Cross-lingual Transfer of Semantic Role Labeling Models

Thesis for obtaining the title of

Doctor of Engineering

of the Faculty of Mathematics and Computer Science

of Saarland University

by

Mikhail Kozhevnikov, M.Sc.

Saarbrücken

November, 2016

Day of Colloquium	24 / 04 / 2017
Dean of the Faculty	Univ.-Prof. Dr. Frank-Olaf Schreyer
Chair of the Committee	Prof. Dr. Manfred Pinkal
Reporters	
First reviewer	Dr. Ivan Titov
Second reviewer	Prof. Dr. Vera Demberg
Academic Assistant	Dr. Daria Stepanova

Abstract

Semantic role labeling is an important step in natural language understanding, offering a formal representation of events and their participants as described in natural language, without requiring the event or the participants to be grounded. Extensive annotation efforts have enabled statistical models capable of accurately analyzing new text in several major languages. Unfortunately, manual annotation for this task is complex and requires training and calibration even for professional linguists, which makes the creation of manually annotated resources for new languages very costly. The process can be facilitated by leveraging existing resources for other languages using techniques such as cross-lingual transfer and annotation projection.

This work addresses the problem of improving semantic role labeling models or creating new ones using cross-lingual transfer methods. We investigate different approaches to adapt to the availability and nature of the existing target-language resources. Specifically, cross-lingual bootstrapping is considered for the case where some annotated data is available for the target language, but using an annotation scheme different from that of the source language. In the more common setup, where no annotated data is available for the target language, we investigate the use of direct model transfer, which requires no sentence-level parallel resources. Finally, for cases where the parallel resources are of limited size or poor quality, we propose a novel method, referred to as feature representation projection, combining the strengths of direct transfer and annotation projection.

Zusammenfassung

Rollensemantische Analyse ist ein wichtiger Teil der Computerlinguistik. Sie bietet eine formale Repräsentation von in natürlichen Sprache beschriebenen Ereignissen und deren Beteiligten, welche selbst nicht unbedingt formal beschrieben sein müssen. Umfangreiche Projekte für rollensemantische Annotation ermöglichen die Entwicklung präziser statistischer Modelle für die Analyse neuer Texte in verschiedenen Sprachen. Für andere Sprachen hingegen besteht leider immer noch Ressourcenmangel, hauptsächlich aufgrund eines Mangels an zeit- und kostenintensiven manuellen Annotationen der semantischen Rollen, die von professionellen Linguisten durchgeführt werden müssen und besonderes Training voraussetzen. Dieser Zeit- und Kostenaufwand kann verringert werden indem man mit Techniken wie Modelltransfer und Annotationsprojektion die bestehende Ressourcen auf andere Sprachen überträgt.

Diese Arbeit beschreibt eine Reihe von Experimenten zu cross-lingualem Transfer von rollensemantischen Analysatoren. Insbesondere betrachten wir drei Fälle:

- cross-linguales Bootstrapping, für den Fall, dass annotierte Ressourcen für Quell- sowie Zielsprache bestehen, diese jedoch unterschiedlichen Annotationsschemata folgen;
- direkten Modelltransfer für den Fall, dass weder annotierten Daten für die Zielsprache noch alinierte Korpora für das Sprachenpaar verfügbar sind;
- ein neuer Ansatz, der die Stärken des direkten Modelltransfers und der Annotationsprojektion kombiniert, für die Fälle, wo alinierte Ressourcen für den Quell- und Zielsprache vorhanden sind, jedoch beschränkt oder von schlechter Qualität.

Contents

Chapter 1

Introduction

Models of shallow semantics have proved useful in a variety of natural language processing tasks, offering an intermediate layer of representation that can help problems such as question answering, event extraction, machine translation and many others. While making high-quality open-domain models available for most languages would be highly desirable, it generally requires a very substantial amount of annotated training data due to the high level of lexical variation, and the construction of annotated resources can be costly.

1.1 Semantic Role Labeling

One of the most popular types of shallow semantic parsing, *semantic role labeling* (SRL) is concerned with representing the propositional aspects of meaning in natural language text. A unit of meaning is usually associated with a *predicate* evoked by a specific word or phrase and its *arguments*, which are said to have different *roles* with respect to the predicate (figure 1.1).

"[They]$_{Agent}$ **threatened** [its life]$_{Theme}$ [with a railway-share]$_{Means}$"

Figure 1.1: Example sentence annotated with semantic roles. The predicate is highlighted in bold and the argument spans are marked with corresponding roles.

Since there are generally many ways to describe the same event or situation (also referred to as *frame*), and the same word can sometimes describe different ones, the predicate needs to be linked to an appropriate inventory to indicate the meaning implied in a given statement, which in turn affects the expected set of participants and their roles. For instance, there would be separate entries in the inventory for the meaning of "threaten" in the example above – *express intention of committing some harmful action*, – and its other sense – *present potential danger*, – as in "the shares **threatened** to decline", where the former can have an a participant labeled AGENT[1], but the latter cannot.

Both the inventory of frames and the number, names and definitions of the roles vary between formalisms. We discuss the history behind various existing options and their respective peculiarities in chapter 2.

[1]We use easily interpretable labels similar to thematic roles for illustration here. This does not necessarily reflect how the sentence would be labeled following the guidelines of existing SRL resources.

Even from this simple example it is clear that the problem of automatically disambiguating the predicate and assigning semantic roles to the participants is highly non-trivial, for instance in this case the model needs to realize whether a given participant can exercise volition.

There is also empirical evidence that training reasonably accurate models for semantic role labeling requires larger vocabulary and larger training corpora than syntax-oriented tasks, and that these models tend to be highly domain- and register-specific.

Accurate semantic role labeling models would enable a certain degree of generalization over different verbalizations of the same meaning, and thereby facilitate language understanding. A considerable amount of work has been dedicated to designing accurate SRL models, but creating a model for a language with little or no annotated resources available remains a challenge. Making use of existing annotated resources for other, related languages can help accomplish this task with considerably less effort, and the ways of profiting from such resources will be the focus of this work.

1.2 Claims

This thesis addresses the problem of bootstrapping or transferring semantic role labeling models across language boundary.

Specifically, we state that the cost of creating a semantic role labeling model for a given language can in certain situations be reduced by leveraging existing resources for other languages. A workable model can even be created using cross-lingual transfer techniques alone, without dedicated target-language annotation, but the exact approach to adopt depends on many factors, in particular the availability of annotated resources for the target language – both syntactic and semantic, – as well as the amount and quality of sentence-level parallel data available for the language pairs in question.

We consider how the roles of aligned arguments in parallel data correspond to each other, if the annotation schemes used for source and target language are different, and show that target-language model can be improved by transferring information from the source-language model through the alignment links. The gains are much higher if prior information on role correspondence for aligned arguments is available.

We further identify strengths and weaknesses of existing cross-lingual transfer techniques by evaluating them on multiple language pairs and propose a new, competitive and flexible approach for this task.

1.3 Relevance

With the wide-spread adoption of natural language processing techniques by the industry, the need for high-quality language understanding tools across languages becomes a matter of high practical importance. Recent efforts in part-of-speech tagging, dependency parsing, named entity recognition and other areas made accurate models available for more languages and attempted to harmonize the output of those models to enable easier internationalization for downstream tasks. We believe that this process will continue, gradually covering various aspects of semantics, from predicate-argument structure to discourse semantics and more abstract formal semantic representations, and would call for further improvements in cross-lingual transfer methods.

The models obtained by cross-lingual bootstrapping or transfer are generally imperfect, but tend to have comparatively broad coverage (assuming both the source-language training corpus and the parallel corpus used for transfer are sufficiently large and diverse) and can serve as a useful starting point in model development. A transferred model can be further improved by using manual correction or employing non-expert annotators to discriminate between correct and incorrect annotations – a generally less demanding task than annotating a sentence from scratch.

Besides the practical benefit of reduced cost of creating a model for a new language, cross-lingual transfer techniques encourage cross-lingual consistency – a desirable property for semantic annotation – and may help design better formalisms.

1.4 Overview

The thesis is structured as follows. We begin with an overview of semantic role labeling research in chapter 2, particularly focusing on the cross-lingual aspects.

Chapter 3, based on Kozhevnikov and Titov (2013b), addresses the question of correspondence between the roles of aligned arguments in a parallel corpus with heterogeneous annotations used for source and target languages.

We show that the correspondence is non-trivial in general, but a usable model thereof can be learned from automatically annotated parallel data, despite the domain shift between annotated corpora and parallel data and correspondingly imperfect annotations on the parallel sentences. The role correspondence model can in turn be used to improve target-language annotations through the transfer of information from a higher-quality source-language model using a *cross-lingual bootstrapping* procedure. We further demonstrate that much more significant improvements could be obtained with a role correspondence model based on prior, external knowledge.

Next, chapter 4, based on Kozhevnikov and Titov (2013a), describes the application of an approach known as *direct model transfer* to the problem of semantic role labeling, which enables the construction of an SRL model for a given language based on an existing model for another language in the absence of sizable word-aligned parallel corpora. We perform a comprehensive evaluation on several language pairs, comparing direct model transfer to a simple annotation projection baseline and a state-of-the-art unsupervised SRL model, and identify scenarios where one or the other is preferable.

Chapter 5 presents an alternative, novel approach to the transfer of semantic role labeling models, introduced in Kozhevnikov and Titov (2014). Based on the representation learning paradigm, the proposed approach combines the strengths of direct model transfer and annotation projection – another well-known cross-lingual transfer technique – to enable efficient utilization of parallel resources, even those of smaller size or questionable quality. The method demonstrates competitive performance compared to direct model transfer and annotation projection baselines and we further elaborate on how it can be adjusted for a particular transfer scenario.

Chapter 6 describes our related work in the direction of leveraging agreement on unlabeled samples for structured prediction in monolingual semi-supervised learning setup, and chapter 7 offers some conclusions and parting thoughts.

Chapter 2

Semantic Role Labeling in Cross-Lingual Setting

2.1 A Brief History of Semantic Roles

Semantic Role Labeling (SRL) is among the most established approaches to shallow semantic parsing, aiming to provide a more abstract representation of predicate structure compared to syntactic analysis, but not attempting to entirely decouple the meaning from the words. It is often said to describe "Who did what to whom", which reflects the focus on events and their participants.

Such representations have proven useful in many natural language processing tasks, including question answering (Narayanan and Harabagiu, 2004; Shen and Lapata, 2007; Kaisser and Webber, 2007), textual entailment (Sammons et al., 2009), machine translation (Boas, 2002; Wu and Fung, 2009; Liu and Gildea, 2010; Gao and Vogel, 2011), summarization (Trandabăţ, 2011) and dialogue systems (van der Plas et al., 2009).

From the point of natural language processing, the primary purpose of SRL is to generalize over various alternative verbalizations of the same event, thus simplifying subsequent analysis. Of course, there are multiple possible levels of granularity and different phenomena to consider. Some only include verbal predicates, for instance, while others handle nominal and adjectival ones as well, and even predicates expressed by prepositional phrases. Other important choices are whether the role fillers should be single words, constituents, spans or some other units, how a predicate type is defined and whether roles are defined per individual predicate, group of predicates or globally, for all predicates.

Here are a few examples of sentences that, in appropriate context (i.e. assuming A and B are both companies), would indicate to a reader that B was acquired by A:

1. Last year, A acquired B.
2. B was bought by A.
3. A bought B for $ 1.7 billion.
4. A reported their acquisition of B.
5. A sells the recently acquired B again.

6. A paid $ 1.7 billion for B.

Here, (3) and (2) illustrate a simple active-passive alternation and can be considered semantically equivalent, if we disregard the extra information such as the time of the deal or the price paid. The same meaning can be expressed by a different verb (1), or using nominal (4) or adjectival (5) predicate.

On the other hand, (6) also suggests that a transaction took place, but since there is no explicit predicate to indicate an acquisition, this inference is generally considered beyond the scope of shallow semantics – the purchase of something would be considered a separate event from the transfer of money in return for the purchase, although some formalisms may incorporate information about the typical order of events or their composition. The interpretations of statements such as "A offered to buy B" or "B may soon be acquired by A" would also be considered to contain the predicate $buy(A, B)$ as a part of the structure. Modality, negation, factuality and other extra-propositional aspects of meaning are generally considered to be outside of the scope of this task, although in some cases modifiers that indicate such aspects are labeled as parts of predicate-argument structure.

2.1.1 Thematic Roles

Concepts similar to thematic roles, such as Θ-roles, deep cases and others, have been considered by many researchers and the motivations for their existence are varied. Most often they are viewed as a way to sub-categorize phrases in a grammar or describe linking between formal-semantic representation of a predicate and the syntactic structure of the corresponding sentence. Depending on the purpose and perspective of the particular researcher, the number and definition of roles have changed.

The general intuition is that most events described by a predicate involve one or more *participants*, usually corresponding to phrases governed by the predicate word or phrase, and that the participants can be categorized according to their relation to the event into groups known as thematic roles. In the following sentence, for example, "bashed" is the predicate with three arguments, which are marked with their relation to the predicate, their *thematic role*.

$$[\text{Someone}]_{Agent}\ \underline{\text{bashed}}\ [\text{him}]_{Theme}\ [\text{with a cudgel}]_{Instrument}.$$

Thematic roles are closely connected with grammatical functions, such as Subject, Direct Object or Indirect Object, but are semantic in nature and must to be defined primarily on the basis of what the predicate entails about a given participant (Dowty, 1991).

The introduction of this concept is usually attributed to Fillmore (1968) and Gruber (1965), although some researchers track similar concepts to as far back as 4th century BCE (Cardona, 1976).

Fillmore (1968) divides arguments into the following six categories, referred to as *deep cases*, and acknowledges, that "Additional cases will surely be needed."

Agentive: the case of the typically animate perceived instigator of the action identified by the verb.

Instrumental: the case of the inanimate force or object causally involved in the action or state identified by the verb.

Dative: the case of the animate being affected by the state or action identified by the verb.

Factitive: the case of the object or being resulting from the action or state identified by the verb, or understood as a part of the meaning of the verb.

Locative: the case which identifies the location or spatial orientation of the state or action identified by the verb.

Objective: the semantically most neutral case, the case of anything representable by a noun whose role in the action or state identified by the verb is identified by the semantic interpretation of the verb itself; conceivably the concept should be limited to things which are affected by the action or state identified by the verb.

Multiple extensions of this list, reinterpretations of certain roles and alternative classifications have been proposed since. The set of thematic relations suggested by Jackendoff (1972), for example, contains ten categories: Agent, Theme, Location, Source, Goal, Experiencer, Percept, Patient, Instrument and Benefactive. Modern resources using concepts similar to thematic roles, such as Extended VerbNet (Kipper et al., 2006b), contain as many as 23 different roles. Following are some points of argument:

Core roles and Modifiers. Andrews (1985) suggests that there are two different categories of thematic roles: participatory, which are central to an event, and circumstantial, which "form part of the setting of the event". The first category includes Agent, Patient, Instrument, Goal, and similar, while the second may contain a potentially infinite set of secondary relations, such as Location or Time of the event. Many formalisms make this distinction, and sometimes secondary roles are ignored completely. In certain grammar-related applications, only arguments directly governed by the verb are considered role-bearing, while prepositional phrases are not.

Fragmentation Cruse (1973) suggests a subdivision of the AGENT role into four different roles to distinguish, for example, an action performed in person and one delegated to a third party. It also appears hard to settle with a single set of spatial and temporal relations.

The fact that so many different definitions of thematic roles have been proposed suggests that there is no clear way to decide which one is correct outside of the context of a particular application. Blake (1930) claims that the roles themselves are "numerous, but not infinite", while others assume that any fixed set would only serve as an approximation. According to Dowty (1991), for example, "total indexing of verbal arguments by thematic role type is almost certainly empirically impossible."

While this argument is of importance in the context of theoretical linguistics, it is, perhaps, less critical from the computational perspective. Here thematic/semantic roles are generally used as a yet another layer of abstraction in applications, a source of information about the content of the text considered. The most important things about such a structure are, therefore, that (a) it is useful for the task at hand and (b) one can derive said structure for a given sentence with reasonable accuracy.

Definition	
During a court appearance, a Sentence, generally a punishment, is imposed on the Convict by a Court, usually represented by a judge.	
List of core roles	
Convict	The Convict is given a Sentence by the Court.
Court	The Court imposes a Sentence on a Convict.
Offense	The illegal action of which the Convict has been found guilty.
Sentence	The Sentence is imposed on the Convict by the Court.
Term_of_Sentence	This frame element denotes the duration of the sentence.
List of non-core roles	
Depictive	The state that the Convict is in at the time of their sentencing.
Place	The location of the sentencing Court.
Time	The Time when the Convict is sentenced.
Type	This frame element denotes the type of sentence.

One or more sentences exemplifying the frame

Frame:Sentence

Court		Convict	Sentence	Offense
[The judge]	sentenced	[Martha Stewart]	[to 2 years in prison]	[for insider trading].

List of *lexical units* evoking this frame
condemn.v, order.v, send up.v, sentence.n, sentence.v

Figure 2.1: FrameNet frame example

Uniqueness of Roles Most agree, that each argument should, ideally, be assigned one and only one role with respect to a given predicate. Chomsky (1981) even uses this as a part of their definition, requiring that each argument of a predicate in the lexicon be assigned a unique θ-role (Θ-criterion). In practice, however, there are examples where this principle leads to counter-intuitive results, in particular where predicate indicates a symmetrical relationship (Dowty, 1991). Whether or not to enforce role uniqueness therefore depends on the inventory of roles used and the task they are applied to.

Event-specific Role Definitions One way to avoid (at least some of) the granularity issues and role ambiguity is to define possible participants for each event type separately. This ensures that the definitions of roles are specific and unambiguous. On the other hand, this also requires us to enumerate every possible event type for a given domain and describe its possible participants, which may constitute a considerable amount of work. And there remains a chance of some potential participants being overlooked.

2.1.2 Frame Semantics

FrameNet (Baker et al., 1998) adopts an intuitive definition of event type. The purpose of the project is to come up with a set of possible events *frames* and describe the participants of each such event, as well as surface forms that may indicate the presence of this event, *evoke the frame*.

The description of frame Sentence, for instance, is presented in figure 2.1.

It may also contain additional information regarding the interaction of this frame with others – the frames are organized in a network with edges indicating various types of interactions between frames. In case of SENTENCE, it is indicated that this is a *subframe* of CRIMINAL_PROCESS and that it is generally preceded by a TRIAL.

To date, this is, perhaps, the richest type of semantic role labeling annotation, for which sizeable annotated resources are available. Were such annotations easily obtainable for new text, this would greatly simplify a large number of tasks. For example, mining a text for *buying* events would[1] more or less boil down to finding all occurrences of the frame COMMERCE_BUY and then extracting the corresponding arguments, such as BUYER, GOODS, SELLER or PRICE. Unfortunately, correctly identifying the frames and participants is not an easy task. Moreover, the existing frame hierarchy does not yet have sufficient coverage (Palmer and Sporleder, 2010) and expansion of this hierarchy is rather costly, primarily due to specialized training required for annotators.

Extending FrameNet

There is one respect in which FrameNet is different from other, more data-driven approaches. The procedure here is to start out with an idea of a frame that is not in the hierarchy thus far, describe lexical units that may evoke the frame, list possible participants and link it to other frames in the hierarchy, then collect and annotate examples by searching a corpus for occurrences of the given frame[2].

This last part can plausibly be crowd-sourced or partly automated and there are several semi-supervised techniques that seek to extend the coverage of frame-semantic parsing models by explicitly finding new lexical items for existing frames using graph-based methods (Fürstenau and Lapata, 2009; Das et al., 2013) or by relying on word representations (Collobert and Weston, 2008; Deschacht and Moens, 2009). Creating new frames is much trickier and seems unlikely to be fully automated anytime soon, although unsupervised techniques may help discover likely groups of predicates.

2.1.3 PropBank

Another prominent SRL resource, PropBank (Palmer et al., 2005a), chooses a different approach. Here each verb is associated with one or more *framesets*. A frameset corresponds to a coarse-grained sense of a verb, meaning that "verb meanings are distinguished as different framesets if they have distinct subcategorization frames" (Babko-Malaya, 2005). For example, verb senses can allow a different number of arguments, or their arguments can have significantly different thematic roles.

A single frameset may look as follows:

Roleset id:	bash.01	to hit, with words or a bat (etc)
Roles:	Arg0:	criticizer, hitter
	Arg1:	entity criticized/hit
Example:		
Bashing [the D.C. government]$_{Arg1}$ is risk-free for [members of Congress]$_{Arg0}$.		

[1] Assuming we also had a working co-reference model and an entity linking model.

[2] A full-text corpus has also been annotated to enable statistical modeling for frame-semantic parsing, as the distribution of frames and frame elements in the annotated example instances is very different from that in continuous text.

Note that the two intuitively distinct senses of "bash" – to hit physically and to criticize – are assigned to the same frameset, as the distinction does not affect the number or roles of the arguments.

The vast majority of verbs in PropBank are associated with only one frameset, but certain distinctions are observed, e.g. frameset RUN.01 indicates running a machine or an organization, whereas RUN.02 corresponds to running as a physical activity.

Light verbs are generally the ones associated with the most framesets. The verb *take*, for example, is associated with 28 different framesets, of which 20 are present in annotated data (the version we consider is the CoNLL 2009 Shared Task corpus) and only 9 are observed ten or more times in that corpus, while the most frequent sense covers over 70% of all occurrences, and such skewed distributions are apparently quite typical.

PropBank distinguishes between *core roles* and *modifiers*. The core roles within a frameset are numbered (e.g. AGR0 and ARG1 above) and modifiers are labeled separately. These are used to label secondary attributes of an event, such as time, place, cause, purpose, location or extent. Interestingly, depending on the predicate, any one of these may become a core role. For example, in STAY.04 (meaning "to remain in a location"), the location argument is promoted to ARG2. It can therefore be said, that modifiers indicate a particular thematic role of an argument, while the mapping of core roles into thematic roles depends on the predicate (frameset) in question. A thematic role, in turn, cannot be associated with a PropBank-style semantic role outside of a context of the predicate, as it may become a core role or a modifier. In some cases, there may even be both a core argument and a modifier bearing the same thematic role in one predicate (Babko-Malaya, 2005):

She put the slipper neatly [by its mate]$_{Arg2\text{-}LOC}$ [at the foot of the bed]$_{ArgM\text{-}LOC}$

The framesets are defined based on examples. An annotator starts by collecting instances of a particular verb from a corpus, then partitioning these instances into framesets. This implies a substantial chance of certain roles being overlooked, if they are not realized in any of the examples, so a predicate could potentially have roles that are not indicated in the corresponding frameset.

As an example, let us consider predicate BASH.01 again. Note that the frameset only indicates two core roles, ARG0 and ARG1, but it is clear that in the following sentence [with a cudgel] does represent an argument and that it should be assigned ARG2, as in related framesets, such as HIT.01.

[Someone]$_{Arg0}$ bashed [him]$_{Arg1}$ [with a cudgel]$_{Arg2}$.

The roles are assigned in order of their prominence on a per-verb basis, starting with ARG0 if the verb in question allows a proto-agent and with ARG1 otherwise. The first two, ARG0 and ARG1, correspond to Proto-Agent and Proto-Patient in the terminology of (Dowty, 1991), the other roles have no consistent interpretation across different predicates, except where closely related verbs are concerned (e.g. HIT.01 and BASH.01 mentioned above or PULL.01 and TUG.01 in the "cause motion" sense).

It is also evident that the inventory of framesets is necessarily incomplete. Assigning numbered role labels to the arguments of predicates not included in the inventory could still be potentially useful, but they would no longer have a clear interpretation.

The Computational Perspective

Even though ARG2-ARG5 mean different things in the context of different predicates, in the applications they are usually predicted by a single model. It may take into account the relevant predicate sense in some fashion, but building a separate classifier for every single frameset would be impractical. Besides, the roles are highly correlated with grammatical functions of the arguments and ARG3-ARG5 cover only about 2.5% of all argument occurrences (see Figure 2.2) in PropBank and therefore their contribution to the overall performance metric is limited.

The PropBank-style semantic roles have been found useful in a variety of applications, from question answering (Kaisser and Webber, 2007; Surdeanu et al., 2011) and information extraction (Surdeanu et al., 2003; Christensen et al., 2010) to machine translation (Wu and Fung, 2009; Liu and Gildea, 2010; Gao and Vogel, 2011) and summarization (Melli et al., 2005).

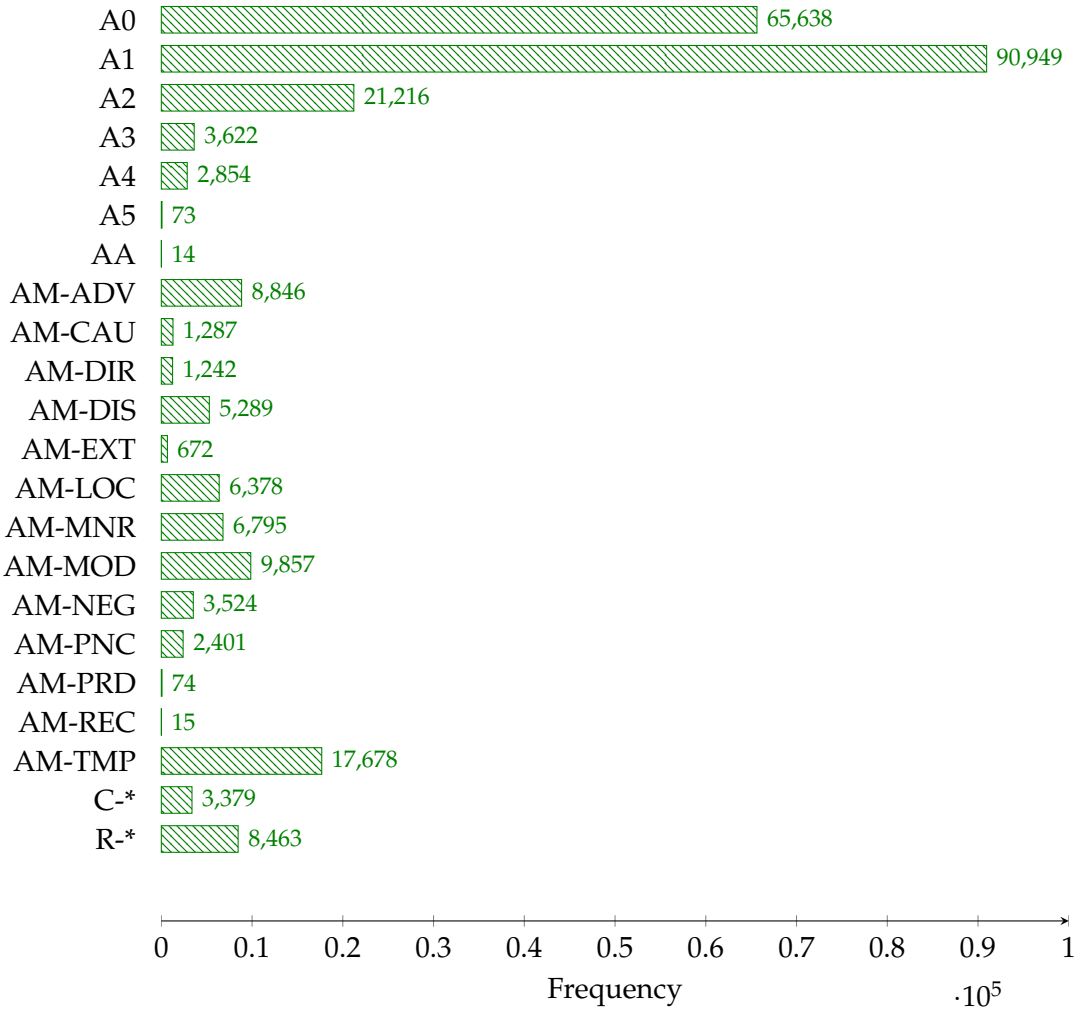


Figure 2.2: Role frequencies in the CoNLL'09 version of PropBank.

give-13.1

members	deal, lend, loan, pass, peddle, refund, render
frame	NP **V** NP PP.recipient
example	"They **lent** a bicycle to me."
syntax	Agent **V** Theme to Recipient
frame	NP **V** NP-Dative NP
example	"They **lent** me a bicycle."
syntax	Agent **V** Recipient Theme
frame	NP **V** NP
example	"I **leased** my house (to somebody)."
syntax	Agent **V** Theme
frame	NP **V** PP.recipient
example	"The bank **lent** to fewer customers."
syntax	Agent **V** to Recipient

Figure 2.3: VerbNet class give-13.1, abridged.

2.1.4 VerbNet and SemLink

VerbNet (Kipper et al., 2006b) is not an SRL corpus as such, but it is closely related to the topic of verbal SRL. This resource describes different syntactic patterns in which a particular class of verbs can be used and indicate the thematic roles of different arguments in each pattern (see figure 2.3).

It is said that "PropBank provides the best training data, VerbNet provides the clearest links between syntax and semantics and FrameNet provides the richest semantics."[3], which indicates the necessity of linking the resources together. The project to establish the mappings is known as SemLink[4] and already has some coverage. Several studies have compared the informativity of the PropBank semantic roles and the VerbNet thematic roles, their robustness and ability to generalize to unseen predicates (Merlo and Van Der Plas, 2009; Zapirain et al., 2008; Yi et al., 2007; Loper et al., 2007). The consensus seems to be that both approaches have their merits. VerbNet roles are more fine-grained and more informative, but also harder to learn and it is suggested that the appropriate ones should be chosen depending on the task at hand. In certain cases the two can be profitably combined (Yi et al., 2007).

2.1.5 Prague Dependency Treebank

Prague Dependency Treebank (PDT) stands apart from other syntactic-semantic resources in that it relies on a somewhat different grammatical formalism. The annotation includes three layers: morphemic, analytical and tectogrammatical. Morphemic annotation roughly corresponds to lemmatization, morphological tagging and part-of-speech tagging. Analytical layer describes surface syntactic structure in a fashion similar to syntactic dependency trees. Tectogrammatical layer includes information about semantic roles, discourse relations, grammatical functions and other phenomena, expressed in a unified

[3]quote from the Semantic Role Labeling Tutorial by Martha Palmer, Shumin Wu and Ivan Titov at NAACL 2013.

[4]`http://verbs.colorado.edu/semlink/`

fashion using *functors*. It is also defined on a specific type of dependency graph, the units of which, however, are not necessarily the input tokens: some nodes are *virtual*, having no corresponding tokens at all, and some of the surface tokens may not be included as a node in the tectogrammatical dependency graph, if they do not carry semantic content of their own. For instance, most prepositions are collapsed, similar to collapsed Stanford dependency structures.

Most semantic roles used in PDT are predicate-independent and rather similar to the VerbNet roles: agent, patient, result, benefactor, origin, cause, means, manner, extent, etc. The dataset used for Czech in CoNLL 2009 Shared Task is, in fact, a conversion of one of the early versions of PDT.

2.1.6 Span-, Constituent- and Dependency-based Formalisms

The notion of predicates and arguments may seem intuitively obvious, but there are considerable variations between different resources. FrameNet allows both words and phrases as lexical units, so VEHICLE_LANDING frame, for instance, can be evoked by [land], [touch down] or [set down]. Every argument (or *Frame Element*) can also be represented by one or more arbitrary spans, provided the spans for different frame elements do not intersect. This makes the annotation flexible, but also makes the problem harder, both for an automatic semantic role labeler and for a human annotator, as argument boundary can sometimes be subjective.

Of the more syntax-bound resources, some (e.g. VerbNet, PropBank, Chinese TreeBank) define an argument as one or more constituents in the phrase structure tree of the sentence. Others (e.g. AnCora, Prague TreeBank) rely on dependency structures instead and associate arguments with subtrees in such structures, as represented by their syntactic heads. The latter is especially suitable where languages with high degree of non-projectivity, such as Czech or German, are concerned, as their syntactic structure is harder to adequately represent by a phrase structure tree.

2.1.7 Less well-known aspects of SRL

Although not explicitly defined that way, SRL remains mostly verb-centric. There are considerable advances in the handling of nominal predicates (Meyers et al., 2004; Padó et al., 2008), but other predicate types remain largely unexplored, although there are indications that jointly modeling different ways of expressing relations in text can be helpful (Srikumar and Roth, 2011).

Other less well-studied questions include *implicit* arguments – i.e. arguments not included in the immediate lexical context of the predicate, but rather found at longer distances in the text (Gerber and Chai, 2012). This appears to be most common for nominal predicates (Gerber et al., 2009), although verbal ones can also be considered to have implicit arguments in certain cases (Roth and Frank, 2016). Other researchers consider implicit arguments to be outside of the scope of SRL and propose to treat it as a special case of anaphora resolution instead (Silberer and Frank, 2012).

Historically, PropBank treated these as external to the task of semantic role labeling, whereas in FrameNet they were considered to be participants, albeit non-realized, and further divided into three subtypes, depending on whether a given frame element is omitted altogether or mentioned elsewhere and whether or not the omission is motivated

by the grammar (Fillmore et al., 2003). The automatic identification and labeling of such arguments is a complex problem and even the agreement between (non-expert) human annotators is limited (Feizabadi and Padó, 2014).

2.2 Supervised SRL

Since the advent of computational linguistics, researchers viewed various representations of language structure and meaning not just as formal ways of describing a phenomenon, but also as a way to represent natural language in a usable form, particularly a form suitable for further processing on modern computers.

Due to the inherent fuzziness of the task definition, it is hard to pinpoint the first automatic analyzer. Most refer to Gildea and Jurafsky (2002) when introducing the concept, but it is clear that certain domain-specific analyzers (Miller et al., 1996) and more traditional grammar-based analyzers involving some notion of semantic roles (Pollard and Sag, 1994) were available earlier.

Gildea and Jurafsky (2002) trained their system on the English FrameNet to identify the constituents bearing a semantic role and assigning the correct role to them, achieving 65% precision and 61% recall on this task. A similar system described in Gildea and Palmer (2002), which stresses the importance of accurate syntactic analysis, was among the first PropBank-based SRL analyzers, and Gildea and Hockenmaier (2003) introduced a closely related CCG-based system.

The subsequent works attempting to improve semantic role labeling performance using various alternative sources of information, learning algorithms, novel features, global constraints, reranking or joint learning are too many to list exhaustively, so we will confine ourselves to listing only the most (subjectively) prominent points.

