Ruprecht-Karls-Universität Heidelberg

Strategies to Address Data Sparseness in Implicit Semantic Role Labeling

Author: Parvin Sadat Feizabadi

Commission Chair: Professor Dr. Jadranka Gvozdavonić

Supervisor and first reviewer: Professor Dr. Sebastian Padó

Secondary reviewer: Professor Dr. Anette Frank

Dissertation submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

in the

Computational Linguistics Department
Faculty of Modern Languages
Ruprecht-Karls-Universität Heidelberg

December 2018

Abstract

Natural language texts frequently contain predicates whose complete understanding requires access to other parts of the discourse. Human readers can retrieve such information across sentence boundaries and infer the implicit piece of information. This capability enables us to understand complicated texts without needing to repeat the same information in every single sentence. However, for computational systems, resolving such information is problematic because computational approaches traditionally rely on sentence-level processing and rarely take into account the extra-sentential context.

In this dissertation, we investigate this omission phenomena, called implicit semantic role labeling. Implicit semantic role labeling involves identification of predicate arguments that are not locally realized but are resolvable from the context. For example, in "What's the matter, Walters? *asked* Baynes sharply.", the ADDRESSEE of the predicate *ask*, Walters, is not mentioned as one of its syntactic arguments, but can be recoverable from the previous sentence. In this thesis, we try to improve methods for the automatic processing of such predicate instances to improve natural language processing applications. Our main contribution is introducing approaches to solve the data sparseness problem of the task. We improve automatic identification of implicit roles by increasing the amount of training set without needing to annotate new instances. For this purpose, we propose two approaches. As the first one, we use crowdsourcing to annotate instances of implicit semantic roles and show that with an appropriate task design, reliable annotation of implicit semantic roles can be obtained from the non-experts without the need to present precise and linguistic definition of the roles to them. As the second approach, we combine seemingly incompatible corpora to solve the problem of data sparseness of ISRL by applying a domain adaptation technique. We show that out of domain data from a different genre can be successfully used to improve a baseline implicit semantic role labeling model, when used with an appropriate domain adaptation technique. The results also show that the improvement occurs regardless of the predicate part of speech, that is, identification of implicit roles relies more on semantic features than syntactic ones. Therefore, annotating instances of nominal predicates, for instance, can help to improve identification of verbal predicates' implicit roles, we well. Our findings also show that the variety of the additional data is more important than its size. That is, increasing a large amount of data does not necessarily lead to a better model.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisor Prof. Dr. Sebastian Padó for his continuous support of my Ph.D. study, his patience and motivation. His guidance helped me in all the time of doing my research and writing this dissertation. I could not have imagined having a better advisor for my Ph.D. study.

Besides my supervisor, I would like to thank the rest of my thesis committee, Prof. Dr. Anette Frank and Prof. Dr. Jadranka Gvozdavonić, for their insightful comments and questions which incented me to widen my research from various perspectives.

I would also like to thank my husband for supporting me spiritually throughout my Ph.D. study and my life in general.

Contents

List of Figures

List of Tables

Part I

Introduction

Chapter 1

Introduction

The key to understand natural language is to understand the meaning of its words, expressions and sentences. Computational semantics is a fairly new interdisciplinary area which combines insights from semantics and computational linguistics to capture such meanings. The main goal of this field is to automate the process of constructing semantic representations for natural language expressions and to use them to perform reasoning and inference (Blackburn and Bos, 2003). Semantic roles are instances of these meaning representations which indicate the participants and properties of the events and the relations among the relevant entities expressed in the sentence. In simple words, semantic roles answer the questions such as WHAT happened? WHO did it? to WHOM was it done? In what MANNER? WHERE? etc. In computational semantics, answering the above questions is the key to the second step, i.e. reasoning.

In this chapter, we briefly introduce semantic roles, the task of automatically identifying these roles, i.e. Semantic Role Labeling (SRL), a more generalized version of SRL which aims at identifying missing participants of an event in the near context, i.e. Implicit Semantic Role Labeling (ISRL [1]) and discuss the focus and contributions of this dissertation.

