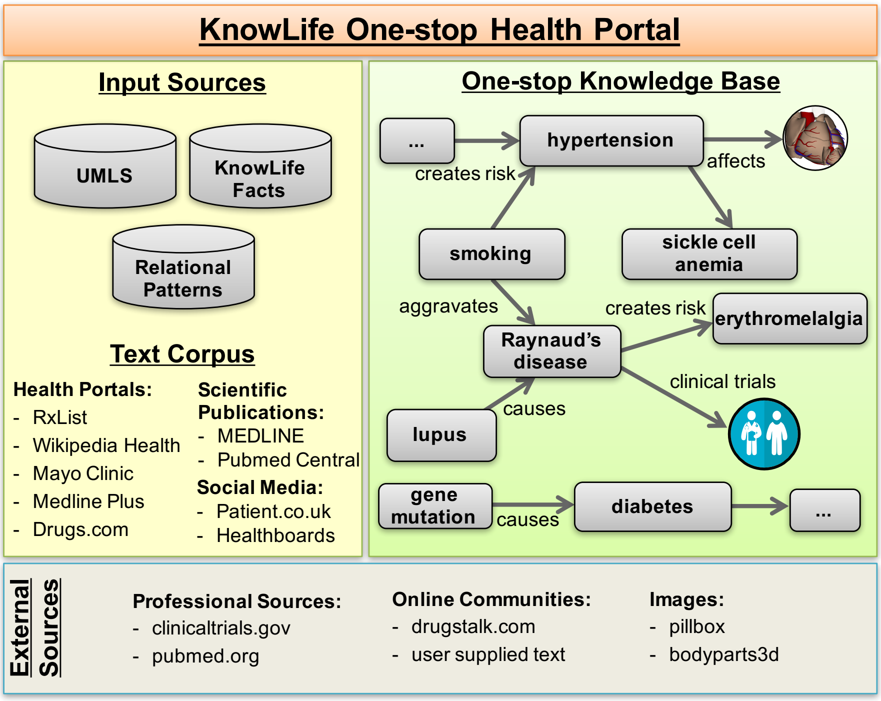


Figure 14: Overview of the *KnowLife One-stop Health Portal*

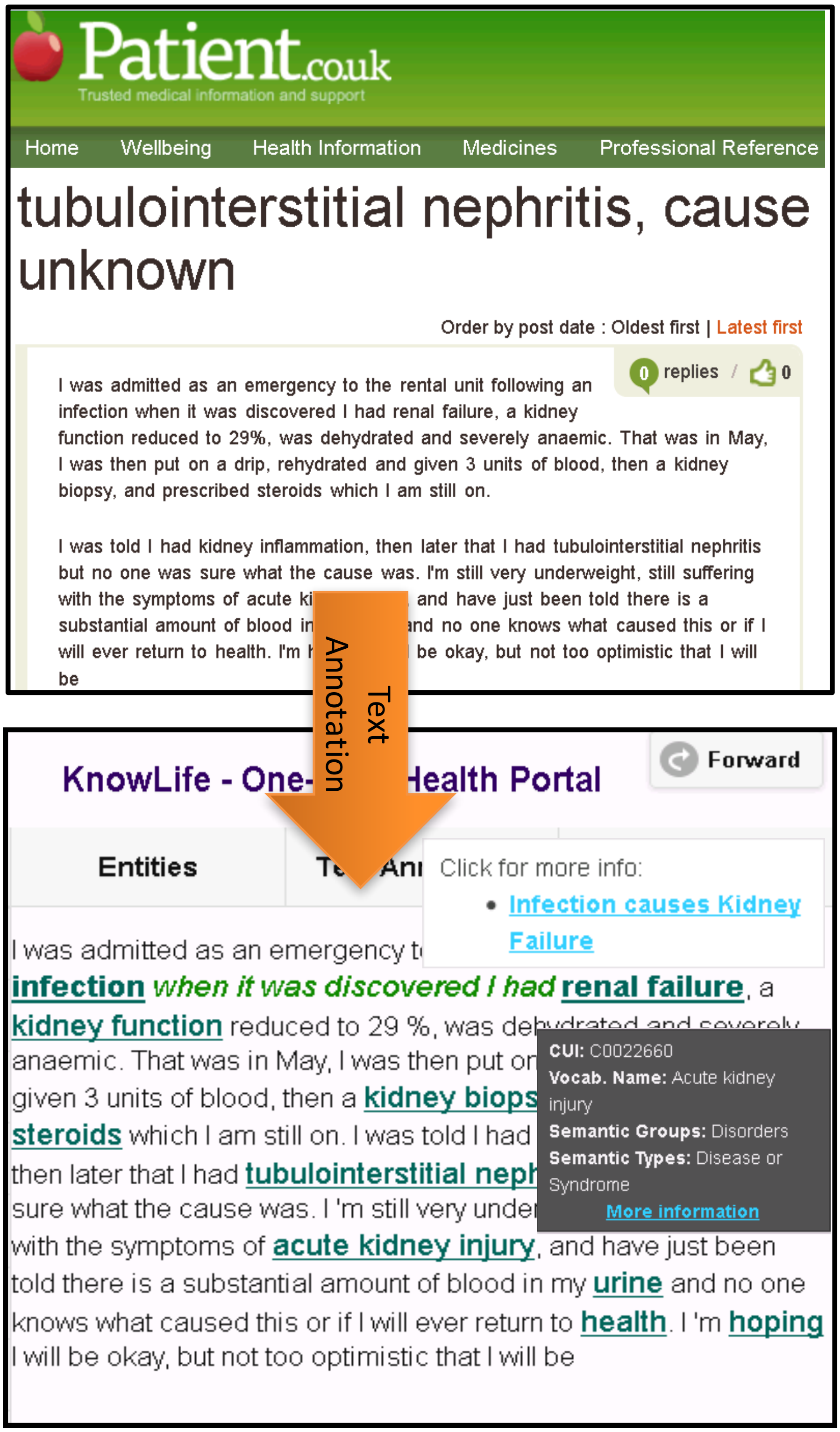
**External Sources:** As an extension to the entities and facts in UMLS and KnowLife, we also include links to external sources to derive more information necessary for a versatile one-stop portal. Here, we tap into additional portals to link to up-to-date discussion threads from social media, such as `Drugtalk.com` or `Patient.co.uk`, and relevant scientific documents, like clinical trials from `Clinicaltrials.gov`, or current scientific publications from the MEDLINE database. Besides more textual information, we also incorporate visual information, such as anatomical 3D visualizations[5], and images to identify drug pills from Pillbox[6].

Similar to Wikipedia and Google Search, we capture all the knowledge in infoboxes to present it to users.

## 6.4 ON-THE-FLY TEXT ANNOTATION

In addition to harvesting knowledge from a large collection of textual sources, our system automatically annotates ad-hoc text on the fly. This allows a user to "speed-read", as the system applies already acquired knowledge to newly seen documents, annotating entities and facts in real time. This functionality is a major extension that distinguishes the portal from other systems. The user can copy-and-paste text or provide a Uniform Resource Locator (URL) for a Web page of interest. In the latter case, we use the CETR tool [194] for

[5] `lifesciencedb.jp/bp3d`
[6] `pillbox.nlm.nih.gov`



tubulointerstitial nephritis, cause unknown

Order by post date : Oldest first | Latest first

I was admitted as an emergency to the rental unit following an infection when it was discovered I had renal failure, a kidney function reduced to 29%, was dehydrated and severely anaemic. That was in May, I was then put on a drip, rehydrated and given 3 units of blood, then a kidney biopsy, and prescribed steroids which I am still on.

I was told I had kidney inflammation, then later that I had tubulointerstitial nephritis but no one was sure what the cause was. I'm still very underweight, still suffering with the symptoms of acute ki and have just been told there is a substantial amount of blood in nd no one knows what caused this or if I will ever return to health. I'm h be okay, but not too optimistic that I will be

Click for more info:

- Infection causes Kidney Failure

I was admitted as an emergency t infection when it was discovered I had renal failure, a kidney function reduced to 29 %, was de anaemic. That was in May, I was then put or given 3 units of blood, then a kidney biops steroids which I am still on. I was told I had then later that I had tubulointerstitial neph sure what the cause was. I 'm still very under with the symptoms of acute kidney injury, and have just been told there is a substantial amount of blood in my urine and no one knows what caused this or if I will ever return to health. I 'm hoping I will be okay, but not too optimistic that I will be

**CUI:** C0022660
**Vocab. Name:** Acute kidney injury
**Semantic Groups:** Disorders
**Semantic Types:** Disease or Syndrome
More information

Figure 15: Text Annotation of an Excerpt from `Patient.co.uk`

removing boilerplate information and casting the HTML input into plain text.

The on-the-fly annotation uses the following steps:

1. **Named Entity Recognition and Disambiguation:** We apply the same dictionary-based method [168] as used by *HighLife* for biomedical entities (see Chapter 5), since it can handle UMLS efficiently and processes entire documents within milliseconds.

2. **Pattern Matching:** Wherever two entities co-occur in a sentence, we rely on salient patterns for relations, learned during knowledge harvesting for *KnowLife* and stored in our KB. We match each pattern against the sentence and collect partial-match results. The results are scored based on their word-level Jaccard overlap with the relation pattern. Patterns with a score above a specified threshold are considered as fact candidates. As an example, consider the excerpt of a discussion from `Patient.co.uk`, shown in Figure 15. The text "when it was discovered I had" matches 7 salient patterns belonging to two different relations: Causes and IsSymptom. Thus, this text becomes a fact candidates for these 2 relations.

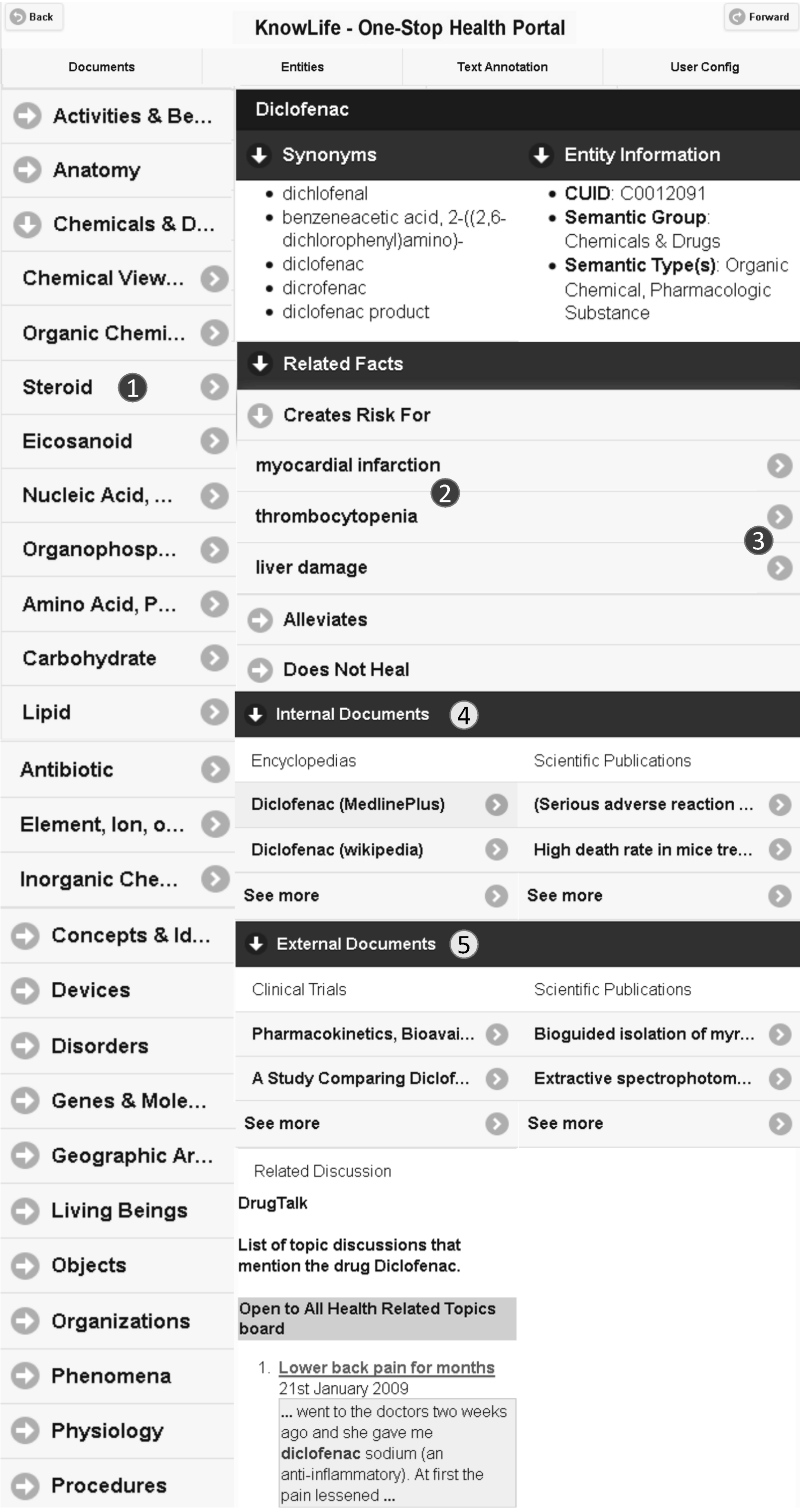
   This pattern matching procedure is efficient for real-time responses, and its decision power is based on previously learned, high-confidence patterns. In contrast to the more elaborate pattern analysis for the knowledge harvesting stage, we avoid expensive computations.

3. **Type Checking:** Using the type signatures of the relations, we can further filter out fact candidates whose arguments have types that are not compatible with the relation. The remaining candidates are accepted as facts and marked in the ad-hoc input text. Returning to our example, the 7 fact candidates are now whittled down to 1 fact: Causes(Infection, Renal Failure). For the other candidate relation, IsSymptom, the type checking prunes this interpretation, as Infection and Renal Failure are of both type Disease, whereas the type signature of IsSymptom would expect a left-hand argument of type Symptom.

## 6.5 DEMO SCENARIOS

The *KnowLife One-stop Health Portal* can be searched and browsed in many ways, supporting both laymen and professionals in knowledge discovery. Figure 16 shows a screenshot of the portal after retrieving an entity, the drug *Diclofenac*. The capability for on-the-fly text annotation has already been shown in Figure 15.