The problem of semantic role labeling attracted some considerable attention, as indicated by the fact that different versions of SRL were featured in the CoNLL shared tasks of

2004: automatically identifying arguments and assigning semantic roles in PropBank (Carreras and Màrquez, 2004)

2005: same as 2004, but also involving automatic predicate identification and disambiguation (Carreras and Màrquez, 2005)

2008: dependency-based SRL, performed jointly with syntactic parsing (Surdeanu et al., 2008)

2009: same as 2008, but for multiple languages (Hajič et al., 2009)

There were also related shared tasks in Senseval-3 (Kwon et al., 2004) and SemEval 2007 (Baker et al., 2007).

The analysis of a sentence is generally performed in two stages, one involving only the predicate (identifying predicates, or, more generally, parts of the sentence evoking a frame, and figuring out what frame they are evoking) and the other handling arguments (identifying role-bearing parts of the sentence with respect to a given predicate and identifying their roles). These are often simply pipelined, but the latter stage is known to provide useful information for the former (Toutanova et al., 2008), so there are also beam

search- and reranking-based approaches attempting to identify better overall solutions, as well as models that perform both stages jointly.

The feature sets used depend on the particular flavor of SRL performed (constituent- vs. dependency- vs. span-based), but various patterns have been found useful over the years. The features proposed by Gildea and Jurafsky (2002) are still used in most applications in some form or another. For dependency-based SRL, which we will largely focus on in this work, these typically include properties of the predicate and of the argument phrase, especially its syntactic head, as well as the path in the dependency tree

- Dependency path between the argument head and the predicate
  - Sequence of dependency relations
  - Tokens on the dependency path and/or their properties, such as POS tags
  - Prefixes, suffixes or individual edges in the path
- The argument phrase, its syntactic head and properties thereof
- Local dependency- or linear context around the argument
- Various properties of the predicate, in particular the voice and subcategorization pattern
- Global (sentence-level) features, typically involving other arguments and larger bits of the dependency tree

In subsequent work, many other sources of information have been suggested, including selectional preference information (Zapirain et al., 2010), named entities (Surdeanu et al., 2003), word sense tagging (Che and Liu, 2010) and morphological features, where applicable.

Constraints have also been found beneficial in identifying arguments and labeling semantic roles (Roth and Yih, 2005), in particular role uniqueness constraints – in PropBank, multiple arguments typically cannot be labeled with the same core role – as well as predicate-specific constraints, which prevent certain roles from being assigned to the arguments of certain predicates, as they are indicated as not applicable in the PropBank description of the corresponding frameset.

Several different types of statistical models have been applied to this task, with the primary focus on capturing progressively more advanced features and interdependencies between individual predicted roles.

Although most approaches to date rely on a pre-specified and often complex feature set, various efforts have been made to simplify or entirely forego the feature engineering step, notably using tree kernels (Moschitti et al., 2008), tensors (Lei et al., 2015) or neural embeddings (FitzGerald et al., 2015; Roth and Lapata, 2016). A number of neural-network-based approaches to SRL have also been proposed, from early work by Henderson et al. (2008) to recent models based on recurrent neural networks (Zhou and Xu, 2015; Swayamdipta et al., 2016; Shi et al., 2016).

Interestingly, ever since Gildea and Hockenmaier (2003) ported PropBank semantic roles[5] to the CCGbank (Hockenmaier and Steedman, 2007), CCG[6]-based semantic role

[5]The projection of roles has later revisited by Boxwell and White (2008).

[6]Combinatory Categorial Grammar (Steedman, 2000).

labeling has existed as a mostly independent field of research (Gildea and Hockenmaier, 2003; Boxwell et al., 2009; Lewis et al., 2015).

2.3 SRL in Low-resource Setting

Automatic assignment of various kinds of semantic roles has been studied for an extended period of time, but most of the early work focused on the supervised setting. As a result, there is a number of models that perform very well on the test section of the relevant corpus, but prove somewhat less accurate on data belonging to a different domain or register (Carreras and Màrquez, 2005; Pradhan et al., 2008).

More recent efforts seem to concentrate on several directions: some focus on interactions between syntax and semantics and joint modeling thereof (Toutanova et al., 2008), others address the question of linking together various SRL resources and other types of lexical semantic resources (Palmer, 2009), or consider the construction of corpora and models for new domains and new languages or refinement of existing models by using various semi-supervised learning techniques.

It appears that due to high degree of lexical variability, semantic role labeling models require considerable amounts of data for training. Existing corpora for some languages contain up to 50 thousand annotated sentences, and even then their coverage is questionable, as corpora are usually domain-specific.

Automatic or semi-automatic construction of models for new languages is of particular interest, as annotators possessed of the necessary skills and fluent in a particular language are hard to find.

This can be addressed in a number of ways. First, one has to decide whether or not to commit to an annotation effort. The costs of such an effort are considerable, regardless of the desired size of the corpus, due to initial cost of creating a new annotation scheme (or adapting an existing one), preparing corresponding annotation guidelines, hiring and training annotators and ensuring sufficient inter-annotator agreement.

Various strategies such as active learning (Roth and Small, 2006; Settles, 2010; Chen et al., 2011), automatic labeling with manual correction (Basile et al., 2012) or crowdsourcing (Fossati et al., 2013; Feizabadi and Padó, 2014; He et al., 2015), can be employed to reduce the cost of actual annotation, but all in all this takes a significant amount of time and funds to accomplish. Whether or not such an investment is justified depends primarily on the intended purpose of the resulting resource, which, in most cases, is to have a model that can automatically assign semantic roles to inputs from a certain distribution with a certain target accuracy, which could in turn be used by a downstream application. Depending on the nature of said application, the specific kind of annotation may or may not be of importance.

If all that is required is some useful notion of structure that can be used to derive features for the task at hand, one may be able to use unsupervised methods and bypass the whole annotation business. Such approaches generally do not assign specific role labels to arguments, but rather cluster the arguments whose roles appear to be similar together, which may be sufficient for certain tasks (Titov and Klementiev, 2011).

Sometimes clustering is not enough, though, in particular if the model to be built on top of the semantic role structure is a rule-based one that expects specific role labels, as might be the case with information extraction (Surdeanu et al., 2003) or

summarization (Melli et al., 2005), for instance.

In such cases the alternative to annotating a new training corpus manually is either to resort to rule-based methods, which, to our knowledge, have not been applied with much success to the task of semantic role labeling, or to leverage existing corpora for other languages.

2.3.1 Unsupervised Approaches

Under certain conditions a sensible SRL-like structure can be induced without using any supervised training at all. Most approaches here start with the assumption that an argument filling a particular semantic role tends to have a specific syntactic function relative to the predicate as well. For instance, a subject of a transitive verb would often have agent-like semantics, and although there is no one-to-one mapping between syntax and semantics that would work for all cases, it can at least constitute a reasonable baseline.

Unsupervised semantic role labeling approaches go further, adding in some amount of lexical information (about both predicate and argument), linear order information and some key characteristics of the predicate, such as the voice, yielding a richer representation of an argument, sometimes referred to as an *argument key*. The predicates and arguments are generally assumed to be given, or to be identifiable with reasonable accuracy be means of heuristics (Lang and Lapata, 2010), although Abend et al. (2009) investigate the possibility of unsupervised argument identification and further attempt to classify these into core and adjunct roles in Abend and Rappoport (2010). The task of an unsupervised semantic role labeler is then to group argument keys coming from different predicate instances such that two keys would be in the same group if and only if they bear the same semantic role.

As with other unsupervised approaches, the evaluation here is not entirely trivial. The usual way of assessing the performance of an unsupervised SRL system is to match the clusters against the gold labeling using metrics known as *purity* (Pu) and *collocation* (Co):

$$Pu = \frac{1}{N} \sum_i \max_j |G_j \cap C_i|$$

$$Co = \frac{1}{N} \sum_j \max_i |G_j \cap C_i|,$$

where C_i is the set of arguments in the i-th induced cluster, G_j is the set of arguments in the j-th gold cluster and N is the total number of arguments. There is a trade-off between the two, akin to precision/recall trade-off in classification problems, in the sense that one can achieve perfect purity by assigning every observed instance to a cluster of its own, or achieve perfect collocation by putting all available instances into a single cluster. Similarly to precision and recall, the two measures are combined using the harmonic mean, which is referred to as *clustering* F_1 (F_1^C) or sometimes simply F_1.

Computing purity, for instance, can be thought of as associating each cluster with the gold role that shares most instances with this cluster, and then computing regular precision given this assignment. Importantly, depending on the task formulation, this "optimistic" assignment may be chosen once or picked anew for each predicate, as is

often done for PropBank-style unsupervised SRL on the basis that the role interpretations are not guaranteed to be consistent across all predicates.

Among the first to attempt unsupervised semantic role labeling were Swier and Stevenson (2004) and Grenager and Manning (2006), both of which propose explicit probabilistic models of syntactic-semantic linking, the former using VerbNet roles as reference and the latter – PropBank.

Further work in this direction includes Bayesian methods with sophisticated priors (Titov and Klementiev, 2012a), graph partitioning methods (Lang and Lapata, 2011), as well as methods leveraging parallel data (Titov and Klementiev, 2012b) or small amounts of labeled data (Titov and Klementiev, 2012c) to guide unsupervised learning. Lorenzo and Cerisara (2012) improves upon the results presented in Titov and Klementiev (2012a) with another Bayesian model, possibly due to explicitly modeling syntactic verb classes.

Garg and Henderson (2012) design a model that accounts for typical role order, and Garg and Henderson (2016) extend that model by also using aligned parallel data, which in this case yields no significant improvements, or small amounts of labeled data, which helps considerably.

Following recent trends, Woodsend and Lapata (2015) and Luan et al. (2016) cast unsupervised SRL as a problem of argument embedding, while Titov and Khoddam (2015) propose a framework based on autoencoders.

For FrameNet, Modi et al. (2012) design a system based on that of Titov and Klementiev (2012a), but inducing frames and roles jointly. O'Connor (2013) tackles a similar problem with a latent-variable generative model.

2.3.2 Semi-supervised Approaches

He and Gildea (2006) were among the first to approach the questions of semi-supervised learning for semantic role labeling, though their conclusions are less than inspiring: initial experiments demonstrated that the type of co-training in question yielded no significant improvement and a simpler self-training approach actually proved harmful.

Further studies showed more promise, however (Zadeh Kaljahi, 2010). Probably the most prominent is that of Fürstenau and Lapata (2009), which proposes a method of heuristically generating surrogate training examples by selecting instances lexically and syntactically similar to those present in the training data. This approach results in a sizable improvement for frames that have very few annotated examples, which is not unrealistic for FrameNet.

Others attempt to improve the performance of SRL models using co-training (Lee et al., 2007; Kaljahi and Baba, 2011) or extend coverage by using word embeddings learned from larger unlabeled corpora (Deschacht and Moens, 2009).

It is also important to mention Collobert et al. (2011), who propose a unified neural architecture for multiple NLP tasks, including POS-tagging, chunking, named entity recognition and SRL, which achieves near state-of-the-art performance without additional feature engineering thanks to multi-task learning.

2.3.3 Annotation Projection

While creating a large manually annotated resource for every new language appears to be prohibitively expensive, one could reuse an existing corpus in a related language as a source of supervision.

One way of leveraging such information is known as annotation projection. The general idea is to align units of text in the two languages (individual tokens, sentences or larger fragments), where the source side is already annotated with the desired information (manually or automatically), and infer the target-side annotation by *projecting* it through the alignment.

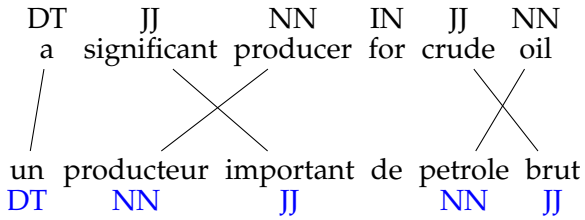


Figure 2.4: Part-of-speech tag projection (Yarowsky et al., 2001). The tags highlighted in blue are obtained by projection.

Depending on the task and the language pair, the degree of similarity between the annotations for a sentence pair may vary, even if the translations are kept as close to the original sentence as possible. The level of agreement is sometimes referred to as *cross-lingual compatibility*. For instance, some words in one language may have to be translated as multiword expressions in another, which would significantly complicate the annotation projection for word-level tasks, such as part-of-speech tagging.

There is, in fact, a deeper issue known as *cross-lingual applicability* – certain elements of annotation may or may not make sense for a given language, e.g. a particular part-of-speech or dependency relation present in the source language may not actually exist in the target one.

Lastly, most annotation projection efforts necessarily rely on existing parallel corpora, most of which comprise translations produced for another purpose. The sources often include translations of parliamentary proceedings (Koehn, 2005; Roukos et al., 1995), books (Tiedemann, 2012), subtitles (Lison and Tiedemann, 2016), technical documentation (Tiedemann, 2009), talk transcripts (Tiedemann, 2012) or religious texts (Resnik et al., 1999; Mayer and Cysouw, 2013), most of which are translated with a view to retain the general meaning, but without taking any special care to avoid slight modifications in syntactic and semantic structure. Such modifications are known as *translation shifts* (Cyrus, 2006) and may have various degrees of severity, from simple syntactic modifications or mild generalizations, where the source statement still formally implies the target one, to trickier cases, where the implication can only be detected given substantial world knowledge external to the document under translation.

With these issues in mind, it is clear that the one-to-one correspondence as seen in 2.4 is more of an exception than a rule. In order to produce high-quality annotation for the target language, projection-based approaches need to work around the various modifications, often by applying extensive post-filtering or heuristics to select examples

where the correspondence between the source and target structures or interest is clearer. Some approaches are edging closer towards weakly-supervised learning than annotation projection, using the source-language annotation as a noisy signal, informative of the expected target-language structure, but no more than that.

Annotation projection methods have been effectively applied to a number of tasks, including

- part-of-speech tagging (Xi and Hwa, 2005; Das and Petrov, 2011; Agić et al., 2015)
- dependency parsing (Ganchev et al., 2009; Smith and Eisner, 2009; Hwa et al., 2005; Agić et al., 2016; Johannsen et al., 2016)
- semantic role labeling (Padó and Lapata, 2009; van der Plas et al., 2011; Annesi and Basili, 2010; Tonelli and Pianta, 2008)
- relation detection (Kim et al., 2010)
- chunking (Yarowsky et al., 2001; Zhu et al., 2014)
- morphology segmentation (Snyder and Barzilay, 2008)
- word sense disambiguation (Diab and Resnik, 2002; Bentivogli and Pianta, 2005)
- predicate disambiguation (van der Plas and Apidianaki, 2014)
- verb classification (Merlo et al., 2002)
- coreference (Grishina and Stede, 2015)
- mention detection (Zitouni and Florian, 2008)
- LFG parsing (Wróblewska and Frank, 2009)
- temporal relation prediction (Spreyer and Frank, 2008)
- discourse relation recognition (Versley, 2010)
- HPSG parse disambiguation (Frermann and Bond, 2012)

Interestingly, it has even been used to propagating morphosyntactic information between old and modern versions of the same language (Meyer, 2011).

2.3.4 Model Transfer

Annotation projection is instrumental in generating training data for low-resource languages, but it crucially depends on high-quality parallel data and leaves open some important questions, for instance, how to handle unaligned units.

A potentially less sensitive to parallel data and alignment quality alternative, which we will refer to as *model transfer*, has been proposed by Zeman and Resnik (2008), based on the parser adaptation technique of McClosky et al. (2006). This approach does not explicitly rely on aligned parallel sentences, instead following the representation learning paradigm – learning to represent units that the model operates on (tokens, in case of a dependency parser) in such a way as to make them indistinguishable to the model. The

shared representation can be the representation native to the source or the target language, or something in between the two.

To illustrate, let us assume that we have an existing dependency parsing model for L_1 and a second language, L_2, for which no manually annotated training data is available. If the two languages are similar in terms of grammar and vocabulary, one could attempt to apply the L_1 model directly to L_2 data. But if the speakers of these languages prefer to use different spelling of certain words (as in the case of Danish and Swedish in Zeman and Resnik (2008)) or even opted for different writing systems (consider German and Yiddish), chances are that the model will fail to recognize a large share of the words.

Depending on the shared representation we choose, possible solution would include representing L_1 words as their L_2 translations or vice versa – a technique sometimes known as *glossing*. Or one could choose some third representation, for instance translate the words in the two languages into a third, pivot language, map words to cross-lingual word clusters (Täckström et al., 2012) or distributed word representations (Klementiev et al., 2012). Note that all the options mentioned above do require *some* knowledge of the correspondence between the source- and target-language tokens, but none of them necessarily need a high-quality parallel data on sentence level.

This example can further be used to illustrate why the intermediate representation might be desirable. If we use the source-language tokens as the shared representation, it is likely that the learned model will learn to rely on some of them as very strong indicators of certain structures, which could be seen as overfitting with respect to the cross-lingual problem. The glossed representation of a target sentence is likely to contain some amount of noise, for instance due to mapping some target-language words into their most frequent translations, not necessarily appropriate in the given context. Using an intermediate representation which might be less expressive, but reproducible with comparable precision for units of both languages, may alleviate this problem.

In the context of dependency parsing, this idea was developed further (Durrett et al., 2012; Søgaard, 2011), augmented to allow for the use of parallel data (McDonald et al., 2011) and even adapted to a multi-source transfer case to enable explicit modeling of syntactic similarities between languages and selective parameter sharing based on this information (Naseem et al., 2012; Täckström et al., 2013).

2.3.5 Evaluation of Unsupervised and Transfer Methods

Evaluation of unsupervised methods in natural language processing in a highly controversial subject – since their objective is to induce some useful linguistic structure without relying on annotated resources for guidance, it is not entirely clear what quality criteria should apply. For lack of better options, one often has to resort to matching the automatically induced structure against manual annotation of the same type. While it is clear that a certain amount may be expected between two useful types of annotation that aim to represent the same phenomena, but exact match is considerably less likely.

In the case of syntactic analysis, for instance, a number of decisions in designing an annotation scheme are known to be more or less arbitrary and to vary from treebank to treebank. Examples include attachment of prepositions, determiners and punctuation, representation of conjunctions and light verb constructions (Zeman et al., 2012). Unsupervised grammar induction approaches evaluated on being able to reproduce the structure considered *correct* by the annotation guidelines of the reference treebank therefore end

up requiring a set of tricks to mimic the aforesaid arbitrary decisions, which boosts the scores, but does not necessarily yield a more useful structure.

Cross-lingual transfer methods face a somewhat similar problem in that we start from the assumption of having no annotated resources for the target language. Yet for evaluation purposes we tend to select language pairs where annotated resources are available on both sides. Ideally these resources should be as similar as possible, so that transferred annotations might be expected to be compatible with the reference ones, but finding out exactly to what point the improved ability to replicate target-level annotations implies better transfer quality is non-trivial. It is important to remember that perfect semantic parallelism is only a simplifying assumption, and even where the meaning is very close the annotation may not match perfectly for a variety of reasons, as we saw in chapter 3.

To get closer to a good quality metric, one could specifically select parallel sentences for the purpose and annotate them ensuring (1) high degree of semantic parallelism and (2) high cross-lingual agreement in terms of annotation (Padó and Lapata, 2009).

An application-based evaluation may also present a viable alternative. Given a fixed model and an evaluation procedure for a well-defined task that requires SRL annotation in the target language as an input, one could measure the performance of a transfer model using that of the resulting target-language model as a proxy. Unfortunately, the relation between such a proxy and the quality of the transferred annotation may also be non-trivial. For instance, certain tasks only care about a subset of semantic roles and ignore the rest.

2.3.6 Alternatives to SRL

While semantic role labeling as described above remains a topic of active research, the formalism itself is regularly challenged as well.

Banarescu et al. (2012), for instance, propose a way of representing shallow semantics in a more lightweight way, referred to as Abstract Meaning Representation (AMR).

AMR does not rely on syntax, thus reducing the cost of annotation (referred to as *sembanking*), and can represent whole-sentence semantics, including non-verbal predicates and even some extra-propositional aspects. In the last few years it has sparked a whole subfield of research into producing AMR representations automatically (Flanigan et al., 2014; Wang et al., 2015), using them for event detection (Kai and Grishman, 2015), summarization (Liu et al., 2015) and entity linking (Pan et al., 2015) among others, and even investigating the feasibility of using AMR as an intermediate representation in machine translation (Xue et al., 2014).

Other attempts to reformulate semantic role labeling aiming for better expressivity, ease of annotation and lower language dependence have been made, such as the Universal Conceptual Cognitive Annotation (Abend and Rappoport, 2013).

Some even propose to get rid of the concept of semantic role altogether (He et al., 2015).

2.3.7 Our Formulation

Semantic role labeling is often regarded as a part of the sentence-level NLP pipeline along with part-of-speech and morphological tagging, syntactic parsing, etc. It is therefore

desirable to make SRL models available for a number of languages. Thus far, corpora have been annotated for several major ones following various formalisms. A series of shared tasks conducted evaluation of different statistical models for this task first on constituent-based SRL (CoNLL 2004-2005) and later on dependency-based SRL (CoNLL 2008-2009). To date, as far as we know, annotated semantic role labeling corpora of one sort or another are available for Catalan, Chinese, Czech, English, German, Japanese, Spanish and Swedish. Several other resources are presently under construction, including ones for Dutch, Korean and Brazilian Portuguese. The cross-lingual aspect, however, has been largely ignored until relatively recently.

When approaching this from cross-lingual perspective, it may at first appear most natural to adopt a FrameNet-style formalism. After all, the hierarchy of frames is language-independent, except for lexical units, which seem relatively easy to populate for a new language. Indeed, this last part can even be largely automated if one leverages other resources (Hartmann and Gurevych, 2013). This is perhaps why FrameNet hierarchy has already been adapted to several other languages, including German (Burchardt et al., 2006), Japanese (Ohara et al., 2003), Chinese (You and Liu, 2005), Spanish (Subirats and Petruck, 2003) and others.

If we consider model transfer or cross-lingual bootstrapping as described in the next chapter, however, there are certain issues. For instance, the coverage of the FrameNet hierarchy is potentially insufficient (Palmer and Sporleder, 2010), which may cause issues when analyzing data belonging to other domains. One should also point out that few sizable full-text corpora are currently available for FrameNet-style semantic role labeling and most of these are stored in a format of their own and we are aware of no unified API for them. There are also certain differences between the corpora. The German one, SALSA (Burchardt et al., 2006), for instance, is constituent-based, while the English FrameNet allows arbitrary spans. Some resources have also extended the hierarchy with frames of their own to account for certain examples not covered by the original frame inventory.

FrameNet parsing in cross-lingual setting does present an interesting field of study and there are issues to be addressed beyond those already tackled by existing research (Padó and Lapata, 2009; Tonelli and Pianta, 2008).

Our choice in favor of PropBank-like formalisms is therefore due to more data and models being readily available for a number of languages, rather than to the paradigm itself.

We will consider different annotation schemes for different languages, but they are all closely related and based largely on the notion the notion of dependency-based semantic role labeling as defined in the CoNLL Shared Task 2009 (Hajič et al., 2009).

In short, we assume that the data is tokenized and lemmatized and labeled with part-of-speech tags, morphological tags (if applicable) and syntactic dependency structures. Single tokens are labeled as predicates and heads of dependency subtrees, identified as arguments of a given predicate are assigned a single semantic role from a language-specific role inventory.

We also confine ourselves to verbal predicates. There have been efforts to handle nominal predicates in much the same fashion (Meyers et al., 2004), but currently the annotation for such predicates is only available in few languages. Addressing the problem of labeling nominal predicates and their arguments using techniques similar to the ones

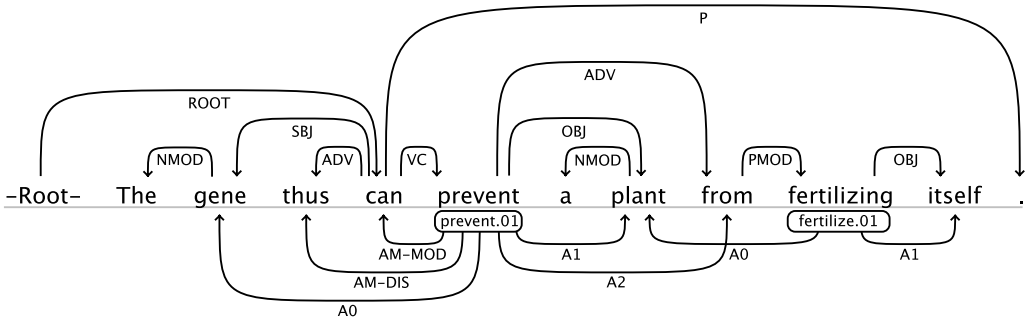


Figure 2.5: Dependency-based semantic role labeling example. The top arcs depict dependency relations, the bottom ones – semantic role structure.

we describe here presents an interesting direction for further research, but for now we will set it aside.

As mentioned above, in PropBank every predicate is associated with a frameset, which describes the interpretations of semantic roles in the context of this particular predicate. However, the role descriptions themselves are of little practical use to a statistical model, so in most cases they are ignored. Some models consider the set of roles described in a frameset as *allowed* for the corresponding predicate and prevent the model from assigning any other core roles. The predicate itself is often included as a feature, but to our knowledge, none of the existing systems have separate models for each predicate. Instead, they assume that semantic roles largely generalize across predicates, which is the case, to a greater or lesser extent, with all datasets we consider.

Chapter 3

Cross-lingual Bootstrapping

Our first approach is a relatively direct application of the idea of learning from similarity on unlabeled samples. As we already mentioned, finding tasks and datasets for the evaluation of such methods appears to be less trivial than might seem likely. Yet there is a kind of corpus that contains pairs of closely related documents, namely sentence-aligned parallel corpora, which are often used in machine translation and a number of other applications.

Most major parallel corpora are bilingual and are usually obtained by collecting texts in one language along with their translations into another language. Usual sources include proceedings of various multilingual organizations (Roukos et al., 1995; Koehn, 2005; Eisele and Chen, 2010), translations of literary works (Skadiņš et al., 2014), Wikipedia (Mohammadi and Ghasem-Aghaee, 2010; Ştefănescu and Ion, 2013), product manuals (Tiedemann, 2009), etc. The necessary processing, such as sentence and word alignment is usually performed automatically, so bilingual parallel corpora are available for a number of language pairs, some of them containing millions of sentence pairs. Monolingual parallel corpora do exist (e.g. Barzilay and McKeown (2001)), but are not as widespread, since their acquisition generally requires more effort and their applications are fewer.

This setup is rather specific in two respects. Firstly, the *related documents* in this case are sentences in different languages, therefore, unlike our previous experiments, different monolingual models will have to be involved. Secondly, there are only two documents in every group, unless we consider parallel corpora for more than two languages. It is possible to jointly align sentences in more than two languages extracted from corpora such as EuroParl (Koehn, 2005), as opposed to creating separate pairwise sentence-aligned corpora and the result may have its uses in studies such as ours, but we leave this topic outside the scope of the present work.

3.1 Motivation

Looking at the CoNLL'09 Shared Task datasets, it may strike one that most of them use annotation schemes of their own. Of the seven languages included, Catalan and Spanish follow the same annotation style, since both datasets were produced within the AnCora project (Taulé et al., 2008), and Chinese is annotated in a fashion similar to English, although, as we will see later, the set of roles is somewhat different.

Yet, as compared to such tasks as part-of-speech tagging or syntactic parsing, where

unifying the annotation schemes for different languages comes at the cost of necessary coarsening of categories (Petrov et al., 2012; McDonald et al., 2013; Zeman et al., 2012), semantic role labeling is a relatively language-independent task. One might therefore expect any reasonable set of semantic roles to be cross-lingually applicable, so the multitude of semantic role inventories is not due to differences between languages, but rather due to the desire of researchers to invent a better one: more informative, more consistent, easier to learn or better characterizing the syntactic-semantic interface.

Whatever the motives, we find that developing language-specific annotated resources in complete isolation creates certain redundancy. We would like to be able to benefit from the growing annotated resources no matter what set of semantic roles is employed. In the following experiment, therefore, we consider an approach which we refer to as cross-lingual bootstrapping. It assumes that we have two corpora for different languages, labeled with semantic roles from different sets, but in both cases following a kind of dependency-based SRL formalism. We further assume that a parallel corpus is available for the two languages considered and attempts to improve the performance of one or both of the models by enforcing a certain role mapping on the aligned instances in parallel data.

There is a piece of work by Zhuang and Zong (2010) that is closely related to this approach, though the objectives differ considerably. The authors of this paper attempt to achieve better decoding accuracy by performing inference on parallel data jointly for English and Chinese. Such inference appears to improve the accuracy one held-out data by a statistically significant margin. A core component of said joint inference step is a model of semantic role correspondence for different predicate pairs that is trained on a parallel corpus, manually annotated with constituency structures and semantic roles. It is unclear, however, whether the results would hold for inference one new, unseen data, drawn from a different domain, nor whether such inference could be used to improve the individual monolingual model.

One observation from this paper that is important to us is that even when parallel sentences are manually labeled using the same set of semantic roles, the mapping for aligned arguments is not necessarily trivial. For instance, the argument labeled A1 ("thing grown") in PropBank predicate GROW.01 generally corresponds to the argument marked A0 in Chinese 增长(ZENGZHANG).01 (see table 3.1), which in English is reserved for the volitional agent causing the growth. Such cases are, apparently, quite frequent, usually caused by minor translation shifts or by differences in the framesets.

Zh\En	A0	A1	A2	AM*	NONE
A0	0	16	0	0	0
A1	0	0	12	0	0
AM*	0	0	4	7	0
NONE	0	0	0	2	10

Table 3.1: Role correspondence for (ZENGZHANG).01 and GROW.01 (Zhuang and Zong, 2010).

Our approach is similar to that of Zhuang and Zong (2010) in the sense that we also model role correspondence, if for a somewhat different purpose.

One crucial difference is that we do not assume the availability of a manually anno-

tated parallel corpus for training. The availability of such a corpus does allow the role correspondence model to be trained in a straightforward supervised fashion, but severely limits the applicability of the proposed approach – to the best of our knowledge, such corpora are currently only available for two language pairs: English-Chinese (Palmer et al., 2005b) and English-Czech (Hajič et al., 2012). It may not seem like much of a difference, but one has to acknowledge that the predicted SRL structures on parallel data are imperfect, be it due to unseen predicates, unseen lexical items or errors in pre-processing and therefore any model learned from automatically labeled parallel data is likely to be less accurate.

We also do not require the annotated data for both languages to use the same set of semantic roles, but still assume that for a given predicate pair there exists a one-to-one mapping of roles. This holds in most cases, but not always. For instance, the Czech corpus we use has multiple different roles for various temporal modifiers, indicating start and end points, duration, frequency or that some other event took place simultaneously with the one described by the predicate, while PropBank includes all these into one category, AM-TMP. Such granularity issues are infrequent, however.

Last but not least, our purpose here is to improve the monolingual models, rather than simply obtain more accurate predictions on parallel data. The two goals are, of course, closely related, but the experiment in Zhuang and Zong (2010) does not provide direct evidence that the technique could be used to improve the accuracy of individual SRL models on unseen monolingual data.

3.2 Approach

The notion of role correspondence model (RCM) is based on the assumption that for a given pair of predicates (p_s, p_t) there exists a one-to-one mapping of roles $R_s \rightarrow R_t$. Given a manually annotated parallel corpus, the problem of learning such a mapping would be quite a standard one. Without one, however, we have to resort to bootstrapping-like iterative procedure, where the parameters of the RCM are estimated simultaneously with those of monolingual SRL models. First the initial models are trained on the available labeled data, then the RCM parameters are estimated from a parallel corpus automatically labeled by these models and then the RCM is used to refine the annotations of parallel data via a joint inference procedure, serving to enforce consistency on the predictions of monolingual models on parallel sentences.

The obtained annotations on the parallel corpus are expected to be of higher quality than the independent predictions of the models, so they can be used to improve the SRL models' performance and/or coverage. We will first verify that joint inference affects the quality of these annotations favorably by training a new pair of monolingual model on the output of the joint inference procedure (which we denote JOINT) and evaluating their performance on a held-out test set. And then we also augment the original training set with these annotations and measure the change in performance.

If one of the languages is relatively poor in resources, the proposed procedure will help propagate information from the stronger model to the weaker one. This scenario is further referred to as *projection*.

If the two models are comparable in their predictive power, we may still be able to benefit from the fact that the roles of certain arguments are indicated more clearly in one

language than in another. We will call this the *symmetric* setup.

More formally, let us consider a pair of languages, α and β, and their corresponding datasets T_α and T_β, annotated with semantic roles, which we will refer to as the *initial* training sets. We also assume that a word-aligned parallel corpus is available for the pair of languages, which we denote P, with the predicates and their respective arguments identified on both sides. Our purpose is to optimize a joint objective $max_{\theta_\alpha,\theta_\beta}\left(L_\alpha(\theta_\alpha, T_\alpha) + L_\beta(\theta_\beta, T_\beta) + L_{RCM}(\theta_\alpha, \theta_\beta, \theta_\Sigma, P)\right)$, where θ_α and θ_β are the parameters of the monolingual SRL models and θ_Σ are the parameters of the agreement function (the RCM). We approximate this using the following iterative procedure:

$$
\begin{aligned}
&\theta_\alpha^1 = argmax_{\theta_\alpha} L_\alpha(\theta_\alpha, T_\alpha) \\
&\theta_\beta^1 = argmax_{\theta_\beta} L_\beta(\theta_\beta, T_\beta) \\
&\forall_{i \in 1..N} \\
&\quad \theta_\Sigma^i = argmax_{\theta_\Sigma} L_{RCM}(\theta_\alpha^i, \theta_\beta^i, \theta_\Sigma, P) \\
&\quad \theta_\alpha^{i+1}, \theta_\beta^{i+1} = argmax_{\theta_\alpha,\theta_\beta}\left(L_\alpha(\theta_\alpha, T_\alpha) + L_\beta(\theta_\beta, T_\beta) + L_{RCM}(\theta_\alpha, \theta_\beta, \theta_\Sigma^i, P)\right)
\end{aligned}
\tag{3.1}
$$

In other words, we train initial monolingual models M_α^1 and M_β^1 on T_α and T_β, respectively, apply them to the two sides of the parallel corpus, resulting in a labeled parallel corpus P^1. Then we collect the semantic role co-occurrence information and train the role correspondence model C^1 on it, then proceed to the joint inference step involving M_α^1, M_β^1 and C^1, resulting in a refined labeling P^2 of the parallel corpus. The two sides of the P^2 are then used to augment the initial training sets, yielding T_α^2 and T_β^2, and new models M_β^2 and M_β^2 are trained on these. The process can then be repeated using M_α^2 and M_β^2 instead of the initial models.