1.1 Semantic Roles

Semantic roles (Fillmore, 1968) are a linguistic concept which identify the entities involved in an event and provide interesting information beyond syntax level. To make

[1] In all this dissertation, ISRL stands for Implicit Semantic Role Labeling. But in the literature it also stands for Incremental Semantic Role Labeling

the importance of semantic roles clear, consider the following examples:

(1) My brother hit me with a stone.

(2) I was hit with a stone by my brother.

(3) The thrown stone by my brother hit me.

In all these examples, the main event is the *hitting* event, though the sentences represent different syntactic forms. In these sentences, the event involves three participants, the hitter, the affected entity and the instrument with which the action is performed. The affected entity fills the direct object and syntactic subject positions in different surface forms of (1) and (2) and the hitter which is the syntactic subject in (1), is represented as a prepositional phrase in (2) and as part of the subject in (3), despite having notationally the same role in all sentences. Annotating these sentences with their semantic roles lets us eliminate the effect of syntactic differences and focus on the semantic commonality between them.

In addition to identifying semantic commonalities, semantic role annotation can help to identify the differences between similar syntactic structures. For example, while syntactic annotation of (4) and (5) are fairly similar and the words *they* in both sentences are labeled as subject, their semantic roles are different. *They* in (4) fills the *AGENT* role, whereas *they* in 5 is an *EXPERIENCER* (for more details about definition of semantic roles cf. Chapter 2).

(4) They broke a chair.

(5) They were sad.

Examples (1), (2) and (3) showed instances of sentences with semantic similarities and different surface forms which talked about the same topic. Examples (4) and (5) indicated instances of sentences with similar syntactic structures but different semantic interpretations. Now, in sentences (6) and (7), we see that even sentences which consider different topics may have common semantic attributes.

(6) The locals eat mainly fish and vegetables.

(7) I broke the car window[2].

[2] All examples in this chapter are adapted from FrameNet (Baker et al., 1998) unless otherwise stated.

In these examples, the subjects, i.e. *the locals* and *I* are both the *doers* of the actions and have the causal responsibility of their actions. Therefore, one can group them together under one category, called AGENT role. In a similar way, *fish and vegetables* and *the car window*, which are inanimate objects directly affected by the action, can fill the same role, called PATIENT. Other common semantic roles include INSTRUMENT, EXPERIENCER, THEME, GOAL, BENEFACTIVE, SOURCE (cf. Chapter 2). Nevertheless, there is no single theory of semantic roles. This topic has been an active area of research for a long time and different theories have been proposed. From the computational linguistics point of view two relevant frameworks exist (FrameNet and PropBank) which are discussed in more details in Chapter 2.

1.2 Semantic Role Labeling

Semantic Role Labeling is a well-defined task whose aim is to identify and label all semantic roles of all predicates in the sentence. In other words, an SRL system identifies syntactic constituents in the sentence which fill semantic roles of the predicate and assigns pre-defined semantic roles to the constituents.

Since this layer of information, semantic information, have been shown to improve different NLP tasks, such as textual entailment (de Salvo Braz et al., 2006), text summarization (Yan and Wan, 2014), question answering (Narayanan and Harabagiu, 2004), and information extraction (Christensen et al., 2010), many studies have addressed the task of automatic semantic role labeling (cf. Chapter 3).

The task was firstly introduced by Gildea and Jurafsky (2002) and today there is a substantial body of work on this topic. One of the main reasons of existence of such a large number of studies is the development of two semantic role annotated corpora, FrameNet and PropBank (cf. Chapter 2). Also, the introduced shared tasks motivated development of many systems. CoNLL 2004 (Carreras and Màrques, 2004) was the first SRL shared task which caused development of ten systems. In this task, the systems utilized partial syntactic information, that is, chunks and clauses forming a tree, and applied different machine learning algorithms to annotate predicates and their semantic roles. The task was followed in ConLL 2005 (Carreras and Màrquez, 2005) by increasing the amount of syntactic and semantic information to boost the performance of machine learning systems.

CoNLL 2008 (Surdeanu et al., 2008) was another shared task which addressed the task from a different perspective and focused on joint parsing of syntactic and semantic

dependencies. This task was extened in its next year's version, CoNLL 2009 (Hajič et al., 2009) to languages other than English.