Users are able to interactively query and explore the rich contents of Knowledge Base (KB) presented in Section 6.3, and will also get a



Diclofenac

Synonyms

- dichlofenal
- benzeneacetic acid, 2-((2,6-dichlorophenyl)amino)-
- diclofenac
- dicrofenac
- diclofenac product

Entity Information

- **CUID**: C0012091
- **Semantic Group**: Chemicals & Drugs
- **Semantic Type(s)**: Organic Chemical, Pharmacologic Substance

Related Facts

Creates Risk For

- myocardial infarction
- thrombocytopenia
- liver damage

Alleviates

Does Not Heal

Internal Documents

| Encyclopedias | Scientific Publications |
|---|---|
| Diclofenac (MedlinePlus) | (Serious adverse reaction ... |
| Diclofenac (wikipedia) | High death rate in mice tre... |
| See more | See more |

External Documents

| Clinical Trials | Scientific Publications |
|---|---|
| Pharmacokinetics, Bioavai... | Bioguided isolation of myr... |
| A Study Comparing Diclof... | Extractive spectrophotom... |
| See more | See more |

Related Discussion

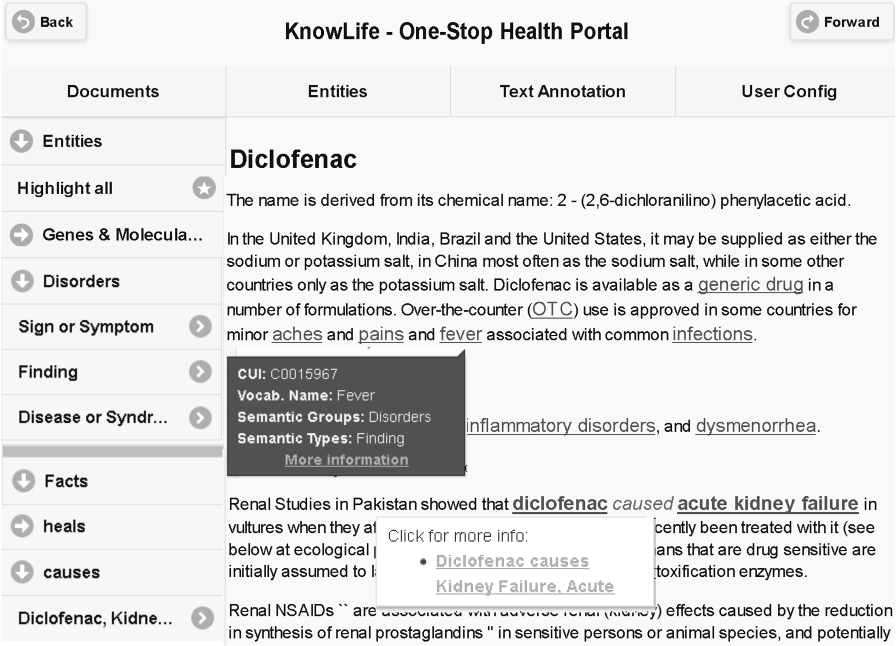
DrugTalk

List of topic discussions that mention the drug Diclofenac.

Open to All Health Related Topics board

1. Lower back pain for months
   21st January 2009
   ... went to the doctors two weeks ago and she gave me **diclofenac** sodium (an anti-inflammatory). At first the pain lessened ...

Figure 16: KnowLife portal for exploring entities: The user can browse over the entity hierarchy (1), jump to related facts (2) or to their textual evidence (3). Relevant internal documents (4) and documents pulled from external sources (5) are listed as well.



Diclofenac

The name is derived from its chemical name: 2 - (2,6-dichloranilino) phenylacetic acid.

In the United Kingdom, India, Brazil and the United States, it may be supplied as either the sodium or potassium salt, in China most often as the sodium salt, while in some other countries only as the potassium salt. Diclofenac is available as a generic drug in a number of formulations. Over-the-counter (OTC) use is approved in some countries for minor aches and pains and fever associated with common infections.

CUI: C0015967
Vocab. Name: Fever
Semantic Groups: Disorders
Semantic Types: Finding
More information

inflammatory disorders, and dysmenorrhea.

Renal Studies in Pakistan showed that diclofenac *caused* acute kidney failure in vultures when they at cently been treated with it (see below at ecological p ans that are drug sensitive are initially assumed to l toxification enzymes.

Click for more info:
- Diclofenac causes Kidney Failure, Acute

Renal NSAIDs `` are ) effects caused by the reduction in synthesis of renal prostaglandins '' in sensitive persons or animal species, and potentially

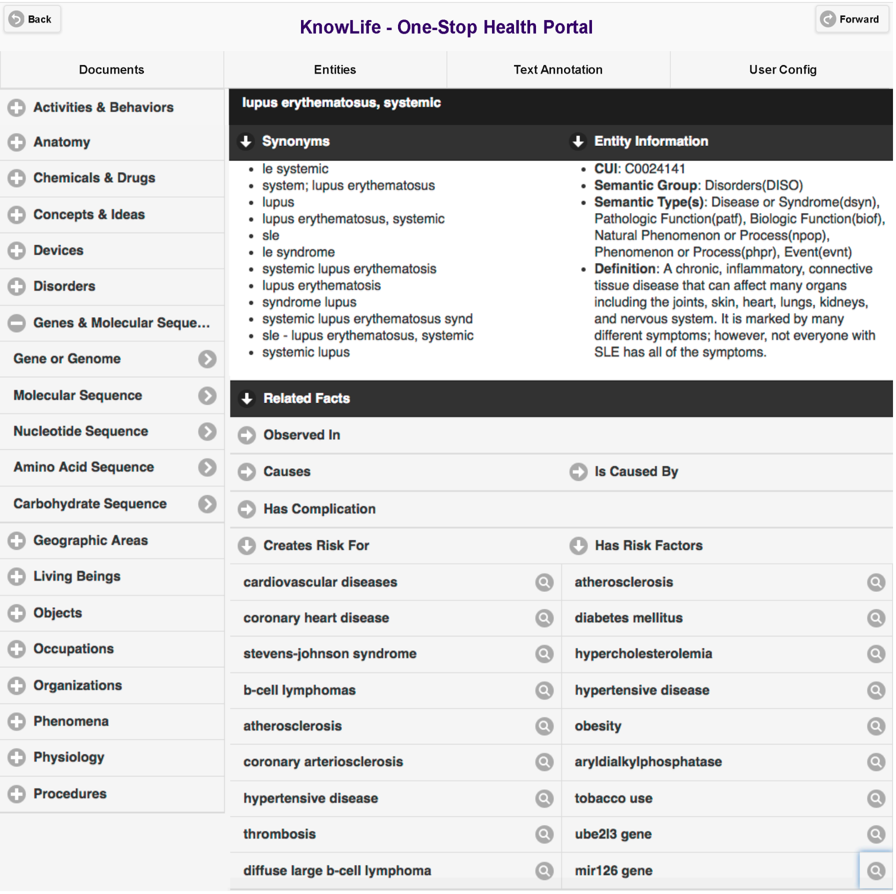
Figure 17: Document about *Diclofenac* Annotated for Speed-reading

deeper understanding of the underlying information extraction and text annotation capabilities.

Below, we discuss two use-case scenarios to illustrate the benefits of the portal for different kinds of end-users.

### *Layman Scenario*

Consider the patient who wrote the post in Figure 15. After she was told that she has either kidney inflammation or tubulointerstitial nephritis, the diagnosis remains unclear. Instead of tediously reading through many health portals, the patient can annotate her own post. This allows her to quickly gain access to relevant information about the recognized entities, as their links springboard her to infoboxes detailing causes, risk factors, symptoms, and more. For instance, she opens the infobox for *Kidney Injury* and notices that *Diclofenac*, a drug she has taken in the past, increases the risk for kidney injury. Therefore, she can explore textual evidence for the fact CreatesRisk(Diclofenac, Kidney Injury). In this case, the system leads her to a Wikipedia article prepared for speed-reading via annotated entities and facts (see Figure 17). Alternatively, she can look at the infobox of *Diclofenac* shown in Figure 16. Here, she can benefit from the experiences of other patients taking this drug by visiting the related discussions section.



lupus erythematosus, systemic

Synonyms

- le systemic
- system; lupus erythematosus
- lupus
- lupus erythematosus, systemic
- sle
- le syndrome
- systemic lupus erythematosis
- lupus erythematosis
- syndrome lupus
- systemic lupus erythematosus synd
- sle - lupus erythematosus, systemic
- systemic lupus

Entity Information

- **CUI**: C0024141
- **Semantic Group**: Disorders(DISO)
- **Semantic Type(s)**: Disease or Syndrome(dsyn), Pathologic Function(patf), Biologic Function(biof), Natural Phenomenon or Process(npop), Phenomenon or Process(phpr), Event(evnt)
- **Definition**: A chronic, inflammatory, connective tissue disease that can affect many organs including the joints, skin, heart, lungs, kidneys, and nervous system. It is marked by many different symptoms; however, not everyone with SLE has all of the symptoms.

| Creates Risk For | Has Risk Factors |
|---|---|
| cardiovascular diseases | atherosclerosis |
| coronary heart disease | diabetes mellitus |
| stevens-johnson syndrome | hypercholesterolemia |
| b-cell lymphomas | hypertensive disease |
| atherosclerosis | obesity |
| coronary arteriosclerosis | aryldialkylphosphatase |
| hypertensive disease | tobacco use |
| thrombosis | ube2l3 gene |
| diffuse large b-cell lymphoma | mir126 gene |

Figure 18: *Lupus* Infobox

*Health Care Professional Scenario*

Physicians and other medical professionals are often more interested in scientific literature, so as to deepen their knowledge in specific areas and stay up-to-date with the latest research. Consider *Lupus*, a challenging disease to diagnose and treat, since there is no standard test and no standard treatment for it. The physician's goal is to quickly obtain relevant scientific information aggregated from multiple articles. To aid diagnosing *Lupus*, the infobox (see Figure 18) lists genes that can be risk factors. By following the links in the infobox, the professional can quickly jump to the latest scientific publications. For treating complex *Lupus* cases, it is also necessary to know about clinical trials, which are directly accessible from the infobox.

## 6.6 CONCLUSION

In this chapter, we presented the first knowledge-based application of this thesis, the *KnowLife One-stop Health Portal*. The web portal is based on a KB, which includes an extensive entity dictionary, informative facts connecting biomedical sub-areas, links to external sources, as well as textual provenance information about relations, facts, and entities. We showed use-case scenarios, showcasing different user needs satisfied by different entrance points to this rich knowledge, e. g. browsing infoboxes, exploring document markup for speed-reading, on-the-fly text annotation, etc. The portal overcomes the limitations of prior work, which is often restricted to a particular biomedical sub-domain or does not offer the same flexibility for exploring and consuming biomedical knowledge and text.

# 7 DEEPLIFE: AN ENTITY-AWARE SEARCH, ANALYTICS AND EXPLORATION PLATFORM

## 7.1 INTRODUCTION

In Chapter 6, we describe the potential of KBs for improving exploratory search systems for navigating and efficiently consuming the ever-growing abundance of biomedical information and health-related contents on the Internet: scientific publications, ontologies and knowledge bases on genes, proteins, drugs, etc., health portals like the one by the Mayo Clinic, online communities where patients and doctors discuss diseases, therapies, drug side effects, etc. However, as with many other systems, the proposed *One-stop Health Portal* has only limited support of finding relevant contents in this wealth of information given a user query, especially when laymen search for specific topics off the mainstream or when experts want high recall on advanced topics from many sources. A typical user approach is to combine keywords with Medical Subject Heading (MeSH) terms when searching PubMed, and to use Google for everything else.

As an example, consider a user who takes asthma medication and plans to go for a 3-month trip to China including rural areas. Which vaccinations are needed, which asthma drugs are not compatible with these vaccines or other drugs that may be needed and purchased during the trip (e. g. diarrhea, sinusitis, influenza)? What is the experience of other travelers? As an example for an expert user's needs, consider a medical student who is investigating the conditions and risk factors under which Zika spreads and causes health problems.

These kinds of users face the following shortcomings of available search engines:

RESTRICTED SEARCH FUNCTIONALITY: The search engines for PubMed or health portals like `Uptodate.com` or `Mayoclinic.org` support only keyword queries with some support for MeSH-like annotations, but lack query functionality that can incorporate hierarchical taxonomies and linkage with knowledge bases. Search over social media sites is even more limited.

LIMITED COVERAGE AND DIVERSITY: Other than Google, all search engines can tap only into one kind of content: either scholarly publications or user-provided social media, but never both. The same holds for intermediate-style contents like health portals.

RESTRICTION TO MOLECULAR ENTITIES: For contents about genes, proteins, pathways, etc., there are structured-data sites that come with richer query and exploration functionality. However, for entities at the level of diseases, therapies, symptoms, risk factors, etc., there are no services of this kind.

LACK OF SUPPORT FOR INTERACTIVE EXPLORATION: Regarding interactive sessions, the only user-friendly support is auto-completion suggestions for user queries. However, these are solely based on the query-and-click history of previous users. This has no awareness of emerging topics in the underlying contents and entity-level background knowledge.

This state of the art for health-related search is in sharp contrast with the state of the art for general-purpose search, say over daily news or general-purpose social media (e. g. , discussing celebrities). Advances in recognizing and disambiguating textual mentions of entities and the linkage to comprehensive knowledge bases like DBpedia, Freebase, Wikidata and YAGO have enabled powerful and user-friendly retrieval systems. Google supports entity-centric search through inter-linkage with the Google Knowledge Graph; Microsoft, Facebook, etc. have similar functionalities. Academic systems such as STICS [71], Broccoli [12] and Semantic Scholar [185] are highly expressive in their capabilities for querying and exploration, with entity-centric auto-completion suggestions and other advanced features. However, none of these covers biomedical or health contents.

This chapter presents a novel system, called *DeepLife*, which provides this kind of user-friendly and expressive support for health-related contents from a wide variety of sources, including scholarly publications, newspaper articles and online communities. Our approach is inspired by the STICS system [71]. However, our content is completely different, and coping with textual mentions of biomedical entities is much harder than recognizing and disambiguating prominent people or places in news articles. In the course of this chapter, we present the system architecture of *DeepLife*, demonstrates its usefulness for various use cases, and discuss how we overcame the aforementioned limitations of prior work and the challenges regarding coverage, scale and usability.

Salient features of *DeepLife* include the following novelties:

- integrating large knowledge bases like UMLS and KnowLife (see Chapter 4) into a search engine over a variety of health-related sources and document feeds,
- providing capabilities for search and exploration based on flexible combinations of keywords (and phrases), biomedical entities, facts, and taxonomic categories,
- supporting users by powerful auto-completion suggestions, interactive query sessions, and basic forms of entity-aware text analytics.

## 7.2 RELATED WORK

In the biomedical domain, the majority of information retrieval systems limit their scopes to PubMed scientific publications. [82] only uses molecular entities for query expansion. The scopes of EVEX [186], Textpresso [116], GoPubMed [48], FACTA+ [184], BioTextQuest+ [130] and CRAB [66] are restricted to genes, proteins, or chemicals. MEDIE [111] and GeneView [182] annotates PubMed articles with various kinds of biomedical entities and events, but both systems do not offer interactive real-time exploration and analytics. Contrary to the systems aforementioned, PolySearch2 [96] goes beyond scientific publications and offers offers 'Given *X*, find all associated *Y*s' queries, where *X* and *Y* are two types, e. g. diseases, toxins, etc, restraining the search together with optionally provided keywords. The goal of the Literome [138] is to facilitate browsing, searching and reasoning over extracted genomic knowledge [137] from PubMed articles by allowing users to search for harvested facts using triple patterns or by traversing the extracted KB step by step. Nonetheless their powerful search capabilities, the last two systems do neither offer an advanced entity-aware search and query interface nor powerful entity analytics.

Besides scientific publications, bio-surveillance systems aggregate and analyze news articles to identify health threats, such as disease outbreaks and food hazards. HealthMap [57] and EpiSpider [79] rely on user created ProMED reports and do not process documents automatically. Contrary to these user-based approaches, Global Health Monitor [49] and the Medical Information System (MedISys) [156] in combination with PULS [174] automatically extract entities and events from relevant medical news. However, the amount of entities both systems can distinguish is limited.