We report the model's performance on a held-out test set, drawn from the same corpus as the corresponding initial training set.

The procedure can be seen as a form of co-training (Blum and Mitchell, 1998) of a pair of monolingual SRL models. In our case, however, the question of the models' agreement is not as trivial as in most applications of co-training, requiring a statistical model of its own.

In the *projection* setup our approach is also similar to self-training with weak supervision coming from the stronger model. It may seem more natural to implement this procedure using perceptron-based classifiers and refining the model directly rather than retraining from scratch each time. We believe, however, that while this is potentially more efficient computationally, it is likely to produce results similar to ours.

Note that although the approach is iterative, we have observed no significant improvements from repeating the procedure, possibly owing to the noise introduced by errors in preprocessing.

3.2.1 Modeling Role Correspondence

It is necessary to distinguish between semantic roles and their interpretation in a particular context. The former can be defined in a variety of ways, depending on the formalism used. In case of FrameNet (Baker et al., 1998), for example, the interpretation of a semantic role (frame element) is explicitly provided for each separate frame, so

a frame and a frame element label together describe the semantics of an argument. PropBank (Palmer et al., 2005a) follows a mixed strategy – the labels for a relatively small set of *core roles* are numbered and their interpretations are provided separately for each predicate (although those of the first two roles, `A0` and `A1`, consistently denote what is known as Proto-Agent and Proto-Patient), while *modifiers* (Merlo and Leybold, 2001) bear labels that are interpreted consistently across all predicates. Other resources, such as VerbNet (Kipper et al., 2006a) or Prague Dependency Treebank (Hajič et al., 2006), use a single set of semantic roles (or *functors*), which are interpretable across different predicates.

From the standpoint of defining the semantic similarity of parallel sentences, the important implication is that we cannot assume that the corresponding arguments should bear the same label, even if the annotation schemes used are compatible. Nor can we write down a single mapping between the roles that will be valid across different predicates (figure 3.1), which motivates the need for a statistical model of semantic role correspondence.

In other words, if we consider a pair of predicate-argument structures that convey the same meaning, the semantics of their corresponding arguments will generally match, but the role labels may not, as in some of the formalisms we consider the interpretation of a semantic role may depend on the predicate used. For example in PropBank terms the argument marked A4 in the context of a predicate `rise.01` will be labeled A1 if we use `reach.01` instead: "The index rose to [42 points]$_{A4}$" vs "The index reached [42 points]$_{A1}$".

We therefore assume the existence of a one-to-one mapping between semantic roles for a given predicate pair.

As the mappings for different predicates are not entirely independent – at least some roles have a consistent interpretation across different predicates, – we choose to build a single model, which relies on features derived from the pair of predicates in question, rather than create a separate model for each predicate pair. The model can then make decisions specific to particular predicates or predicate pairs, where sufficient data has been observed or back off to a generic mapping where there is not enough data.

For the purpose of this study, we choose to separately model the probability of a target role, given the source one and the necessary contextual information and vice versa. These two components are referred to as *projection models* and realized as a pair of linear classifiers.

3.2.2 Joint Inference

The joint inference would have been simplest if the arguments were classified independently. This assumption is too restrictive, though, since the interdependencies between the arguments can be used to improve the accuracy of semantic role labeling (Roth and Yih, 2005).

Projection Setup

In the projection setup we assume that the model for one of the languages, which we will henceforth refer to as *source*, is much better informed than the one for the other language, referred to as *target*, so we only have to propagate the information one way. The scoring

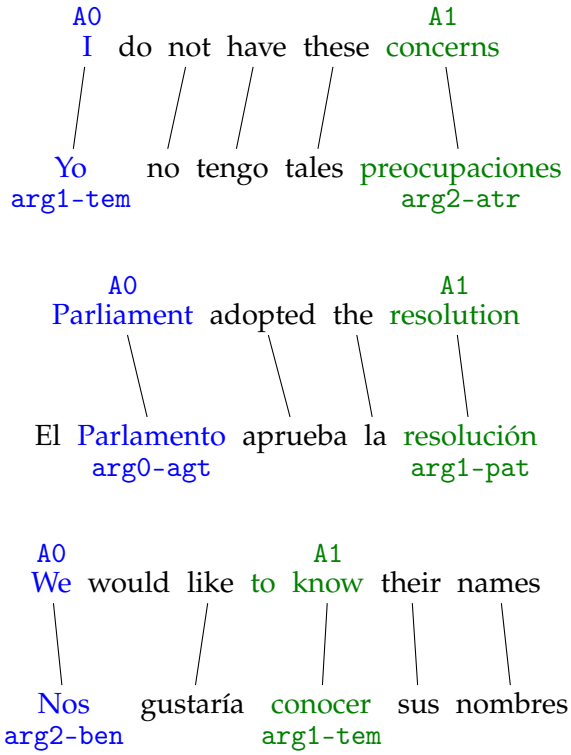


Figure 3.1: Predicate-specific role mapping. Note that `A0` corresponds to `arg0-agt`, `arg1-tem` or `arg2-ben`, and `A1` – to `arg1-tem`, `arg1-pat` or `arg2-atr`, depending on the predicate.

functions of these two models will be denoted f_s and f_t, respectively, and that of the projection models from source to target and from target to source – f_{st} and f_{ts}. Source and target sentences are denoted S_s and S_t, and aligned predicates in these sentences – p_s and p_t.

Since we assume that the parameters of the source-language model are fixed, then so are its predictions r_s on the source side and the joint inference step will consist in determining the assignment r_t of roles for the target side:

$$r_t = argmax_{r_t} \left(\lambda_t f_t(r_t, S_t, p_t) + \lambda_{st} f_{st}(r_t, r_s, p_s, p_t) \right), \tag{3.2}$$

where λ coefficients denote the respective weights of the models, which we will discuss below.

In the general case, optimizing such an objective is computationally expensive and may require approximations. Assuming that the models involved can incorporate a bias towards or away from a certain prediction on a given argument, one may use dual decomposition (Johnson, 2008), decoupling the r_t variables into r_t and r_{st} and maximizing $\lambda_t f_t(r_t, S_t, p_t) + \lambda_{st} f_{st}(r_{st}, r_s, p_s, p_t)$ under the condition that $r_t = r_{st}$ and then relaxing this constraint, replacing it with a soft one using slack variables δ, resulting in the following objective:

$$min_\delta max_{r_t, r_{st}} L'_1(r_t, r_{st}, \delta) =$$
$$\lambda_t f_t(r_t, S_t, p_t) + \lambda_{st} f_{st}(r_{st}, r_s, p_s, p_t) + \sum_i \sum_{r \in R_t} \delta^{i,r} \left(I(r_t^i = r) - I(r_{st}^i = r) \right),$$

where i indexes aligned argument pairs and I is an indicator function. This is equivalent to

$$min_\delta max_{r_t, r_{st}} L'_1(r_t, r_{st}, \delta) = min_\delta \Big(max_{r_t} g_t(r_t, S_t, p_t, \delta) + max_{r_{st}} g_{st}(r_{st}, r_s, p_s, p_t, \delta) \Big), \quad (3.3)$$

where

$$\begin{aligned} g_t(r_t, S_t, p_t, \delta) &= \lambda_t f_t(r_t, S_t, p_t) + \sum_i \sum_{r \in R_t} \delta^{i,r} I(r_t^i = r) \\ g_{st}(r_{st}, r_s, p_s, p_t, \delta) &= \lambda_{st} f_{st}(r_{st}, r_s, p_s, p_t) - \sum_i \sum_{r \in R_t} \delta^{i,r} I(r_{st}^i = r) \end{aligned} \quad (3.4)$$

are the augmented objectives of the two component models, incorporating bias factors on various possible predictions. The dual objective 3.3 can then be minimized with respect to δ using subgradient descent algorithm (Sontag et al., 2011), which essentially increases the penalty for a given prediction by one of the models if it does not match that of the other model until the two agree. If the consensus is achieved, the predicted role assignment is the global maximum of the sum of the objectives.

There are other possible approximations, of course. And in our specific case the problem is further simplified by the fact that our projection model is factorized over arguments and can produce weights for every possible role being assigned to every argument, so we can essentially pre-calculate the biases it would induce on the monolingual model and run the inference for the target-side model with that set of biases.

Taking this into account we can also be sure that the dual decomposition procedure described above will always converge, given appropriate update rate and a sufficient number of iterations, – the fixed weights offered by the projection model will eventually be overcome by the bias factors. Even limiting the number of iterations to 1000, however, the described approximation demonstrates 99% convergence rate and we observe no significant differences in subsequent evaluation using exact inference or the above approximation.

Symmetric Setup

If the models have comparable accuracy, the problem becomes somewhat more complex, as we have to propagate information in both directions, and exact inference appears to be too expensive.

It is possible to use dual decomposition with this problem directly using three separate components, two for the monolingual models and one for the RCM, which would have to make its own predictions for the semantic roles on both sides without conditioning on the predictions of the monolingual models. This calls for a different kind of model than the one we use, however, – a model that would rely on a (possibly simplified) feature representation of the source and target arguments to jointly predict their labels.

The approximation we use instead does not exactly fit the dual decomposition framework, but appears to perform reasonably in practice. We use subgradient descent as in the inference procedure for projection setup, but in both directions simultaneously, interleaving gradient descent steps and allowing the projection models to access the updated predictions of the monolingual models. This results in a block gradient descent

algorithm with the following updates:

$$
\begin{aligned}
r_t^{n+1} &= argmax_{r_t} g_t(r_t, S_t, p_t, \delta_t^n) \\
r_s^{n+1} &= argmax_{r_t} g_s(r_s, S_s, p_s, \delta_s^n) \\
r_{st}^{n+1} &= argmax_{r_{st}} g_{st}(r_{st}, r_s^n, p_s, p_t, \delta_t^n) \\
r_{ts}^{n+1} &= argmax_{r_{ts}} g_{ts}(r_{ts}, r_t^n, p_t, p_s, \delta_s^n) \\
\forall_i \forall_{r \in R_s} \delta_s^{n+1,i,r} &= \delta_s^{n,i,r} + \gamma_s(n)(I(r_{ts}^{n,i} = r) - I(r_s^{n,i} = r)) \\
\forall_i \forall_{r \in R_t} \delta_t^{n+1,i,r} &= \delta_t^{n,i,r} + \gamma_t(n)(I(r_{st}^{n,i} = r) - I(r_t^{n,i} = r)),
\end{aligned}
\tag{3.5}
$$

where $\gamma_s(n) = \gamma_t(n) = \frac{\gamma_0}{n+1}$ is the update rate function for step n, and g_s and g_{ts} are defined as in (3.4), but with the direction reversed.

The algorithm also demonstrates convergence similar to that of the projection version, although it lacks the optimality guarantees.

3.3 Experimental Setup

We evaluate our approach on four language pairs, namely English vs German, Spanish, Czech and Chinese, which we will denote `en-de`, `en-es`, `en-cz` and `en-zh` respectively.

3.3.1 Parallel Data

The parallel data for the first three language pairs is drawn from Europarl v6 (Koehn, 2005) and from MultiUN (Eisele and Chen, 2010) for English-Chinese. We applied Stanford Tokenizer for English, tokenizer scripts (Koehn, 2005) provided with the Europarl corpus to German, Spanish and Czech, and Stanford Chinese Segmenter (Chang et al., 2008) to Chinese, then performed POS-tagging, morphology tagging (where applicable) and dependency parsing using MATE-tools (Bohnet, 2010).

Word alignments were acquired using GIZA++ (Och and Ney, 2003) with its standard settings. Predicate identification on the parallel data was done using the supervised classifiers of the monolingual SRL systems, except for German, where a simple heuristic had to be used instead, as only some of the predicates are marked in the training data, which makes it hard to train a supervised classifier. Following van der Plas et al. (2011), we then retain only those sentences where all identified predicates were aligned.

In the experiments we used 50 thousand predicate pairs in each case, as increasing the amount further did not yield noticeable benefits, while increasing the running time.

3.3.2 Annotated Data

The CoNLL'09 (Hajič et al., 2009) datasets were used as a source of annotated data for all languages. Only verbal predicates were considered and predicted syntax was used in evaluation.

We consider subsets of the training data in order to emulate the scenario with a resource-poor language. Due to the different sources the datasets were derived from, sentences contain different proportions of annotated predicates depending on the language. The German corpus, for example, contains about 6 times fewer argument labels per sentence than the English one. We will therefore indicate the sizes of the datasets

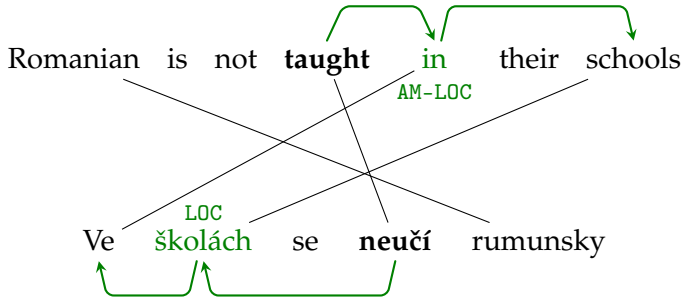


Figure 3.2: Role correspondence in parallel sentences, an example. Note that in the English sentence, the preposition "in" is marked as a locative modifier and in the Czech one it is the lexical head, "školách".

used in the number of argument labels they contain, referred to as *instances*, rather than the number of predicates or sentences. The corpus for English, for example, contains 6.2 such instances per sentence on average.

We use the 20 thousand instances of the available data as the training corpus for each language and split the rest equally between the development and the test set. The secondary ("out-of-domain") test sets are preserved as they are.

In dependency-based SRL, only heads of syntactic constituents are marked with semantic roles. The heads of corresponding arguments may or may not align, however, even if the arguments are lexically very similar, because their syntactic structure may differ. In general, one would have to identify the whole phrase for each argument and take into account the links between constituents, rather than single words (Padó and Lapata, 2005). As reconstructing the constituents from the dependency tree is non-trivial (Hwang et al., 2010), we are using a heuristic to address the most common version of this problem, i.e. a preposition or an auxiliary verb being an argument head. In such a case we also take into account any alignment links involving the head's immediate descendants. See figure 3.2 for an example.

3.3.3 Implementation

Our system is based on that of Björkelund et al. (2009). It is a pipeline system comprised of a set of binary or multiclass linear classifiers. Both here and in the projection model, the classifiers are trained using Liblinear (Fan et al., 2008). Logistic regression mode was used here, since we need accurate probability estimates in order to incorporate the bias in the joint inference phase.

We also employed constraints on role labels, as this appears to be more reliable in a low-resource setting we consider than the reranker the original system employed. Punyakanok et al. (2005) list a number of constraints that are useful for constituent-based PropBank-style semantic role labeling (Carreras and Màrquez, 2005).

1. Arguments cannot overlap with the predicate.

2. Arguments cannot exclusively overlap with the clauses.

3. If a predicate is outside a clause, its arguments cannot be embedded in that clause.

4. No overlapping or embedding arguments.

5. No duplicate argument classes for core arguments, such as A0–A5 and AA.

6. If there is an R-*arg* argument, then there has to be an *arg* argument. That is, if an argument is a reference to some other argument arg, then this referenced argument must exist in the sentence. This constraint is directly derived from the definition of R-arg arguments.

7. If there is a C-arg argument, then there has to be an arg argument; in addition, the C-arg argument must occur after arg. This is stricter than the previous rule because the order of appearance also needs to be considered. Similarly, this constraint is directly derived from the definition of C-arg arguments.

8. Given the predicate, some argument classes are illegal (e.g. predicate stalk can take only A0 or A1). This information can be found in PropBank Frames.

Constituents

These constraints are said to contribute considerably to the performance of a model. Since we are working with dependency-based SRL, we have to make certain modifications. In particular, in the absence of phrase structure information we ignore 2 and 3, 4 boils down to "one label per argument" rule, which is enforced implicitly, and 1 becomes "predicate cannot be an argument of itself", which most models can learn without introducing an explicit constraint.

Role Uniqueness

As Punyakanok et al. (2005) mention, there are exceptions to the core role uniqueness constraint (5), but only in about 0.3% cases overall and no more than 0.8% for each individual role. The exceptions usually have to do with conjunction-like structures or enumerations:

- "a repertoire$_{A1}$ that$_{R\text{-}A1}$ **ranges** from$_{A3}$ light classical to$_{A4}$ light jazz to$_{A4}$ light pop"
- "consumer prices$_{A1}$ **surged** by$_{A2}$ 5% and wholesale prices$_{A1}$ by$_{A2}$ 1.3%"
- "unemployment$_{A0}$ has **reached** 27.6%$_{A1}$ in$_{AM\text{-}LOC}$ Azerbaijan, 25.7%$_{A1}$ in$_{AM\text{-}LOC}$ Tadzhikistan, 22.8%$_{A1}$ in$_{AM\text{-}LOC}$ Uzbekistan, 18.8%$_{A1}$ in$_{AM\text{-}LOC}$ Turkmenia, ..."

We observe that there is also a considerable number of cases where role uniqueness is violated purely due to annotation errors, particularly where symmetrical relations are concerned.

- "a letter$_{A1}$ **attached** to$_{A1}$ grand jury's report"
- "ads$_{A1}$ were **tied** in with$_{*A1}$ pitches"
- "costs$_{A1}$ rose 1.2% **matching** the second-quarter pace$_{*A1}$"

Note that while in case of `tie.01` and `match.01` the second argument should, according to the corresponding frames, be labeled $A2$, for `attach.01` it is explicitly stated that multiple $A1$ arguments may be present in the same sentence.

Since the share of cases where the constraint does not hold is small enough and without the constraint the model tends to assign duplicate labels quite often, we enforce this as a hard constraint for core roles in English and Chinese.

Referential Arguments

Here the authors state that the constraint on referential arguments (6) follows directly from their definition. One should point out that referential arguments are not a part of the original PropBank annotation, but were produced as a part of conversion procedure, originally developed for CoNLL Shared Task 2004. In the introduction to the shared task (Carreras and Màrquez, 2004), it is indeed stated that "references (R-) are arguments representing arguments realized in other parts of the sentence. The role of a reference is the same as the role of the referenced argument." We observe, however, that this constraint is often violated within the corpus we use: over 12% predicates with R-A# arguments (or 18%, if we count non-core roles as well) do not have a corresponding A# argument and in about 2% of cases A# is not present anywhere in the given sentence (which is, strictly speaking, irrelevant, as argument roles have little meaning outside the context of a given predicate). Manual inspection also shows that referential arguments often have no counterparts in a given sentence, especially where they are realized by question words (such as *what, how, when, where*), as in "He believes in what$_{R\text{-}A1}$ he$_{A0}$ **plays**." We therefore ignore this constraint.

Continued Arguments

As in the case of referential arguments, these are not separate semantic roles, but rather a way of representing discontinuous arguments. We use this constraint (7) without modification.

Frames

The last constraint (8) assumes that a predicate can have only those core roles that are mentioned in the corresponding frame file. This, as we point out in section 2.1.3, holds for the occurrences of predicates in the corpus itself, but may not necessarily be true for new data we address. Besides, working in a low-resource setup it appears unrealistic to assume the existence of a resource listing every possible predicate and its roles.

We observe that in a random sample of 40 verbs from EuroParl that are not present in the CoNLL 2009 Shared Task corpus as verbal predicates, four are not in the PropBank inventory[1] at the time of this writing. On one hand, this seems to indicate that PropBank is rapidly developing and covering more and more domains. On the other hand, if we assume that we deal with a resource-poor language, we must expect far worse coverage from the frame inventory, if one is available at all.

[1] according to the unified verb index interface on `http://verbs.colorado.edu/verb-index/index.php`

abdicate	allot	authorise	bathe	better
brainwash	burrow	collude	compliment	debilitate
deforest	deliberate	delude	deluge	devolve
disabuse	disentangle	effect	engender	fiddle
forestall	**habituate**	**hassle**	leach	legalise
leverage	menace	misquote	network	oppress
overfish	overrun	recollect	relegate	**remaster**
sensitise	tidy	transcend	traumatise	utilise

Figure 3.3: A sample of 40 verbs from EuroParl that are not present in the CoNLL 2009 Shared Task corpus as verbal predicates. The instances that, to our knowledge, do not have a PropBank frameset at the time of this writing are highlighted in bold.

3.3.4 The Projection Model

Each projection model is realized by a single linear classifier applied to each argument pair independently. It relies on features derived from the source semantic role and source and target predicates, and predicts the semantic role for the argument in the target sentence.

```
FORMPAIR=A3-went-ging
LEMMAPAIR=A3-go-gehen
FORMSRC=A3-went
FORMTGT=A3-ging
LEMMASRC=A3-go
LEMMATGT=A3-gehen
LABEL=A3
```

Figure 3.4: Projection model features example.

The features include the source semantic role and its conjunctions with (lowercased) forms and lemmata of the source and target predicates. For example, assuming the source semantic role is `A3` and the source and target predicates are `went` and `ging` (past tense of "gehen", German), the features would be as shown in figure 3.4.

3.3.5 Parameters

In case of projection there are two parameters, λ_{st} and λ_t, – the weights of the component models in the objective. Only their relative values matter (except in the choice of γ_0), so we set λ_t to 1 and only tune the weight of the role correspondence model.

In the symmetric setup, the objective takes the form $L(r_t, r_s) = \lambda_t f_t(r_t, S_t, p_t) + \lambda_{st} f_{st}(r_t, r_s, p_s, p_t) + \lambda_s f_s(r_s, S_s, p_s) + \lambda_{ts} f_{ts}(r_s, r_t, p_t, p_s)$. Since we assume that the two monolingual models here have comparable performance, we do not tune their relative weights, setting both λ_s and λ_t to 1.

We also use the same weight for both projection models, $\lambda_{st} = \lambda_{ts}$, and this value plays an important role – it basically indicates how strongly we insist on the role correspondence models' correctness. If this weight is set to 0, the RCM will accept the initial predictions the monolingual models make, and if it is set to a sufficiently

large value, the predictions of the monolingual models will be biased until they match the mapping suggested by the RCM. The optimal weight will therefore depend on the language pair, the sizes of the initial training sets and the RCM used. We use the value of 0.7 in all projection experiments and 0.5 in the symmetric setup, however, as excessive tuning may be undesirable in the low-resource setting.

3.3.6 Domains

One important factor in the understanding of the evaluation figures presented is the fact that sources of annotated and parallel data belong to different domains. The former usually contains some sort of newswire text – Wall Street Journal in case of English (Palmer et al., 2005a), Xinhua newswire, Hong Kong news and Sinorama news magazine for Chinese (Palmer et al., 2005b), etc.

Parallel data, on the other hand, comes from the proceedings of European Parliament (Koehn, 2005) and United Nations (Eisele and Chen, 2010), which are quite different. For example, the sentences in the latter domain often start with someone being addressed, either by name or by title, which can hardly be expected to occur as often in a newspaper or a magazine article.

3.4 Evaluation

As is well-known, the performance of many statistical tools drops significantly outside the domain they were trained on, and the preprocessing and SRL models used here are no exception (Pradhan et al., 2008), which results in relatively low quality of the initial predictions on the parallel text. The low argument identification performance, in particular, is presumably due to inaccurate dependency parses, on which it heavily relies. Several approaches have been proposed to improve the accuracy of dependency parsers and other tools on out-of-domain data, such as Choi and Palmer (2011), but they are often non-trivial to apply or require additional resources, so we have not considered using them here. In some cases, though seldom, sources of parallel data belonging to the same domain as the annotated training data can be obtained.

Another concern is that the performance of a model trained on automatically labeled parallel data as measured on a test set we use may not reflect the quality of these annotations. To assess the resulting model's coverage, it would be interesting to evaluate it on data outside the original domain, so we consider the out-of-domain (OOD) test sets as provided for the CoNLL Shared Task 2009 where available.

Perhaps the most interesting one of these is the German OOD test set, which is drawn from Europarl (as is the parallel data we use). It was originally annotated with syntactic dependency trees and semantic structure in the SALSA format (Burchardt et al., 2006) for Padó and Lapata (2005), and then converted into a PropBank-like form for the CoNLL Shared Task 2009 (Hajič et al., 2009). The OOD test set for English is drawn from the Brown corpus (Francis and Kucera, 1967) and the one for Czech – from a Czech translation of Wall Street Journal articles (Hajič et al., 2012).

In the notation introduced above, the self-training baseline model (Self) is trained on P^0_β, the joint model (Joint) – on P^1_β and the combined model (Comb) – on T^1_β.

The first question we are interested in is how the joint inference affects the quality

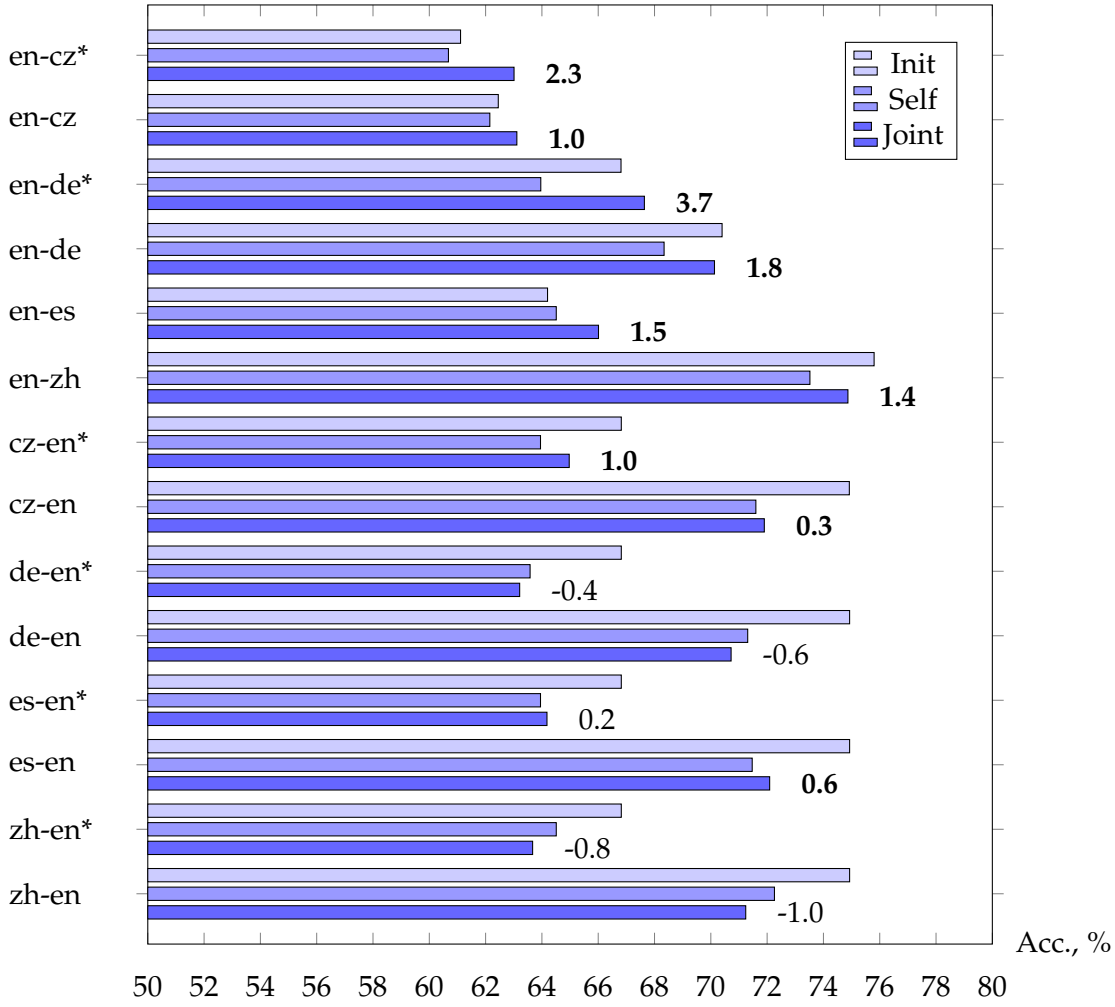


Figure 3.5: Projection setup results: self-training baseline, refined model and the difference in their performance. The numbers next to the bar indicate the improvement of Joint model over the self-training baseline. Asterisk indicates out-of-domain test set, statistically significant improvements are highlighted in bold.

of the automatically obtained annotations on the parallel data. To answer this, we will run the monolingual models independently and jointly, then train models on the output of these two procedures and compare their performance on a test set. Note that we do not add the initial training data at this point, so the initial model scores are provided for reference, rather than as a baseline.

3.4.1 Projection Setup

A small initial training set of 600 instances was used here for the target language here and the full training set (20000 instances) for the source one. λ_{st} was set to 0.7 in all experiments in this section.

In figure 3.5, we present the accuracy of the model trained on the output of the joint inference (JOINT) against that of the self-training baseline (SELF). The Δ_{SELF} column contains the difference between the two. Note that the SELF model is trained on the

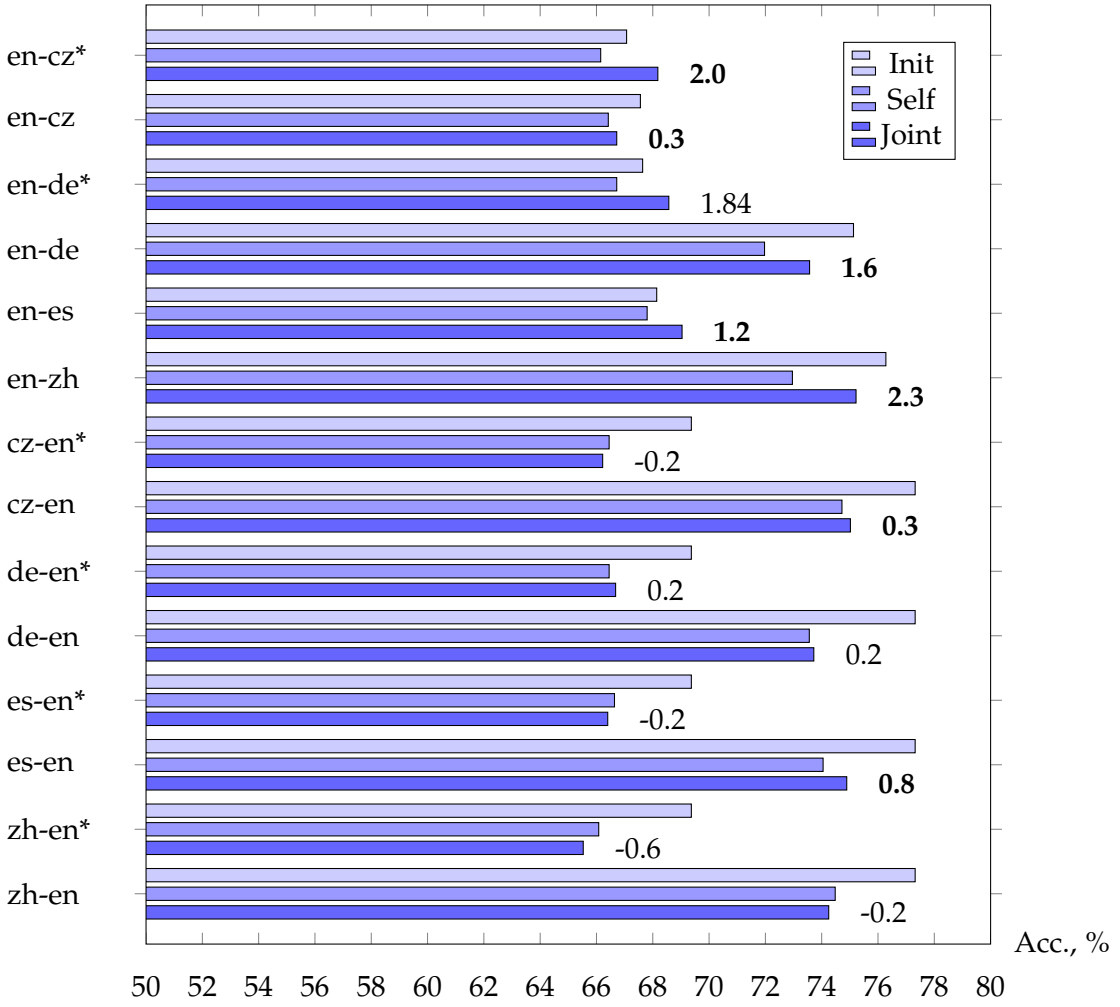


Figure 3.6: Comparing JOINT model against the self-training baseline in symmetric setup. Asterisk indicates out-of-domain test set, statistically significant improvements are highlighted in bold.

parallel data automatically annotated using monolingual SRL models (not mixed with the initial training set), since we are interested in the effect of joint inference on the quality of the annotation obtained. Where the improvement is positive and statistically significant with $p < 0.005$ according to the permutation test (Good, 2000), they are highlighted in bold.

We can see that the refined model (JOINT) outperforms the self-training baseline in most cases by a moderate, but statistically significant margin, which indicates that the joint inference does improve the quality of annotations on the parallel corpus.

The slightly higher improvement on the German OOD test set supports our hypothesis that the procedure enhances the performance of the model on parallel data, as the data for this test set is also drawn from the Europarl corpus. The result is statistically significant with $p < 0.05$ (higher p-value is primarily due to the smaller size of the evaluation set).

Figure 3.7 shows how the performance of the JOINT model changes with the size of the initial training set. The improvements are smaller for en-cz, en-de and en-zh, but they

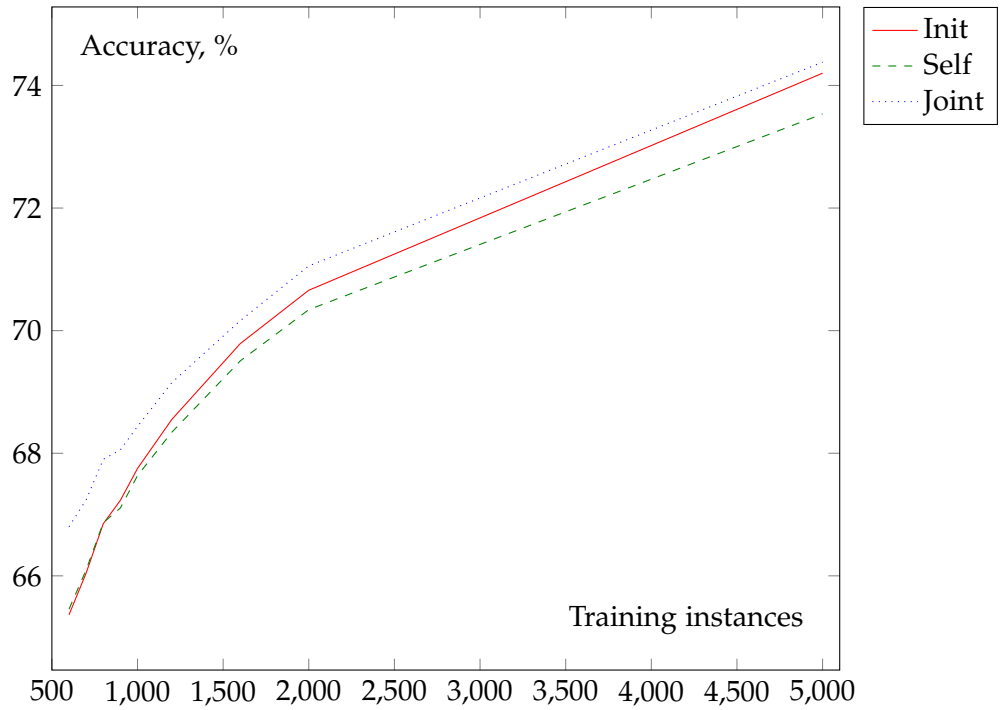


Figure 3.7: English-Spanish projection setup, model performance as a function of the size of the initial training set.

are also statistically significant for initial training sets of up to 2000 instances. Projection to English from other languages performs worse.

3.4.2 Combining

In practice, automatically obtained annotations are usually combined with the existing labeled data. For this purpose, the initial training set is replicated so as to constitute 0.3 (an empirically chosen value that appears to work well in most experiments) of the size of the automatically labeled dataset. We compare the performance of the model trained on the resulting dataset (Comb) with that of the Joint model and the initial models. The results are presented in table 3.9. We omit projection from other languages to English, since the Joint model there fails to outperform the initial model and we do not expect to benefit from adding the automatically annotated data to the initial training set in this case.

The procedure yields a small, but statistically significant improvements for English-Czech and English-German, and has very little effect otherwise. It should be noted that the approach of simply pooling the training data is rather ad hoc, and certain types of models would allow for more elegant and more efficient ways to do this. Models trained via stochastic gradient descent, for instance, could conceivably be post-trained directly on the newly available data, possibly incorporating an additional regularization term to ensure the parameters do not stray too far from those of the initial model.

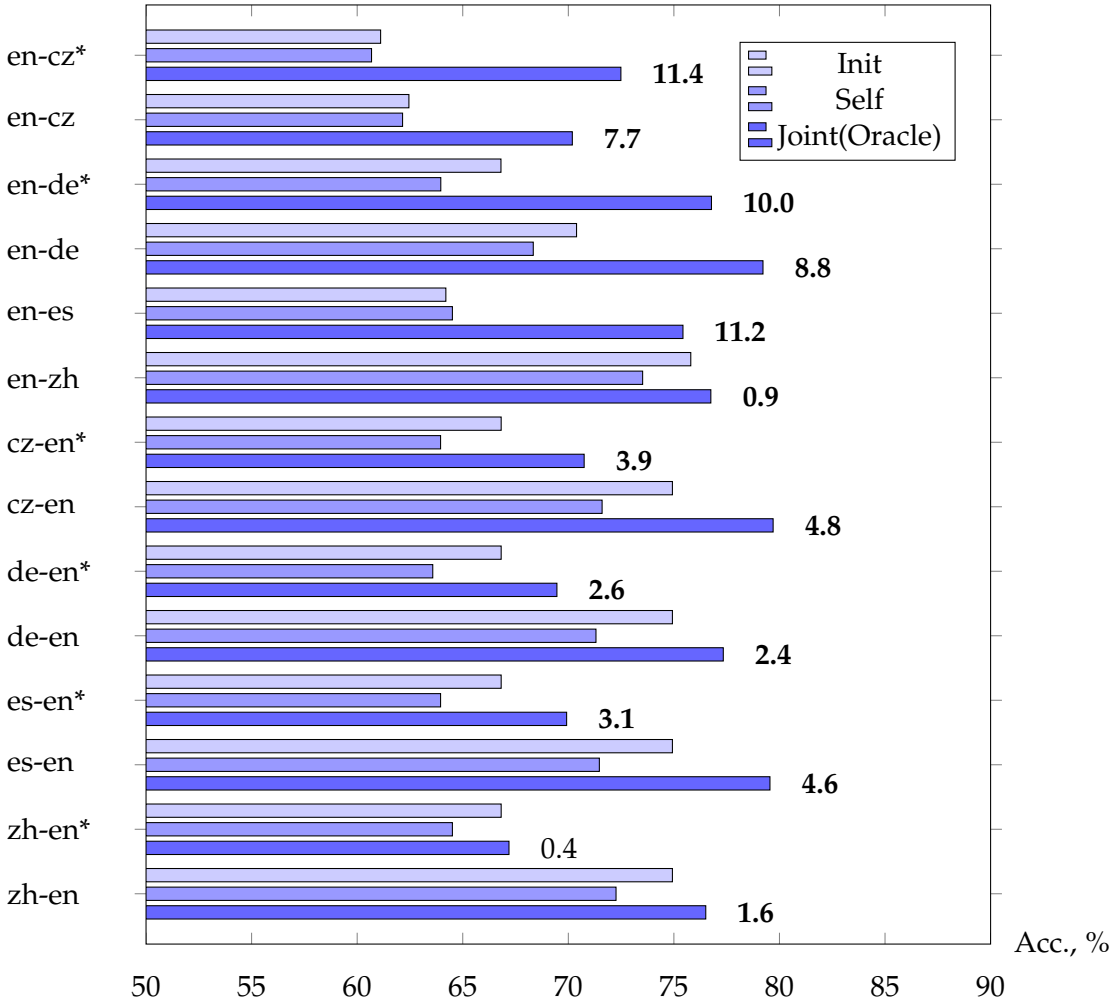


Figure 3.8: Performance using oracle RCM. The numbers next to the bars indicate the improvement from using the joint model over the initial one in absolute accuracy. Asterisk indicates out-of-domain test set, statistically significant improvements are highlighted in bold.

3.4.3 Symmetric Setup

In the symmetric setup evaluation, we use a slightly larger initial training set of 1400 instances for both source and target language. The projection model weight is set to 0.5. Figure 3.6 shows the accuracy of the JOINT model and the SELF baseline.

Note that here, unlike section 3.4.1, the joint inference is run once and then a model is trained for each language and evaluated on the corresponding test set. The results support our intuition that joint inference helps improve the quality of the resulting annotations, at least in some cases, even though all we are doing is essentially enforcing consistency on the semantic role correspondence across languages.

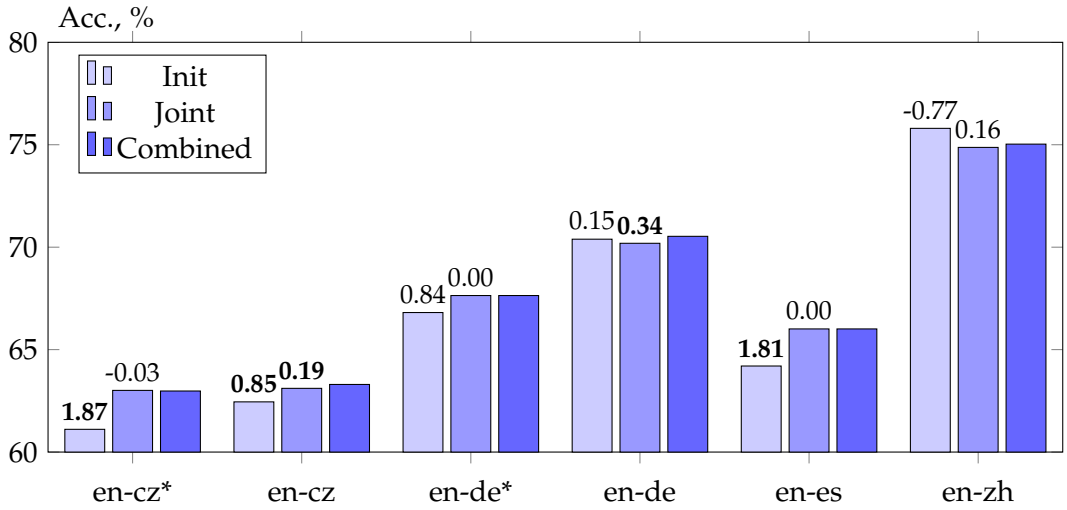


Figure 3.9: The effect of adding automatically obtained annotation to the initial training set. Asterisk indicates out-of-domain test set. The numbers indicate the improvement of the combined system (rightmost bar) over the respective baseline, and are highlighted in bold if the difference is statistically significant.

3.4.4 Oracle RCM

It would be useful to know to what extent the performance of the role correspondence model affects the quality of the output (and thus the performance of the resulting model). The RCM we use is rather simplistic, and we believe it can be substantially improved for any given language pair by incorporating prior knowledge and/or using external sources of information. In order to estimate the potential impact of such improvements, we simulate a better informed projection model, giving it access to the predictions of more accurate monolingual models on the parallel data – those trained on the full training set, rather than the initial training set used in this particular experiment. The result, referred to as *oracle* RCM, is evaluated in the projection setup in figure 3.8.

The more informed RCM clearly makes a big difference, yielding significant improvements over the self-training baseline in all cases, and up to 11 points in absolute accuracy compared to the initial model.

3.5 Role Correspondence Model Experiments

As we observe above, the role correspondence model has limited accuracy. This may be accounted for by several factors:

- **argument misalignment:** where the two arguments do not, in fact, have the same meaning and therefore should not bear the corresponding role
- **role granularity:** as we pointed out earlier, in some cases the target semantic roles are more fine-grained and the source roles may not be sufficiently informative to predict the target ones
- **unseen predicates:** if the roles are predicate-specific, it may be hard to deduce the correct mapping if the particular pair of predicates has not been seen previously

- **translation shifts:** this includes cases where two arguments roughly correspond to each other, but due to rephrasing in translation they have somewhat different semantic roles

- **interpretation variations:** even if roles are not predicate-specific, their interpretation may be somewhat different in the two languages (this may actually be seen as a type of mild translation shift as well)

Let us use the full models to label the two sides of the parallel corpus, train a role correspondence model on that and consider the cross-validation score of the resulting classifier. The results are as shown in figure 3.10. Given that only one of the datasets here relies on semantic roles that are not expressly predicate-specific, it is not surprising that using predicates as features improves the scores considerably.

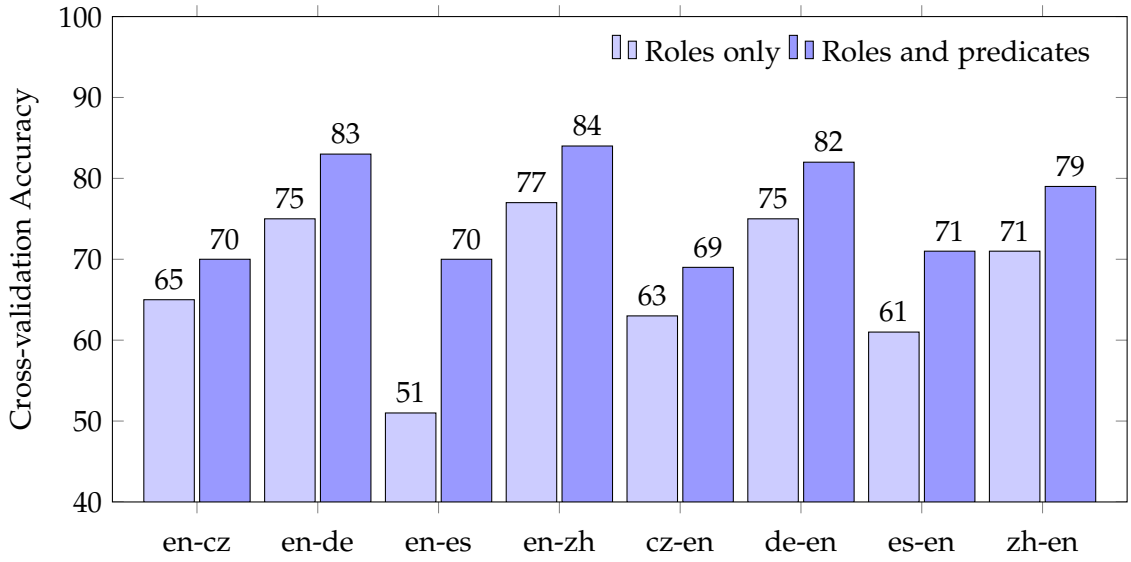


Figure 3.10: RCM cross-validation scores for different language pairs with and without using predicates as features.

To eliminate two of these factors, we consider the role correspondence on a manually annotated parallel resource, Prague Czech-English Dependency Treebank 2.0 (Hajič et al., 2012). The set of roles used for both sides is the same and is largely not predicate-specific, which excludes the granularity issues and limits the effect of unseen predicate pair problem. In this case a model taking into account only source roles, as may be expected, learns the identity mapping, which yield approximately 87% accuracy. A model that also can observe source and target predicate lemmata gets up to 90%.

3.6 Discussion

Our conclusion from this study, put simply, is that while predicate-argument structures in translated sentences are certainly related and knowing them for one sentence can help analyze its counterpart, various factors render the specific approach we consider here impractical. Firstly, the predicted syntactic structure of sentences in a parallel corpus is often inaccurate, which leads to unreliable argument correspondence information.

We also observe that in the parallel data we use a large share of predicted arguments are unaligned (even taking into account only aligned predicates), which is only partly accounted for by inaccurate argument identification. Perhaps we could benefit more from such approaches using more sophisticated argument matching or, possibly, by considering semantic representations not directly bound to specific tokens.

Secondly, the relation between semantic roles of corresponding arguments is less straightforward that it may seem at first, especially where semantic roles of different types are concerned. As we mentioned earlier, even if a manually aligned parallel corpus annotated with semantic roles from the same inventory is used, the accuracy of a role correspondence model reaches only about 90%.

Lastly, on closer inspection of the various datasets and annotation schemes, it appears to us that neither of the semantic role inventories in question is noticeably language-specific, though the associated resources (such as the PropBank frameset inventory) often are. We are thus led to believe that the multiety of dependency-based SRL annotation cannot be fully accounted for by the differences in the subject languages, and that in the future we might benefit from reusing existing annotation schemes insofar as possible, rather than inventing a new one every time.

Chapter 4

Direct Model Transfer

In this chapter we describe application of direct model transfer to dependency-based semantic role labeling.

Though the general approach is well established and reasonably straightforward, to our knowledge, this is the first time it is used in conjunction with semantic role labeling.

4.1 Motivation

Annotation projection works well given sufficient amounts of high-quality parallel data (as in most applications, by parallel data we mean word-aligned translated sentences). Since sentence boundary detection, tokenization, sentence alignment and word alignment are usually performed automatically, they are subject to errors and the sentences themselves can undergo various modifications in translation.

Admittedly, word-alignment errors are fairly uncommon, especially when the intersection heuristic known to favor high-confidence alignments is applied. This comes at the price of having some missing alignment links which in turn hinders projection-based approaches.

Semantic role labelling typically requires some additional annotations, such as part-of-speech tags or dependency trees, which are in turn subject to prediction errors. Syntactic parsing, in particular, is rather error-prone (see, for example, figure 4.2), especially where sentences from a different domain/register are concerned. Parliamentary proceedings constituting Europarl, in particular, contain significant amounts of direct speech, imperative statements and questions, which are less common in newswire text typically used to train syntactic parsers.

Direct model transfer, on the other hand, does not explicitly rely on parallel data. Instead, it acts on the assumption that the instances from source and target languages are similar enough in the chosen shared representation that they can be treated uniformly by the model.

The choice of the shared representation is therefore of utmost importance. Ideally, it should be picked such as to preserve as much relevant information as possible, while omitting the aspects that are likely to be language-specific.

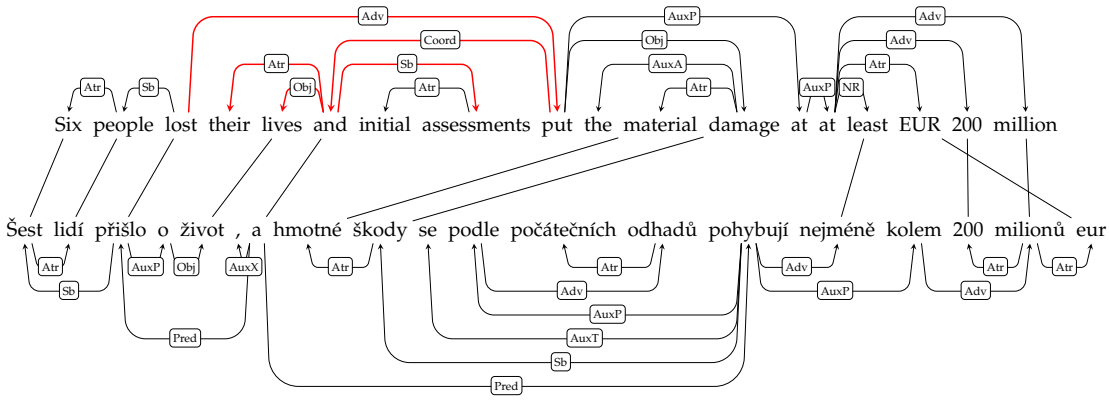


Figure 4.1: An example of bi-sentence containing a parsing error. The wrong edges are highlighted in red.

4.2 Direct Model Transfer for Dependency Parsing

Among the major applications of direct model transfer are part-of-speech tagging and dependency parsing (Zeman and Resnik, 2008; McDonald et al., 2011).

The idea of this approach consists in modifying a source-language model to make it directly applicable to a new language, which involves constructing a shared feature representation across the two languages. McDonald et al. (2011) successfully apply this idea to the transfer of dependency parsers, using part-of-speech tags as the shared representation of words. A later extension of Täckström et al. (2012) enriches this representation with cross-lingual word clusters, considerably improving the performance.

In the case of SRL, a shared representation that is purely syntactic is likely to be insufficient, since structures with different semantics may be realized by the same syntactic construct. For instance, "walking with a stone" is not the same as "hitting with a stone", although their dependency structure should be identical.

We therefore make use of recently introduced cross-lingual word representations, such as the cross-lingual clustering (Täckström et al., 2012) and cross-lingual distributed word representations of Klementiev et al. (2012), as well as simple gloss translations.

4.3 Model Transfer

As mentioned above, the idea of this approach is to abstract the model away from a particular source language and apply it to a new one. This setup requires that we use the same feature representation for both languages. The features generally used in semantic role labeling are often language-specific. In dependency-based SRL the features used for classification usually include lexical and syntactic properties of the predicate word, argument word and its parent, siblings and children in the dependency tree, as well as those of preceding and following words.

Much of this information is specific to a particular language and dataset – different part-of-speech tag and dependency relation inventories and, of course, the word types themselves. Some of this information can be mapped into a common space, though usually at the cost of considerable coarsening.

Depending on the pair of languages concerned and how structurally similar they are,

different levels of abstraction may become necessary, as a particular semantic role can be indicated by a particle, position of the argument word or its inflection, or combination thereof. For instance,

- English, position: "The dog chases the cat"
- Hungarian, inflection: "A kutya kergeti a macskát"
- Japanese, particle: 「犬は猫を追う」

There are some attempts at representing these various signals in a unified fashion (Tsarfaty, 2013), but the problem is non-trivial and may require considerable manual effort as well as in-depth knowledge of the languages concerned. Most direct transfer approaches, ours included, assume a comparatively high level of similarity and ignore structural issues, assuming that the correspondence is comparatively straightforward.

Further, in this work we only use the attributes available in all languages concerned to define the shared representation: part-of-speech tags, syntactic dependency structures and representations of word types.

4.3.1 Word Types

We assume the word types from the two languages in question to be from distinct sets. This is not necessarily true, especially of closely related languages, due to borrowing and other phenomena. There are also named entities and other foreign (with respect to one or both of the languages) words that may be left untranslated, as well as numbers, punctuation and other non-word tokens. Proper handling of these phenomena would require a module capable of automatically identifying such occurrences in each of the languages to avoid confusion of tokens whose surface representations match incidentally. For example, both English and Czech have a token "a", but in the latter case it is a conjunction (corresponding to English "and"), rather than an article. For the purpose of this work, we assume all tokens to be language-specific.

Word types from the two languages have to be mapped into a common feature space. We would like this mapping to preserve as much relevant information as possible and assign closely related words from both languages similar representations. That is, for a given word $w \in L_1$, its representation should be

1. be close to $w' \in L_1$, if replacing w with w' in most cases will preserve the meaning of the utterance
2. be close to $v \in L_2$, if w is likely to be translated into v
3. preserve useful information, so two words that are not close in the above mentioned sense are clearly distinguishable in said representation

Gloss.

The most straightforward way of addressing this is to represent each source-language word as itself (for instance, using the traditional one-hot encoding) and map target-language words into their likely translations.

This representation essentially preserves all word-type information available (item 3) and implements item 2 more or less directly. We can expect it to work well enough, except where highly polysemous words or certain function words that have no natural counterpart in the source language (e.g. particles, determiners or prepositions) are concerned. On the downside, it does not generalize well over related words in the source language – if we observe "blue" in the training data, but not, for instance, "teal", any word that translates into the latter will not yield any useful information.

The mapping itself can be derived from a bilingual dictionary, if available, or estimated based on a parallel corpus. We follow the latter path, collecting co-occurrence rates and mapping each target-language word into the source-language word it is aligned to most often. The result is, admittedly, somewhat noisy, but seems to be useful as a feature in our experiments.

Note that it is also possible to use target-language words as a shared representation, but intuitively one may expect such a mapping to introduce a certain amount of noise, thereby hindering the training of the source-language model.

Cross-lingual clusters.

One way to provide better generalization at the cost of some coarsening is to use cross-lingual clusters (Täckström et al., 2012), rather than separate words.

These can be obtained, for instance, by first clustering source-side words using a regular monolingual clustering algorithm, then *projecting* these clusters to the target side through word alignment links and again running clustering on the target side using the projected assignment as a constraint.

This has an edge over the gloss approach, as far as generalization among the source-language words is concerned, and is likely to yield much fewer unseen items, but at the cost of losing some information.

We did not rerun the procedure, but instead used the cross-lingual clusters provided by the authors. Unfortunately, these are available only in one granularity – with 256 clusters per language pair, – which, as we show later, may be insufficient for our purposes.

Distributed Word Representations.

Cross-lingual distributed word representations offer an alternative to hard clustering in that each word is represented by a real-valued vector, rather than a single id, which allows this representation to capture multiple notions of similarity. Intuitively, using such a representation of word type information may be expected to benefit model transfer (Turian et al., 2010). However, at the time this work was performed, the only published method of deriving cross-lingual word representation was the multi-task-learning-based (MTL) approach proposed by Klementiev et al. (2012). We attempted to make use of said representations, but observed no significant improvements compared to using the feature groups described above and will omit these experiments in further discussion for the sake of conciseness.

One could speculate why using cross-lingual vector representations fails to yield an improvement in our case. First of all, we note that the proposed MTL training procedure is rather computationally expensive, which caused the authors to limit the vocabulary

to the 3000 most frequent tokens and subsample the available training data, which may have a negative impact on the utility of the resulting word representations for our task.

4.3.2 Syntactic Information

As mentioned above, we ignore certain types of information, such as morphological tagging, and concentrate on basic attributes that can be found in any language and represented with reasonable consistently across languages in question.

Identifying representations that generalize better over less closely related languages is an open problem (Tsarfaty, 2013), which is, unfortunately, largely beyond the scope of this work.

Part-of-speech Tags.

We map part-of-speech tags into the universal tagset following Petrov et al. (2012). This may have a negative effect on the performance of a monolingual model, since most part-of-speech tagsets are more fine-grained than the universal POS tags considered here. For example, Penn Treebank inventory contains 36 tags and the universal POS tagset – only 12. On the other hand, the finer-grained POS tags often reflect more language-specific phenomena and may not be very useful (except for very closely related languages) in the cross-lingual setting.

The universal part-of-speech tags used in evaluation are derived from gold-standard annotation for all languages except French, where predicted ones were used instead (due to unavailability of gold-standard annotations for this dataset).

Dependency Structure.

Another important aspect of syntactic information is the dependency structure and dependency relations between modifiers and their respective heads. Most dependency relation inventories are language-specific, and finding a shared representation for them is a challenging problem[1]. One could map dependency relations into a simplified form that would be shared between languages, as it is done for part-of-speech tags in Petrov et al. (2012). The extent to which this would be useful, however, depends on the similarity of syntactic-semantic interfaces of the languages in question.

In this work we discard the dependency relation labels where the inventories do not match and only consider the unlabeled syntactic dependency graph. Note that some discrepancies, such as variations in attachment order, may be present even there, but this does not appear to be the case with the datasets we use for evaluation. If a target language is poor in resources, one can obtain a dependency parser for the target language by means of cross-lingual model transfer (Zeman and Resnik, 2008). We later compare the results based on the original dependency structures and the ones obtained by means of cross-lingual model transfer.

[1]The problem has since been addressed to some extent, in particular by McDonald et al. (2013) (cf. `http://universaldependencies.org`) and Zeman et al. (2014) (cf. `http://ufal.mff.cuni.cz/hamledt`).

4.3.3 Feature Selection

Compatibility of feature representations is necessary but not sufficient for successful model transfer. We have to make sure that the features we use are predictive of similar outcomes (in our case – the same semantic role) in the two languages as well.

Depending on the pair of languages in question, different aspects of the feature representation will retain or lose their predictive power. We can be reasonably certain that the identity of an argument word is predictive of its semantic role in any language, but it might or might not be true of, for example, the word directly preceding the argument word. It is therefore important to prevent the model from capturing overly specific aspects of the source language, which we do by confining the model to first-order features. We also avoid feature selection, which, performed on the source language, is unlikely to help the model to better generalize to the target one. The experiments confirm that feature selection and the use of second-order features degrade the performance of the transferred model.

4.3.4 Feature Groups

For each word, we use its part-of-speech tag, cross-lingual cluster id, word identity (glossed, when evaluating on the target language) and its dependency relation to its parent. Features associated with an argument word include the attributes of the predicate word, the argument word, its parent, siblings and children, and the words directly preceding and following it. Also included are the sequences of part-of-speech tags and dependency relations on the path between the predicate and the argument.

Since we are also interested in the impact of different aspects of the feature representation, we divide the features into groups as summarized in table 4.1 and evaluate their respective contributions to the performance of the model. If a feature group is enabled – the model has access to the corresponding source of information. For example, if only `POS` group is enabled, the model relies on the part-of-speech tags of the argument, the predicate and the words to the right and left of the argument word. If `Synt` is enabled too, it also uses the POS tags of the argument's parent, children and siblings.

Word order information constitutes an implicit group that is always available. It includes the `Position` feature, which indicates whether the argument is located to the left or to the right of the predicate, and allows the model to look up the attributes of the words directly preceding and following the argument word. The model we compare against the baselines uses all applicable feature groups (`Deprel` is only used in EN-CZ and CZ-EN experiments with original syntax).

`POS`	part-of-speech tags
`Synt`	unlabeled dependency graph
`Cls`	cross-lingual word clusters
`Gloss`	glossed word forms
`Deprel`	dependency relations

Table 4.1: Feature groups.

4.4 Evaluation

4.4.1 The Model

We use a simple linear model for the following experiments with the feature extractors based on those in Björkelund et al. (2009), which we modified to accommodate the cross-lingual cluster features. The classifiers are trained using Liblinear (Fan et al., 2008).

4.4.2 Datasets

Evaluation of the cross-lingual model transfer requires a rather specific kind of dataset. Namely, the data in both languages has to be annotated with the same set of semantic roles following the same (or compatible) guidelines, which is seldom the case. Indeed, scarcity of such data is a primary reason for investigating direct transfer approaches. We have identified three language pairs for which such resources are available: English-Chinese, English-Czech and English-French.

English-Chinese

The evaluation datasets for English and Chinese are those from the CoNLL Shared Task 2009 (Hajič et al., 2009) (henceforth CoNLL-ST). Their annotation in the CoNLL-ST is not identical, but the guidelines for "core" semantic roles are similar (Kingsbury et al., 2004), so we evaluate only on core roles here.

English-Czech

The data for the second language pair is drawn from the Prague Czech-English Dependency Treebank 2.0 (Hajič et al., 2012), which we converted to a format similar to that of CoNLL-ST[2]. The original annotation uses the tectogrammatical representation (Hajič, 2002) and an inventory of semantic roles (or *functors*), most of which are interpretable across various predicates.

Since tectogrammatical adopted by PCEDT is substantially different from dependency-based SRL we are working with, the conversion is necessarily lossy.

PCEDT, based on and partly overlapping with the Prague Dependency Treebank we briefly discussed in section 2.1.5, contains several connected layers of annotation:

- M-layer (morphemic), containing lexical and morphological information
- A-layer (analytical), describing syntactic structure of a sentence using something closely resembling a dependency tree.
- T-layer (tectogrammatical), describing shallow semantics of a sentence. Some of the tectogrammatical functors in T-layer serve the same purpose as semantic roles in PropBank, others are closer to syntax.

The layers are connected, with the relation between morphemic and analytical layers being relatively straightforward in that each node in one corresponds to exactly one node in the other, so one could think of the information contained in the morphemic layer as

[2] see `http://www.ml4nlp.de/code-and-data/treex2conll`

annotations on the nodes of the analytical tree. The alignment between analytical and tectogrammatic layers is trickier, however, since T-layer consists of a separate tree, only partially aligned with the A-layer one (see figure 4.4.2).

The conversion procedure is largely based on the script used by the CoNLL 2009 shared task organizers. Unfortunately, the format PCEDT is available in is different and said script is not directly applicable. We have therefore implemented a similar procedure for this new format in python[3].

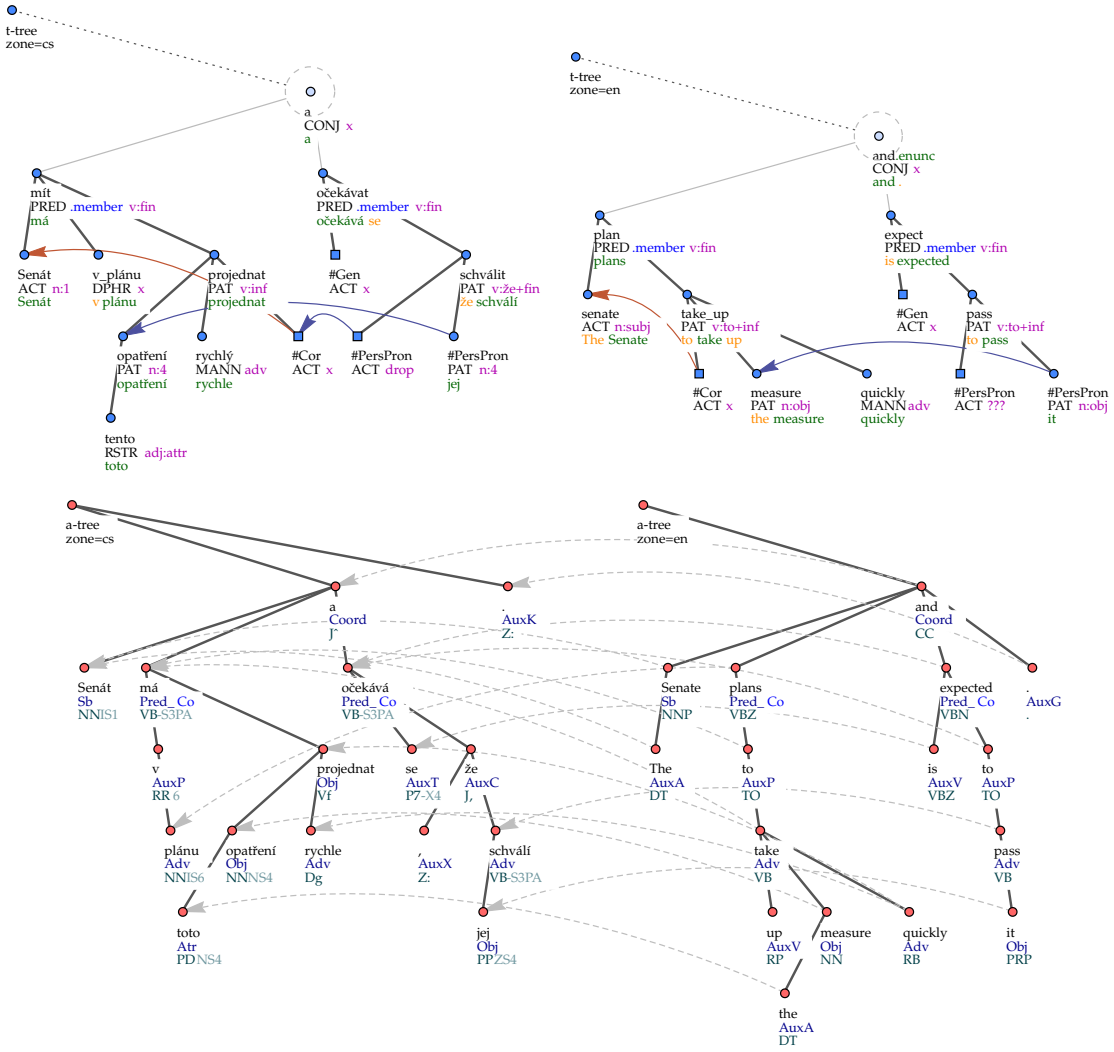


Figure 4.2: Example sentence pair from Prague Czech-English Dependency Treebank.
EN: The Senate plans to take up the measure quickly and is expected to pass it.
CZ: Senát má v plánu toto opatření rychle projednat a očekává se, že jej schválí.

The conversion is mostly straightforward, treating the A-layer tree as a dependency structure and porting over a subset of T-layer functors as semantic roles. The only tricky items concern generated nodes and handling of co-reference and conjunctions.

Figure 4.4.2 contains the bisentence converted from the structure in figure 4.4.2.