The interactive exploration system Life-iNet [150] system performs NERD using the entire UMLS dictionary and distantly supervised fact harvesting on scientific publications and articles from the Wikipedia health portal. However, the system does not provide any search functionality based on entities, keywords, etc., and is instead targeted for relation-based exploration, where user can query the knowledge with triple patterns, explore the extracted factual knowledge using a network visualization, or digest distinctive entity summarizations for user-provided type queries. The same applies to Literome [138], which also aims to facilitate browsing, and reasoning over extracted genomic knowledge [137] from PubMed articles by allowing users to search for harvested facts using triple patterns or by traversing the extracted KB step by step. Contrary to *DeepLife*, DELVE [68] is not a full-fledged search engine, but a modular faceted browser for exploring Pubmed search result, which provides different kinds of visualization, such as word clouds, phrase nets, and word trees, which can also leverage information from KB. Released a year after *DeepLife*,

| Genre | Sources | Documents | Entity Occurrences | Distinct Entities |
|---|---|---|---|---|
| Clinical Trials | 2 | 16,476 | 49,170 | 8,962 |
| Encyclopedic Articles | 44 | 11,139 | 405,795 | 16,505 |
| News | 121 | 76,534 | 3,058,111 | 38,295 |
| Scientific Publications | 15 | 19,884,225 | 214,531,153 | 453,647 |
| Social Media | 1 | 9,473 | 117,421 | 4,433 |
| **Total** | **182** | **19,997,847** | **218,161,650** | **454,620** |

Table 24: Input Corpus Snapshot on June 1st, 2016

Semedico [54] provides similar functionality, but relies on a smaller entity dictionary compared to our systems.

Pang et al. [128] emphasize the need for better exploratory search capabilities for health content, but they do not consider semantic assets, like entities or a KB.

## 7.3 DEEPLIFE'S KNOWLEDGE BASE

Knowledge Bases (KBs) store facts about entities, their properties, and the relationships between entities. A fact is a triple consisting of two entities $e_1$, $e_2$, which serve as left- and right-hand arguments of a relation $R$, denoted by $R(e_1, e_2)$. We augment and integrate two large knowledge bases to generate *DeepLife*'s KB covering the entire spectrum of biomedical entities, together with an extensive type system featuring salient facts.

**UMLS:** As entity dictionary, we again rely on UMLS. For *DeepLife* the UMLS semantic type system is too shallow, i. e. it only assigns 127 types to more than 3 million entities. Therefore, we generate our own type system by automatically augmenting UMLS with type hierarchies from its source vocabularies. For each vocabulary, we compute its entity coverage in our text corpus depending on the entities' semantic types. The hierarchy of the vocabulary with the highest coverage for a particular semantic type is then used, i. e. for genes the Gene Ontology (GO) is used, for anatomical entities the Foundational Model of Anatomy (FMA) and for drugs and diseases the Medical Subject Heading (MeSH)

**KnowLife:** To integrate the cross-domain facts from KnowLife (see Chapter 4) into our system, we represent them as types. For all facts, $\mathsf{R}(e_1, e_2)$, we create a new type by using the relation $\mathsf{R}$ and the right-hand argument $e_2$ as type name. For example, for all left-hand arguments $e_1$ appearing in facts such as $\mathsf{isRiskFactor}(e_1, \mathsf{Asthma})$ we create the type $\mathsf{RiskFactorsforAsthma}$.

Altogether, *DeepLife*'s knowledge base covers 3.2 million entities with around 12.8 million entity names and synonyms, 64,568 custom types from source vocabularies and 136,437 fact types.

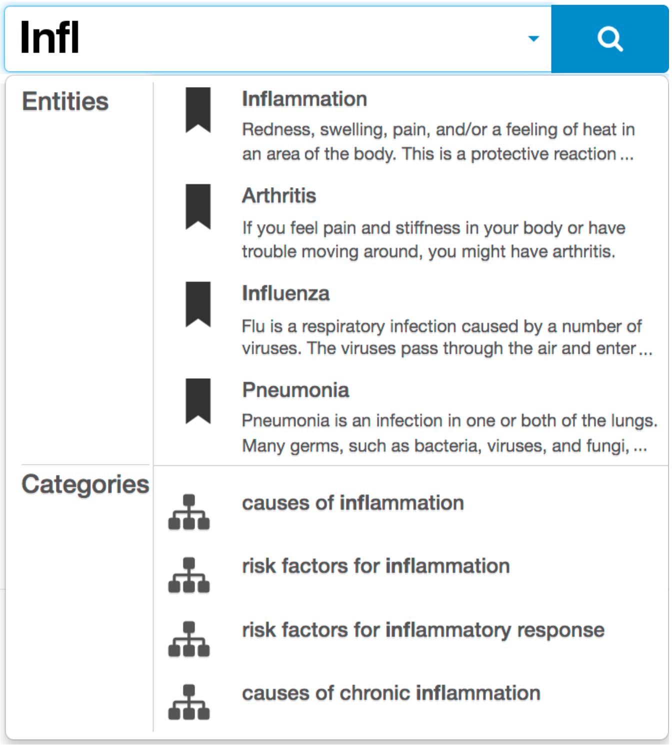
## 7.4 ENTITY RECOGNITION AND DISAMBIGUATION

*DeepLife* has indexed 19,997,847 documents and extracted 218,161,650 entities from a continuous stream of 182 RSS feeds spanning five text genres. As Table 24 shows, this constantly growing and diverse corpus covers the full spectrum of biomedical information on the web. Clinical trials and scientific publications describe research findings and target professionals. For this purpose, *DeepLife* includes the entire MEDLINE collection and all clinical trials from `ClinicalTrials.gov`. Encyclopaedic articles are educational resources providing insights to laymen. Social media, such as patient discussion forums, are mainly used to share experiences and to receive advice. By including news articles, our system is always up-to-date on the latest health topics, such as disease outbreaks or lifestyle information.

### *Entity Recognition*

To process incoming articles in real-time and to stay up-to-date, our system applies an agile entity recognition method. As introduced in Section 2.2.1, Stanford's CoreNLP toolkit is used to split sentences, tokenize words and determine part-of-speech tags. A word phrase chunker from OpenNLP[1] is used to generate an initial set of noun chunk candidates. We extend this set by applying a rule-based approach, e. g. splitting or merging prepositional phrases, conjunctions, as well as proper and common nouns. Candidates are then matched against the entity names in UMLS using string-similarity, giving preference to the longest possible matching chunk. To efficiently handle the large dictionary and volume of candidates, we use our own method which is based on Locality Sensitive Hashing (LSH) with min-wise independent permutations (MinHash) to quickly find matching candidates [168]. LSH probabilistically reduces the high-dimensional space of all character-level 3-grams, while MinHash quickly estimates the similarity between two sets of 3-grams. A successful match provides us also with the entity's semantic type.

[1] `opennlp.apache.org`



Infl

Entities

Inflammation
Redness, swelling, pain, and/or a feeling of heat in an area of the body. This is a protective reaction ...

Arthritis
If you feel pain and stiffness in your body or have trouble moving around, you might have arthritis.

Influenza
Flu is a respiratory infection caused by a number of viruses. The viruses pass through the air and enter ...

Pneumonia
Pneumonia is an infection in one or both of the lungs. Many germs, such as bacteria, viruses, and fungi, ...

Categories

causes of inflammation

risk factors for inflammation

risk factors for inflammatory response

causes of chronic inflammation

Figure 19: Entity/Category Auto-completion

*Entity Disambiguation*

The entity type information is used to disambiguate between multiple entity candidates matched to the same noun chunk in the input text. In the first filtering step, we reduce the number of entity candidates by only retaining the most specifically typed entities according to the UMLS semantic type system. Taking into account that UMLS provides a ranked list of entities for every possible name, we further disambiguate between the remaining candidates by determining the highest ranked entities. In case two entities share the same rank, we determine their popularities by the number of occurrences in different UMLS source vocabularies and prefer the more popular entity. As shown in Table 24, our system has currently extracted 218,161,650 mentions of 454,620 distinct entities.

## 7.5 ENTITY-AWARE AUTO-COMPLETION

Formulating queries with *DeepLife* is user-friendly and responsive. Providing an entity auto-completion which combines prefix matching with entity popularity, the system lets users easily explore and navigate through an extensive amount of entities and categories. For a user-provided prefix, the system retrieves entity and category candidates, where any token of their name or synonyms matches the prefix. These candidates are then ranked by corpus statistics which the system constantly updates. For example, Figure 19 depicts *Arthritis* as the second suggestion, because its synonym *Joint Inflammation* matches the prefix, and because of its high prevalence in the corpus.

## 7.6 entity and category search

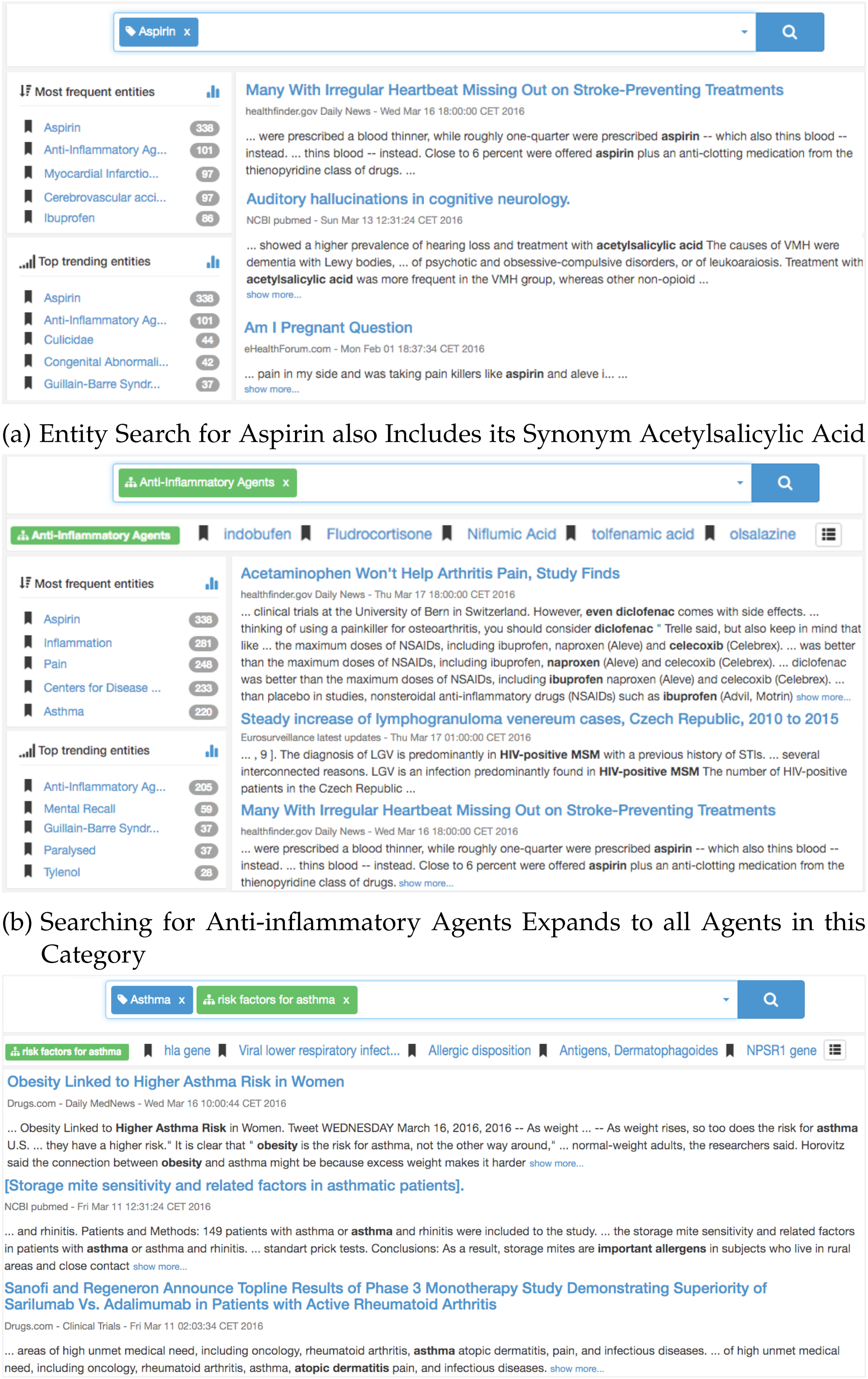
The entity-based search of our system excels over traditional keyword-based search. It increases recall, since the system automatically includes all synonyms of an entity, as well as precision, since the disambiguation removes unwanted occurrences. For example, as depicted in Figure 20a, if users search for *Aspirin*, documents mentioning its synonym *Acetylsalicylic Acid* are also retrieved. An important feature of our system is the possibility to search for categories of entities. This allows users to broaden their search request to all entities of the same type, i. e. entities which share common attributes or features. For example, to search for all "aspirin like drugs" which share therapeutic properties, one can search for the category *Anti-inflammatory Agents* (see Figure 20b). The system automatically determines all entities within the category via *DeepLife*'s type system to retrieve relevant documents. Figure 20b also highlights *DeepLife*'s diverse set of sources. The search results cover news, publications, as well as discussions. To tap into specific sources, users can easily customize queries with search filters.

## 7.7 cross-domain combined search

*DeepLife*'s knowledge base empowers our system to provide an intuitive method for searching facts by combining category and entity search. This is especially useful for layman users. Consider a user who is suffering from asthma and is interested in finding all risk factors triggering the disease. In this case, using the category *Risk Factors for Asthma* generated from facts together with the entity *Asthma* as search query, the system retrieves all documents mentioning *Asthma* with its risk factors (see Figure 20c). Displaying the individual risk factors (e. g. *HLA Gene, Viral Lower Respiratory Infection, etc.*) as an expansion of the category provides immediate insights and facilitates further exploration.

## 7.8 analytics

Our system offers interactive entity-based analytics to spot trends and topic shifts. Such analyses benefit from the improved recall and precision aspects aforementioned. Statistics, based on entity occurrences in documents over time, are computed and visualized. For example in Figure 21, entity occurrences of *Zika* and related countries in our corpus (Y-Axis) are visualized over time (X-Axis). Users can zoom



| Most frequent entities | |
|---|---|
| Aspirin | 338 |
| Anti-Inflammatory Ag... | 101 |
| Myocardial Infarctio... | 97 |
| Cerebrovascular acci... | 97 |
| Ibuprofen | 86 |

| Top trending entities | |
|---|---|
| Aspirin | 338 |
| Anti-Inflammatory Ag... | 101 |
| Culicidae | 44 |
| Congenital Abnormali... | 42 |
| Guillain-Barre Syndr... | 37 |

Many With Irregular Heartbeat Missing Out on Stroke-Preventing Treatments

healthfinder.gov Daily News - Wed Mar 16 18:00:00 CET 2016

... were prescribed a blood thinner, while roughly one-quarter were prescribed **aspirin** -- which also thins blood -- instead. ... thins blood -- instead. Close to 6 percent were offered **aspirin** plus an anti-clotting medication from the thienopyridine class of drugs. ...

Auditory hallucinations in cognitive neurology.