Note that the syntactic annotation of English and Czech in PCEDT is quite similar (to the extent permitted by the difference in the structure of the two languages), so we can

[3]see `http://www.ml4nlp.de/code-and-data/treex2conll`

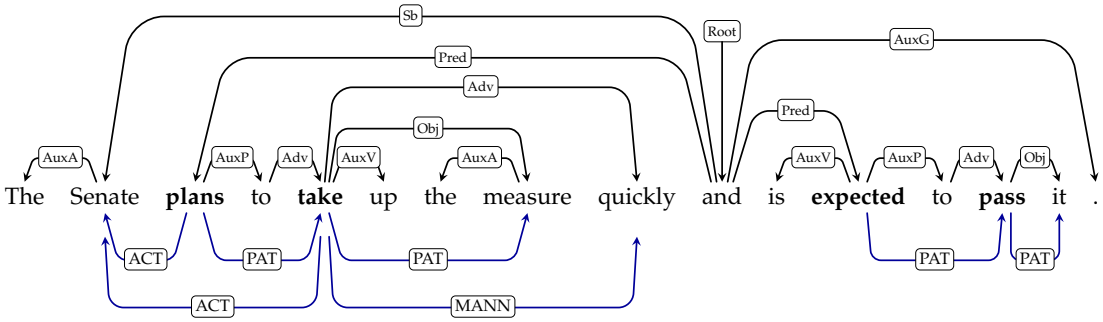


Figure 4.3: An English sentence converted from PCEDT.

also use the dependency relations in our experiments.

English-French

For English-French, the English CoNLL-ST dataset was used as a source and the model was evaluated on the manually annotated dataset from van der Plas et al. (2011). The latter contains one thousand sentences from the French part of the Europarl (Koehn, 2005) corpus, annotated with semantic roles following an adapted version of PropBank (Palmer et al., 2005a) guidelines. The authors perform annotation projection from English to French, using a joint model of syntax and semantics and employing heuristics for filtering. We use a model trained on the output of this projection system as one of the baselines. The evaluation dataset is relatively small in this case, so we perform the transfer only one-way, from English to French.

4.4.3 Preprocessing

The part-of-speech tags in all datasets were replaced with the universal POS tags of Petrov et al. (2012). For Czech, we have augmented the mappings to account for the tags that were not present in the datasets from which the original mappings were derived. Namely, tag "t" is mapped to "VERB" and "Y" – to "PRON".

We use parallel data to construct a bilingual dictionary used in word mapping, as well as in the projection baseline. For English-Czech and English-French, the data is drawn from Europarl (Koehn, 2005), for English-Chinese – from MultiUN (Eisele and Chen, 2010). The word alignments were obtained using GIZA++ (Och and Ney, 2003) and the intersection heuristic.

4.4.4 Syntactic Transfer

In the low-resource setting, we cannot always rely on the availability of an accurate dependency parser for the target language. If one is not available, the natural solution would be to use cross-lingual model transfer to obtain it.

Unfortunately, the models presented in the previous work, such as Zeman and Resnik (2008), McDonald et al. (2011) and Täckström et al. (2012), were not made available, so we did our best to reproduce the direct transfer algorithm of McDonald et al. (2011), using Malt parser (Nivre, 2008) and a similar set of features. Note, however, that we did not

reimplement the projected transfer algorithm and used the default training procedure instead of perceptron-based learning suggested by the authors.

Features:

1. POS tags and CLC ids for the top two tokens on the stack
2. POS tags and CLC ids for the three input tokens
3. bigrams of POS tags and CLC ids within and across (1) and (2)
4. combination of POS tag and CLC id for each of the first three input tokens and top two tokens on the stack
5. dependency relation of the top token on the stack to its rightmost and leftmost dependents and its head
6. dependency relation of the first input token to its leftmost dependent

The full list of features can be found in figure A in the appendix.

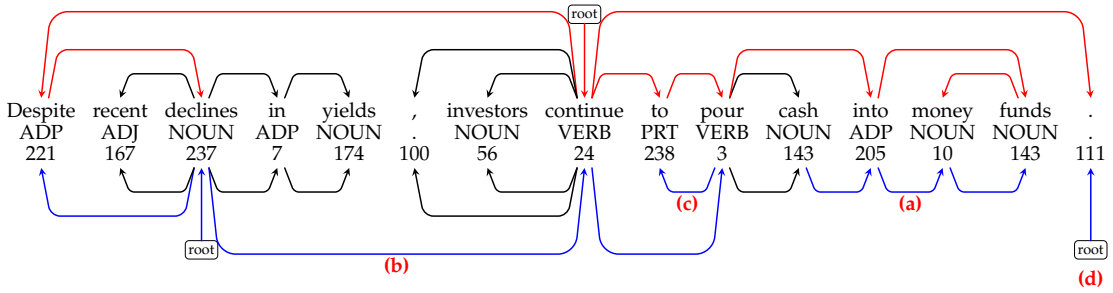


Figure 4.4: Original (above) and transferred (below) unlabeled dependency trees. We highlight the edges present only in the gold parse in red and those only present in the transferred syntactic tree in blue.

Figure 4.4.4 shows an example of a transferred syntactic parse compared to the gold syntactic parse (we use an English sentence and the output of a syntactic parser transferred from Czech for demonstration). Overall, the dependency structure does seem to contain some useful information. Some typical errors we observe here are due to lack of information or discrepancies in the annotation between the source and the target language.

For example,

(a) in Czech, any modifier preceding the head noun would typically be adjectivized, so it is reasonable for the model to assume the first noun in a noun phrase would be the head.

(b) given the coarse representation we use, the model often has trouble discriminating between possible roots.

(c) "to" has its own part-of-speech tag in PTB, which gets mapped into PRT (particle), rather than ADP (adposition), making the situation rather complicated. Interestingly, the Czech word it translates into, "v" in this context, is consistently marked with tag R in PDT, which is mapped into universal POS category PRON (pronoun).

(d) in PDT (or, rather, in its dependency version), terminal punctuation is typically attached directly to the root, rather than to the main verb – one of the many arbitrary decisions that differentiate dependency treebanks without reflecting actual language specificity.

The scores vary considerably across language pairs and within a language pair, depending on the direction of transfer. Their values are shown in table 4.2. It is clear that the dependency structure thus obtained is only a rough approximation and could be improved by using more sophisticated model transfer techniques or even simple feature selection. However, it serves our purpose as we will demonstrate that this is a useful source of information for SRL model transfer and, besides, even more advanced approaches are unlikely to perform very well transferring syntax between such languages as Czech and English, given the inherent difference in their structure.

Setup	UAS, %
EN-ZH	35
ZH-EN	42
EN-CZ	36
CZ-EN	39
EN-FR	67

Table 4.2: Syntactic transfer accuracy, unlabeled attachment score (percent). Note that in case of French we evaluate against the output of a supervised system, since manual annotation is not available for this dataset. This score does not reflect the true performance of syntactic transfer.

We will refer to the syntactic annotations that were provided with the datasets as *original*, as opposed to the annotations obtained by means of syntactic transfer (*transferred*).

4.4.5 Baselines

Unsupervised Baseline:

We are using a version of the unsupervised semantic role induction system of Titov and Klementiev (2012a) adapted to the shared feature representation considered in order to make the scores comparable with those of the transfer model and, more importantly, to enable evaluation on transferred syntax. Note that the original system, tailored to a more expressive language-specific syntactic representation and equipped with heuristics to identify active/passive voice and other phenomena, achieves higher scores than those we report here.

Projection Baseline:

The projection baseline we use for English-Czech and English-Chinese is a straightforward one: we label the source side of a parallel corpus using the source-language model, then identify those verbs on the target side that are aligned to a predicate, mark them as predicates and propagate the argument roles in the same fashion. A model is then trained on the resulting training data and applied to the test set.

For English-French we instead use the output of a fully featured projection model of van der Plas et al. (2011), published in the CLASSiC project.

4.4.6 Evaluation Measures

We use the F_1 measure as a metric for the argument identification stage and accuracy as an aggregate measure of argument classification performance. When comparing to the unsupervised SRL system the clustering evaluation measures are used instead. These are purity and collocation

$$Pu = \frac{1}{N}\sum_i \max_j |G_j \cap C_i|$$

$$Co = \frac{1}{N}\sum_j \max_i |G_j \cap C_i|,$$

where C_i is the set of arguments in the i-th induced cluster, G_j is the set of arguments in the jth gold cluster and N is the total number of arguments. We report the harmonic mean of the two (Lang and Lapata, 2011) and denote it F_1^c to avoid confusion with the supervised metric.

4.5 Results

In order to ensure that the results are consistent, the test sets, except for the French one, were partitioned into five equal parts (of 5 to 10 thousand sentences each, depending on the dataset) and the evaluation performed separately on each one. All evaluation figures for English, Czech or Chinese below are the average values over the five subsets. In case of French, the evaluation dataset is too small to split it further, so instead we ran the evaluation five times on a randomly selected 80% sample of the evaluation data and averaged over those. In both cases the results are consistent over the subsets, the standard deviation does not exceed 0.5% for the transfer system and projection baseline and 1% for the unsupervised system.

4.5.1 Argument Identification

We summarize the results in figure 4.5. Argument identification is known to rely heavily on syntactic information, so it is unsurprising that it proves inaccurate when transferred syntax is used. Our simple projection baseline suffers from the same problem. Even with original syntactic information available, the performance of argument identification is not very satisfactory. We therefore set the problem of argument identification aside and in the following experiments consider argument classification based on the gold argument positions.

Most unsupervised SRL approaches assume that the argument identification is performed by some external means, for example heuristically (Lang and Lapata, 2011). Such heuristics or unsupervised approaches to argument identification (Abend et al., 2009) can also be used in the present setup.

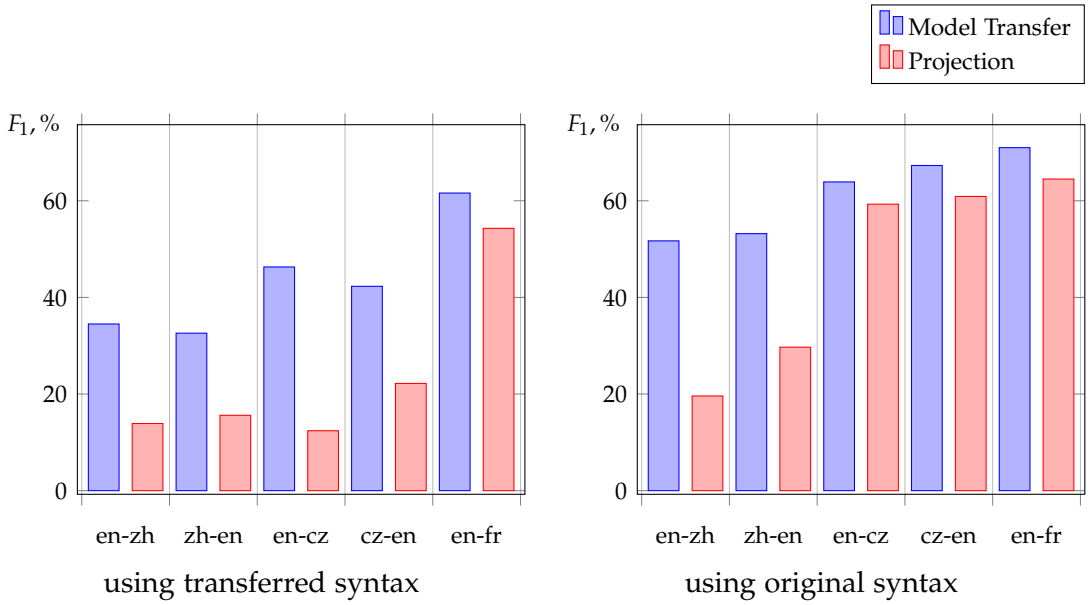


Figure 4.5: Argument identification, transferred model vs. projection baseline, F_1.

4.5.2 Argument Classification

In the following tables, TRANS column contains the results for the transferred system, UNSUP – for the unsupervised baseline and PROJ – for projection baseline. We highlight in bold the higher score where the difference exceeds twice the maximum of the standard deviation estimates of the two results.

Figures 4.7 and 4.6 present the unsupervised evaluation results. Note that the unsupervised model performs as well as the transferred one or better where the original syntactic dependencies are available. In the arguably more realistic scenario with transferred syntax, however, the transferred model proves more accurate.

In figures 4.8 and 4.9 we compare the transferred system with the projection baseline using original and transferred syntax, respectively. The scores vary rather strongly between language pairs due to both the difference in the annotation scheme used and the degree of relatedness between the languages. The drop in performance when transferring the model to another language is large in every case, though, see table 4.3.

Setup	Target	Source
EN-ZH	71.7	87.1
ZH-EN	66.1	86.2
EN-CZ	59.0	80.1
CZ-EN	61.0	75.4
EN-FR	63.0	82.5

Table 4.3: Model accuracy on the source and target language using original syntax. The source language scores for English vary between language pairs because of the difference in syntactic annotation and role subset used.

We also include the scores for individual labels for EN-CZ transfer with original syntax

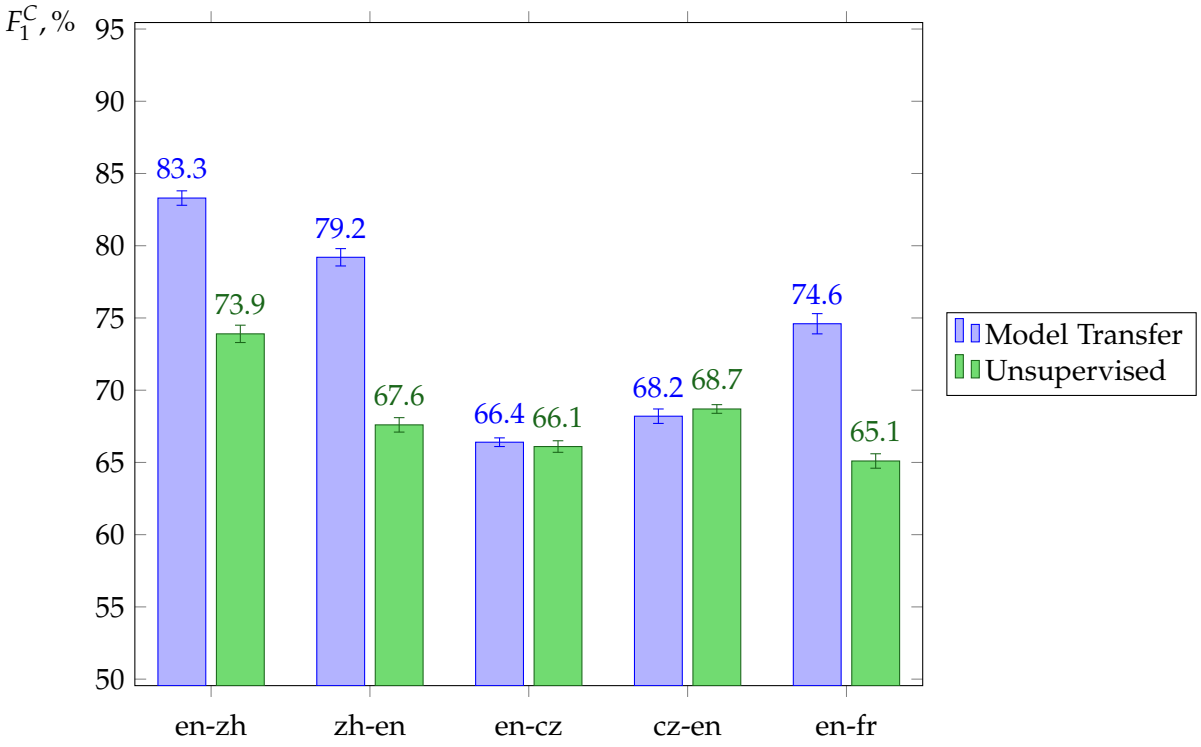


Figure 4.6: Argument classification, transferred model vs. unsupervised baseline, transferred syntax.

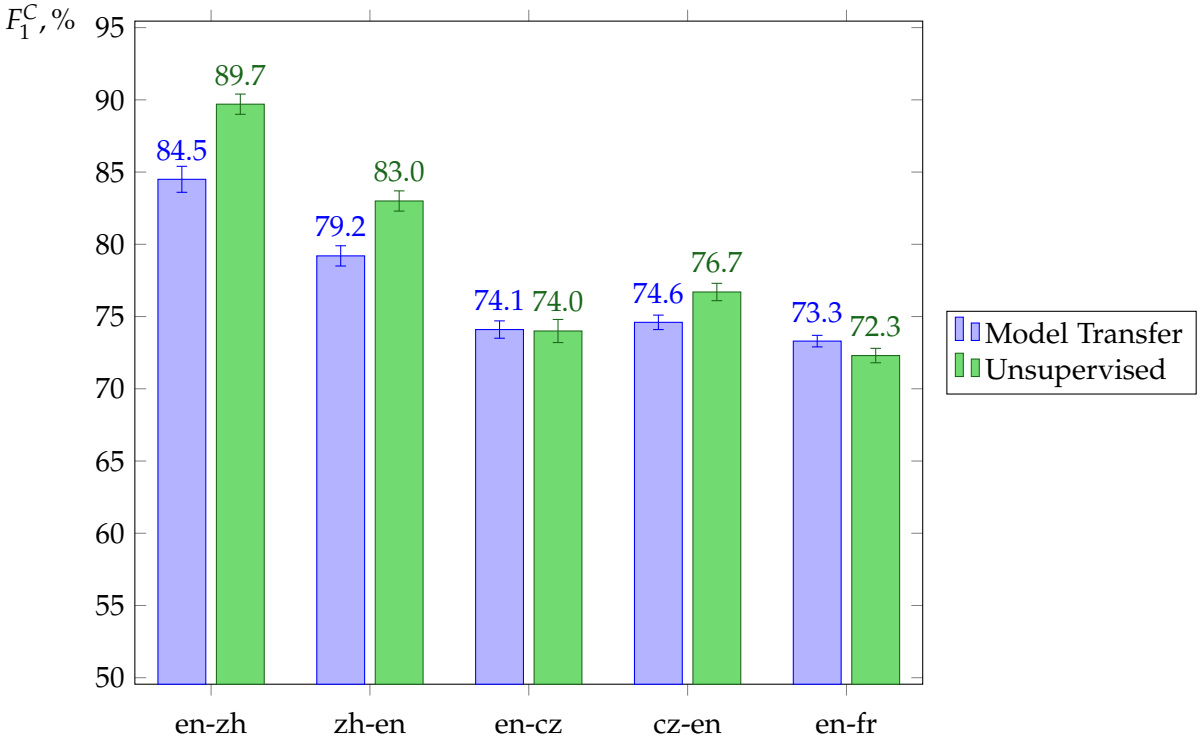


Figure 4.7: Argument classification, transferred model vs. unsupervised baseline, original syntax.

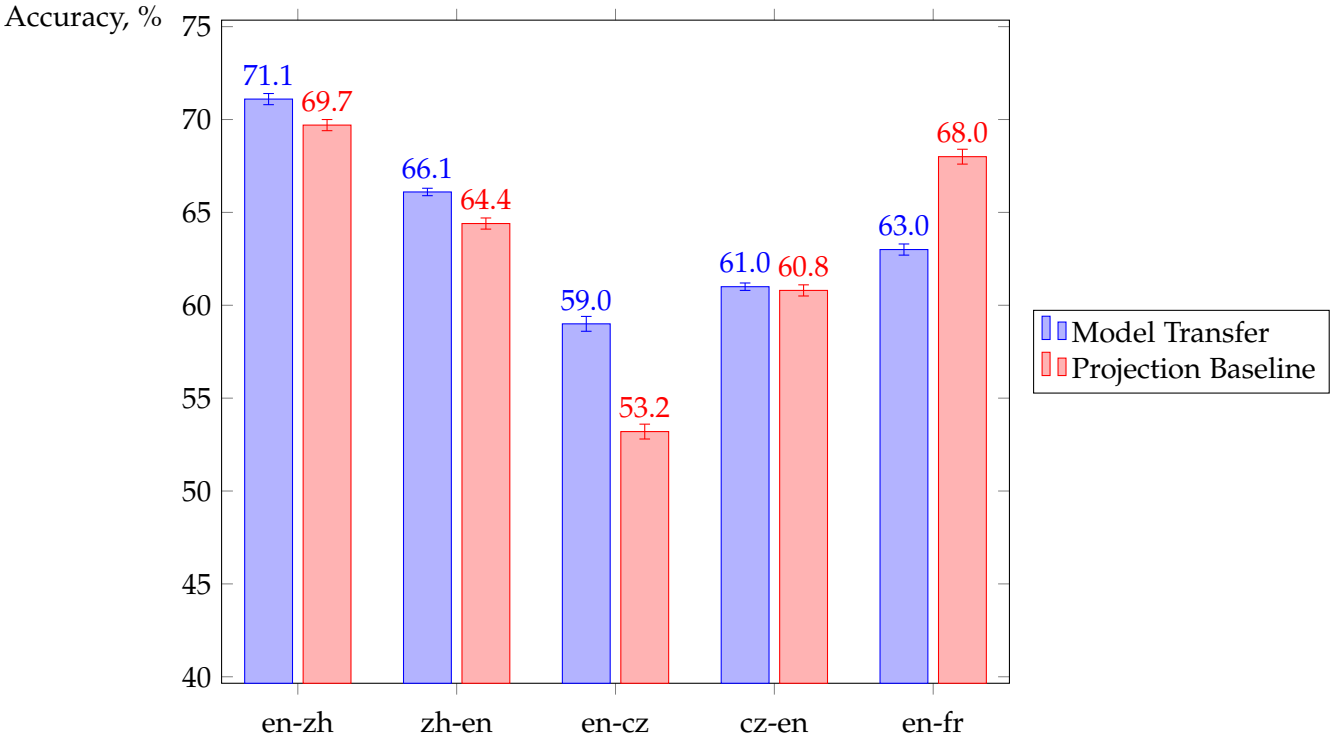


Figure 4.8: Argument classification with original syntax, supervised evaluation.

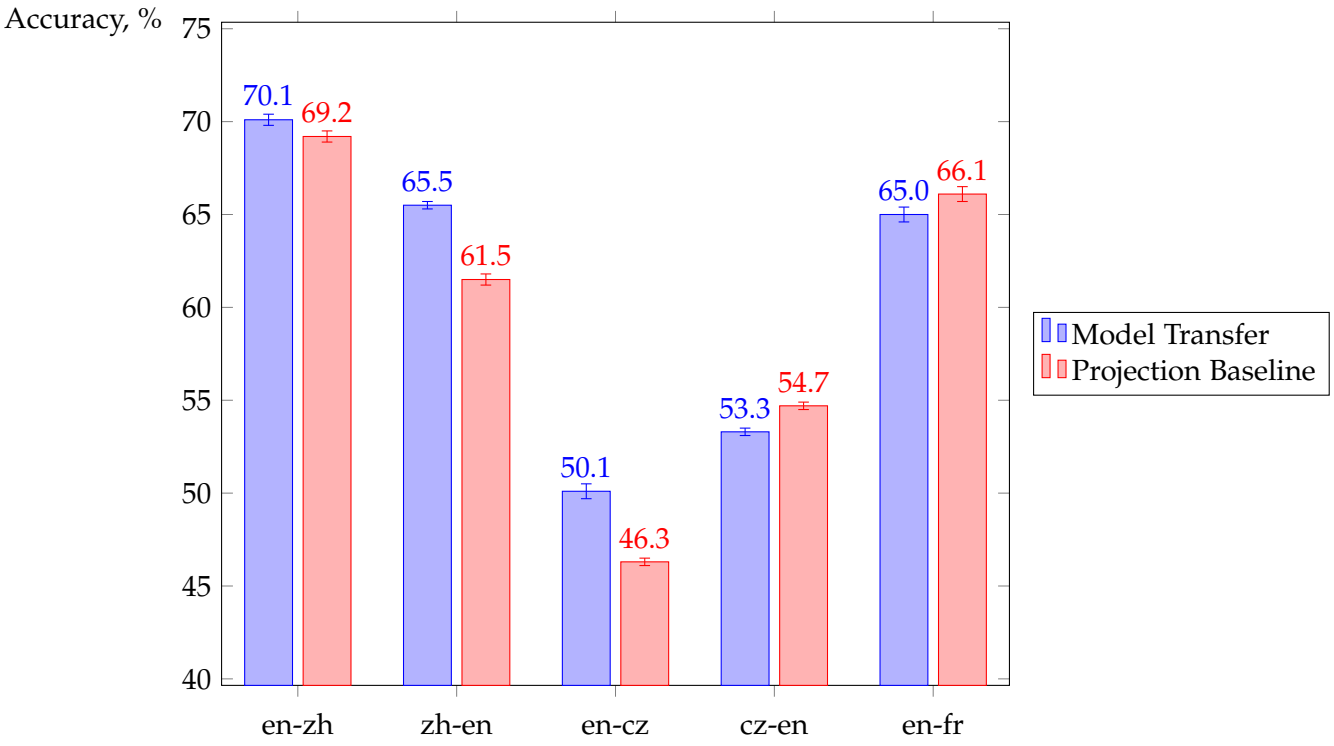


Figure 4.9: Argument classification with transferred syntax, supervised evaluation.

in figure 4.10. The model provides meaningful predictions for most of these, despite somewhat low overall accuracy.

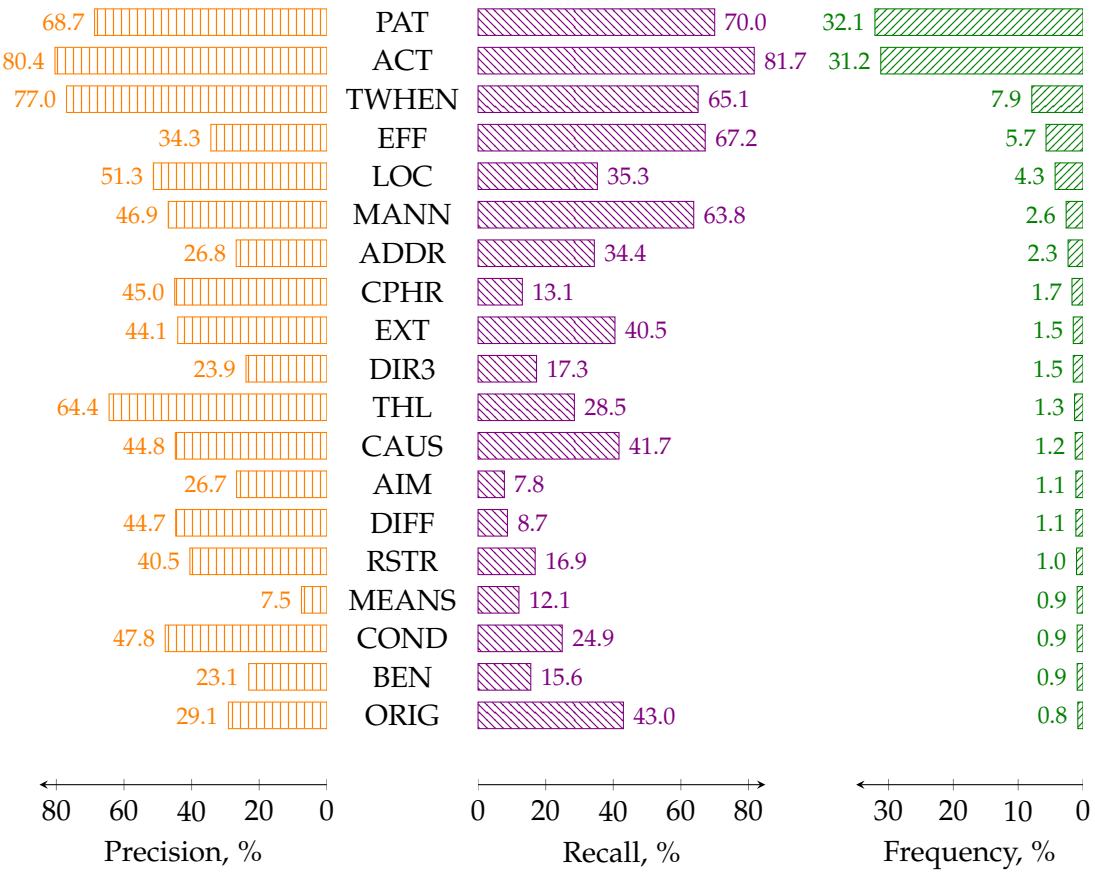


Figure 4.10: Per-role precision and recall for EN-CZ transfer (with original syntax).

Most of the labels are self-explanatory: Patient (PAT), Actor (ACT), Time (TWHEN), Effect (EFF), Location (LOC), Manner (MANN), Addressee (ADDR), Extent (EXT), Cause (CAUS), Aim (AIM), Difference (DIFF), Means (MEANS), Condition (COND), Benefactor (BEN), etc. For the less obvious ones we refer the reader to the description provided by the corpus authors, see `http://ufal.mff.cuni.cz/pcedt2.0/en/functors.html`.For example THL stands for duration (Time, How Long), CPHR marks the nominal part of a complex predicate, as in "to have [a baby]$_{CPHR}$", DIR3 indicates destination, and ORIG indicates origin in non-locative sense – a source of something or a material something is made of.

4.5.3 Additional Experiments

We now evaluate the contribution of different aspects of the feature representation to the performance of the model. Figure 4.11 contains the results for English-French.

The fact that the model performs slightly better with transferred syntax may be explained by two factors. Firstly, as we already mentioned, the original syntactic annotation is also produced automatically. Secondly, in the model transfer setup it is more important how closely the syntactic-semantic interface on the target side resembles that on the source side than how well it matches the "true" structure of the target language, and in

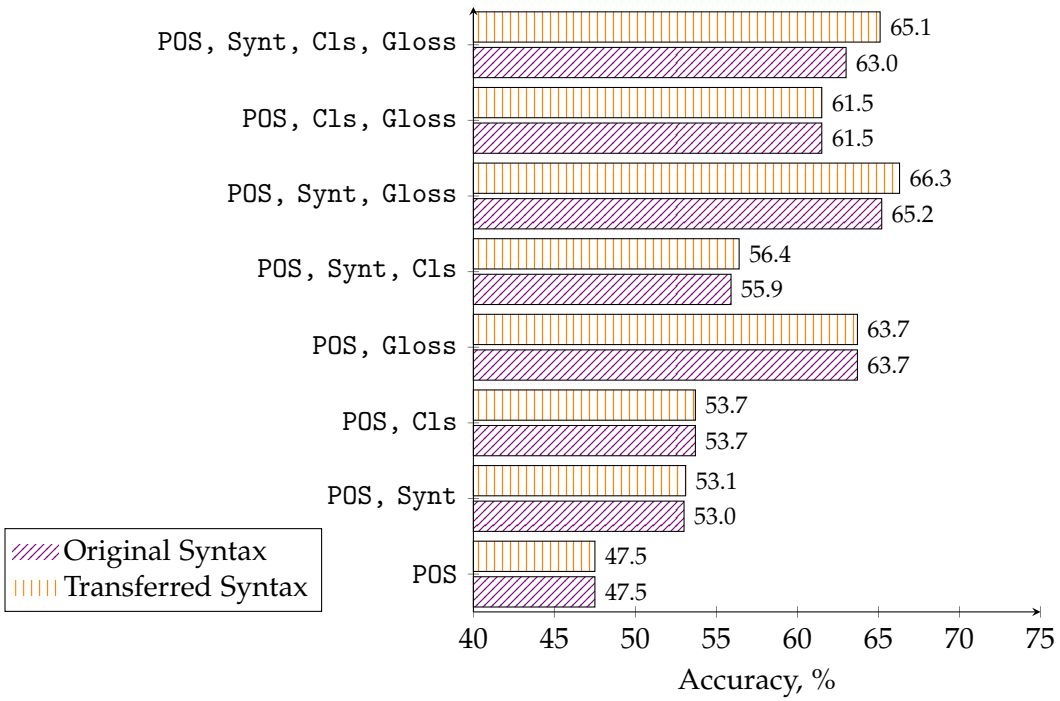


Figure 4.11: EN-FR model transfer accuracy with different feature subsets, using original and transferred syntactic information.

this respect a transferred dependency parser may have an advantage over one trained on target-language data.

The high impact of the `Gloss` features here may be partly attributed to the fact that the mapping is derived from the same corpus as the evaluation data – Europarl (Koehn, 2005) – and partly by the similarity between English and French in terms of word order, usage of articles and prepositions. The moderate contribution of the cross-lingual cluster features are likely due to the insufficient granularity of the clustering for this task.

For more distant language pairs, the contributions of individual feature groups are less interpretable, so we only highlight a few observations. First of all, both EN-CZ and CZ-EN benefit noticeably from the use of the original syntactic annotation, including dependency relations, but not from the transferred syntax, most likely due to the low syntactic transfer performance. Both perform better when lexical information is available, although the improvement is not as significant as in the case of French – only up to 5%.

The situation with Chinese is somewhat complicated in that adding lexical information here fails to yield an improvement in terms of the metric considered. This is likely due to the fact that we consider only the core roles, which can usually be predicted with high accuracy based on syntactic information alone.

4.6 Conclusion

We have considered the cross-lingual model transfer approach as applied to the task of semantic role labeling and observed that for closely related languages it performs comparably to annotation projection approaches. It allows one to quickly construct an

SRL model for a new language without manual annotation or language-specific heuristics, provided an accurate model is available for one of the related languages along with a certain amount of parallel data for the two languages. While annotation projection approaches require sentence- and word-aligned parallel data and crucially depend on the accuracy of the syntactic parsing and SRL on the source side of the parallel corpus, cross-lingual model transfer can be performed using only a bilingual dictionary.

Unsupervised SRL approaches have their advantages, in particular when no annotated data is available for any of the related languages and there is a syntactic parser available for the target one, but the annotation they produce is not always sufficient. In applications such as Information Retrieval it is preferable to have precise labels, rather than just clusters of arguments, for example.

Also note that when applying cross-lingual model transfer in practice, one can improve upon the performance of the simplistic model we use for evaluation, for example by picking the features manually, taking into account the properties of the target language. Domain adaptation techniques can also be employed to adjust the model to the target language.

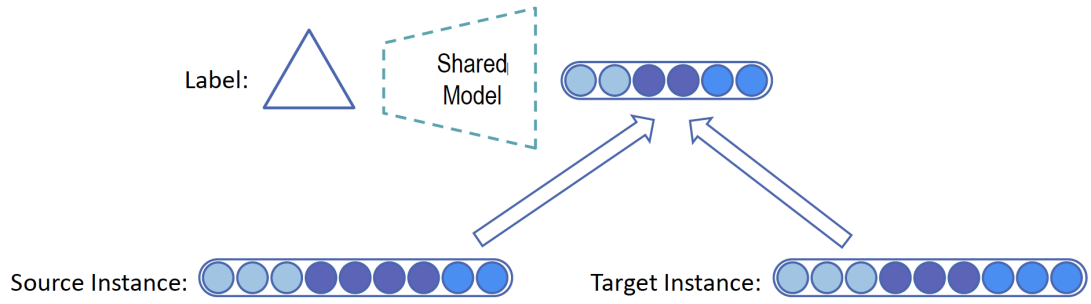
Chapter 5

Feature Representation Projection

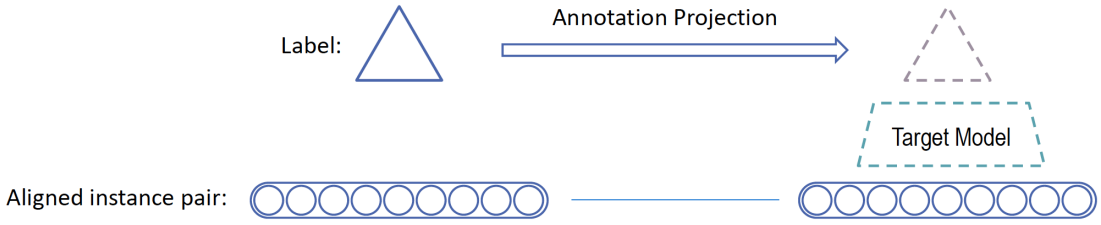
5.1 Motivation

Let us recapitulate our discussion of cross-lingual transfer methods in sections 2.3.3 and 2.3.4.

Direct model transfer attempts to find a shared feature representation for samples from different languages, usually generalizing and abstracting away from language-specific representations. Once this is achieved, instances from both languages can be mapped into this space and a model trained on the source-language data directly applied to the target language. If parallel data is available, it can be further used to enforce model agreement on this data to adjust for discrepancies between the two languages, for example by means of *projected transfer* (McDonald et al., 2011), or refined using heuristics or linguistically-motivated constraints.



The shared feature representation depends on the task in question, but most often each aspect of the original feature representation is handled separately. Word types, for example, may be replaced by cross-lingual word clusters (Täckström et al., 2012) or cross-lingual distributed word representations (Klementiev et al., 2012). Part-of-speech tags, which are often language-specific, can be converted into universal part-of-speech tags (Petrov et al., 2012) and morpho-syntactic information can also be represented in a unified way (Zeman et al., 2012; McDonald et al., 2013; Tsarfaty, 2013). Unfortunately, the design of such representations and corresponding conversion procedures is by no means trivial.

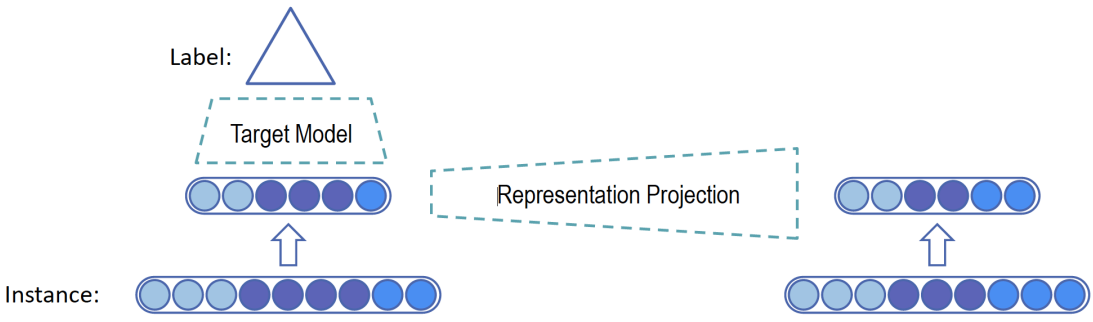


Annotation projection, on the other hand, does not require any changes to the feature representation. Instead, it operates on translation pairs, usually on sentence level, applying the available source-side model to the source sentence and transferring the resulting annotations through the word alignment links to the target one. The quality of predictions on source sentences depends heavily on the quality of parallel data and the domain it belongs to (or, rather, the similarity between this domain and that of the corpus the source-language model was trained on). The transfer itself also introduces errors due to translation shifts (Cyrus, 2006) and word alignment errors, which may lead to inaccurate predictions. These issues are generally handled using heuristics and filtering, for example based on alignment coverage (Padó and Lapata, 2006; van der Plas et al., 2011).

5.1.1 Approach

In this chapter we propose an alternative approach to cross-lingual transfer (Kozhevnikov and Titov, 2014), which we will refer to as *feature representation projection* (FRP), based on representation learning. FRP occupies a middle ground between annotation projection and direct transfer and can be seen as a compromise between the two.

It is similar to direct transfer in that it also relies on a shared feature representation. Instead of designing this representation manually, however, we create compact monolingual feature representations for source and target languages separately and automatically estimate the mapping between the two from parallel data.



This allows us to make use of language-specific annotations and account for the interplay between different types of information. For example, a certain preposition attached to a token in the source language might map into a morphological tag in the target language, which would be hard to handle for traditional direct model transfer other than using some kind of refinement procedure involving parallel data. Note also

that any such refinement procedure applicable to direct transfer would likely work for FRP as well.

Compared to annotation projection, our approach may be expected to be less sensitive to parallel corpus size and quality, since we do not have to commit to a particular prediction on a given instance from parallel data. We also believe that FRP may profit from using other sources of information about the correspondence between source and target feature representations, such as dictionary entries, and thus have an edge over annotation projection in those cases where the amount of parallel data available is limited.

5.2 Evaluation

Similarly to chapter 4, we evaluate the proposed approach on the task of dependency-based semantic role labeling (SRL) (Hajič et al., 2009), focusing specifically on the role assignment step cast as a multiclass classification problem.

Following the work described in the preceding chapter, we conducted a more detailed analysis of the evaluation setup and refined it accordingly here.

5.2.1 Cleaning up Parallel Data

We found that many of the errors we observed were due to low-quality syntactic annotation of the parallel data and certain issues with the parallel corpus itself. In particular, we found that the target side of the parallel corpus contained a small number of sentences in other languages, e.g. for English-Czech including English, French, Italian, Spanish, Greek, etc. The syntactic analysis and SRL annotations were also found to be rather inaccurate for certain types of sentences that the model may not have been exposed to in the training data, specifically direct speech, remarks in parentheses, sentences lacking terminal punctuation. Some examples:

- Žádost o ochranu parlamentní imunity: viz zápis
  - Request for the defence of parliamentary immunity: see Minutes
- (‘ συνεδρίαση αρχίζει στις 9.40 π.μ.)
  - (“ meeting will begin at 9:40 a.m.)
- (NL) Prezident Obama dodržel slovo.
  - (NL) President Obama kept his word.

Some of these cases have been eliminated using a small set of regular expressions, keeping only sentences starting with a capital letter and ending in a valid terminal punctuation mark and eliminating anything that contains characters from other character sets.

The latter is a very simple heuristic, of course, since character sets of many European languages overlap significantly. Indeed, the Czech one is a strict superset of the English one. A better way to handle language mixing would be to employ a statistical language identification model (Cavnar et al., 1994), but even that may not work perfectly, especially for single sentences (Baldwin and Lui, 2010). Fortunately, language mixing is only occasional in the dataset in question and its impact is limited.

5.2.2 Evaluation Protocol

Our intuition is that given a parallel data set of sufficient size and quality, one may expect methods that rely on parallel data to eventually achieve superior performance. Therefore in this chapter we make a set of experiments with subsets of the parallel data of different size, namely 1, 2, 5, 10, 20, 50 and 100 thousand aligned argument instances and plot the performance of the models depending on it. To ensure reliability, the models are evaluated on 20 random samples of parallel data of each size and averaged over those.

5.2.3 Language Pairs

In addition to the language pairs used in the direct transfer experiments, here we include English-Spanish as well. The monolingual corpus for Spanish is derived from AnCora (Taulé et al., 2008), which has a more fine-grained set of labels than PropBank – each core argument is marked with both a numeric role similar to the PropBank ones and a thematic role. The two are concatenated and generally treated as a unit, e.g. `arg1-pat` (core argument A1, Patient) or `argM-loc` (modifier, Location). A simple mapping A.1 is applied to the Spanish corpus to make the role sets compatible.

Note that for English-Chinese, we also include the non-core roles, which are, though named differently, appear to be largely compatible. The role labels are mapped into a shared inventory according to the semi-automatically constructed mapping in table A.2.

5.3 Approach

We consider a pair of languages (L^s, L^t) and assume that an annotated training set $D_T^s = \{(x^s, y)\}$ is available in the source language as well as a parallel corpus of instance pairs $D^{st} = \left\{\left(x^s, x^t\right)\right\}$ and a target dataset $D_E^t = \left\{x^t\right\}$ that needs to be labeled.

Intermediate compact monolingual feature representations ω_1^s and ω_1^t are defined, as well as transformations M_s and M_t to map source and target samples x^s and x^t from their original representations, ω_0^s and ω_0^t, into the intermediate representations.

The aligned instances found in the parallel data are transformed into the intermediate feature representations

$$\bar{D}^{st} = \left\{\left(x_1^s, x_1^t\right)\right\} = \left\{\left(M_s(x^s), M_t(x^t)\right)\right\},$$

and the transformation between the source and target intermediate feature representations M_{ts} is learned from the resulting set of pairs.

$$M_{ts} = argmax_M \sum_{(x_1^s, x_1^t \in \bar{D}^{st})} \left\| x_1^s - M(x_1^t) \right\|_2$$

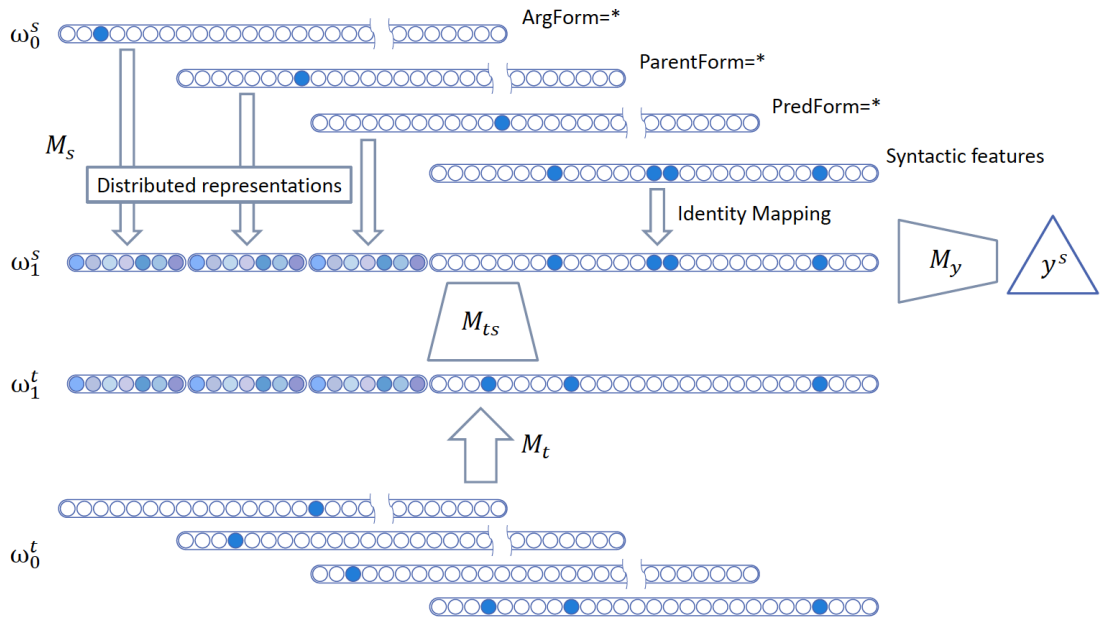
A classification model M_y is trained on the source-language training data, taking source-side intermediate representation as input:

$$\bar{D}_T^s = \{(x_1^s, y)\} = \{(M_s(x^s), y), (x^s, y) \in D_T^s\}.$$

The target language model M_y' is then produced as a composition of M_y, M_{ts} and M_t, so the label for a target-language instance $x^t \in D_E^t$ can be obtained as

$$y^t = M_y(M_{ts}(M_t(x^t))).$$

The following diagram illustrates the approach graphically.



5.3.1 Feature Representation

Our objective is to make the intermediate feature representation sufficiently compact that the mapping between source and target feature spaces could be reliably estimated from a limited amount of parallel data, while preserving, insofar as possible, the information relevant for classification.

We start out with a generic feature set, covering the properties – word type, part-of-speech tag, dependency relation to the respective head and morphological tags, if any – of the argument head token, the predicate token and the parent token of the argument in the dependency tree. Following prior works, a position feature, indicating whether a given argument appears before or after the predicate, is also included.

Compared to the features used in direct model transfer experiments in chapter 4, we omit the features involving the argument's siblings in the dependency tree and the tokens preceding and following the argument token. This is done for technical reasons, namely to limit the dimensionality of the input feature vector.

This feature representation, ω_0, is sparse – using one-hot encoding the feature vectors in our experiments have between 20 and 60 thousand components. Estimating the mapping directly between ω_0^s and ω_0^t is both computationally expensive and likely inaccurate – using one-hot encoding the feature vectors in our experiments would have between 20 and 60 thousand components. Fortunately, there is a number of ways to make this representation more compact. We start with the most obvious one, namely representing word types with neural word embeddings produced by the skip-gram model (Mikolov et al., 2013a). This corresponds to M_s and M_t above and significantly reduces the dimension of the feature space, making direct estimation of the mapping practical. We will refer to this representation as ω_1.

In the present work we keep all syntactic features in one-hot encoding, since they are comparatively few – Czech has the most at just over three hundred, mostly due to the morphological tags – and to our knowledge there are no readily available ways to make this information even more compact.

5.3.2 Learning Better Monolingual Representations

To go further, one can, for instance, apply dimensionality reduction techniques to obtain a more compact intermediate representation by eliminating redundancy, or define auxiliary tasks and produce a vector representation useful for those tasks. One may speculate that this would be especially helpful if the amount of monolingual data available for either language significantly exceeds the size of the available parallel corpus. We have conducted initial experiments using simple neural autoencoders with and without additional hidden layers and with different activation functions, albeit without introducing additional monolingual corpora into the equation, but were unable to attain significant dimensionality reduction without a noticeable loss in informativeness. It is possible that more recent deep learning techniques could yield better results here.

In general, the intermediate representation needs only retain the information that is important for the particular task. For the source language it is possible to train a supervised model which would rely on a compact latent representation (such as a hidden layer in a network) and reuse that latent representation as ω_1 in the present approach.

For the target language we assume no SRL-annotated training data to be available, but one may consider using other, related tasks to derive a useful intermediate representation, or even use a simple context-prediction setup similar to the skip-gram model. The shift towards learning reusable dense vector representations of various language units was shown to benefit many NLP applications, and such representations can be derived for predicates and arguments in the context of semantic role labeling as well. In fact, such are already produced as a side-product of unsupervised semantic role labeling (Woodsend and Lapata, 2015; Luan et al., 2016).

5.3.3 Baselines

As mentioned above we compare the performance of this approach to that of direct transfer and annotation projection baselines.

The shared feature representation for direct transfer is derived from ω_0 by replacing language-specific part-of-speech tags with universal ones (Petrov et al., 2012) and adding cross-lingual word clusters (Täckström et al., 2012) to word types. The word types themselves are left as they are in the source language and replaced with their gloss translations in the target one (Zeman and Resnik, 2008). For reference, we also provide the score of the direct transfer model relying on transferred, rather than gold syntax, produced as described in section 4.4.4.

The annotation projection baseline implementation is straightforward. The source-side instances from a parallel corpus are labeled using a classifier trained on source-language training data and transferred to the target side. The resulting annotations are then used to train a target-side classifier for evaluation.

5.3.4 Word Embeddings

The 250-dimensional monolingual word representations for ω_1 are obtained using WORD2VEC tool applied to all available monolingual data – a concatenation of the labeled corpus and the relevant side of the parallel corpus. In Mikolov et al. (2013b) the authors consider embeddings of up to 800 dimensions, but we would not expect to benefit as

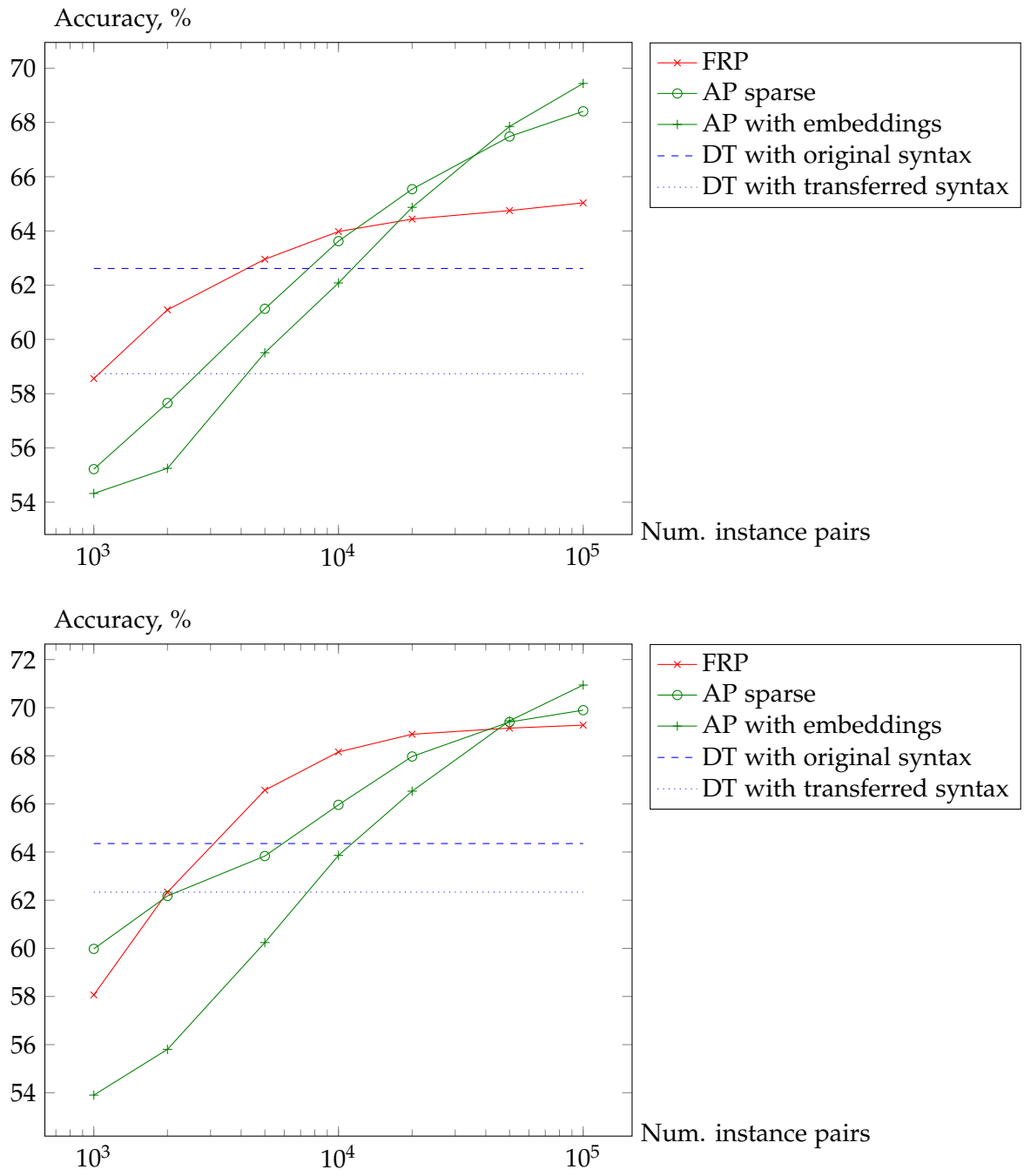


Figure 5.1: Results for English-Czech (top) and Czech-English (bottom).

much from larger vectors since we are using a much smaller corpus to train them.

A possible extension of the proposed model would be to use cross-lingual word embeddings produced by one of the many promising models that became available recently (Gouws et al., 2015; Lauly et al., 2014).

5.3.5 Projection Model

The projection model, M_{ts}, maps a dense vector representation of instances from the target language into a similar dense vector representation for the source language. In

other words, one would like to learn a transformation between the two vector spaces from the available aligned instances, which would generalize well to unseen examples at decoding time. Depending on the type of transformation chosen, a variety of methods could be employed to learn the parameters. We decided in favor of a neural network for its flexibility and used the PYLEARN2 toolkit (Goodfellow et al., 2013), based on THEANO (Bergstra et al., 2010), to conduct the experiments.

Having considered networks with up to two hidden layers and a number of different training regimes, we observed that learning a non-linear mapping between two spaces with up to 1000 dimensions each is non-trivial, especially when presented with fewer training samples. The model ended up with very different mappings depending on the subset of parallel data presented to it, often of poor quality, therefore in the experiments below we are learning a linear mapping only.

In the initial experiments, we used the same toolkit to learn the source-side classification model M_y and to decode on the target side. Unfortunately, at the time the support for sparse features was not available – a dense matrix was generated based on the inputs – which made it impossible to run the baseline experiments without truncating the feature vocabulary to a few thousand items, which may be seen as unfair towards the baselines using a sparse input representation. Consequently, here train the classification model and decode using LIBLINEAR (Fan et al., 2008) instead, i.e. using an L_2-regularized max-margin classifier capable of handling sparse feature spaces.

Another concern regarding the results of the initial experimental presented in Kozhevnikov and Titov (2014) had to do with how much simply using the word embeddings rather than one-hot encoding would contribute towards the performance of the model. To evaluate this, we introduce an additional baseline, same as the regular annotation projection baseline, but relying on the word embeddings that FRP model uses.

We observe no consistent improvement from using the embeddings in the annotation projection model, possibly due to the limited quality of the embeddings themselves – the training corpus for them is comparatively small (3 to 8 million tokens depending on the language) and dominated by the Europarl data belonging to a highly specialized register.

5.3.6 Data

The evaluation is performed on four different language pairs, English-Czech, English-Chinese, English-Spanish and English-French. The labeled data is drawn from the CoNLL 2009 shared task (Hajič et al., 2009) dataset and, in the case of Czech, from the Prague Czech-English Dependency Treebank 2.0 (Hajič et al., 2012). For French, the manually corrected evaluation set from van der Plas et al. (2011) is used as the test set. Since this is comparatively small – one thousand sentences – we reserve the whole dataset for testing and only evaluate transfer from English to French, not the other way around. For other languages, 30 thousand labeled instances are reserved for testing and the rest is used as training data.

Similar to the previous chapter, the parallel data is drawn from Europarl (Koehn, 2005) and MultiUN (Eisele and Chen, 2010), aligned using GIZA++ (Och and Ney, 2003) and pre-processed using MATE-TOOLS (Björkelund et al., 2009; Bohnet, 2010).

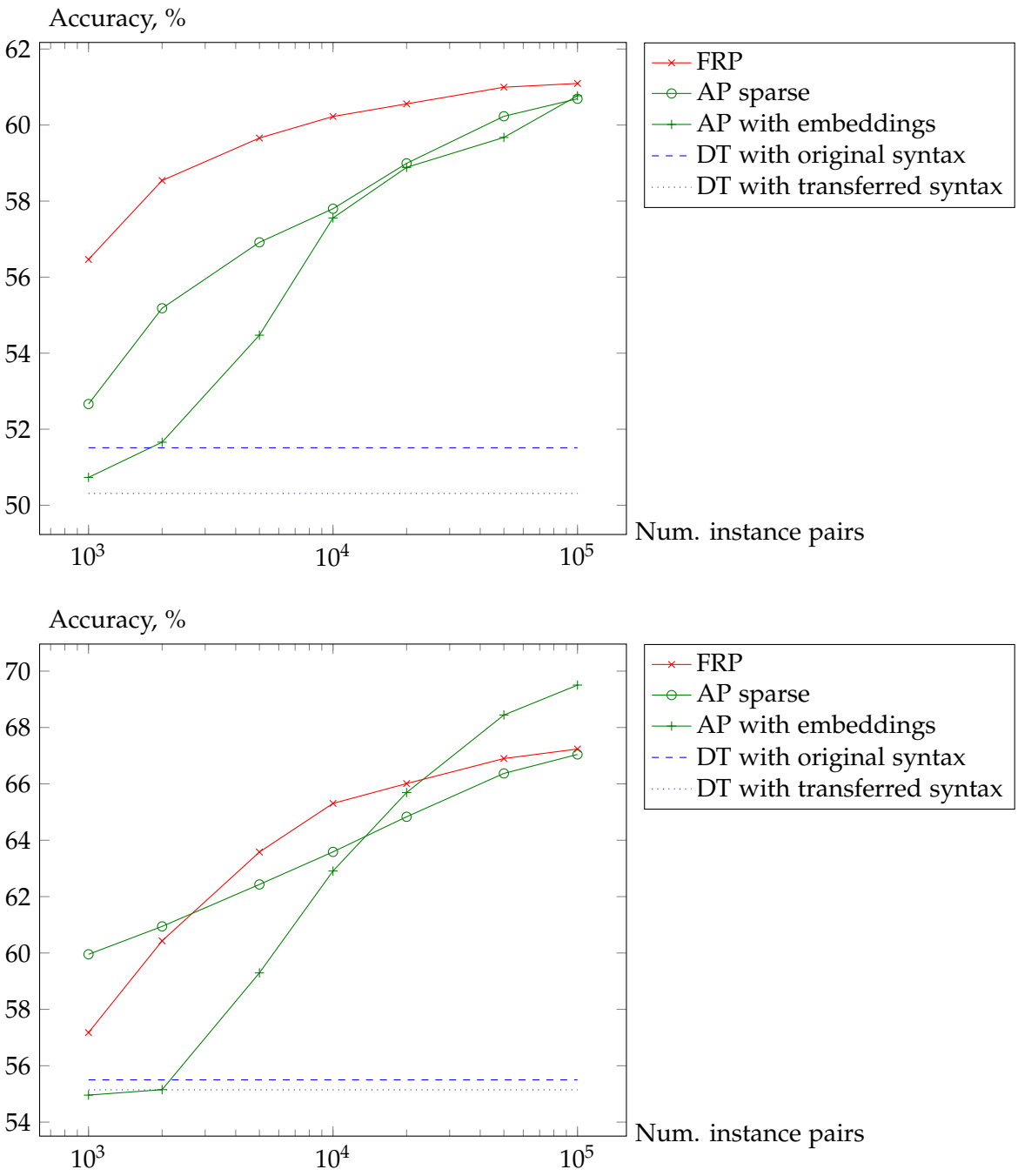


Figure 5.2: Results for English-Spanish (top) and Spanish-English (bottom).

5.4 Results

The accuracy of the proposed model and the baselines given varying amount of parallel data is reported in figures 5.1-5.4. The training set for each language is fixed. We denote the two baselines AP (annotation projection), using sparse word representation or word embeddings and DT (direct transfer), using gold or transferred syntax.

The number of parallel instances in these experiments is shown on a logarithmic scale, the actual values considered are 1, 2, 5, 10, 20, 50 and 100 thousand pairs.

Note that we report only a single value for direct transfer, since this approach does not

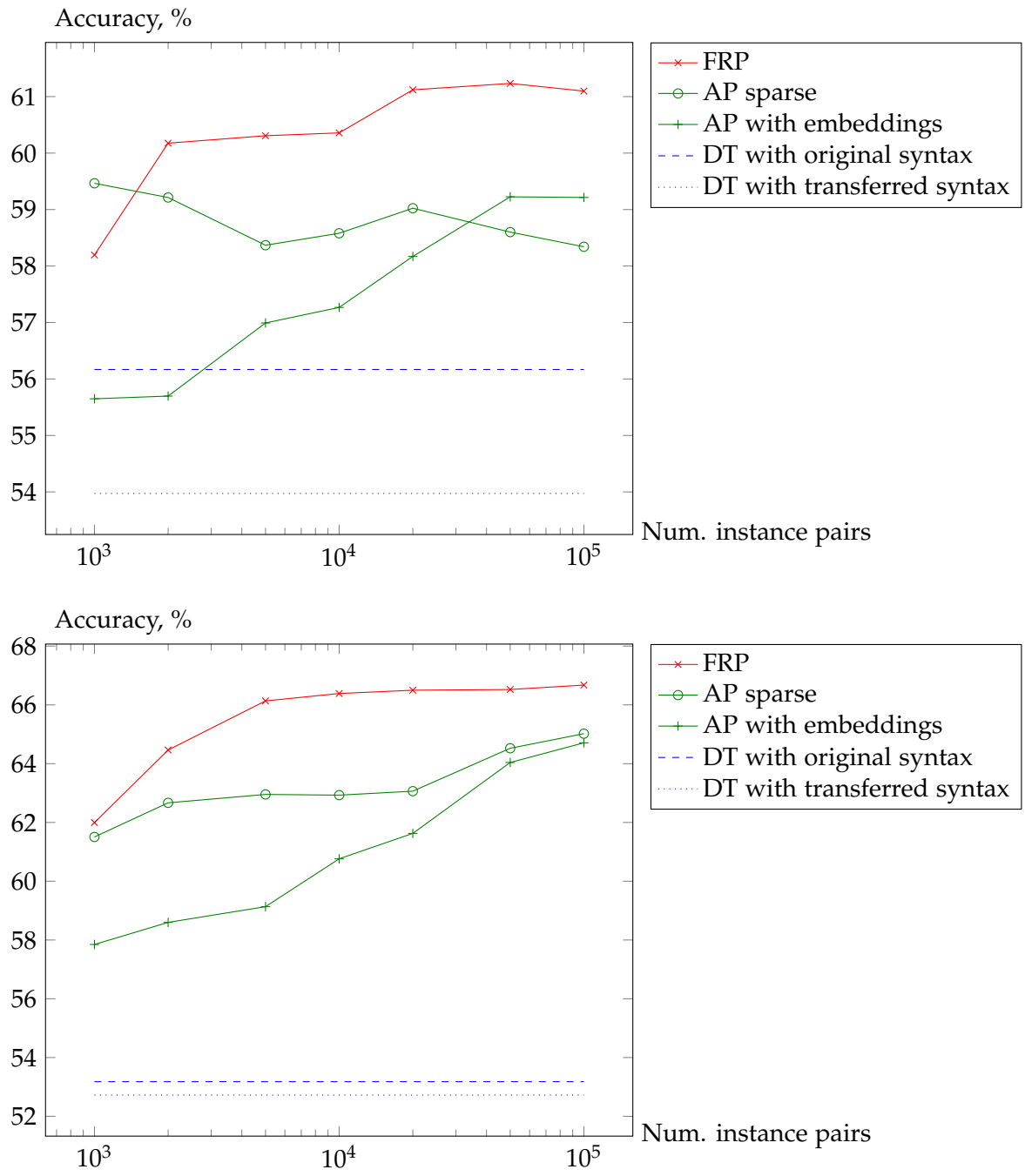


Figure 5.3: Results for English-Chinese (top) and Chinese-English (bottom).

explicitly rely on parallel data. Although some of the features – namely, gloss translations and cross-lingual clusters – used in direct transfer are, in fact, derived from parallel data, they generally do not require well-formed parallel sentences and can potentially be learned from alternative sources.

We observe that the performance of both annotation projection and FRP improves with the size of the parallel corpus, as expected, but seems to get flatter as we approach the maximum available parallel dataset size. It would be natural to expect it to plateau eventually, but for FRP in particular this could also be caused by the intentionally limited, fixed capacity of the model, which may explain the superior performance of the

annotation projection baseline given larger parallel datasets in figure 5.1. We further note that, as we conjectured, both annotation projection and FRP tend to outperform direct transfer given a sufficient amount of (cleaned-up) parallel data.

Given a very large parallel corpus one may benefit from using higher-dimensional embeddings or additional features in the source and target representations.

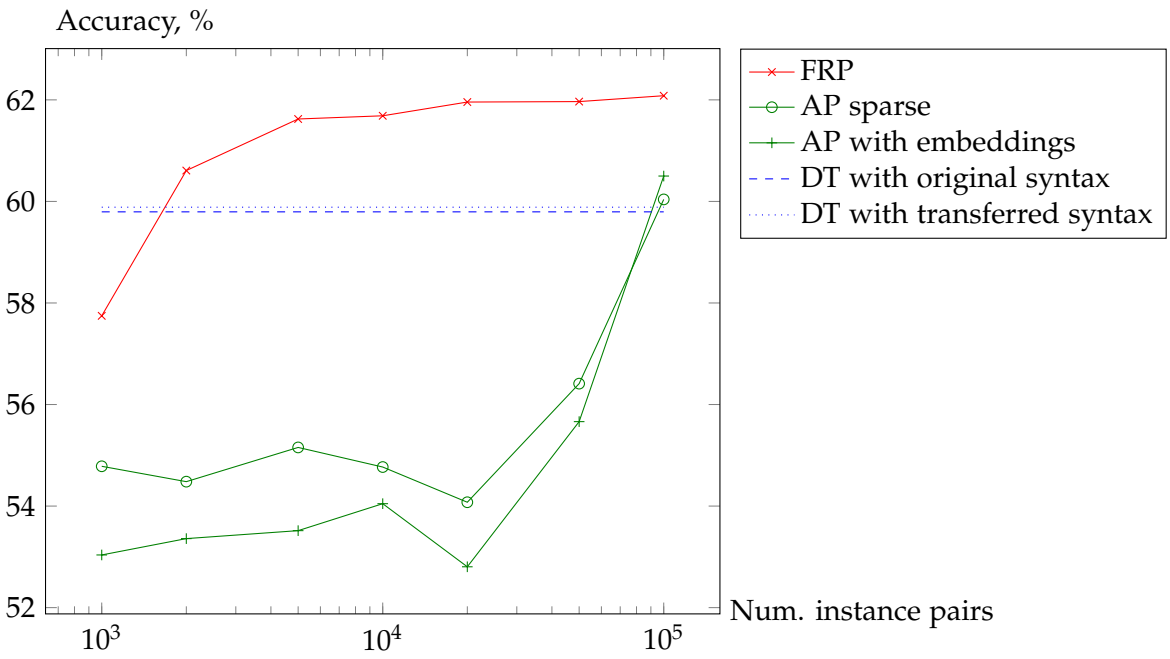


Figure 5.4: Results for English-French.