NCBI pubmed - Sun Mar 13 12:31:24 CET 2016

... showed a higher prevalence of hearing loss and treatment with **acetylsalicylic acid** The causes of VMH were dementia with Lewy bodies, ... of psychotic and obsessive-compulsive disorders, or of leukoaraiosis. Treatment with **acetylsalicylic acid** was more frequent in the VMH group, whereas other non-opioid ...
show more...

Am I Pregnant Question

eHealthForum.com - Mon Feb 01 18:37:34 CET 2016

... pain in my side and was taking pain killers like **aspirin** and aleve i... ...
show more...

(a) Entity Search for Aspirin also Includes its Synonym Acetylsalicylic Acid

| Most frequent entities | |
|---|---|
| Aspirin | 338 |
| Inflammation | 281 |
| Pain | 248 |
| Centers for Disease ... | 233 |
| Asthma | 220 |

| Top trending entities | |
|---|---|
| Anti-Inflammatory Ag... | 205 |
| Mental Recall | 59 |
| Guillain-Barre Syndr... | 37 |
| Paralysed | 37 |
| Tylenol | 28 |

Acetaminophen Won't Help Arthritis Pain, Study Finds

healthfinder.gov Daily News - Thu Mar 17 18:00:00 CET 2016

... clinical trials at the University of Bern in Switzerland. However, **even diclofenac** comes with side effects. ... thinking of using a painkiller for osteoarthritis, you should consider **diclofenac** " Trelle said, but also keep in mind that like ... the maximum doses of NSAIDs, including ibuprofen, naproxen (Aleve) and **celecoxib** (Celebrex). ... was better than the maximum doses of NSAIDs, including ibuprofen, **naproxen** (Aleve) and celecoxib (Celebrex). ... diclofenac was better than the maximum doses of NSAIDs, including **ibuprofen** naproxen (Aleve) and celecoxib (Celebrex). ... than placebo in studies, nonsteroidal anti-inflammatory drugs (NSAIDs) such as **ibuprofen** (Advil, Motrin) show more...

Steady increase of lymphogranuloma venereum cases, Czech Republic, 2010 to 2015

Eurosurveillance latest updates - Thu Mar 17 01:00:00 CET 2016

... , 9 ]. The diagnosis of LGV is predominantly in **HIV-positive MSM** with a previous history of STIs. ... several interconnected reasons. LGV is an infection predominantly found in **HIV-positive MSM** The number of HIV-positive patients in the Czech Republic ...

Many With Irregular Heartbeat Missing Out on Stroke-Preventing Treatments

healthfinder.gov Daily News - Wed Mar 16 18:00:00 CET 2016

... were prescribed a blood thinner, while roughly one-quarter were prescribed **aspirin** -- which also thins blood -- instead. ... thins blood -- instead. Close to 6 percent were offered **aspirin** plus an anti-clotting medication from the thienopyridine class of drugs. show more...

(b) Searching for Anti-inflammatory Agents Expands to all Agents in this Category

Asthma x
risk factors for asthma x
risk factors for asthma
hla gene
Viral lower respiratory infect...
Allergic disposition
Antigens, Dermatophagoides
NPSR1 gene

Obesity Linked to Higher Asthma Risk in Women

Drugs.com - Daily MedNews - Wed Mar 16 10:00:44 CET 2016

... Obesity Linked to **Higher Asthma Risk** in Women. Tweet WEDNESDAY March 16, 2016, 2016 -- As weight ... -- As weight rises, so too does the risk for **asthma** U.S. ... they have a higher risk." It is clear that " **obesity** is the risk for asthma, not the other way around," ... normal-weight adults, the researchers said. Horovitz said the connection between **obesity** and asthma might be because excess weight makes it harder show more...

[Storage mite sensitivity and related factors in asthmatic patients].

NCBI pubmed - Fri Mar 11 12:31:24 CET 2016

... and rhinitis. Patients and Methods: 149 patients with asthma or **asthma** and rhinitis were included to the study. ... the storage mite sensitivity and related factors in patients with **asthma** or asthma and rhinitis. ... standart prick tests. Conclusions: As a result, storage mites are **important allergens** in subjects who live in rural areas and close contact show more...

Sanofi and Regeneron Announce Topline Results of Phase 3 Monotherapy Study Demonstrating Superiority of Sarilumab Vs. Adalimumab in Patients with Active Rheumatoid Arthritis

Drugs.com - Clinical Trials - Fri Mar 11 02:03:34 CET 2016

... areas of high unmet medical need, including oncology, rheumatoid arthritis, **asthma** atopic dermatitis, pain, and infectious diseases. ... of high unmet medical need, including oncology, rheumatoid arthritis, asthma, **atopic dermatitis** pain, and infectious diseases. show more...

(c) Combined Category and Entity Search for Documents containing Asthma and its Risk Factors

Figure 20: Entity and Category Search

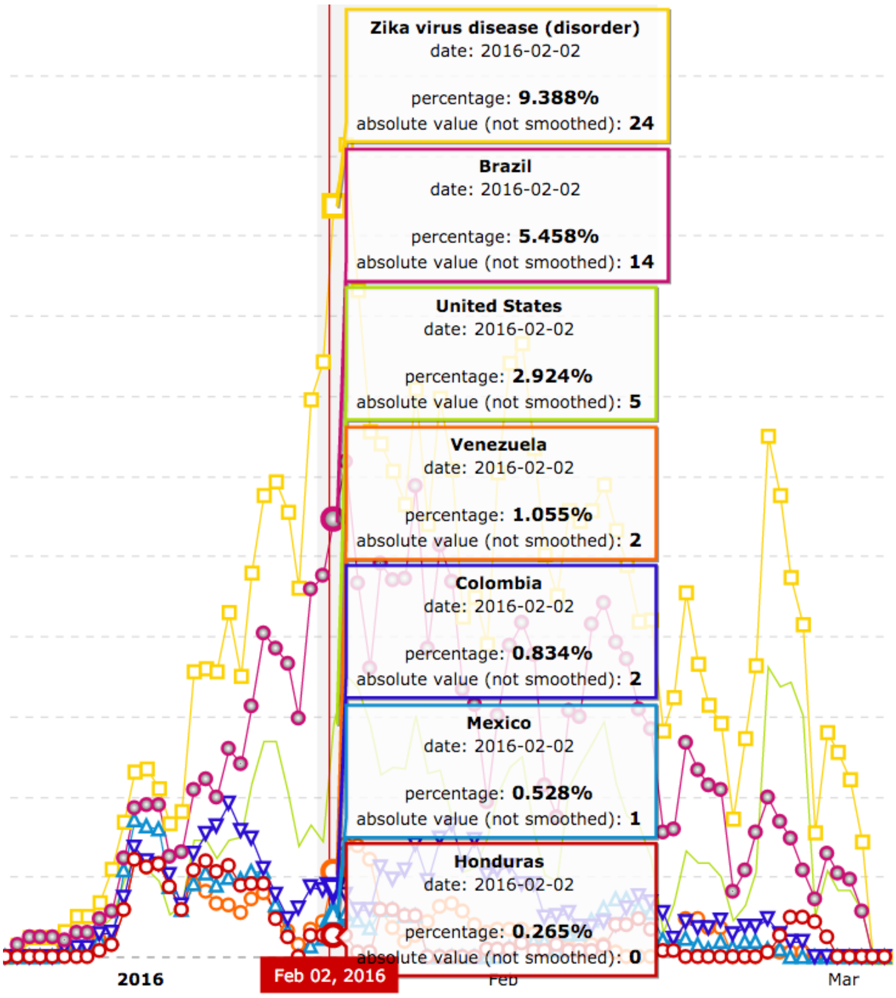


Figure 21: Countries co-occurring with Zika

into specific time frames and explore documents the statistics are based on. Not only can entities be tracked individually, the analytics can also be constrained on one entity of main interest, i. e. only those documents in which this entity appears. In the same example in Figure 21, to gather insights about countries affected by the virus, the user set *Zika virus disease* as the main entity to compute analytics based on documents where *Zika* and a particular country were mentioned.

## 7.9 CONCLUSION

We presented the entity-aware search system *DeepLife* in this chapter, which integrates large KBs and harnesses entity linking methods to overcome limitations of existing search engines. Starting with expedited query formulation using entity-aware auto-completion, *DeepLife* lets users formulate powerful search queries in terms of keywords and phrases, biomedical entities, and taxonomic categories, to improve search over all kinds of biomedical and health texts. It also provides functionality for entity-aware text analytics over health-centric contents. To showcase salient features we described different use-case scenarios benefiting from *DeepLife* capabilities.

# Part IV

# CONCLUSION

# 8 SUMMARY & FUTURE WORK

Fact harvesting is a key component for the construction of extensive Knowledge Bases (KBs) out of large text corpora. In this thesis, we presented two methods which focussed on constructing a versatile and expressive KB for the biomedical domain.

The first method, called *KnowLife*, is a versatile, largely automated, and scalable approach to construct a comprehensive KB for the life sciences. The approach follows the fact-pattern duality paradigm and uses statistics-based pattern matching for high recall with logics-based consistency reasoning for high precision of accepted facts. Experiments showed, that the best configured KB contains more than 500,000 facts at a precision of 93% for 13 relations covering genes, organs, diseases, symptoms, treatments, as well as environmental and lifestyle risk factors.

The second method of this thesis, coined *HighLife*, generalized the fact-pattern duality paradigm of *KnowLife* and other previous methods to higher-arity cases. Besides harvesting higher-arity facts, a key novelty is the integration of facts with missing arguments, by first extending pattern-learning to higher-arity partial patterns and facts, and secondly by applying reasoning not only over the consistency, but also over the composability of partial fact candidates into full facts.

Furthermore, we described two use-case systems leveraging KBs for exploring, analyzing and searching over large text collections on health.

The first system, named *KnowLife One-stop Health Portal*, demonstrated the benefits of versatile KBs for entity-relationship-oriented search and exploration over large amounts of content. We presented different ways how different kinds of textual content can be linked to a KB for sophisticated exploration.

The second system was *DeepLife*, an entity-aware search and analytics platforms, which harnesses entity linking methods to annotate articles from a KB in real-time. Based on this knowledge, *DeepLife* provides powerful capabilities for search, and analytics based on flexible combinations of keywords (and phrases), entities, relational facts, and taxonomic categories.

## OUTLOOK

Besides the contributions of this thesis, there are still further aspects worth exploring in the future:

FACT RANKING: Currently, our systems lack the capability to sophisticatedly rank facts. For example, when users query for symptoms of a particular disease, a layman usually seeks to retrieve the most common ones first, while a professional with extensive prior knowledge would be rather interested in characteristics of asymptomatic cases. To provide such context-dependent rankings, we need to quantify the prevalence of facts in our KB, e. g. quantify rare vs. common symptoms of disease, and personalize search results based on on a user's search history. The amount of mentions of a fact within a text corpus can be a signal for a prevalence measure, but it will not be sufficient for a sophisticated solution. For example, the mention frequency of common flu symptoms will be rather low in scientific publications, since they are probably well-known and thus not interesting for scientific studies. Linguistic cues are another signal to infer prevalence, e. g. phrases like "main symptoms of" vs "rare symptom of". An integrated approach combining mention frequencies and linguistic cues conditioned on the source of a fact could be worthwhile to investigate as possible solution.

DATABASE EVIDENCE: One type of information we did not consider in this thesis are case studies or patient records. There is limited work, trying to build KBs from such record-based datasets. Most notable Rotmensch et al. [159] tap into electronic health records for harvesting knowledge. An integrated approach leveraging statistical weights from record analyses fused with linguistic evidence extracted from text corpora would be an interesting future direction to investigate.

# Part V

# APPENDIX

a

# APPENDIX

## A.1 KNOWLIFE CONSISTENCY RULES

Pattern-fact Duality

$\mathrm{Express}(P,R) \wedge \mathrm{Occurs}(P,X,Y)$

$\wedge \mathrm{Type}(X,\mathrm{DOM}) \wedge \mathrm{Type}(Y,\mathrm{RAN})$

$\wedge \mathrm{Domain}(R,\mathrm{DOM}) \wedge \mathrm{Range}(R,\mathrm{RAN})$

$\Rightarrow R(X,Y)$

Mutual Exclusion

$\mathrm{Aggravates}(X,Y) \Rightarrow \neg \mathrm{Treats}(X,Y)$

$\mathrm{Aggravates}(X,Y) \Rightarrow \neg \mathrm{Alleviates}(X,Y)$

$\mathrm{Aggravates}(X,Y) \Rightarrow \neg \mathrm{ReducesRisk}(X,Y)$

$\mathrm{Causes}(X,Y) \Rightarrow \neg \mathrm{Treats}(X,Y)$

$\mathrm{Causes}(X,Y) \Rightarrow \neg \mathrm{Alleviates}(X,Y)$

$\mathrm{Causes}(X,Y) \Rightarrow \neg \mathrm{ReducesRisk}(X,Y)$

$\mathrm{CreatesRisk}(X,Y) \Rightarrow \neg \mathrm{Treats}(X,Y)$

$\mathrm{CreatesRisk}(X,Y) \Rightarrow \neg \mathrm{Alleviates}(X,Y)$

$\mathrm{CreatesRisk}(X,Y) \Rightarrow \neg \mathrm{ReducesRisk}(X,Y)$

$\mathrm{CreatesRisk}(X,Y) \Rightarrow \neg \mathrm{IsSymptom}(X,Y)$

$\mathrm{SideEffect}(X,Y) \Rightarrow \neg \mathrm{Diagnoses}(Y,X)$

$\mathrm{SideEffect}(X,Y) \Rightarrow \neg \mathrm{Treats}(Y,X)$

$\mathrm{SideEffect}(X,Y) \Rightarrow \neg \mathrm{Alleviates}(Y,X)$

$\mathrm{SideEffect}(X,Y) \Rightarrow \neg \mathrm{ReducesRisk}(Y,X)$

$\mathrm{Contraindicates}(X,Y) \Rightarrow \neg \mathrm{Treats}(X,Y)$

$\mathrm{Contraindicates}(X,Y) \Rightarrow \neg \mathrm{Alleviates}(X,Y)$

$\mathrm{Contraindicates}(X,Y) \Rightarrow \neg \mathrm{ReducesRisk}(X,Y)$

## A.2 HIGHLIFE CONSISTENCY RULES

N-ary Pattern-fact Duality

$\mathrm{TreePattern}(P,R) \wedge \mathrm{Occurs}(P,X_1,X_2) \wedge \mathrm{Sig}(R,T_1,T_2) \wedge$

$\mathrm{Type}(X_1,T_1) \wedge \mathrm{Type}(X_2,T_2) \Rightarrow R(X_1,X_2)$

$\mathrm{TreePattern}(P,R) \wedge \mathrm{Occurs}(P,X_1,X_2,X_3) \wedge \mathrm{Sig}(R,T_1,T_2,T_3) \wedge$

$\mathrm{Type}(X_1,T_1) \wedge \mathrm{Type}(X_2,T_2) \wedge \mathrm{Type}(X_3,T_3) \Rightarrow R(X_1,X_2,X_3)$

$\mathrm{TreePattern}(P,R) \wedge \mathrm{Occurs}(P,X_1,X_2,X_3,X_4) \wedge \mathrm{Sig}(R,T_1,T_2,T_3,T_4) \wedge$

$\mathrm{Type}(X_1,T_1) \wedge \mathrm{Type}(X_2,T_2) \wedge \mathrm{Type}(X_3,T_3) \wedge \mathrm{Type}(X_4,T_4)$

$\Rightarrow R(X_1,X_2,X_3,X_4)$

Corresponding rules for arities higher than 4 are left out to ensure greater clarity.