Consistently with previous experiments, the direct transfer approach tends to yield better results when relying on the original unlabeled syntactic dependency information, compared to the dependencies obtained by direct transfer.

Note also that the comparative performance of the two variants of annotation projection – the one using sparse word type features and the one using word embeddings – varies considerably between language pairs and the sparse version seems to performs slightly better on average. It is therefore safe to assume that the performance differences between FRP and the regular annotation projection baseline cannot be attributed solely to the use of word embeddings in the former.

5.5 Conclusions

We observe that the performance of this method is competitive with that of established cross-lingual transfer approaches and its application requires very little manual adjustment – particularly no explicit shared feature representation design. It may have some advantage when presented with a limited amount of parallel data, or parallel data of poor quality.

We have demonstrated the general viability of the approach, but there are many directions to explore further.

The real test for any cross-lingual transfer approach is to apply it in a real-world annotation scenario involving an actual resource-poor language. Such studies do exist,

but they generally focus on a single transfer method and the results across different studies are quite hard to compare.

5.6 Possible Extensions

A primary argument if favor of the proposed approach is its flexibility, allowing for many possible improvements.

In addition to the already mentioned variations using pre-learned monolingual representations in section 5.3.2, one can consider adapting existing techniques from both annotation projection and direct transfer, many of which will be applicable to this problem. Certain filters commonly used in annotation projection can be applied here directly, for instance filtering out sentence pairs where some predicates are unaligned.

If little parallel data is available, one can simplify the learning by leveraging techniques from direct transfer, for instance converting the dependency annotations on both sides to universal dependencies (McDonald et al., 2013) or HamleDT-style dependency annotation (Zeman et al., 2014).

In structured prediction tasks, such as dependency parsing (or semantic role labeling, should one take into account the inter-argument constraints), a technique similar to projected transfer (McDonald et al., 2011) could be of use.

Finally, one can adapt FRP to the multi-source transfer scenario, should compatible resources be available for multiple source languages, although in this case a separate projection model will be necessary for each language pair.

5.6.1 Alternative Sources of Information

The amount of parallel data available for many language pairs is growing steadily. However, cross-lingual transfer methods are often applied in cases where parallel resources are scarce or of poor quality and must be used with care. In such situations an ability to use alternative sources of information may be crucial. Potential sources of such information include dictionary entries or information about the mapping between certain elements of syntactic structure, for example a known part-of-speech tag mapping.

The available parallel data itself may also be used more comprehensively – aligned arguments of aligned predicates, for example, constitute only a small part of it, while the mapping of vector representations of individual tokens is likely to be the same for all aligned words, so we could start by learning a cross-lingual word embedding model (Gouws et al., 2015; Hermann and Blunsom, 2013; Luong et al., 2015) on all aligned tokens, and use it to initialize the transformation model.

Chapter 6

Learning from Agreement in Monolingual Setting

This work started out with the idea of improving semantic parsing models by learning from similarity or *agreement* between instances, specifically considering the semi-supervised learning scenario where in addition to the annotated training data one has access to a number of groups of instances which are known to have compatible interpretations.

Much of human knowledge, whether absolute or transient, is put into words many times by different people. Consider for instance (1) news, where a given event is observed and described by multiple reporters, then picked up and reformulated again by social media, (2) scientific or educational information, which needs to be presented to different audiences at different level of detail, or (3) instructions such as cooking recipes, which get passed on from person to person with little or no modification, but likely described slightly differently every time.

The variations in the wording may sometimes appear trivial to a human reader, since it is generally obvious to them whether or not two fragments of text express the same meaning, but for machines – and sometimes human readers with imperfect command of the language, such as non-native speakers – recognizing paraphrases is a very hard problem (Dolan et al., 2004; Blacoe and Lapata, 2012). In second language learning, reading or listening comprehension is often evaluated based on the learner's ability to recognize textual entailment, of which paraphrasing is a special case.

Taking the parallel with language acquisition further, when a learner expresses lack of understanding of a certain statement, our first impulse is to rephrase the statement in different, simpler terms, so paraphrasing data is a natural source of information for semantic parsing models.

To somewhat formalize the idea, one may consider training a model to recognize meaning representation y in a natural-language utterance x on a corpus of utterances explicitly annotated with such meaning representations $D_a = \{(x, y)\}_{1..N}$ and another corpus containing groups of utterances expressing some meaning that is unknown to us, but consistent among the utterances within a given group $D_p = \{(x_1, x_2, x_3, \ldots)\}_{1..M}$.

For classification tasks data of this sort could be incorporated into training by using a self-training-like approach with the additional constraint of assigning all items in a group the same label, for example by choosing an example within each group where model's uncertainty is lowest and considering the group to be labeled according to this

sample. For structured prediction tasks this is complicated by the fact that it is often hard to measure the model's confidence and that there can be no majority agreement – the model can produce similar, but distinct interpretations for each item in a group, without a natural way of identifying one of those interpretations as *correct*.

Besides, in practice different verbalizations will rarely have absolutely identical semantics – one would expect them to be consistent with respect to the essentials, but different people describing the same entity, event or process would tend to verbalize different aspects of its semantics and may often disagree as to the finer details. Indeed, one could question if it is at all possible to have two distinct documents with exactly the same semantics – every tiny variation in the wording might subtly alter the meaning, – and whether or not the semantics of two documents can be considered identical depends mostly on how fine-grained our model of semantics happens to be. One could say that paraphrasing is in the eye of the beholder.

In practice, given a fixed level of granularity, it would generally benefit a model to learn the different ways of verbalizing a given aspect of the semantic representation, and we believe that groups of documents with consistent hidden semantics can be used to this end.

6.1 Aligning Semantics and Verbalizations

We applied the idea of learning from agreement to the task of semantic parsing in the weather forecast domain, introduced in Liang et al. (2009).

Here the semantics are represented as a set of database records, with each field storing the expected value of a particular aspect of the weather at a given point in time, e.g. "`.id:3 .type:windDir .date:2009-02-08 .label:Tonight @time:17-30 @mode:SSW`", in this case indicating that wind from south-south west was expected. A single forecast contains 36 records in total (see example in figure A.1).

Each forecast is accompanied by a short textual description, generated from the database record, e.g.

```
Mostly cloudy ,
with a low around 30 .
South wind between 7 and 9 mph .
```

as well as an alignment between the textual description and the record, in this case

```
0 5
1 0
2 3 2,
```

indicating that the first line of the textual description corresponds to the 5th record, the second – to the 0th one, and the last line covers records 2 and 3. Note that the model only describes a subset of records and within each record some information can be omitted as well (for instance the maximum temperature in the example above).

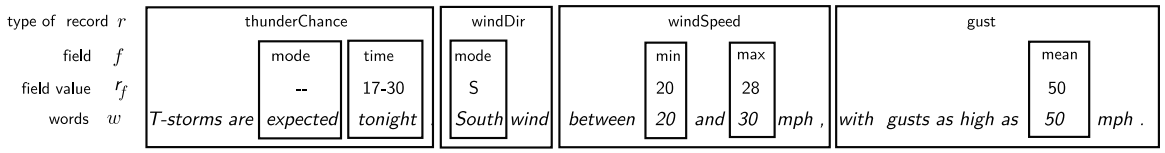


Figure 6.1: A hierarchical segmentation of a text fragment into spans corresponding to individual records and fields.

6.1.1 Approach

The structure of the domain implies a simple generative story – for each record, a subset of records is selected for verbalization and for records containing several values a subset of values is selected in addition (for instance, one can choose to only mention the peak temperature, but not the mean one). A textual description of the weather conditions is then produced, mentioning the selected values in a chosen order. We will refer to these descriptions as *documents*. Note that given the same world state, selecting different fields would result in documents that are not paraphrases of each other. Strictly speaking, we cannot guarantee that there is *any* overlap at all, only that no two documents will contradict each other if they do mention the same aspect of the world state. Hence we refer to these documents as *non-contradictory*.

Liang et al. (2009) learn semantic alignment models capable of establishing the correspondence between the textual representation and the world state without relying on explicit alignments at training time. In Titov and Kozhevnikov (2010)[1] we adopt this setup to evaluate the proposed approach: learning from both documents associated with a known world state and groups of documents corresponding to the same hidden world state.

The objective of the original model (Liang et al., 2009) was to segment a textual description, simultaneously linking each segment to the appropriate aspect of the world state (figure 6.1). Importantly, the world state would be available both at training and inference time, so it does not need to be generated. In our scenario, however, the groups of non-contradictory documents correspond to a hidden world state, which needs to be modeled in order for us to benefit from this signal. Figure 6.2 shows the generative model for this domain, covering both the world state and the textual representation.

6.1.2 Data Collection

For evaluation, we augmented the dataset with alternative weather forecasts for a small subset of records. Each rater was shown a table representation of the weather conditions (see figure 6.3) and asked to produce a description, mentioning the information of general interest, such as the expected temperature and sky cover, as well as any other aspects worth mentioning – precipitation, if expected, strong winds or unusually low or high humidity.

We collected a total of 391 descriptions from two raters using an Amazon Mechanical Turk HIT sandbox, together with the texts from the original dataset yielding 259 groups – 650 texts with an average of 2.5 texts per group.

[1]Please note that the author is not the primary contributor on that paper.

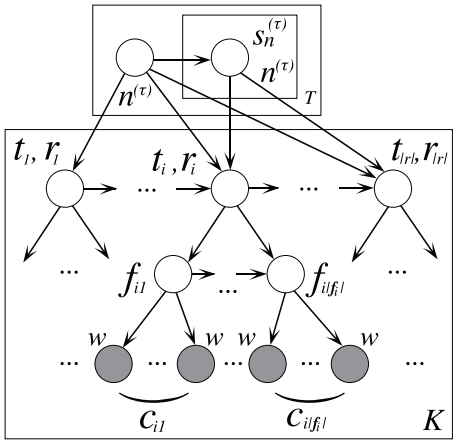


Figure 6.2: The generative model of a weather forecast $\{w\}$ given the shared world state s. t denotes a time of day, r is a record index and f is a field index inside a record.

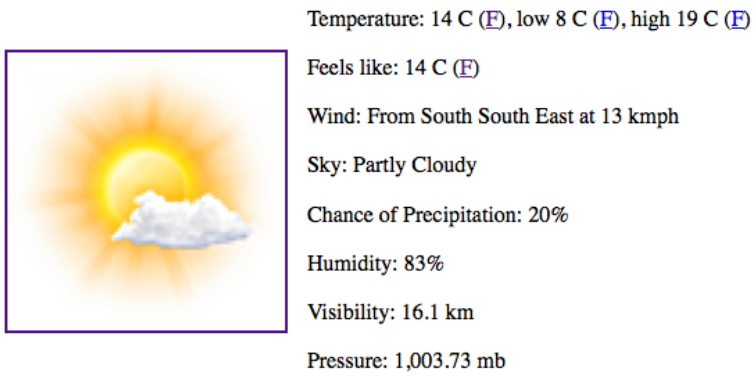


Figure 6.3: Representation of the weather conditions as shown to the annotators. For simplicity we display the conditions for a single timestamp.

The overlap in terms of the verbalized fields between documents within a group was about 35%, and 60% of fields were only mentioned in a single document. See an example in figure 6.4.

For the test set, we selected 50 documents from the original dataset and 50 more produced by each of the two raters. The newly added documents were manually aligned to the corresponding fields.

6.1.3 The Model

The proposed approach exploits the labeled data in much the same fashion as Liang et al. (2009), using Expectation Maximization (Dempster et al., 1977). For the groups of non-contradictory documents, we need to generate a world state, which is non-trivial due to the relative complexity of the semantic space.

Jointly parsing multiple documents turns out to be rather expensive computationally, so instead an iterative procedure is proposed, selecting one document at a time and

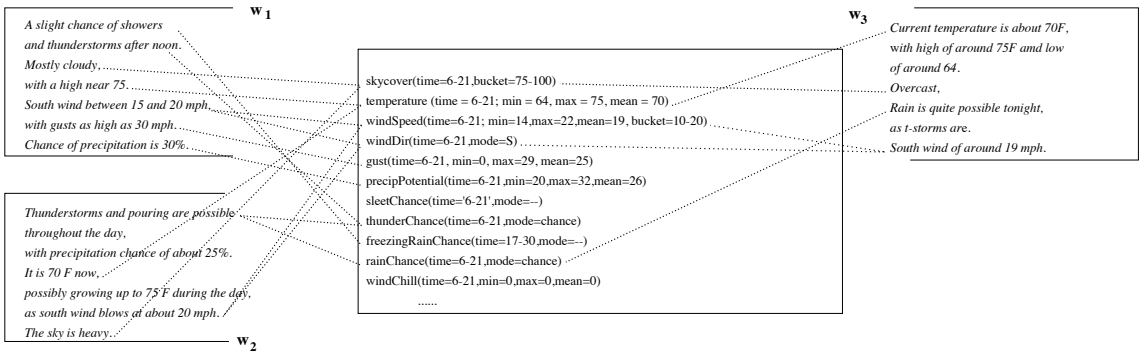


Figure 6.4: An example of three non-contradictory weather forecasts and their alignment to the semantic representation. Note that the semantic representation (the block in the middle) is not observable at training time.

inferring the part of the world state it describes. To ensure consistency, the interpretation of remaining documents is conditioned on the state inferred so far. Since the order in which the documents are parsed also affects the outcome, at each step the algorithm selects a document such that the world state following this step would maximize the probability of the remaining documents.

This approach yields a noticeable improvement in the overall scores for alignment prediction, covering 55% of the gap between the semi-supervised baseline and the supervised upper bound:

	P	R	F_1
Supervised baseline	63.3	52.9	57.6
Semi-supervised baseline	68.8	69.4	69.1
Proposed approach	**78.8**	**69.5**	**73.9**
Supervised upper bound	69.4	88.6	77.9

Qualitative evaluation also shows that the word forms unseen in labeled training data – such as "sun", "cloudiness" or "gaps" – generally get assigned to correct clusters by the model leveraging groups of non-contradictory documents, with a small number of errors caused by correlation between different aspect of the world state. For example, words such as "rainfall" or "inch" get aligned to the `sky cover` field, rather than the precipitation field, because the two are strongly correlated. However, this is likely to be largely mitigated by using larger amounts of data covering more diverse weather conditions.

6.2 Weather Forecast Parsing Experiments

Given the encouraging results described above for the semantic alignment problem, we decided to look into bootstrapping an actual parser from the same type of data, capable of recovering the world state based on a description. The fact that the alignment model can benefit from the non-contradictory document groups already implies, if indirectly, that the model is capable of inferring a reasonable approximation of the world state during training.

As mentioned already, a single weather forecast in the dataset created by Liang et al. (2009) contains 36 records, some with multiple fields, but only an average of 6 records is verbalized in any given document. The alignment scenario has a certain advantage here, in that some parts of the world state can be safely ignored, although the can make the inference more expensive computationally. Predicting the actual world state given a document that only covers one in six records is somewhat trickier, as is the subsequent evaluation. One could imagine a generative model of the world state that would condition records that have not been verbalized on those that have (e.g. if the forecast says it is going to rain it is likely that the atmospheric pressure is low), plus some sensible prior (e.g. if precipitation is not mentioned, there is likely to be none).

To mitigate this issue, for the parsing setup we reduce the world state representation to a handful of fields mentioned most often, namely the high, mean and low temperature, wind speed and direction, sky cover and chances of rain or snow.

We used a CRF (Lafferty et al., 2001) model with factors covering label unigrams and bigrams for each token, similar to a regular linear chain CRF, with additional factors connecting tokens aligned to a given field to the value of that field. The model was trained using stochastic gradient descent, with the expectations computed using a custom Metropolis-Hastings sampling procedure – sampling the world state and computing expectations for the labeling using the traditional forward-backward algorithm. At decoding time, the same sampling procedure was used in simulated annealing setting (Kirkpatrick et al., 1983) .

Unfortunately, using groups of non-contradictory documents in this setup yields only moderate gains in terms of labeling accuracy, which translate into even smaller improvements in terms of the ability to accurately predict the world state. We attribute this primarily to the fact that accurate prediction of the world state in this domain does not necessarily require a very accurate labeling of the textual representation, since only a few key tokens need to be labeled correctly to reconstruct the correct world state. In fact, for some values one may not need to refer to the document at all due to the strong correlations between the different aspects of the world state we mentioned in the preceding section. One could say the task is simply too easy.

6.3 Conclusions

The general approach of learning from documents with non-contradictory semantics can be useful for certain applications, especially where such data occurs naturally. However, most semantic parsing tasks considered at the time would use smaller, and often highly specialized types of text (e.g. NFL recaps or Robocup sportcasting (Liang et al., 2009)), so one would likely need to construct the dataset from scratch, as we did for the weather forecasts. Natural sources of such data usually contain larger documents with complex semantics, which are hard to represent formally, although one could use a similar approach in the context of shallower analysis, for instance in event extraction models (Doddington et al., 2004). In fact, some event extraction systems make use of a similar signal – automatically identified groups of documents with overlapping and presumably compatible semantics – in a fully unsupervised setting (Krause et al., 2015).

Based on these considerations we decided to shift focus to transfer- and agreement-based learning for shallower semantics, namely semantic role labeling (SRL), which

became the subject of this work.

Fürstenau and Lapata (2009) already demonstrated that SRL models can be improved by using sentences with similar semantic structure, even if the instances are selected automatically, so having groups of sentences that are known to have consistent semantics can be expected to yield an improvement as well. However, there are few manually curated datasets available (Vila et al., 2010; Dolan and Brockett, 2005) and these tend to contain up to a few thousand sentences, which makes the practical applicability of approaches relying on these corpora questionable. Hence the focus of our work is on the cross-lingual setting, where parallel data tends to occur naturally and is readily available for many language pairs.

Chapter 7

Conclusions

We have presented three approaches aiding cross-lingual transfer for semantic role labeling models. Specifically,

Cross-lingual Bootstrapping We investigate the possibility of improving a semantic role labeling model for a language, where only a small annotated corpus is available, by enforcing agreement between the source- and target-language models on parallel sentences. In experiments with four language pairs we observe improvements under certain conditions, but the idea appears to have limited practical applicability in cases where substantially different annotation schemes are used for the two languages, as we demonstrate the role correspondence to be non-trivial to learn, at least for the languages and formalisms in our experiments. If a role correspondence model can be provided externally, the quality of the target-language model can be improved substantially through the bootstrapping process.

Direct Model Transfer To our knowledge, this is the first application of direct model transfer in the context of semantic role labeling. We compare this approach to another established cross-lingual technique, annotation projection, and a state-of-the-art unsupervised SRL baseline. The general technique shows competitive performance under the experimental conditions and we further investigate the contributions of various sources of information and the effect of using syntactic annotation obtained by syntactic transfer.

Feature Representation Projection A novel approach to cross-lingual transfer is proposed, based on representation learning and aiming to combine the strengths of annotation projection and direct transfer. We demonstrate the viability of the approach in experiments with four language pairs and propose a number of ways in which it can be refined further. As a part of this work, we also compare the performance of the proposed approach to that of annotation projection and direct transfer as a function of the size of the parallel corpus.

7.1 Multilingual Language Processing

Despite the growing interest in multilingual NLP, most research in computational linguistics and natural language processing still focuses on a single language, sometimes

borrowing from existing work for other languages with more developed infrastructure, but rarely taking a holistic, multilingual perspective from the start. Attempts to link various resources in different languages together, reconcile the annotation differences and make them more cross-lingually consistent usually come as an afterthought.

For certain tasks, this bottom-up approach can make a lot of sense. Consider dependency parsing: different languages have quite different syntactic structure and therefore it is reasonable to maintain language-specific sets of dependency relation markers and design annotation guidelines for each language more or less from scratch (although even here many of the variations between language-specific formalisms are quite arbitrary). Developing a unified way to represent syntactic dependency structures across languages and procedures to derive these from the existing language-specific resources is a separate, non-trivial effort, which for dependency parsing has started only a few years ago (McDonald et al., 2013; Zeman et al., 2014). This unification also comes at the cost of some considerable coarsening, so the universal dependency annotations are potentially less informative for downstream tasks.

For higher-level, semantic annotation, on the other hand, language-specificity appears less necessary. At least some degree of cross-lingual applicability and compatibility would be desirable, and in certain cases these can even be seen as additional quality criteria for the proposed formalism.

The variations we observe in SRL formalisms and datasets for different languages, for instance, appear to largely be byproducts of researcher's desire to refine and improve what others have come up with, not some specific properties of the language concerned.

Going forward, it seems fairly clear that as a community we would benefit from maintaining some level of consistency across languages. Yet even now we see examples of new formalisms being designed in a way that makes them inherently hard to internationalize. Abstract Meaning Representation, or AMR (Banarescu et al., 2012) is one such example, based as it is on the system of PropBank framesets. Xue et al. (2014) offer an interesting analysis of cross-lingual compatibility of AMR structures, but such comparison, as well as any future attempt to apply this representation in a cross-lingual setting is made harder by its reliance on language-specific inventories.

We hope to see the notion of multilinguality built directly into the design of the semantic formalisms to come.

Appendices

Appendix A

Appendix

```
.id:0 .type:temperature .date:2009-02-08 .label:Tonight @time:17-30 #min:32 #mean:35 #max:44
.id:1 .type:windChill .date:2009-02-08 .label:Tonight @time:17-30 #min:0 #mean:0 #max:0
.id:2 .type:windSpeed .date:2009-02-08 .label:Tonight @time:17-30 #min:7 #mean:8 #max:9 @mode-bucket-0-20-2:0-10
.id:3 .type:windDir .date:2009-02-08 .label:Tonight @time:17-30 @mode:SSW
.id:4 .type:gust .date:2009-02-08 .label:Tonight @time:17-30 #min:0 #mean:0 #max:0
.id:5 .type:skyCover .date:2009-02-08 .label:Tonight @time:17-30 @mode-bucket-0-100-4:50-75
.id:6 .type:skyCover .date:2009-02-08 .label:Tonight @time:17-21 @mode-bucket-0-100-4:50-75
.id:7 .type:skyCover .date:2009-02-08 .label:Tonight @time:17-26 @mode-bucket-0-100-4:50-75
.id:8 .type:skyCover .date:2009-02-08 .label:Tonight @time:21-30 @mode-bucket-0-100-4:50-75
.id:9 .type:skyCover .date:2009-02-08 .label:Tonight @time:26-30 @mode-bucket-0-100-4:50-75
.id:10 .type:precipPotential .date:2009-02-08 .label:Tonight @time:17-30 #min:9 #mean:10 #max:14
.id:11 .type:thunderChance .date:2009-02-08 .label:Tonight @time:17-30 @mode:-
.id:12 .type:thunderChance .date:2009-02-08 .label:Tonight @time:17-21 @mode:-
.id:13 .type:thunderChance .date:2009-02-08 .label:Tonight @time:17-26 @mode:-
.id:14 .type:thunderChance .date:2009-02-08 .label:Tonight @time:21-30 @mode:-
.id:15 .type:thunderChance .date:2009-02-08 .label:Tonight @time:26-30 @mode:-
.id:16 .type:rainChance .date:2009-02-08 .label:Tonight @time:17-30 @mode:-
.id:17 .type:rainChance .date:2009-02-08 .label:Tonight @time:17-21 @mode:-
.id:18 .type:rainChance .date:2009-02-08 .label:Tonight @time:17-26 @mode:-
.id:19 .type:rainChance .date:2009-02-08 .label:Tonight @time:21-30 @mode:-
.id:20 .type:rainChance .date:2009-02-08 .label:Tonight @time:26-30 @mode:-
.id:21 .type:snowChance .date:2009-02-08 .label:Tonight @time:17-30 @mode:-
.id:22 .type:snowChance .date:2009-02-08 .label:Tonight @time:17-21 @mode:-
.id:23 .type:snowChance .date:2009-02-08 .label:Tonight @time:17-26 @mode:-
.id:24 .type:snowChance .date:2009-02-08 .label:Tonight @time:21-30 @mode:-
.id:25 .type:snowChance .date:2009-02-08 .label:Tonight @time:26-30 @mode:-
.id:26 .type:freezingRainChance .date:2009-02-08 .label:Tonight @time:17-30 @mode:-
.id:27 .type:freezingRainChance .date:2009-02-08 .label:Tonight @time:17-21 @mode:-
.id:28 .type:freezingRainChance .date:2009-02-08 .label:Tonight @time:17-26 @mode:-
.id:29 .type:freezingRainChance .date:2009-02-08 .label:Tonight @time:21-30 @mode:-
.id:30 .type:freezingRainChance .date:2009-02-08 .label:Tonight @time:26-30 @mode:-
.id:31 .type:sleetChance .date:2009-02-08 .label:Tonight @time:17-30 @mode:-
.id:32 .type:sleetChance .date:2009-02-08 .label:Tonight @time:17-21 @mode:-
.id:33 .type:sleetChance .date:2009-02-08 .label:Tonight @time:17-26 @mode:-
.id:34 .type:sleetChance .date:2009-02-08 .label:Tonight @time:21-30 @mode:-
.id:35 .type:sleetChance .date:2009-02-08 .label:Tonight @time:26-30 @mode:-
```

Table A.1: An example weather forecast record from the dataset of Liang et al. (2009).

```
InputColumn(POSTAG, Stack[0])
InputColumn(POSTAG, Stack[1])
InputColumn(POSTAG, Input[0])
InputColumn(POSTAG, Input[1])
InputColumn(POSTAG, Input[2])
Merge(InputColumn(POSTAG, Input[1]), InputColumn(POSTAG, Input[0]))
Merge(InputColumn(POSTAG, Input[2]), InputColumn(POSTAG, Input[1]))
Merge(InputColumn(POSTAG, Input[0]), InputColumn(POSTAG, Stack[0]))
Merge(InputColumn(POSTAG, Input[1]), InputColumn(POSTAG, Stack[0]))
Merge(InputColumn(POSTAG, Input[2]), InputColumn(POSTAG, Stack[0]))
Merge(InputColumn(POSTAG, Input[0]), InputColumn(POSTAG, Stack[1]))
Merge(InputColumn(POSTAG, Input[1]), InputColumn(POSTAG, Stack[1]))
Merge(InputColumn(POSTAG, Input[2]), InputColumn(POSTAG, Stack[1]))
Merge(InputColumn(POSTAG, Stack[1]), InputColumn(POSTAG, Stack[0]))
InputColumn(CLC, Stack[0])
InputColumn(CLC, Stack[1])
InputColumn(CLC, Input[0])
InputColumn(CLC, Input[1])
InputColumn(CLC, Input[2])
Merge(InputColumn(CLC, Input[1]), InputColumn(CLC, Input[0]))
Merge(InputColumn(CLC, Input[2]), InputColumn(CLC, Input[1]))
Merge(InputColumn(CLC, Input[0]), InputColumn(CLC, Stack[0]))
Merge(InputColumn(CLC, Input[1]), InputColumn(CLC, Stack[0]))
Merge(InputColumn(CLC, Input[2]), InputColumn(CLC, Stack[0]))
Merge(InputColumn(CLC, Input[0]), InputColumn(CLC, Stack[1]))
Merge(InputColumn(CLC, Input[1]), InputColumn(CLC, Stack[1]))
Merge(InputColumn(CLC, Input[2]), InputColumn(CLC, Stack[1]))
Merge(InputColumn(CLC, Stack[1]), InputColumn(CLC, Stack[0]))
Merge(InputColumn(CLC, Input[0]), InputColumn(POSTAG, Input[0]))
Merge(InputColumn(CLC, Input[1]), InputColumn(POSTAG, Input[1]))
Merge(InputColumn(CLC, Input[2]), InputColumn(POSTAG, Input[2]))
Merge(InputColumn(CLC, Stack[0]), InputColumn(POSTAG, Stack[0]))
Merge(InputColumn(CLC, Stack[1]), InputColumn(POSTAG, Stack[1]))
OutputColumn(DEPREL, Stack[0])
OutputColumn(DEPREL, ldep(Stack[0]))
OutputColumn(DEPREL, rdep(Stack[0]))
OutputColumn(DEPREL, ldep(Input[0]))
```

Table A.2: Features used in syntactic transfer in the format used by the Malt Parser. CLC indicates the cross-lingual cluster ID associated with a given token.

arg0-agt	A0
arg0-cau	A0
arg0-exp	A0
arg0-null	A0
arg0-pat	A0
arg0-src	A0
arg1-ext	A1
arg1-loc	A1
arg1-null	A1
arg1-pat	A1
arg1-tem	A1
argL-null	A1
arg2-atr	A2
arg2-ben	A2
arg2-efi	A2
arg2-exp	A2
arg2-ext	A2
arg2-ins	A2
arg2-loc	A2
arg2-null	A2
arg2-tem	A2
argM-ins	A2
arg3-ben	A3
arg3-des	A3
arg3-ein	A3
arg3-exp	A3
arg3-fin	A3
arg3-ins	A3
arg3-loc	A3
arg3-null	A3
arg3-ori	A3
arg4-des	A4
arg4-efi	A4
argM-adv	AM-ADV
argM-cau	AM-CAU
argM-ext	AM-EXT
argM-loc	AM-LOC
argM-atr	AM-MNR
argM-mnr	AM-MNR
argM-fin	AM-PNC
argM-tmp	AM-TMP

Figure A.1: Role mapping for English-Spanish. The original ones are derived from AnCora and are mapped here to be compatible with the PropBank-based ones.

A0	A0
A1	A1
A2	A2
A3	A3
A4	A4
A5	A5
AA	A0
AM	AM
AM-ADV	ADV
AM-CAU	ADV
AM-DIR	DIR
AM-DIS	DIS
AM-EXT	EXT
AM-LOC	LOC
AM-MNR	MNR
AM-MOD	_
AM-NEG	_
AM-PNC	PNC
AM-PRD	PRD
AM-PRT	_
AM-REC	_
AM-TMP	TMP

A0	A0
A1	A1
A2	A2
A3	A3
A4	A4
A5	A5
ADV	ADV
ARGM	AM
ASP	_
BNF	ADV
C-A0	A0
C-A1	A1
C-A2	A2
C-A3	A3
C-ADV	ADV
C-ARGM-DIS	DIS
C-TMP	TMP
CND	ADV
CRD	_
DGR	EXT
DIR	DIR
DIS	DIS
EXT	EXT
FRQ	TMP
LOC	LOC
MNR	MNR
PRD	PRD
PRP	PNC
PSE	_
PSR	_
QTY	_
TMP	TMP
TPC	ADV
VOC	_

Figure A.2: Role label mapping for English-Chinese. The sets of roles are partially compatible, but the naming is different and some roles have no clear counterparts. We map compatible roles from both sets into a smaller, simplified role inventory and omit the remaining ones. The mapping for the English corpus is on the left and for Chinese on the right.

Bibliography

Omri Abend and Ari Rappoport. Fully unsupervised core-adjunct argument classification. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 226–236. Association for Computational Linguistics, 2010.

Omri Abend and Ari Rappoport. Universal conceptual cognitive annotation (UCCA). In *ACL (1)*, pages 228–238, 2013.

Omri Abend, Roi Reichart, and Ari Rappoport. Unsupervised argument identification for semantic role labeling. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, ACL '09, pages 28–36, Suntec, Singapore, 2009. Association for Computational Linguistics. ISBN 978-1-932432-45-9.

Željko Agić, Dirk Hovy, and Anders Søgaard. If all you have is a bit of the Bible: Learning pos taggers for truly low-resource languages. In *The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2015)*, 2015.

Željko Agić, Anders Johannsen, Barbara Plank, Héctor Alonso Martínez, Natalie Schluter, and Anders Søgaard. Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics*, 4:301–312, 2016.

Avery Andrews. The major functions of the noun phrase. *Language typology and syntactic description*, 1:62–154, 1985.

Paolo Annesi and Roberto Basili. Cross-lingual alignment of FrameNet annotations through hidden Markov models. In *Proceedings of the 11th international conference on Computational Linguistics and Intelligent Text Processing*, CICLing'10, pages 12–25. Springer-Verlag, 2010.

Olga Babko-Malaya. Guidelines for PropBank framers. *Unpublished manual, September*, 2005.

Collin Baker, Michael Ellsworth, and Katrin Erk. Semeval'07 task 19: Frame semantic structure extraction. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 99–104, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1621474.1621492`.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The Berkeley FrameNet project. In *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational*

Linguistics and Seventeenth International Conference on Computational Linguistics (ACL-COLING'98), pages 86–90, Montreal, Canada, 1998.

Timothy Baldwin and Marco Lui. Language identification: The long and the short of the matter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 229–237. Association for Computational Linguistics, 2010.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. Abstract meaning representation (amr) 1.0 specification. In *Parsing on Freebase from Question-Answer Pairs." In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle: ACL*, pages 1533–1544, 2012. URL `http://www.aclweb.org/anthology/W13-2322`.

Regina Barzilay and Kathleen R McKeown. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 50–57. Association for Computational Linguistics, 2001.

Valerio Basile, Johan Bos, Kilian Evang, and Noortje Venhuizen. Developing a large semantically annotated corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3196–3200, Istanbul, Turkey, 2012. URL `http://www.let.rug.nl/bos/pubs/BasileBosEvangVenhuizen2012LREC.pdf`.

Luisa Bentivogli and Emanuele Pianta. Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor corpus. *Natural Language Engineering*, 11(03):247–261, 2005.

James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio. Theano: a CPU and GPU math expression compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*, Austin, TX, 2010.

Anders Björkelund, Love Hafdell, and Pierre Nugues. Multilingual semantic role labeling. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 43–48, Boulder, Colorado, June 2009. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W09-1206`.

William Blacoe and Mirella Lapata. A comparison of vector-based representations for semantic composition. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 546–556. Association for Computational Linguistics, 2012.

Frank R Blake. A semantic analysis of case. *Language*, pages 34–49, 1930.

Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the Workshop on Computational Learning Theory (COLT 98)*, 1998. URL `citeseer.nj.nec.com/blum98combining.html`.

Hans Christian Boas. Bilingual FrameNet dictionaries for machine translation. In *LREC*, 2002.

Bernd Bohnet. Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 89–97, Beijing, China, August 2010. URL http://www.aclweb.org/anthology/C10-1011.

Stephen A Boxwell and Michael White. Projecting PropBank roles onto the CCGBank. In *LREC*, 2008.

Stephen A Boxwell, Dennis Mehay, and Chris Brew. Brutus: a semantic role labeling system incorporating CCG, CFG, and dependency features. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 37–45. Association for Computational Linguistics, 2009.

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Pado, and Manfred Pinkal. The SALSA corpus: a German corpus resource for lexical semantics. In *Proceedings of LREC 2006*, Genoa, Italy, 2006.

Georgio R. Cardona. *Panini, A Survey of Research*. De Gruyter Mouton, 1976. URL http://www.degruyter.com/view/product/47863.

Xavier Carreras and Lluís Màrquez. Introduction to the CoNLL-2004 shared task: Semantic role labeling. In Hwee Tou Ng and Ellen Riloff, editors, *HLT-NAACL 2004 Workshop: Eighth Conference on Computational Natural Language Learning (CoNLL-2004)*, pages 89–97, Boston, Massachusetts, USA, May 6 - May 7 2004. Association for Computational Linguistics.

Xavier Carreras and Lluís Màrquez. Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 152–164, Ann Arbor, Michigan, 2005. URL http://www.aclweb.org/anthology/W/W05/W05-0620.

William B Cavnar, John M Trenkle, et al. N-gram-based text categorization. *Ann Arbor MI*, 48113(2):161–175, 1994.

Pi-Chuan Chang, Michel Galley, and Christopher D. Manning. Optimizing Chinese word segmentation for machine translation performance. In *Proceedings of the Third Workshop on Statistical Machine Translation*, StatMT '08, pages 224–232, Columbus, Ohio, 2008. Association for Computational Linguistics. ISBN 978-1-932432-09-1.

Wanxiang Che and Ting Liu. Using word sense disambiguation for semantic role labeling. In *Universal Communication Symposium (IUCS), 2010 4th International*, pages 167–174. IEEE, 2010.

Chenhua Chen, Alexis Palmer, and Caroline Sporleder. Enhancing active learning for semantic role labeling via compressed dependency trees. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 183–191, Chiang Mai, Thailand, November 2011. Asian Federation of Natural Language Processing. URL http://www.aclweb.org/anthology/I11-1021.

Jinho D Choi and Martha Palmer. Getting the most out of transition-based dependency parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 687–692. Association for Computational Linguistics, 2011.

Noam Chomsky. *Lectures on Government and Binding*. Foris, Dordrecht, 1981.

Janara Christensen, Stephen Soderland, Oren Etzioni, et al. Semantic role labeling for open information extraction. In *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*, pages 52–60. Association for Computational Linguistics, 2010.

R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12: 2493–2537, 2011.

Ronan Collobert and Jason Weston. A unified architecture for natural language processing: deep neural networks with multitask learning. In *Proceedings of the Twenty-Fifth International Conference (ICML 2008)*, pages 160–167, Helsinki, Finland, 2008.

D. A. Cruse. Some thoughts on agentivity. *Journal of Linguistics*, 9:11–23, 3 1973. ISSN 1469-7742. doi: 10.1017/S0022226700003509.

Lea Cyrus. Building a resource for studying translation shifts. *CoRR*, abs/cs/0606096, 2006.

Dipanjan Das and Slav Petrov. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 600–609, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P11-1061`.

Dipanjan Das, Desai Chen, André FT Martins, Nathan Schneider, and Noah A Smith. Frame-semantic parsing. 2013.

A. P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39(1):1–38, 1977.

Koen Deschacht and Marie-Francine Moens. Semi-supervised semantic role labeling using the Latent Words Language Model. In *EMNLP*, 2009.

Mona Diab and Philip Resnik. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 255–262, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073126. URL `http://dx.doi.org/10.3115/1073083.1073126`.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie Strassel, and Ralph M Weischedel. The automatic content extraction (ACE) program-tasks, data, and evaluation. In *LREC*, volume 2, page 1, 2004.

Bill Dolan, Chris Quirk, and Chris Brockett. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the Conference on Computational Linguistics (COLING)*, pages 350–356, 2004.

William B Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Proc. of IWP*, 2005.

David Dowty. Thematic proto-roles and argument selection. *Language*, 67(3):547–619, 1991.

Greg Durrett, Adam Pauls, and Dan Klein. Syntactic transfer using a bilingual lexicon. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1–11, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/D12-1001`.

Andreas Eisele and Yu Chen. MultiUN: A multilingual corpus from United Nation documents. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA), 2010. ISBN 2-9517408-6-7.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9: 1871–1874, 2008.

Parvin Sadat Feizabadi and Sebastian Padó. Crowdsourcing annotation of non-local semantic roles. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 226–230, Gothenburg, Sweden, April 2014. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/E14-4044`.

C. J. Fillmore, C. R. Johnson, and M. R. L. Petruck. Background to FrameNet. *International Journal of Lexicography*, 16:235–250, 2003.

Charles Fillmore. The case for case. In Emmon Bach and R. Harms, editors, *Universals in Linguistic Theory*. Holt, Rinehart, and Winston, New York, 1968.

Nicholas FitzGerald, Oscar Täckström, Kuzman Ganchev, and Dipanjan Das. Semantic role labeling with neural network factors. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2015.

Jeffrey Flanigan, Sam Thomson, Jaime G Carbonell, Chris Dyer, and Noah A Smith. A discriminative graph-based parser for the abstract meaning representation. 2014.

Marco Fossati, Claudio Giuliano, and Sara Tonelli. Outsourcing FrameNet to the crowd. In *ACL*, pages 742–747, 2013.

S. Francis and H. Kucera. *Computing Analysis of Present-day American English*. Brown University Press, Providence, RI, 1967.

Lea Frermann and Francis Bond. Cross-lingual parse disambiguation based on semantic correspondence. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 125–129, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P12-2025.

Hagen Fürstenau and Mirella Lapata. Graph alignment for semi-supervised semantic role labeling. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 11–20, Singapore, 2009.

Kuzman Ganchev, Jennifer Gillenwater, and Ben Taskar. Dependency grammar induction via bitext projection constraints. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 369–377, Suntec, Singapore, August 2009. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P/P09/P09-1042.

Qin Gao and Stephan Vogel. Corpus expansion for statistical machine translation with semantic role label substitution rules. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 294–298, Portland, Oregon, USA, 2011. URL http://www.aclweb.org/anthology/P11-2051.

Nikhil Garg and James Henderson. Unsupervised semantic role induction with global role ordering. In *ACL*, 2012.

Nikhil Garg and James Henderson. A bayesian model of multilingual unsupervised semantic role induction. *CoRR*, abs/1603.01514, 2016. URL http://arxiv.org/abs/1603.01514.

Matt Gerber, Joyce Y. Chai, and Adam Meyers. The role of implicit argumentation in nominal SRL. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 146–154, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-41-1. URL http://dl.acm.org/citation.cfm?id=1620754.1620776.

Matthew Gerber and Joyce Y Chai. Semantic role labeling of implicit arguments for nominal predicates. *Computational Linguistics*, 38(4):755–798, 2012.

Daniel Gildea and Julia Hockenmaier. Identifying semantic roles using combinatory categorial grammar. In *Proceedings of EMNLP-2003*, Sapporo, Japan, 2003.

Daniel Gildea and Daniel Jurafsky. Automatic labelling of semantic roles. *Computational Linguistics*, 28(3):245–288, 2002.

Daniel Gildea and Martha Palmer. The necessity of parsing for predicate argument recognition. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, pages 239–246, Philadelphia, PA, 2002.

P. Good. *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. Springer, 2000.

Ian J. Goodfellow, David Warde-Farley, Pascal Lamblin, Vincent Dumoulin, Mehdi Mirza, Razvan Pascanu, James Bergstra, Frédéric Bastien, and Yoshua Bengio. Pylearn2: a machine learning research library. *CoRR*, abs/1308.4214, 2013.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. Bilbowa: Fast bilingual distributed representations without word alignments. In *International Conference on Machine Learning (ICML)*, 2015.

Trond Grenager and Christopher D. Manning. Unsupervised discovery of a statistical verb lexicon. In *Proceedings of EMNLP*, 2006.

Yulia Grishina and Manfred Stede. Knowledge-lean projection of coreference chains across languages. *ACL-IJCNLP 2015*, page 14, 2015.

Jeffrey Gruber. *Studies in Lexical Relation*. MIT Press, Cambridge, MA, 1965.

J. Hajič, J. Panevová, E. Hajičová, P. Sgall, P. Pajas, J. Štěpánek, J. Havelka, M. Mikulová, Z. Žabokrtskỳ, and M. Ševčíková-Razímová. Prague dependency treebank 2.0. *LDC*, 2006.

Jan Hajič. Tectogrammatical representation: Towards a minimal transfer in machine translation. In Robert Frank, editor, *Proceedings of the 6th International Workshop on Tree Adjoining Grammars and Related Frameworks (TAG+6)*, pages 216—226, Venezia, 2002. Universita di Venezia.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 1–18, Boulder, Colorado, 2009. URL `http://www.aclweb.org/anthology/W09-1201`.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. Announcing Prague Czech-English dependency treebank 2.0. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

Silvana Hartmann and Iryna Gurevych. FrameNet on the way to Babel: Creating a bilingual FrameNet using Wiktionary as interlingual connection. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1363–1373, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P13-1134`.

Luheng He, Mike Lewis, and Luke Zettlemoyer. Question-answer driven semantic role labeling: Using natural language to annotate natural language. In *Proceedings of the*

2015 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 643–653, 2015.

Shan He and Daniel Gildea. Self-training and co-training for semantic role labeling: Primary report. Technical report, University of Rochester, 2006. URL `ftp://anon.cs.rochester.edu/pub/papers/ai/07.tr891.Self-training_and_Co-training_for_Sem_Role_Labeling.pdf`.

James Henderson, Paola Merlo, Gabriele Musillo, and Ivan Titov. A latent variable model of synchronous parsing for syntactic and semantic dependencies. In *CoNLL 2008: Proc. Twelfth Conf. on Computational Natural Language Learning*, pages 178–182, Manchester, UK, 2008.

Karl Moritz Hermann and Phil Blunsom. Multilingual distributed representations without word alignment. *arXiv preprint arXiv:1312.6173*, 2013.

Julia Hockenmaier and Mark Steedman. CCGbank: a corpus of CCG derivations and dependency structures extracted from the penn treebank. *Computational Linguistics*, 33 (3):355–396, 2007.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. Bootstrapping parsers via syntactic projection across parallel text. *Natural Language Engineering*, 11(3):311–325, 2005.

Jena D. Hwang, Rodney D. Nielsen, and Martha Palmer. Towards a domain independent semantics: Enhancing semantic representation with construction grammar. In *Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics*, pages 1–8, Los Angeles, California, June 2010. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W10-0801`.

Ray Jackendoff. *Semantic Interpretation in Generative Grammar*. MIT Press, Cambridge, MA, 1972.

Anders Johannsen, Željko Agić, and Anders Søgaard. Joint part-of-speech and dependency projection from multiple sources. In *54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, 2016.

Jason K Johnson. *Convex relaxation methods for graphical models: Lagrangian and maximum entropy approaches*. PhD thesis, Massachusetts Institute of Technology, 2008.

Xiang Li Thien Huu Nguyen Kai and Cao Ralph Grishman. Improving event detection with abstract meaning representation. *ACL-IJCNLP 2015*, page 11, 2015.

Michael Kaisser and Bonnie Webber. Question answering based on semantic roles. In *ACL Workshop on Deep Linguistic Processing*, 2007.

Rasoul Samad Zadeh Kaljahi and Mohd Sapiyan Baba. Investigation of co-training views and variations for semantic role labeling. *ROBUS 2011*, page 41, 2011.

Seokhwan Kim, Minwoo Jeong, Jonghoon Lee, and Gary Geunbae Lee. A cross-lingual annotation projection approach for relation detection. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 564–571, Beijing, China, 2010. Association for Computational Linguistics.

Paul Kingsbury, Nianwen Xue, and Martha Palmer. Propbanking in parallel. In *In Proceedings of the Workshop on the Amazing Utility of Parallel and Comparable Corpora, in conjunction with LREC'04*, 2004.

K. Kipper, A. Korhonen, N. Ryant, and M. Palmer. Extensive classifications of English verbs. In *Proceedings of the 12th EURALEX International Congress*, 2006a.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. Extending VerbNet with novel verb classes. *Proceedings of LREC*, 2006(2.2):1, 2006b.

Scott Kirkpatrick, C Daniel Gelatt, Mario P Vecchi, et al. Optimization by simmulated annealing. *science*, 220(4598):671–680, 1983.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. Inducing crosslingual distributed representations of words. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, Bombay, India, 2012.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand, 2005. AAMT. URL `http://mt-archive.info/MTS-2005-Koehn.pdf`.

Mikhail Kozhevnikov and Ivan Titov. Cross-lingual transfer of semantic role labeling models. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1190–1200, Sofia, Bulgaria, August 2013a. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P13-1117`.

Mikhail Kozhevnikov and Ivan Titov. Bootstrapping semantic role labelers from parallel data. *Atlanta, Georgia, USA*, page 317, 2013b. URL `http://www.aclweb.org/website/old_anthology/S/S13/S13-1.pdf#page=343`.

Mikhail Kozhevnikov and Ivan Titov. Cross-lingual model transfer using feature representation projection. In *ACL (2)*, pages 579–585. Association for Computational Linguistics, 2014. URL `http://anthology.aclweb.org/P/P14/P14-2095.pdf`.

Sebastian Krause, Enrique Alfonseca, Katja Filippova, and Daniele Pighin. Idest: Learning a distributed representation for event patterns. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL'15)*, pages 1140–1149, 2015.

Namhee Kwon, Michael Fleischman, and Eduard Hovy. Senseval automatic labeling of semantic roles using maximum entropy models. In *Senseval-3*, pages 129–132, Barcelona, Spain, 2004.

John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conf. on Machine Learning*, pages 282–289. Morgan Kaufmann, San Francisco, CA, 2001.

Joel Lang and Mirella Lapata. Unsupervised induction of semantic roles. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of*

the Association for Computational Linguistics, pages 939–947, Los Angeles, California, June 2010. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/N10-1137.

Joel Lang and Mirella Lapata. Unsupervised semantic role induction via split-merge clustering. In *Proc. of Annual Meeting of the Association for Computational Linguistics (ACL)*, 2011.

Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C Raykar, and Amrita Saha. An autoencoder approach to learning bilingual word representations. In *Advances in Neural Information Processing Systems*, pages 1853–1861, 2014.

J.Y. Lee, Y.I. Song, and H.C. Rim. Investigation of weakly supervised learning for semantic role labeling. In *Advanced Language Processing and Web Information Technology, 2007. ALPIT 2007. Sixth International Conference on*, pages 165–170. IEEE, 2007.

Tao Lei, Yuan Zhang, Ro Moschitti, and Regina Barzilay. High-order lowrank tensors for semantic role labeling. In *In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACLHLT 2015*, 2015.

Mike Lewis, Luheng He, and Luke Zettlemoyer. Joint A* CCG parsing and semantic role labelling. In *Empirical Methods in Natural Language Processing*, 2015.

Percy Liang, Michael I. Jordan, and Dan Klein. Learning semantic correspondences with less supervision. In *ACL-IJCNLP*, 2009.

Pierre Lison and Jörg Tiedemann. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *LREC*, 2016.

Ding Liu and Daniel Gildea. Semantic role features for machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, Beijing, China, 2010.

Fei Liu, Jeffrey Flanigan, Sam Thomson, Norman Sadeh, and Noah A Smith. Toward abstractive summarization using semantic representations. 2015. URL http://repository.cmu.edu/cgi/viewcontent.cgi?article=1155&context=lti.

E. Loper, Szu ting Yi, and Martha Palmer. Combining lexical resources: Mapping between PropBank and VerbNet. In *Proceedings of the IWCS*, 2007.

Alejandra Lorenzo and Christophe Cerisara. Unsupervised frame based semantic role induction: application to French and English. In *Proceedings of the ACL 2012 Joint Workshop on Statistical Parsing and Semantic Processing of Morphologically Rich Languages*, pages 30–35, Jeju, Republic of Korea, July 2012. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W12-3404.

Yi Luan, Yangfeng Ji, Hannaneh Hajishirzi, and Boyang Li. Multiplicative representations for unsupervised semantic role induction. In *ACL 2016*. Association for Computational Linguistics, 2016. URL http://www.boyangli12.co/paper/luan-acl2016.pdf.

Thang Luong, Hieu Pham, and Christopher D Manning. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159, 2015.

Thomas Mayer and Michael Cysouw. Creating a massively parallel Bible corpus. *Oceania*, 135(273):40, 2013.

David McClosky, Eugene Charniak, and Mark Johnson. Reranking and self-training for parser adaptation. In *Proc. of the Annual Meeting of the ACL and the International Conference on Computational Linguistics*, Sydney, Australia, 2006.

Ryan McDonald, Slav Petrov, and Keith Hall. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 62–72, Edinburgh, United Kingdom, 2011. Association for Computational Linguistics. ISBN 978-1-937284-11-4.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P13-2017`.

Gabor Melli, Yang Wang, Yudong Liu, Mehdi M Kashani, Zhongmin Shi, Baohua Gu, Anoop Sarkar, and Fred Popowich. Description of Squash, the SFU question answering summary handler for the DUC-2005 summarization task. *safety*, 1:14345754, 2005.

Paola Merlo and Matthias Leybold. Automatic distinction of arguments and modifiers: the case of prepositional phrases. In *Proceedings of the Fifth Computational Natural Language Learning Workshop (CoNLL-2001)*, pages 121–128, Toulouse, France, 2001.

Paola Merlo and Lonneke Van Der Plas. Abstraction and generalisation in semantic role labels: PropBank, VerbNet or both? In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 288–296. Association for Computational Linguistics, 2009.

Paola Merlo, Suzanne Stevenson, Vivian Tsang, and Gianluca Allaria. A multi-lingual paradigm for automatic verb classification. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, pages 207–214, Philadelphia, PA, 2002.

Roland Meyer. New wine in old wineskins?–Tagging old Russian via annotation projection from modern translations. *Russian Linguistics*, 35(2):267(15), 2011.

A. Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman. The nombank project: An interim report. In A. Meyers, editor, *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 24–31, Boston, Massachusetts, USA, 2004.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013a.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168, 2013b.

Scott Miller, David Stallard, Robert Bobrow, and Richard Schwartz. A fully statistical approach to natural language interfaces. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 55–61. Association for Computational Linguistics, 1996.

Ashutosh Modi, Ivan Titov, and Alexandre Klementiev. Unsupervised induction of frame-semantic representations. In *Proceedings of the NAACL-HLT 2012 Workshop on Inducing Linguistic Structure*, Montreal, Canada, June 2012.

Mehdi Mohammadi and Nasser Ghasem-Aghaee. Building bilingual parallel corpora based on Wikipedia. In *Computer Engineering and Applications (ICCEA), 2010 Second International Conference on*, volume 2, pages 264–268. IEEE, 2010.

Alessandro Moschitti, Daniele Pighin, and Roberto Basili. Tree kernels for semantic role labeling. *Computational Linguistics*, 34(2):193–224, 2008.

Srini Narayanan and Sanda Harabagiu. Question answering based on semantic structures. In *Proceedings of the 20th international conference on Computational Linguistics*, page 693. Association for Computational Linguistics, 2004.

Tahira Naseem, Regina Barzilay, and Amir Globerson. Selective sharing for multilingual dependency parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 629–637, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P12-1066`.

Joakim Nivre. Algorithms for deterministic incremental dependency parsing. *Comput. Linguist.*, 34(4):513–553, December 2008. ISSN 0891-2017. doi: 10.1162/coli.07-056-R1-07-027.

Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1), 2003.

Brendan O'Connor. Learning frames from text with an unsupervised latent variable model. *arXiv preprint arXiv:1307.7382*, 2013.

Kyoko Hirose Ohara, Seiko Fujii, Hiroaki Saito, Shun Ishizaki, Toshio Ohori, and Ryoko Suzuki. The Japanese FrameNet project: A preliminary report. In *Proceedings of pacific association for computational linguistics*, pages 249–254. Citeseer, 2003.

Sebastian Padó and Mirella Lapata. Cross-linguistic projection of role-semantic information. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 859–866, Vancouver, British Columbia, Canada, 2005. URL `http://www.aclweb.org/anthology/H/H05/H05-1108`.

Sebastian Padó and Mirella Lapata. Optimal constituent alignment with edge covers for semantic projection. In *Proc. 44th Annual Meeting of Association for Computational Linguistics and 21st International Conf. on Computational Linguistics*, ACL-COLING 2006, pages 1161–1168, Sydney, Australia, 2006. URL `http://www.aclweb.org/anthology/P/P06/P06-1146`.

Sebastian Padó and Mirella Lapata. Cross-lingual annotation projection for semantic roles. *Journal of Artificial Intelligence Research*, 36:307–340, 2009.

Sebastian Padó, Marco Pennacchiotti, and Caroline Sporleder. Semantic role assignment for event nominalisations by leveraging verbal data. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 665–672, Manchester, UK, 2008.

Alexis Palmer and Caroline Sporleder. Evaluating FrameNet-style semantic parsing: the role of coverage gaps in FrameNet. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 928–936, Beijing, China, 2010. Association for Computational Linguistics.

Martha Palmer. Semlink: Linking PropBank, VerbNet and FrameNet. In *Proceedings of the generative lexicon conference*, pages 9–15, 2009.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31:71–105, 2005a.

Martha Palmer, Nianwen Xue, Olga Babko-Malaya, Jinying Chen, and Benjamin Snyder. A parallel Proposition Bank II for Chinese and English. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, CorpusAnno '05, pages 61–67, Ann Arbor, Michigan, 2005b. Association for Computational Linguistics.

Xiaoman Pan, Taylor Cassidy, Ulf Hermjakob, Heng Ji, and Kevin Knight. Unsupervised entity linking with abstract meaning representation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics–Human Language Technologies*, 2015. URL `http://nlp.cs.rpi.edu/paper/amrel.pdf`.

Slav Petrov, Dipanjan Das, and Ryan McDonald. A universal part-of-speech tagset. In *Proceedings of LREC*, May 2012.

Carl Pollard and Ivan A Sag. *Head-driven phrase structure grammar*. University of Chicago Press, 1994.

Sameer S. Pradhan, Wayne Ward, and James H. Martin. Towards robust semantic role labeling. *Computational Linguistics*, 34(2):289–310, 2008.

Vasin Punyakanok, Peter Koomen, Dan Roth, and Wen tau Yih. Generalized inference with multiple semantic role labeling systems. In *Proceedings of CoNLL-2005*, Ann Arbor, MI USA, 2005.

Philip Resnik, Mari Broman Olsen, and Mona Diab. The Bible as a parallel corpus: Annotating the 'book of 2000 tongues'. *Computers and the Humanities*, 33(1-2):129–153, 1999.

Dan Roth and Kevin Small. Margin-based active learning for structured output spaces. In *European Conference on Machine Learning*, pages 413–424. Springer, 2006.

Dan Roth and Wen-tau Yih. Integer linear programming inference for conditional random fields. In *ICML*, pages 736–743, 2005. doi: 10.1145/1102351.1102444.

Michael Roth and Anette Frank. Inducing implicit arguments from comparable texts: A framework and its applications. *Computational Linguistics*, 2016.

Michael Roth and Mirella Lapata. Neural semantic role labeling with dependency path embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, Berlin, Germany, 2016. To appear.

Salim Roukos, David Graff, and Dan Melamed. Hansard French/English. *Linguistic Data Consortium, Philadelphia*, 1995.

Mark Sammons, Vinod Vydiswaran, Tim Vieira, Nikhil Johri, Ming wei Chang, Dan Goldwasser, Vivek Srikumar, Gourab Kundu, Yuancheng Tu, Kevin Small, Joshua Rule, Quang Do, and Dan Roth. Relation alignment for textual entailment recognition. In *Text Analysis Conference (TAC)*, 2009.

Burr Settles. Active learning literature survey. *Computer Sciences Technical Report*, 1648, 2010.

Dan Shen and Mirella Lapata. Using semantic roles to improve question answering. In *EMNLP*, 2007.

Peng Shi, Zhiyang Teng, and Yue Zhang. Exploiting mutual benefits between syntax and semantic roles using neural network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 968–974, Austin, Texas, November 2016. Association for Computational Linguistics. URL `https://aclweb.org/anthology/D16-1098`.

Carina Silberer and Anette Frank. Casting implicit role linking as an anaphora resolution task. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 1–10. Association for Computational Linguistics, 2012.

Raivis Skadiņš, Jörg Tiedemann, Roberts Rozis, and Daiga Deksne. Billions of parallel words for free: Building and using the EU Bookshop corpus. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2014)*, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).

David A Smith and Jason Eisner. Parser adaptation and projection with quasi-synchronous grammar features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 822–831. Association for Computational Linguistics, 2009.

Benjamin Snyder and Regina Barzilay. Cross-lingual propagation for morphological analysis. In *Proceedings of the 23rd national conference on Artificial intelligence*, 2008. ISBN 978-1-57735-368-3.

Anders Søgaard. Data point selection for cross-language adaptation of dependency parsers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 2 of *HLT '11*, pages 682–686, Portland, Oregon, 2011. Association for Computational Linguistics. ISBN 978-1-932432-88-6.

David Sontag, Amir Globerson, and Tommi Jaakkola. Introduction to dual decomposition for inference. In Suvrit Sra, Sebastian Nowozin, and Stephen J. Wright, editors, *Optimization for Machine Learning*. MIT Press, 2011.

Kathrin Spreyer and Anette Frank. Projection-based acquisition of a temporal labeller. *Proceedings of IJCNLP 2008*, 2008.

Vivek Srikumar and Dan Roth. A joint model for extended semantic role labeling. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 129–139. Association for Computational Linguistics, 2011.

Mark Steedman. *The Syntactic Process*. MIT Press, Cambridge, 2000.

Dan Ştefănescu and Radu Ion. Parallel-Wiki: A collection of parallel sentences extracted from Wikipedia. In *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2013)*, pages 24–30, 2013.

Carlos Subirats and Miriam Petruck. Surprise: Spanish FrameNet. In *Proceedings of CIL*, volume 17, page 188, 2003.

Mihai Surdeanu, Sanda Harabagiu, John Williams, and Paul Aarseth. Using predicate-argument structures for information extraction. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 45–52. Sapporo, Japan, 2003.

Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proceedings of the 12th Conference on Computational Natural Language Learning (CoNLL-2008)*, 2008.

Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. Learning to rank answers to non-factoid questions from web collections. *Computational Linguistics*, 37(2):351–383, 2011.

Swabha Swayamdipta, Miguel Ballesteros, Chris Dyer, and Noah A Smith. Greedy, joint syntactic-semantic parsing with stack LSTMs. *arXiv preprint arXiv:1606.08954*, 2016.

Robert Swier and Suzanne Stevenson. Unsupervised semantic role labelling. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-04)*, pages 95–102, Barcelona, Spain, 2004.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proc. of the Annual Meeting of the North American Association of Computational Linguistics (NAACL)*, pages 477–487, Montréal, Canada, 2012.

Oscar Täckström, Ryan McDonald, and Joakim Nivre. Target language adaptation of discriminative transfer parsers. 2013. URL `http://www.ryanmcd.com/papers/targetNAACL2013.pdf`.

Mariona Taulé, M. Antònia Martì, and Marta Recasens. Ancora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, 2008. ISBN 2-9517408-4-0. http://www.lrec-conf.org/proceedings/lrec2008/.

Jörg Tiedemann. News from OPUS-A collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing*, volume 5, pages 237–248, 2009.

Jörg Tiedemann. Parallel data, tools and interfaces in OPUS. In *LREC*, pages 2214–2218, 2012.

Ivan Titov and Ehsan Khoddam. Unsupervised induction of semantic roles within a reconstruction-error minimization framework. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1–10, Denver, Colorado, May–June 2015. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/N15-1001`.

Ivan Titov and Alexandre Klementiev. A Bayesian model for unsupervised semantic parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1445–1455, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P11-1145`.

Ivan Titov and Alexandre Klementiev. A Bayesian approach to unsupervised semantic role induction. In *Proc. of European Chapter of the Association for Computational Linguistics (EACL)*, 2012a. URL `http://ivan-titov.org/papers/eacl12.pdf`.

Ivan Titov and Alexandre Klementiev. Crosslingual induction of semantic roles. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Jeju Island, South Korea, July 2012b. Association for Computational Linguistics.

Ivan Titov and Alexandre Klementiev. Semi-supervised semantic role labeling: Approaching from an unsupervised perspective. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, Bombay, India, December 2012c.

Ivan Titov and Mikhail Kozhevnikov. Bootstrapping semantic analyzers from non-contradictory texts. In *ACL*, 2010.

Sara Tonelli and Emanuele Pianta. Frame information transfer from English to Italian. In *Proceedings of LREC 2008*, 2008.

Kristina Toutanova, Aria Haghighi, and Christopher D. Manning. A global joint model for semantic role labeling. *Computational Linguistics*, 34(2):161–191, 2008. doi: 10.1162/coli.2008.34.2.161.

Diana Trandabăţ. Using semantic roles to improve summaries. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 164–169. Association for Computational Linguistics, 2011.

Reut Tsarfaty. A unified morpho-syntactic scheme of stanford dependencies. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 578–584, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P13-2103.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. Word representations: A simple and general method for semi-supervised learning. In *ACL*, 2010.

Lonneke van der Plas and Marianna Apidianaki. Cross-lingual word sense disambiguation for predicate labelling of french. In *Proceedings of the 21st TALN (Traitement Automatique des Langues Naturelles) conference*, pages 46–55, 2014.

Lonneke van der Plas, James Henderson, and Paola Merlo. Domain adaptation with artificial data for semantic parsing of speech. In *Proc. 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 125–128, Boulder, Colorado, 2009. URL http://www.aclweb.org/anthology/N/N09/N09-2032.

Lonneke van der Plas, Paola Merlo, and James Henderson. Scaling up automatic cross-lingual semantic role annotation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, HLT '11, pages 299–304, Portland, Oregon, USA, 2011. Association for Computational Linguistics.

Yannick Versley. Discovery of ambiguous and unambiguous discourse connectives via annotation projection. In *Proceedings of Workshop on Annotation and Exploitation of Parallel Corpora (AEPC)*, pages 83–82, 2010.

Marta Vila, Horacio Rodríguez, and M Antònia Martí. WRPA: A system for relational paraphrase acquisition from Wikipedia. *Procesamiento del lenguaje natural*, 45:11–19, 2010.

Chuan Wang, Nianwen Xue, Sameer Pradhan, and Sameer Pradhan. A transition-based algorithm for AMR parsing. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 366–375, 2015. URL http://www.cs.brandeis.edu/~xuen/publications/wang-2015-naacl.pdf.

Kristian Woodsend and Mirella Lapata. Distributed representations for unsupervised semantic role labeling. 2015.

Alina Wróblewska and Anette Frank. Cross-lingual projection of LFG F-structures: Building an F-structure bank for Polish. In *Eighth International Workshop on Treebanks and Linguistic Theories*, page 209, 2009.

Dekai Wu and Pascale Fung. Can semantic role labeling improve SMT? In *Proceedings of 13th Annual Conference of the European Association for Machine Translation (EAMT 2009)*, Barcelona, 2009.

Chenhai Xi and Rebecca Hwa. A backoff model for bootstrapping resources for non-english languages. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 851–858, Vancouver, British Columbia, Canada, October 2005. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/H/H05/H05-1107.

Nianwen Xue, Ondrej Bojar, Jan Hajic, Martha Palmer, Zdenka Urešová, and Xiuhong Zhang. Not an interlingua, but close: Comparison of English AMRs to Chinese and Czech. 2014. URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/384_Paper.pdf.

David Yarowsky, Grace Ngai, and Richard Wicentowski. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first international conference on Human language technology research*, pages 1–8. Association for Computational Linguistics, 2001.

Szu-ting Yi, Edward Loper, and Martha Palmer. Can semantic roles generalize across genres? In *Proceedings of the Human Language Technologies 2007 (NAACL-HLT'07)*, pages 548–555, Rochester, New York, 2007.

Liping You and Kaiying Liu. Building Chinese FrameNet database. In *2005 International Conference on Natural Language Processing and Knowledge Engineering*, pages 301–306. IEEE, 2005.

Rasoul Samad Zadeh Kaljahi. Adapting self-training for semantic role labeling. In *Proceedings of the ACL 2010 Student Research Workshop*, ACLstudent '10, pages 91–96, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1858913.1858929.

B. Zapirain, E. Agirre, L. L. Màrquez, and M. Surdeanu. Improving semantic role classification with selectional prefrences. In *Proceedings of the Meeting of the North American chapter of the Association for Computational Linguistics (NAACL 2010)*, Los Angeles, 2010.

Beñat Zapirain, Eneko Agirre, and Lluís Màrquez. Robustness and generalization of role sets: PropBank vs. VerbNet. In *Proceedings of ACL-08: HLT*, pages 550–558, Columbus, Ohio, 2008. URL http://www.aclweb.org/anthology/P/P08/P08-1063.

Daniel Zeman and Philip Resnik. Cross-language parser adaptation between related languages. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, pages 35–42, Hyderabad, India, January 2008. Asian Federation of Natural Language Processing. URL http://www.aclweb.org/anthology/I/I08/I08-0108.

Daniel Zeman, David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Zdeněk Žabokrtský, and Jan Hajič. HamleDT: To parse or not to parse? In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

Daniel Zeman, Ondřej Dušek, David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Zdeněk Žabokrtskỳ, and Jan Hajič. HamleDT: Harmonized multi-language dependency treebank. *Language Resources and Evaluation*, 48(4):601–637, 2014.

Jie Zhou and Wei Xu. End-to-end learning of semantic role labeling using recurrent neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1127–1137, Beijing, China, July 2015. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P15-1109`.

Ling Zhu, Derek F Wong, and Lidia S Chao. Unsupervised chunking based on graph propagation from bilingual corpus. *The Scientific World Journal*, 2014, 2014.

Tao Zhuang and Chengqing Zong. Joint inference for bilingual semantic role labeling. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 304–314, Cambridge, Massachusetts, 2010. Association for Computational Linguistics.

Imed Zitouni and Radu Florian. Mention detection crossing the language barrier. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2008.