### A.2.1 *Biomedical Domain Consistency Constrains*

Dosage Equality
$\mathit{Treats}(X_1,X_2,X_3,X_4,X_5) \wedge \mathit{Treats}(X_1,X_2,Y_3,Y_4,X_5) \Rightarrow \mathit{equals}(X_3,Y_3)$

Mutual Exclusion
$\mathit{Causes}(X_1,X_2,X_3,X_4) \Rightarrow \neg \mathit{Treats}(X_1,X_2,X_3,X_4,X_5)$
$\mathit{Causes}(X_1,X_2,X_3,X_4) \Rightarrow \neg \mathit{ReducesRisk}(X_1,X_2,X_3,X_4)$

### A.2.2 *News Domain Consistency Constrains*

Price Equality
$\mathit{CompanyAcquired}(X_1,X_2,X_3,X_4,X_5) \wedge$
$\mathit{CompanyAcquired}(X_1,X_2,Y_3,Y_4,Y_5)$
$\Rightarrow \mathit{equals}(X_5,Y_5)$

Buyer Equality
$\mathit{CompanyAcquired}(X_1,X_2,X_3,X_4,X_5) \wedge$
$\mathit{CompanyAcquired}(Y_1,X_2,X_3,Y_4,Y_5)$
$\Rightarrow \mathit{equals}(X_1,Y_1)$

Time Equality
$\mathit{CompanyAcquired}(X_1,X_2,X_3,X_4,X_5) \wedge$
$\mathit{CompanyAcquired}(X_1,X_2,Y_3,Y_4,Y_5)$
$\Rightarrow \mathit{equals}(X_3,Y_3)$

Mutual Exclusion (Argument Swapping)
$\mathit{CompanyAcquired}(X_1,X_2,X_3,X_4,X_5)$
$\Rightarrow \neg \mathit{CompanyAcquired}(X_2,X_1 Y_3,Y_4,Y_5)$

$\mathit{CompanyAcquired}(X_1,X_2,X_3,X_4,X_5)$
$\Rightarrow \neg \mathit{CompanyAcquired}(X_2,X_3 Y_3,Y_4,Y_5)$

Location Part Of Equivalence
$\mathit{AthleteWonAward}(X_1,X_2,X_3,X_4,X_5,X_6) \wedge$
$\mathit{AthleteWonAward}(X_1,X_2,X_3,Y_4,Y_5,Y_6)$
$\Rightarrow \mathit{partOf}(X_4,Y_4)$

Time Equality
$\mathit{AthleteWonAward}(X_1,X_2,X_3,X_4,X_5,X_6) \wedge$
$\mathit{AthleteWonAward}(X_1,X_2,X_3,X4,Y_5,Y_6)$
$\Rightarrow \mathit{equals}(X_5,Y_5)$

Event Equality
$\mathit{AthleteWonAward}(X_1,X_2,X_3,X_4,X_5,X_6) \wedge$
$\mathit{AthleteWonAward}(X_1,X_2,Y_3,X4,X_5,Y_6)$
$\Rightarrow \mathit{equals}(X_3,Y_3)$

`Type Of Sports Equality`
$AthleteWonAward(X_1, X_2, X_3, X_4, X_5, X_6) \wedge$
$AthleteWonAward(X_1, Y_2, Y_3, Y_4, Y_5, Y_6)$
$\Rightarrow \mathrm{equals}(X_6, Y_6)$

# BIBLIOGRAPHY

[1] Eugene Agichtein and Luis Gravano. "Snowball: Extracting Relations from Large Plain-text Collections." In: *Proceedings of the ACM Conference on Digital Libraries*. DL '00. San Antonio, Texas, USA: ACM, 2000, pp. 85–94.

[2] Antti Airola, Sampo Pyysalo, Jari Björne, Tapio Pahikkala, Filip Ginter, and Tapio Salakoski. "All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning." In: *BMC Bioinformatics* 9.11 (2008), S2.

[3] Dean Allemang and James Hendler. *Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL*. 2nd ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011.

[4] Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. "Leveraging Linguistic Structure For Open Domain Information Extraction." In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. ACL '15. Beijing, China: Association for Computational Linguistics, 2015, pp. 344–354.

[5] Cecilia N. Arighi et al. "BioCreative III interactive task: an overview." In: *BMC Bioinformatics* 12.8 (2011), S4.

[6] Alan R Aronson and François-Michel Lang. "An overview of MetaMap: historical perspective and recent advances." In: *Journal of the American Medical Informatics Association* 17.3 (2010), p. 229.

[7] Lora Aroyo and Chris Welty. "Measuring Crowd Truth for Medical Relation Extraction." In: *AAAI Fall Symposium Series*. Menlo Park, CA, USA: AAAI Press, 2013.

[8] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. "Gene Ontology: tool for the unification of biology." In: *Nature Genetics* 25.1 (2000), pp. 25–29.

[9] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. "The Berkeley FrameNet Project." In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. ACL '98. Montreal, Quebec, Canada: Association for Computational Linguistics, 1998, pp. 86–90.

[10] Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. "Open Information Extraction from the Web." In: *Proceedings of the International Joint Conference on Artifical Intelligence*. IJCAI '07. Hyderabad, India: Morgan Kaufmann Publishers Inc., 2007, pp. 2670–2676.

[11] Hannah Bast, Buchhold Björn, and Elmar Haussmann. "Semantic Search on Text and Knowledge Bases." In: *Foundations and Trends in Information Retrieval* 10.2-3 (2016), pp. 119–271.

[12] Hannah Bast and Björn Buchhold. "An Index for Efficient Semantic Full-text Search." In: *Proceedings of the ACM International Conference on Information & Knowledge Management*. CIKM '13. San Francisco, California, USA: ACM, 2013, pp. 369–378.

[13] Soumia Lilia Berrahou, Patrice Buche, Juliette Dibie, and Mathieu Roche. "Xart System: Discovering and Extracting Correlated Arguments of N-ary Relations from Text." In: *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*. WIMS '16. N&#238;mes, France: ACM, 2016, 8:1–8:12.

[14] Jari Björne and Tapio Salakoski. "Generalizing Biomedical Event Extraction." In: *Proceedings of the Workshop on Biomedical Natural Language Processing*. BioNLP '11. Portland, Oregon: Association for Computational Linguistics, 2011, pp. 183–191.

[15] Jari Björne and Tapio Salakoski. "TEES 2.2: Biomedical Event Extraction for Diverse Corpora." In: *BMC Bioinformatics* 16.16 (2015), S4.

[16] Jari Björne, Juho Heimonen, Filip Ginter, Antti Airola, Tapio Pahikkala, and Tapio Salakoski. "Extracting Complex Biological Events with Rich Graph-based Feature Sets." In: *Proceedings of the Workshop on Biomedical Natural Language Processing*. BioNLP '09. Boulder, Colorado: Association for Computational Linguistics, 2009, pp. 10–18.

[17] Christian Blaschke, Miguel A Andrade, Christos A Ouzounis, and Alfonso Valencia. "Automatic extraction of biological information from scientific text: protein-protein interactions." In: *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*. ISMB '99. Menlo Park, CA, USA: AAAI Press, 1999, pp. 60–67.

[18] Olivier Bodenreider. "The Unified Medical Language System (UMLS): integrating biomedical terminology." In: *Nucleic Acids Research* 32.suppl_1 (2004), pp. D267–D270.

[19] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. "Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge." In: *Proceedings of the ACM SIGMOD International Conference on Management of*

*Data*. SIGMOD '08. Vancouver, Canada: ACM, 2008, pp. 1247–1250.

[20] R. Brachman and H. Levesque. *Knowledge Representation and Reasoning*. The Morgan Kaufmann Series in Artificial Intelligence. Elsevier Science, 2004.

[21] Àlex Bravo, Montserrat Cases, Núria Queralt-Rosinach, C. Ferran Sanz, and Laura Inés Furlong. "A knowledge-driven approach to extract disease-related biomarkers from the literature." In: *Journal of BioMed Research International* 2014 (2014).

[22] Àlex Bravo, Janet Piñero, Núria Queralt-Rosinach, Michael Rautschka, and Laura I. Furlong. "Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research." In: *BMC Bioinformatics* 16.1 (2015), p. 55.

[23] Sergey Brin. "Extracting Patterns and Relations from the World Wide Web." In: *International Workshop on The World Wide Web and Databases*. WebDB '98. London, UK, UK: Springer-Verlag, 1999, pp. 172–183.

[24] Andrei Z Broder, Moses Charikar, Alan M Frieze, and Michael Mitzenmacher. "Min-Wise Independent Permutations." In: *Journal of Computer and System Sciences* 60.3 (2000), pp. 630–659.

[25] Markus Bundschus, Mathaeus Dejori, Martin Stetter, Volker Tresp, and Hans-Peter Kriegel. "Extraction of semantic biomedical relations from text using conditional random fields." In: *BMC Bioinformatics* 9.1 (2008), p. 207.

[26] Razvan C. Bunescu and Raymond J. Mooney. "A Shortest Path Dependency Kernel for Relation Extraction." In: *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. HLT '05. Vancouver, British Columbia, Canada: Association for Computational Linguistics, 2005, pp. 724–731.

[27] Razvan Bunescu, Ruifang Ge, Rohit J. Kate, Edward M. Marcotte, Raymond J. Mooney, Arun K. Ramani, and Yuk Wah Wong. "Comparative experiments on learning information extractors for proteins and their interactions." In: *Artificial Intelligence in Medicine* 33.2 (2005). Information Extraction and Summarization from Medical Documents, pp. 139–155.

[28] John D Burger, Emily Doughty, Ritu Khare, Chih-Hsuan Wei, Rajashree Mishra, John Aberdeen, David Tresner-Kirsch, Ben Wellner, Maricel G Kann, Zhiyong Lu, and Lynette Hirschman. "Hybrid curation of gene–mutation relations combining automated extraction and crowdsourcing." In: *Database: The Journal of Biological Databases and Curation* 2014 (2014), 1—13.

[29] Ekaterina Buyko, Erik Faessler, Joachim Wermter, and Udo Hahn. "Event Extraction from Trimmed Dependency Graphs." In: *Proceedings of the Workshop on Biomedical Natural Language Processing*. BioNLP '09. Boulder, Colorado: Association for Computational Linguistics, 2009, pp. 19–27.

[30] Rui Cai, Xiaodong Zhang, and Houfeng Wang. "Bidirectional Recurrent Convolutional Neural Network for Relation Classification." In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. ACL '16. Berlin, Germany: Association for Computational Linguistics, 2016, pp. 756–765.

[31] Andrew Carlson, Justin Betteridge, Richard C. Wang, Estevam R. Hruschka Jr., and Tom M. Mitchell. "Coupled Semi-supervised Learning for Information Extraction." In: *Proceedings of the ACM International Conference on Web Search and Data Mining*. WSDM '10. New York, New York, USA: ACM, 2010, pp. 101–110.

[32] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. "Toward an Architecture for Never-ending Language Learning." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI '10. Atlanta, Georgia: AAAI Press, 2010, pp. 1306–1313.

[33] Yee Seng Chan and Dan Roth. "Exploiting Background Knowledge for Relation Extraction." In: *Proceedings of the International Conference on Computational Linguistics*. COLING '10. Beijing, China: Association for Computational Linguistics, 2010, pp. 152–160.

[34] Moses S. Charikar. "Similarity Estimation Techniques from Rounding Algorithms." In: *Proceedings of the Annual Symposium on Theory of Computing*. STOC '02. Montreal, Quebec, Canada: ACM, 2002, pp. 380–388.

[35] Lauren E. Charles-Smith, Tera L. Reynolds, Mark A. Cameron, Mike Conway, Eric H. Y. Lau, Jennifer M. Olsen, Julie A. Pavlin, Mika Shigematsu, Laura C. Streichert, Katie J. Suda, and Courtney D. Corley. "Using Social Media for Actionable Disease Surveillance and Outbreak Management: A Systematic Literature Review." In: *PLOS ONE* 10.10 (2015), pp. 1–20.

[36] Rachel Chasin, Anna Rumshisky, Ozlem Uzuner, and Peter Szolovits. "Word sense disambiguation in the clinical domain: a comparison of knowledge-rich and knowledge-poor unsupervised methods." In: *Journal of the American Medical Informatics Association* 21.5 (2014), p. 842.

[37] Yun Chi, Yirong Yang, and Richard R. Muntz. "Canonical Forms for Labelled Trees and Their Applications in Frequent Subtree Mining." In: *Knowledge and Information Systems* 8.2 (2005), pp. 203–234.

[38] Jihye Choi, Youngtae Cho, Eunyoung Shim, and Hyekyung Woo. "Web-based infectious disease surveillance systems and public health perspectives: a systematic review." In: *BMC Public Health* 16.1 (2016), p. 1238.

[39] Hong-Woo Chun, Yoshimasa Tsuruoka, Jin-Dong Kim, Rie Shiba, Naoki Nagata, Teruyoshi Hishiki, and Jun'ichi Tsujii. "Extraction of gene-disease relations from Medline using domain dictionaries and machine learning." In: *Pacific Symposium on Biocomputing*. Vol. 11. 2006, pp. 4–15.

[40] James Clarke, Vivek Srikumar, Mark Sammons, and Dan Roth. "An NLP Curator (or: How I Learned to Stop Worrying and Love NLP Pipelines)." English. In: *Proceedings of the International Conference on Language Resources and Evaluation*. LREC '12. ACL Anthology Identifier: L12-1388. Istanbul, Turkey: European Language Resources Association (ELRA), 2012, pp. 3276–3283.

[41] Mark Craven and Johan Kumlien. "Constructing Biological Knowledge Bases by Extracting Information from Text Sources." In: *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, 1999, pp. 77–86.

[42] Aron Culotta and Jeffrey Sorensen. "Dependency Tree Kernels for Relation Extraction." In: *Proceedings of the Annual Meeting on Association for Computational Linguistics*. ACL '04. Barcelona, Spain: Association for Computational Linguistics, 2004.

[43] Luciano Del Corro and Rainer Gemulla. "ClausIE: Clause-based Open Information Extraction." In: *Proceedings of the International Conference on World Wide Web*. WWW '13. Rio de Janeiro, Brazil: ACM, 2013, pp. 355–366.

[44] Laura Dietz, Alexander Kotov, and Edgar Meij. "Utilizing Knowledge Graphs in Text-centric Information Retrieval." In: *Proceedings of the ACM International Conference on Web Search and Data Mining*. WSDM '17. Cambridge, United Kingdom: ACM, 2017, pp. 815–816.

[45] Jing Ding, Daniel Berleant, Dan Nettleton, and Eve Wurtele. "Mining MEDLINE: abstracts, sentences, or phrases?" In: *Proceedings of the Pacific Symposium on Biocomputing*. Vol. 7. World Scientific, 2002, pp. 326–337.

[46] George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. "The Automatic Content Extraction (ACE) Program Tasks, Data, and Evaluation." In: *Proceedings of the International Conference on Language Resources and Evaluation*. LREC '04. ACL Anthology Identifier: L04-1011. Lisbon, Portugal: European Language Resources Association (ELRA), 2004, pp. 837–840.

[47] John Domingue, Dieter Fensel, and James A. Hendler. *Handbook of Semantic Web Technologies*. 1st. Springer Publishing Company, Incorporated, 2011.

[48] Andreas Doms and Michael Schroeder. "GoPubMed: exploring PubMed with the Gene Ontology." In: *Nucleic Acids Research* 33.suppl_2 (2005), W783.

[49] Son Don, Ai Kawazoe, and Nigel Collier. "Global Health Monitor – A Web-based System for Detecting and Mapping Infectious Diseases." In: *Proceedings of the International Joint Conference on Natural Language Processing*. Vol. 2. IJCNLP' 08. 2008, pp. 951–956.

[50] Maximilian Dylla, Iris Miliaraki, and Martin Theobald. "A Temporal-probabilistic Database Model for Information Extraction." In: *Proceedings of the VLDB Endowment* 6.14 (2013), pp. 1810–1821.

[51] Patrick Ernst, Amy Siu, and Gerhard Weikum. "KnowLife: a versatile approach for constructing a large knowledge graph for biomedical sciences." In: *BMC Bioinformatics* 16.1 (2015), p. 157.

[52] Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. "Unsupervised Named-entity Extraction from the Web: An Experimental Study." In: *Artificial Intelligence* 165.1 (2005), pp. 91–134.

[53] Anthony Fader, Stephen Soderland, and Oren Etzioni. "Identifying Relations for Open Information Extraction." In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '11. Edinburgh, United Kingdom: Association for Computational Linguistics, 2011, pp. 1535–1545.

[54] Erik Faessler and Udo Hahn. "Semedico: A Comprehensive Semantic Search Engine for the Life Sciences." In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. ACL '17. Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 91–96.

[55] Paolo Ferragina and Ugo Scaiella. "Fast and Accurate Annotation of Short Texts with Wikipedia Pages." In: *IEEE Software* 29.1 (2012), pp. 70–75.

[56] David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A. Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John Prager, Nico Schlaefer, and Chris Welty. "Building Watson: An Overview of the DeepQA Project." In: *AI MAGAZINE* 31.3 (2010), pp. 59–79.

[57] Clark Freifeld, Kenneth Mandl, Ben Reis, and John Brownstein. "HealthMap: Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media Reports." In: *Journal of the American Medical Informatics Association* 15.2 (2008), pp. 150–157.

[58] Diego García-Olano, Marta Arias Vicente, and Josep Larriba Pey. "Instant Espresso: Interactive Analysis of Relationships in Knowledge Graphs." In: *Proceedings of the Workshop on Data Science for Social Good*. SoGood '16. Riva del Garda, Italy, 2016, pp. 1–17.

[59] Sahil Garg, Aram Galstyan, Ulf Hermjakob, and Daniel Marcu. "Extracting Biomolecular Interactions Using Semantic Parsing of Biomedical Text." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI '16. Phoenix, Arizona: AAAI Press, 2016, pp. 2718–2726.

[60] Guillermo Garrido, Anselmo Peñas, Bernardo Cabaleiro, and Álvaro Rodrigo. "Temporally Anchored Relation Extraction." In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. ACL '12. Jeju Island, Korea: Association for Computational Linguistics, 2012, pp. 107–116.

[61] Kiril Gashteovski, Rainer Gemulla, and Luciano Del Corro. "MinIE: Minimizing Facts in Open Information Extraction." In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '17. Copenhagen, Denmark: Association for Computational Linguistics, 2017, pp. 2620–2630.

[62] Daniel Gildea and Daniel Jurafsky. "Automatic Labeling of Semantic Roles." In: *Computational Linguistics* 28.3 (2002), pp. 245–288.

[63] Benjamin M. Good and Andrew I. Su. "Crowdsourcing for bioinformatics." In: *Bioinformatics* 29.16 (2013), p. 1925.

[64] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[65] Ralph Grishman, Silja Huttunen, and Roman Yangarber. "Information extraction for enhanced access to disease outbreak reports." In: *Journal of Biomedical Informatics* 35.4 (2002), pp. 236–246.

[66] Yufan Guo, Diarmuid Ó Séaghdha, Ilona Silins, Lin Sun, Johan Högberg, Ulla Stenius, and Anna Korhonen. "CRAB 2.0: A text mining tool for supporting literature review in chemical cancer risk assessment." In: *Proceedings of the International Conference on Computational Linguistics*. COLING '14. Stroudsburg, PA, USA: Association for Computational Linguistics, 2014, pp. 76–80.

[67] Nathan Harmston, Wendy Filsell, and Michael P. H. Stumpf. "Which species is it? Species-driven gene name disambiguation using random walks over a mixture of adjacency matrices." In: *Bioinformatics* 28.2 (2012), p. 254.

[68] Daniel R Harris, Ramakanth Kavuluru, Jerzy W Jaromczyk, and Todd R Johnson. "Rapid and Reusable Text Visualization and Exploration Development with DELVE." In: *Proceedings of the AMIA Joint Summits on Translational Science* 2017 (2017), 139——148.

[69] Marti A. Hearst. "Automatic Acquisition of Hyponyms from Large Text Corpora." In: *Proceedings of the International Conference on Computational Linguistics*. Vol. 2. COLING '92. Nantes, France: Association for Computational Linguistics, 1992, pp. 539–545.

[70] Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó. Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. "SemEval-2010 Task 8: Multi-way Classification of Semantic Relations Between Pairs of Nominals." In: *Proceedings of the International Workshop on Semantic Evaluation*. SemEval '10. Los Angeles, California: Association for Computational Linguistics, 2010, pp. 33–38.

[71] Johannes Hoffart, Dragan Milchevski, and Gerhard Weikum. "STICS: Searching with Strings, Things, and Cats." In: *Proceedings of the ACM SIGIR International Conference on Research; Development in Information Retrieval*. SIGIR '14. Gold Coast, Queensland, Australia: ACM, 2014, pp. 1247–1248.

[72] Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. "Robust Disambiguation of Named Entities in Text." In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '11. Edinburgh, United Kingdom: Association for Computational Linguistics, 2011, pp. 782–792.

[73] Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, and Gerhard Weikum. "YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia." In: *Artificial Intelligence* 194 (2013), pp. 28–61.

[74] Raphael Hoffmann, Congle Zhang, and Daniel S. Weld. "Learning 5000 Relational Extractors." In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. ACL '10. Uppsala, Sweden: Association for Computational Linguistics, 2010, pp. 286–295.

[75] Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. "Knowledge-based Weak Supervision for Information Extraction of Overlapping Relations." In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. HLT '11. Portland, Oregon: Association for Computational Linguistics, 2011, pp. 541–550.

[76] Lawrence Hunter, Zhiyong Lu, James Firby, William A. Baumgartner, Helen L. Johnson, Philip V. Ogren, and K Bretonnel Cohen. "OpenDMAP: An open source, ontology-driven concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-type-specific gene expression." In: *BMC Bioinformatics* 9.1 (2008), p. 78.

[77] David S. Johnson. "Approximation algorithms for combinatorial problems." In: *Journal of Computer and System Sciences* 9.3 (1974), pp. 256–278.

[78] Nanda Kambhatla. "Combining Lexical, Syntactic, and Semantic Features with Maximum Entropy Models for Extracting Relations." In: *Proceedings of the Annual Meeting on Association for Computational Linguistics*. ACL '04. Barcelona, Spain: Association for Computational Linguistics, 2004.

[79] Mikaela Keller, Michael Blench, Herman Tolentino, Clark C Freifeld, Kenneth D Mandl, Abla Mawudeku, Gunther Eysenbach, and John S Brownstein. "Use of Unstructured Event-Based Reports for Global Infectious Disease Surveillance." In: *Emerging Infectious Disease Journal* 15.5 (2009), p. 689.

[80] Jin-Dong Kim, Tomoko Ohta, and Jun'ichi Tsujii. "Corpus annotation for mining biomedical events from literature." In: *BMC Bioinformatics* 9.1 (2008), p. 10.

[81] Jin-Dong Kim, Jung-jae Kim, Xu Han, and Dietrich Rebholz-Schuhmann. "Extending the evaluation of Genia Event task toward knowledge base construction and comparison to Gene Regulation Ontology task." In: *BMC Bioinformatics* 16.10 (2015), S3.

[82] Jung-jae Kim, Piotr Pezik, and Dietrich Rebholz-Schuhmann. "MedEvi: Retrieving textual evidence of relations between biomedical concepts from Medline." In: *Bioinformatics* 24.11 (2008), pp. 1410–1412.

[83] Corinna Kolářik, Martin Hofmann-Apitius, Marc Zimmermann, and Juliane Fluck. "Identification of new drug classification terms in textual resources." In: *Bioinformatics* 23.13 (2007), p. i264.

[84] Martin Krallinger, Jose Izarzugaza, Carlos Rodriguez-Penagos, and Alfonso Valencia. "Extraction of human kinase mutations from literature, databases and genotyping studies." In: *BMC Bioinformatics* 10.8 (2009), S1.

[85] Sebastian Krause, Hong Li, Hans Uszkoreit, and Feiyu Xu. "Large-Scale Learning of Relation-extraction Rules with Distant Supervision from the Web." In: *Proceedings of the International Conference on The Semantic Web*. ISWC '12. Boston, MA: Springer-Verlag, 2012, pp. 263–278.

[86] Sebastian Krause, Leonhard Hennig, Andrea Moro, Dirk Weissenborn, Feiyu Xu, Hans Uszkoreit, and Roberto Navigli. "Sargraphs: A language resource connecting linguistic knowledge with semantic relations from knowledge graphs." In: *Web Semantics: Science, Services and Agents on the World Wide Web* 37 (2016), pp. 112–131.

[87] Erdal Kuzey, Jilles Vreeken, and Gerhard Weikum. "A Fresh Look on Knowledge Bases: Distilling Named Events from News." In: *Proceedings of the ACM International Conference on Conference on Information and Knowledge Management*. CIKM '14. Shanghai, China: ACM, 2014, pp. 1689–1698.

[88] Erdal Kuzey and Gerhard Weikum. "Extraction of Temporal Facts and Events from Wikipedia." In: *Proceedings of the Temporal Web Analytics Workshop*. TempWeb '12. Lyon, France: ACM, 2012, pp. 25–32.

[89] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, and Christian Bizer. "DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia." In: *Semantic Web* 6.2 (2015), pp. 167–195.

[90] Gaël Lejeune, Romain Brixtel, Antoine Doucet, and Nadine Lucas. "Multilingual event extraction for epidemic detection." In: *Artificial Intelligence in Medicine* 65.2 (2015). Intelligent healthcare informatics in big data era, pp. 131–143.

[91] Gondy Leroy and Hsinchun Chen. "Genescene: An ontology-enhanced integration of linguistic and co-occurrence based relations in biomedical texts." In: *Journal of the American Society for Information Science and Technology* 56.5 (2005), pp. 457–468.

[92] Hong Li, Sebastian Krause, Feiyu Xu, Andrea Moro, Hans Uszkoreit, and Roberto Navigli. "Improvement of N-ary Relation Extraction by Adding Lexical Semantics to Distant-Supervision Rule Learning." In: *Proceedings of the International Conference on Agents and Artificial Intelligence*. ICAART 2015. Lisbon, Portugal: SCITEPRESS - Science and Technology Publications, Lda, 2015, pp. 317–324.

[93] Qi Li and Heng Ji. "Incremental Joint Extraction of Entity Mentions and Relations." In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. ACL '14. Stroudsburg, PA, USA: Association for Computational Linguistics, 2014, pp. 402–412.

[94] Haibin Liu, Lawrence Hunter, Vlado Kešelj, and Karin Verspoor. "Approximate Subgraph Matching-Based Literature Mining for Biomedical Events and Relations." In: *PLOS ONE* 8.4 (2013), pp. 1–16.

[95] Haibin Liu, Karin Verspoor, Donald C. Comeau, Andrew D. MacKinlay, and W. John Wilbur. "Optimizing graph-based patterns to extract biomedical events from the literature." In: *BMC Bioinformatics* 16.16 (2015), S2.

[96] Yifeng Liu, Yongjie Liang, and David Wishart. "PolySearch2: a significantly improved text-mining system for discovering associations between human diseases, genes, drugs, metabolites, toxins and more." In: *Nucleic Acids Research* 43.W1 (2015), W535–W542.

[97] Emily K. Mallory, Ce Zhang, Christopher Ré, and Russ B. Altman. "Large-scale extraction of gene interactions from full-text literature using DeepDive." In: *Bioinformatics* 32.1 (2016), pp. 106–113.

[98] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. "The Stanford CoreNLP Natural Language Processing Toolkit." In: *Proceedings of Annual Meeting of the Association for Computational Linguistics*. ACL '14. Baltimore, Maryland: Association for Computational Linguistics, 2014, pp. 55–60.

[99] Volker Markl, Peter J. Haas, Marcel Kutsch, Nimrod Megiddo, Utkarsh Srivastava, and Tam Minh Tran. "Consistent Selectivity Estimation via Maximum Entropy." In: *The VLDB Journal* 16.1 (2007), pp. 55–76.

[100] Mausam, Michael Schmitz, Robert Bart, Stephen Soderland, and Oren Etzioni. "Open Language Learning for Information Extraction." In: *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natu-*

*ral Language Learning*. EMNLP-CoNLL '12. Jeju Island, Korea: Association for Computational Linguistics, 2012, pp. 523–534.

[101] Mausam Mausam. "Open Information Extraction Systems and Downstream Applications." In: *Proceedings of the International Joint Conference on Artificial Intelligence*. IJCAI '16. New York, New York, USA: AAAI Press, 2016, pp. 4074–4077.

[102] David McClosky, Sebastian Riedel, Mihai Surdeanu, Andrew McCallum, and Christopher D. Manning. "Combining joint models for biomedical event extraction." In: *BMC Bioinformatics* 13.11 (2012), S9.

[103] Alexa T. McCray, Anita Burgun, and Oliver Bodenreider. "Aggregating UMLS semantic types for reducing conceptual complexity." In: *Studies in Health Technology and Informatics* 84.0 1 (2001), pp. 216–220.

[104] Ryan McDonald, Fernando Pereira, Seth Kulick, Scott Winters, Yang Jin, and Pete White. "Simple Algorithms for Complex Relation Extraction with Applications to Biomedical IE." In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. ACL '05. Ann Arbor, Michigan: Association for Computational Linguistics, 2005, pp. 491–498.

[105] Filipe Mesquita, Jordan Schmidek, and Denilson Barbosa. "Effectiveness and Efficiency of Open Relation Extraction." In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '13. Seattle, Washington, USA: Association for Computational Linguistics, 2013, pp. 447–457.

[106] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. "Distant Supervision for Relation Extraction Without Labeled Data." In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. ACL '09. Suntec, Singapore: Association for Computational Linguistics, 2009, pp. 1003–1011.

[107] T. Mitchell et al. "Never-ending Learning." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI '15. Austin, Texas: AAAI Press, 2015, pp. 2302–2310.

[108] Makoto Miwa, Rune Sætre, Yusuke Miyao, and Jun'ichi Tsujii. "Protein–protein interaction extraction by leveraging multiple kernels and parsers." In: *International Journal of Medical Informatics* 78.12 (2009). Mining of Clinical and Biomedical Text and Data Special Issue, pp. 39–46.

[109] Makoto Miwa, Rune Sætre, Jin-Dong Kim, and Jun'ichi Tsujii. "Event Extraction with complex event classification using rich features." In: *Journal of Bioinformatics and Computational Biology* 8.1 (2010), pp. 131–146.

[110] Makoto Miwa, Paul Thompson, John McNaught, Douglas B. Kell, and Sophia Ananiadou. "Extracting semantically enriched events from biomedical literature." In: *BMC Bioinformatics* 13.1 (2012), p. 108.

[111] Yusuke Miyao, Tomoko Ohta, Katsuya Masuda, Yoshimasa Tsuruoka, Kazuhiro Yoshida, Takashi Ninomiya, and Jun'ichi Tsujii. "Semantic Retrieval for the Accurate Identification of Relational Concepts in Massive Textbases." In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. ACL '06. Sydney, Australia: Association for Computational Linguistics, 2006, pp. 1017–1024.

[112] Raymond J. Mooney and Razvan C. Bunescu. "Subsequence Kernels for Relation Extraction." In: *Advances in Neural Information Processing Systems*. Ed. by Y. Weiss, P. B. Schölkopf, and J. C. Platt. NIPS '05. MIT Press, 2006, pp. 171–178.

[113] Alessandro Moschitti. "Making Tree Kernels Practical for Natural Language Learning." In: *Conference of the European Chapter of the Association for Computational Linguistics*. EACL '06. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006, pp. 113–120.

[114] Dana Movshovitz-Attias and William W. Cohen. "Bootstrapping Biomedical Ontologies for Scientific Text Using NELL." In: *Proceedings of the Workshop on Biomedical Natural Language Processing*. BioNLP '12. Montreal, Quebec, Canada: Association for Computational Linguistics, 2012, pp. 11–19.

[115] Subhabrata Mukherjee, Gerhard Weikum, and Cristian Danescu-Niculescu-Mizil. "People on Drugs: Credibility of User Statements in Health Communities." In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '14. New York, NY, USA: ACM, 2014, pp. 65–74.

[116] Hans-Michael Müller, Eimear E Kenny, and Paul W Sternberg. "Textpresso: An Ontology-Based Information Retrieval and Extraction System for Biological Literature." In: *PLOS Biology* 2.11 (2004).

[117] Mark A. Musen, Natasha F. Noy, Nigam H. Shah, Patricia L. Whetzel, Christopher G. Chute, Margaret-Anne Story, Barry Smith, and the NCBO team. "The National Center for Biomedical Ontology." In: *Journal of the American Medical Informatics Association* 19.2 (2012), pp. 190–195.

[118] Ndapandula Nakashole, Martin Theobald, and Gerhard Weikum. "Scalable Knowledge Harvesting with High Precision and High Recall." In: *Proceedings of the ACM International*

*Conference on Web Search and Data Mining*. WSDM '11. Hong Kong, China: ACM, 2011, pp. 227–236.

[119] Victoria Nebot Romero, Min Ye, Mario Albrecht, Jae-Hong Eom, and Gehard Weikum. "DIDO: A Disease-determinants Ontology from Web Sources." In: *Proceedings of the International Conference on World Wide Web*. WWW '11. Hyderabad, India: ACM, 2011, pp. 237–240.

[120] Victoria Nebot and Rafael Berlanga. "Exploiting semantic annotations for open information extraction: an experience in the biomedical domain." In: *Knowledge and Information Systems* 38.2 (2014), pp. 365–389.

[121] Claire Nédellec. "Learning language in logic-genic interaction extraction challenge." In: *Proceedings of the Learning Language in Logic Workshop*. Vol. 7. LLL '05. 2005, pp. 31–37.

[122] Mariana Neves. "An analysis on the entity annotations in biological corpora." In: *F1000Research* 3.96 (2014).

[123] Nhung TH Nguyen, Makoto Miwa, Yoshimasa Tsuruoka, Takashi Chikayama, and Satoshi Tojo. "Wide-coverage relation extraction from MEDLINE using deep syntax." In: *BMC Bioinformatics* 16.1 (2015), p. 107.

[124] Rolf Niedermeier and Peter Rossmanith. "New Upper Bounds for Maximum Satisfiability." In: *Journal of Algorithms* 36.1 (2000), pp. 63–88.

[125] Natasha Noy and Alan Rector. *Defining N-ary Relations on the Semantic Web*. W3C Working Group Note. World Wide Web Consortium, 2006.

[126] Martha Palmer, Daniel Gildea, and Paul Kingsbury. "The Proposition Bank: An Annotated Corpus of Semantic Roles." In: *Computational Linguistics* 31.1 (2005), pp. 71–106.

[127] Martha Palmer, Daniel Gildea, and Nianwen Xue. "Semantic Role Labeling." In: *Synthesis Lectures on Human Language Technologies* 3.1 (2010), pp. 1–103.

[128] Patrick Cheong-Iao Pang, Karin Verspoor, Jon Pearce, and Shanton Chang. "Better Health Explorer: Designing for Health Information Seekers." In: *Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction*. OzCHI '15. Parkville, VIC, Australia: ACM, 2015, pp. 588–597.

[129] Patrick Cheong-Iao Pang, Karin Verspoor, Jon Pearce, and Shanton Chang. "Finding and Exploring Health Information with a Slider-Based User Interface." In: *Studies in Health Technology and Informatics* 227 (2016), pp. 106–112.

[130] Nikolas Papanikolaou, Georgios A. Pavlopoulos, Evangelos Pafilis, Theodosios Theodosiou, Reinhard Schneider, Venkata P. Satagopam, Christos A. Ouzounis, Aristides G. Eliopoulos, Vasilis J. Promponas, and Ioannis Iliopoulos. "BioTextQuest+ : a knowledge integration platform for literature mining and concept discovery." In: *Bioinformatics* 30.22 (2014), pp. 3249–3256.

[131] Ankur P. Parikh, Hoifung Poon, and Kristina Toutanova. "Grounded Semantic Parsing for Complex Knowledge Extraction." In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, Colorado: Association for Computational Linguistics, 2015, pp. 756–766.

[132] Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. "Cross-Sentence N-ary Relation Extraction with Graph LSTMs." In: *Transactions of the Association of Computational Linguistics* 5 (2017), pp. 101–115.

[133] Yifan Peng and Zhiyong Lu. "Deep learning for extracting protein-protein interactions from biomedical literature." In: *Proceedings of the Workshop on Biomedical Natural Language Processing*. BioNLP '17. Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 29–38.

[134] Conrad Plake, Torsten Schiemann, Marcus Pankalla, Jörg Hakenberg, and Ulf Leser. "AliBaba: PubMed as a graph." In: *Bioinformatics* 22.19 (2006), pp. 2444–2445.

[135] Hoifung Poon and Pedro Domingos. "Unsupervised Semantic Parsing." In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '09. Singapore: Association for Computational Linguistics, 2009, pp. 1–10.

[136] Hoifung Poon and Pedro Domingos. "Unsupervised Ontology Induction from Text." In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. ACL '10. Uppsala, Sweden: Association for Computational Linguistics, 2010, pp. 296–305.

[137] Hoifung Poon, Kristina Toutanova, and Chris Quirk. "Distant Supervision for Cancer Pathway Extraction from Text." In: *Proceedings of the Pacific Symposium on Biocomputing*. Singapore: World Scientific Publishing Company, 2015, pp. 120–131.

[138] Hoifung Poon, Chris Quirk, Charlie DeZiel, and David Heckerman. "Literome: PubMed-scale genomic knowledge base in the cloud." In: *Bioinformatics* 30.19 (2014), pp. 2840–2842.

[139] Soujanya Poria, Erik Cambria, Grégoire Winterstein, and Guang-Bin Huang. "Sentic patterns: Dependency-based rules for concept-level sentiment analysis." In: *Knowledge-Based Systems* 69 (2014), 45–63.

[140] Vasin Punyakanok, Dan Roth, and Wen-tau Yih. "The Importance of Syntactic Parsing and Inference in Semantic Role Labeling." In: *Computational Linguistics* 34.2 (2008), pp. 257–287.

[141] Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Björne, Jorma Boberg, Jouni Järvinen, and Tapio Salakoski. "BioInfer: a corpus for information extraction in the biomedical domain." In: *BMC Bioinformatics* 8.1 (2007), p. 50.

[142] Sampo Pyysalo, Tomoko Ohta, Makoto Miwa, Han-Cheol Cho, Jun'ichi Tsujii, and Sophia Ananiadou. "Event extraction across multiple levels of biological organization." In: *Bioinformatics* 28.18 (2012), p. i575.

[143] Longhua Qian and Guodong Zhou. "Tree kernel-based protein–protein interaction extraction from biomedical literature." In: *Journal of Biomedical Informatics* 45.3 (2012), pp. 535–543.

[144] Longhua Qian, Guodong Zhou, Fang Kong, Qiaoming Zhu, and Peide Qian. "Exploiting Constituent Dependencies for Tree Kernel-based Semantic Relation Extraction." In: *Proceedings of the International Conference on Computational Linguistics*. COLING '08. Manchester, United Kingdom: Association for Computational Linguistics, 2008, pp. 697–704.

[145] Changqin Quan, Meng Wang, and Fuji Ren. "An Unsupervised Text Mining Method for Relation Extraction from Biomedical Literature." In: *PLOS ONE* 9.7 (2014), pp. 1–8.

[146] Benjamin L. Ranard, Yoonhee P. Ha, Zachary F. Meisel, David A. Asch, Shawndra S. Hill, Lance B. Becker, Anne K. Seymour, and Raina M. Merchant. "Crowdsourcing—Harnessing the Masses to Advance Health and Medicine, a Systematic Review." In: *Journal of General Internal Medicine* 29.1 (2014), pp. 187–203.

[147] Lev Ratinov and Dan Roth. "Design Challenges and Misconceptions in Named Entity Recognition." In: *Proceedings of the Conference on Computational Natural Language Learning*. CoNLL '09. Boulder, Colorado: Association for Computational Linguistics, 2009, pp. 147–155.

[148] Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. "Local and Global Algorithms for Disambiguation to Wikipedia." In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Vol-*

*ume 1*. HLT '11. Portland, Oregon: Association for Computational Linguistics, 2011, pp. 1375–1384.

[149] Xiang Ren, Zeqiu Wu, Wenqi He, Meng Qu, Clare R. Voss, Heng Ji, Tarek F. Abdelzaher, and Jiawei Han. "CoType: Joint Extraction of Typed Entities and Relations with Knowledge Bases." In: *Proceedings of the International Conference on World Wide Web*. WWW '17. Perth, Australia: International World Wide Web Conferences Steering Committee, 2017, pp. 1015–1024.

[150] Xiang Ren, Jiaming Shen, Meng Qu, Xuan Wang, Zeqiu Wu, Qi Zhu, Meng Jiang, Fangbo Tao, Saurabh Sinha, David Liem, Peipei Ping, Richard Weinshilboum, and Jiawei Han. "Life-iNet: A Structured Network-Based Knowledge Exploration and Analytics System for Life Sciences." In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. ACL '17. Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 55–60.

[151] Sebastian Riedel and Andrew McCallum. "Fast and Robust Joint Models for Biomedical Event Extraction." In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '11. Edinburgh, United Kingdom: Association for Computational Linguistics, 2011, pp. 1–12.

[152] Sebastian Riedel, Limin Yao, and Andrew McCallum. "Modeling Relations and Their Mentions Without Labeled Text." In: *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*. ECML PKDD'10. Barcelona, Spain: Springer-Verlag, 2010, pp. 148–163.

[153] Thomas C. Rindflesch and Marcelo Fiszman. "The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text." In: *Journal of Biomedical Informatics* 36.6 (2003), pp. 462–477.

[154] Thomas C Rindflesch, Bisharah Libbus, Dimitar Hristovski, Alan R Aronson, and Halil Kilicoglu. "Semantic Relations Asserting the Etiology of Genetic Diseases." In: *AMIA Annual Symposium Proceedings*. Vol. 2003. Bethesda, MD, USA: American Medical Informatics Association, 2003, pp. 554–558.

[155] Thomas C. Rindflesch, Halil Kilicoglu, Marcelo Fiszman, Graciela Rosemblat, and Dongwook Shin. "Semantic MEDLINE: An Advanced Information Management Application for Biomedicine." In: *Inf. Serv. Use* 31.1-2 (2011), pp. 15–21.

[156] Agnès Rortais, Jenya Belyaeva, Monica Gemo, Erik van der Goot, and Jens Linge. "MedISys: An early-warning system for the detection of (re-)emerging food- and feed-borne hazards." In: *Food Research Internat.* 43.5 (2010), pp. 1553–1556.

[157] Barbara Rosario and Marti A. Hearst. "Classifying Semantic Relations in Bioscience Texts." In: *Proceedings of the Annual Meeting on Association for Computational Linguistics*. ACL '04. Barcelona, Spain: Association for Computational Linguistics, 2004.

[158] Cornelius Rosse and José L.V. Mejino. "A reference ontology for biomedical informatics: the Foundational Model of Anatomy." In: *Journal of Biomedical Informatics* 36.6 (2003), pp. 478–500.

[159] Maya Rotmensch, Yoni Halpern, Abdulhakim Tlimat, Steven Horng, and David Sontag. "Learning a Health Knowledge Graph from Electronic Medical Records." In: *Scientific Reports* 7.1 (2017), p. 5994.

[160] Cicero dos Santos, Bing Xiang, and Bowen Zhou. "Classifying Relations by Ranking with Convolutional Neural Networks." In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2015, pp. 626–634.

[161] Lynn Marie Schriml, Cesar Arze, Suvarna Nadendla, Yu-Wei Wayne Chang, Mark Mazaitis, Victor Felix, Gang Feng, and Warren Alden Kibbe. "Disease Ontology: a backbone for disease semantic integration." In: *Nucleic Acids Research* 40.D1 (2012), pp. D940–D946.

[162] Isabel Segura-Bedmar, Paloma Martínez, and César de Pablo-Sánchez. "Extracting drug-drug interactions from biomedical texts." In: *BMC Bioinformatics* 11.5 (2010), P9.

[163] Stephan Seufert, Patrick Ernst, Srikanta J. Bedathur, Sarath Kumar Kondreddi, Klaus Berberich, and Gerhard Weikum. "Instant Espresso: Interactive Analysis of Relationships in Knowledge Graphs." In: *Proceedings of the International Conference on World Wide Web*. WWW '16. Montreal, Quebec, Canada: International World Wide Web Conferences Steering Committee, 2016, pp. 251–254.

[164] Dafna Shahaf and Carlos Guestrin. "Connecting Two (or Less) Dots: Discovering Structure in News Articles." In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 5.4 (2012), 24:1–24:31.

[165] Wei Shen, Jianyong Wang, and Jiawei Han. "Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions." In: *IEEE Transactions on Knowledge and Data Engineering* 27.2 (2015), pp. 443–460.

[166] Jaeho Shin, Sen Wu, Feiran Wang, Christopher De Sa, Ce Zhang, and Christopher Ré. "Incremental Knowledge Base Construction Using DeepDive." In: *Proceedings of the VLDB Endowment* 8.11 (2015), pp. 1310–1321.

[167] Amy Siu. "Knowledge-driven Entity Disambiguation in Biomedical Text." PhD thesis. Saarbrücken: Saarland University, 2017.

[168] Amy Siu, Dat Ba Nguyen, and Gerhard Weikum. "Fast entity recognition in biomedical text." In: *Proceedings of the Workshop on Data Mining for Healthcare*. DMH '13. New York, NY, USA: ACM, 2013.

[169] Barry Smith. "Ontology." In: *The Blackwell Guide to the Philosophy of Computing and Information*. Blackwell Publishing Ltd, 2008, pp. 153–166.

[170] Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. "Semantic Compositionality Through Recursive Matrix-vector Spaces." In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '12. Jeju Island, Korea: Association for Computational Linguistics, 2012, pp. 1201–1211.

[171] Andreas Spitz, Satya Almasian, and Michael Gertz. "EVELIN: Exploration of Event and Entity Links in Implicit Networks." In: *Proceedings of the International Conference on World Wide Web*. WWW '17. Perth, Australia: International World Wide Web Conferences Steering Committee, 2017, pp. 273–277.

[172] Vivek Srikumar and Dan Roth. "Modeling Semantic Relations Expressed by Prepositions." In: *Transactions of the Association for Computational Linguistics* 1 (2013), pp. 231–242.

[173] Padmini Srinivasan, Xiao-Ning Zhang, Roxane Bouten, and Caren Chang. "Ferret: a sentence-based literature scanning system." In: *BMC Bioinformatics* 16.1 (2015), p. 198.

[174] Ralf Steinberger, Flavio Fuart, Erik van der Goot, Clive Best, Peter von Etter, and Roman Yangarber. "Text Mining from the Web for Medical Intelligence." In: *Mining Massive Data Sets for Security*. Vol. 19. IOS Press, 2008, pp. 295–310.

[175] Fabian M. Suchanek, Georgiana Ifrim, and Gerhard Weikum. "Combining Linguistic and Statistical Analysis to Extract Relations from Web Documents." In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and*

*Data Mining*. KDD '06. Philadelphia, PA, USA: ACM, 2006, pp. 712–717.

[176] Fabian M. Suchanek, Mauro Sozio, and Gerhard Weikum. "SOFIE: A Self-organizing Framework for Information Extraction." In: *Proceedings of the International Conference on World Wide Web*. WWW '09. Madrid, Spain: ACM, 2009, pp. 631–640.

[177] Fabian M. Suchanek and Gerhard Weikum. "Knowledge harvesting from text and Web sources." In: *Proceedings of the IEEE International Conference on Data Engineering*. Washington, DC, USA: IEEE Computer Society, 2013, pp. 1250–1253.

[178] Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. "Evaluating temporal relations in clinical text: 2012 i2b2 Challenge." In: *Journal of the American Medical Informatics Association* 20.5 (2013), pp. 806–813.

[179] Mihai Surdeanu and Ji Heng. "Overview of the English Slot Filling Track at the TAC2014 Knowledge Base Population Evaluation." In: *Proceedings of the TAC-KBP 2014 Workshop*. 2014.

[180] György Szarvas, Veronika Vincze, Richárd Farkas, György Móra, and Iryna Gurevych. "Cross-Genre and Cross-Domain Detection of Semantic Uncertainty." In: *Computational Linguistics* 38.2 (2012), pp. 335–367.

[181] Partha Pratim Talukdar, Derry Wijaya, and Tom Mitchell. "Coupled Temporal Scoping of Relational Facts." In: *Proceedings of the ACM International Conference on Web Search and Data Mining*. WSDM '12. Seattle, Washington, USA: ACM, 2012, pp. 73–82.

[182] Philippe Thomas, Johannes Starlinger, Alexander Vowinkel, Sebastian Arzt, and Ulf Leser. "GeneView: a comprehensive semantic search engine for PubMed." In: *Nucleic Acids Research* 40.W1 (2012), W585.

[183] Manabu Torii, Cecilia N. Arighi, Qinghua Wang, Cathy H. Wu, and K. Vijay-Shanker. "Text Mining of Protein Phosphorylation Information Using a Generalizable Rule-Based Approach." In: *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*. BCB '13. Wshington DC, USA: ACM, 2013, pp. 201–210.

[184] Yoshimasa Tsuruoka, Makoto Miwa, Kaisei Hamamoto, Jun'ichi Tsujii, and Sophia Ananiadou. "Discovering and visualizing indirect associations between biomedical concepts." In: *Bioinformatics* 27.13 (2011), pp. i111–i119.

[185] Marco Valenzuela, Vu Ha, and Oren Etzioni. "Identifying Meaningful Citations." In: *Proceedings of the Workshop on Scholarly Big Data at AAAI*. 2015.

[186] Sofie Van Landeghem, Kai Hakala, Samuel Rönnqvist, Tapio Salakoski, Yves Van de Peer, and Filip Ginter. "Exploring Biomolecular Literature with EVEX: Connecting Genes through Events, Homology, and Indirect Associations." In: *Advances in Bioinformatics* 2012 (2012), p. 12.

[187] Sofie Van Landeghem, Jari Björne, Chih-Hsuan Wei, Kai Hakala, Sampo Pyysalo, Sophia Ananiadou, Hung-Yu Kao, Zhiyong Lu, Tapio Salakoski, Yves Van de Peer, and Filip Ginter. "Large-Scale Event Extraction from Literature with Multi-Level Gene Normalization." In: *PLOS ONE* 8.4 (2013), pp. 1–12.

[188] Denny Vrandečić and Markus Krötzsch. "Wikidata: A Free Collaborative Knowledgebase." In: *Communications of the ACM* 57.10 (2014), pp. 78–85.

[189] Vinod V.G. Vydiswaran, Cheng Xiang Zhai, and Dan Roth. "Gauging the Internet Doctor: Ranking Medical Claims Based on Community Knowledge." In: *Proceedings of the Workshop on Data Mining for Medicine and Healthcare*. DMMH '11. San Diego, California, USA: ACM, 2011, pp. 42–51.

[190] Chang Wang and James Fan. "Medical Relation Extraction with Manifold Models." In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Baltimore, Maryland: Association for Computational Linguistics, 2014, pp. 828–838.

[191] Linlin Wang, Zhu Cao, Gerard de Melo, and Zhiyuan Liu. "Relation Classification via Multi-Level Attention CNNs." In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. ACL '16. Stroudsburg, PA, USA: Association for Computational Linguistics, 2016, pp. 1298–1307.

[192] Yafang Wang, Bin Yang, Lizhen Qu, Marc Spaniol, and Gerhard Weikum. "Harvesting Facts from Textual Web Sources by Constrained Label Propagation." In: *Proceedings of the ACM International Conference on Information and Knowledge Management*. CIKM '11. Glasgow, Scotland, UK: ACM, 2011, pp. 837–846.

[193] Tuangthong Wattarujeekrit, Parantu K. Shah, and Nigel Collier. "PASBio: predicate-argument structures for event extraction in molecular biology." In: *BMC Bioinformatics* 5.1 (2004), p. 155.

[194] Tim Weninger, William H. Hsu, and Jiawei Han. "CETR: Content Extraction via Tag Ratios." In: *Proceedings of the International Conference on World Wide Web*. WWW '10. Raleigh, North Carolina, USA: ACM, 2010, pp. 971–980.

[195] Patricia L. Whetzel, Natalya F. Noy, Nigam H. Shah, Paul R. Alexander, Csongor Nyulas, Tania Tudorache, and Mark A. Musen. "BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications." In: *Nucleic Acids Research* 39.suppl_2 (2011), W541–W545.

[196] M Whirl-Carrillo, E M McDonagh, J M Hebert, L Gong, K Sangkuhl, C F Thorn, R B Altman, and T E Klein. "Pharmacogenomics Knowledge for Personalized Medicine." In: *Clinical Pharmacology & Therapeutics* 92.4 (2012), pp. 414–417.

[197] Aaron Steven White, Drew Reisinger, Keisuke Sakaguchi, Tim Vieira, Sheng Zhang, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. "Universal Decompositional Semantics on Universal Dependencies." In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '16. Austin, Texas, USA: Association for Computational Linguistics, 2016, pp. 1713–1723.

[198] Ryen W. White, Rave Harpaz, Nigam H. Shah, William DuMouchel, and Eric Horvitz. "Toward enhanced pharmacovigilance using patient-generated data on the Internet." In: *Clinical Pharmacology & Therapeutics* 96 (2014).

[199] Antony J. Williams, Lee Harland, Paul Groth, Stephen Pettifer, Christine Chichester, Egon L. Willighagen, Chris T. Evelo, Niklas Blomberg, Gerhard Ecker, Carole Goble, and Barend Mons. "Open PHACTS: semantic interoperability for drug discovery." In: *Drug Discovery Today* 17.21 (2012), pp. 1188–1198.

[200] Adam Wright, Elizabeth S. Chen, and Francine L. Maloney. "An automated technique for identifying associations between medications, laboratory results and problems." In: *Journal of Biomedical Informatics* 43.6 (2010), pp. 891 –901.

[201] Fei Wu and Daniel S. Weld. "Autonomously Semantifying Wikipedia." In: *Proceedings of the ACM Conference on Conference on Information and Knowledge Management*. CIKM '07. Lisbon, Portugal: ACM, 2007, pp. 41–50.

[202] Fei Wu and Daniel S. Weld. "Open Information Extraction Using Wikipedia." In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. ACL '10. Uppsala, Sweden: Association for Computational Linguistics, 2010, pp. 118–127.

[203] Minguang Xiao and Cong Liu. "Semantic Relation Classification via Hierarchical Recurrent Neural Network with Attention." In: *Proceedings of the International Conference on Computational Linguistics*. COLING '16. Osaka, Japan: The COLING Organizing Committee, 2016, pp. 1254–1263.

[204] Chenyan Xiong, Russell Power, and Jamie Callan. "Explicit Semantic Ranking for Academic Search via Knowledge Graph Embedding." In: *Proceedings of the International Conference on World Wide Web*. WWW '17. Perth, Australia: International World Wide Web Conferences Steering Committee, 2017, pp. 1271–1279.

[205] Rong Xu, Li Li, and QuanQiu Wang. "dRiskKB: a large-scale disease-disease risk relationship knowledge base constructed from biomedical text." In: *BMC Bioinformatics* 15.1 (2014), p. 105.

[206] Rong Xu and QuanQiu Wang. "Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing." In: *BMC Bioinformatics* 14.1 (2013), p. 181.

[207] Yan Xu, Lili Mou, Ge Li, Yunchuan Chen, Hao Peng, and Zhi Jin. "Classifying Relations via Long Short Term Memory Networks along Shortest Dependency Paths." In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '15. Lisbon, Portugal: Association for Computational Linguistics, 2015, pp. 1785–1794.

[208] Mohamed Yahya, Steven Whang, Rahul Gupta, and Alon Halevy. "ReNoun: Fact Extraction for Nominal Attributes." In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '14'. Doha, Qatar: Association for Computational Linguistics, 2014, pp. 325–335.

[209] Cong Yu, Denilson Barbosa, and Haixun Wang. "Shallow Information Extraction for the Knowledge Web." In: *Proceedings of the IEEE International Conference on Data Engineering*. ICDE '13. Washington, DC, USA: IEEE Computer Society, 2013, pp. 1264–1267.

[210] Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. "Kernel Methods for Relation Extraction." In: *Journal of Machine Learning Research* 3 (2003), pp. 1083–1106.

[211] Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. "Relation Classification via Convolutional Deep Neural Network." In: *Proceedings of International Conference on Computational Linguistics*. COLING '14. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, 2014, pp. 2335–2344.

[212] Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. "Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks." In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portu-

gal: Association for Computational Linguistics, 2015, pp. 1753–1762.

[213] Lei Zhang, Michael Färber, and Achim Rettinger. "XKnowSearch!: Exploiting Knowledge Bases for Entity-based Cross-lingual Information Retrieval." In: *Proceedings of the ACM International on Conference on Information and Knowledge Management*. CIKM '16. Indianapolis, Indiana, USA: ACM, 2016, pp. 2425–2428.

[214] Yaoyun Zhang, Buzhou Tang, Min Jiang, Jingqi Wang, and Hua Xu. "Domain adaptation for semantic role labeling of clinical text." In: *Journal of the American Medical Informatics Association* 22.5 (2015), pp. 967–979.

[215] Zhehuan Zhao, Zhihao Yang, Hongfei Lin, Jian Wang, and Song Gao. "A Protein-protein Interaction Extraction Approach Based on Deep Neural Network." In: *International Journal of Data Mining and Bioinformatics* 15.2 (2016), pp. 145–164.

[216] GuoDong Zhou, Jian Su, Jie Zhang, and Min Zhang. "Exploring Various Knowledge in Relation Extraction." In: *Proceedings of the Annual Meeting on Association for Computational Linguistics*. ACL '05. Ann Arbor, Michigan: Association for Computational Linguistics, 2005, pp. 427–434.

[217] GuoDong Zhou, Min Zhang, DongHong Ji, and QiaoMing Zhu. "Hierarchical learning strategy in semantic relation extraction." In: *Information Processing & Management* 44.3 (2008), pp. 1008–1021.

[218] Guodong Zhou, Longhua Qian, and Jianxi Fan. "Tree kernel-based semantic relation extraction with rich syntactic and semantic information." In: *Information Sciences* 180.8 (2010), pp. 1313–1325.

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF DEFINITIONS

# ACRONYMS

| | |
|---|---|
| KB | Knowledge Base |
| UMLS | Unified Medical Language System |
| YAGO | Yet Another Great Ontology |
| MeSH | Medical Subject Heading |
| GO | Gene Ontology |
| NERD | Named Entity Recognition and Disambiguation |
| NLP | Natural Language Processing |
| DOM | Document Object Model |
| IE | Information Extraction |
| PPI | Protein-Protein Interactions |
| SPO | Subject-Predicate-Object |
| RDF | Resource Description Framework |
| IRI | International Resource Identifier |
| URL | Uniform Resource Locator |
| W3C | World Wide Web Consortium |
| LOD | Linked Open Data |
| NCBO | National Center for Biomedical Ontology |
| AI | Artificial Intelligence |
| POS | Parts of Speech |
| NERD | Named Entity Recognition and Disambiguation |
| BioNERD | Biomedical Named Entity Recognition and Disambiguation |
| FMA | Foundational Model of Anatomy |
| LSH | Locality Sensitive Hashing |
| NED | Named Entity Disambiguation |
| NER | Named Entity Recognition |

| | |
|---|---|
| QA | Question Answering |
| IR | Information Retrieval |
| KBC | Knowledge Base Construction |
| OIE | Open Information Extraction |
| DIPRE | Dual Iterative Pattern Relation Expansion |
| NELL | Never-Ending Language Learning system |
| NLM | United States National Library of Medicine |
| CVT | Compound Value Type |
| MID | Machine Identifiers |
| LSTM | Long Short Term Memory |
| CNN | Convolutional Neural Network |
| ACE | Automatic Content Extraction |
| SVM | Support Vector Machine |
| PLOS | Public Library of Science |
| TEES | Turku Event Extraction System |
| SRL | Semantic Role Labeling |