The participants of the shared tasks and also many later studies (e.g. Das et al. (2010a); Das and Smith (2011); Srikumar and Roth (2011); Täckström et al. (2015)) developed supervised systems relying on FrameNet (Baker et al., 1998) and PropBank (Palmer et al., 2005) (cf. Chapter 2), though some studies have also applied unsupervised learning techniques (e.g. Swier and Stevenson (2004) and Lang and Lapata (2010)) (for more details on related work cf. Chapter 3).

Though a direct comparison between all SRL systems is not possible due to the different training and test sets, it can be said that the top systems perform with a high F-score of 0.70-0.80 (e.g. Johansson and Nugues (2008) in CoNLL 2008 and Srikumar and Roth (2011) and Täckström et al. (2015) among the later studies). Most of these studies treat the task as a classification task and use different machine learning techniques, among which Maximum Entropy(ME) and support Vector Machines (SVM) are the most widely used learning algorithms.

The steps performed in a typical SRL system are as follows: firstly, the predicates are determined, then the system identifies which roles are required for the target predicate and in the next step, spans of text are annotated as semantic roles. The implementation of these steps, however, is various among the systems. For example, Riedel and Meza-Ruiz (2008) performed predicate and semantic roles identification and classification jointly, whereas Ciaramita et al. (2008) implemented the three components with a pipeline architecture. It is not easy to determine which approach performs better. In CoNLL 2008, for instance, the top five systems in the closed challenge (where systems had to be trained only with the information contained in the given training corpus) were systems with pipeline architectures, but in the open challenge (where the systems were allowed to use extra training sets) joint learning systems performed slightly better, but not significantly. That is, the joint modeling of the task is not trivial (Surdeanu et al., 2008). Using the pipeline architecture has been followed by many later studies, as well (e.g. Titov and Klementiev (2012b) and Titov and Klementiev (2012a)).

As mentioned, current semantic role labelers perform fairly well, but they still suffer from some shortcomings. One problem is the error propagation problem of the pipeline architecture. That is, predicates' sense identification is not less important than semantic role classification sub-task. In addition, semantic roles are closely related to syntactic structures and SRL systems rely on syntactic parsers and utilize many features which are extracted from the parse trees. Therefore, errors in the parse trees can decrease the

performance of the SRL systems, as well. Furthermore, around 10% of semantic roles do not exactly match the syntactic constituents (Màrquez et al., 2008). In addition, current systems focus on verb and nominal predicates, while sentences can express relations using other lexical items, such as prepositions (Srikumar and Roth, 2011).

Furthermore, since typical semantic role labelers rely on the syntactic structures, they search for the semantic roles in the "local" context, where locality is a concept defined, in the first instance, as a predicate's predicate-argument structure (e.g. *Jones* in (8)), and then extended by its modifiers (e.g. *in September last year* in (8)), and adjoined constituents to its parent (or grandparent, great grandparent, etc.) (e.g. *Both planning and control* in (9)).

(8) [AGENT Jones] *arrived* in Paris [TIME in September last year].

(9) [GOAL Both planning and control] are difficult to *achieve* in this form of production.

1.3 Implicit Semantic Roles

In recent years, a more generalized version of semantic role labeling has been introduced which focuses on implicit semantic roles, DNIs (cf. Chapter 2)(Ruppenhofer et al., 2010), i.e. semantic roles which are not locally realized but can be retrieved from the context. More formally, implicit semantic roles are defined as those roles which are "neither instantiated as direct dependents of the target predicates nor displaced through long-distance dependency or co-instantiation constructions" (Ruppenhofer et al., 2010).

To better understand implicit roles, consider (10) indicates the annotation of a typical SRL system for the given sentence.

(10) [AGENT The annual visitors to Nepal] have also *brought* [THEME another problem] – litter.

In this case the AGENT role (i.e. the answer to WHO did the action), and the THEME role (i.e. the answer to WHAT was brought) are answered by the current annotation. But the predicate *brought* indicates a movement which calls for a SOURCE and a DESTINATION.

While the SOURCE is not mentioned at all, a human reader can infer *to Nepal* as the DESTINATION of the movement. *To Nepal*, in this case, is an instance of *Implicit Semantic Roles* whose recognition is out of the scope of the state-of-the-art SRL systems. An ideal SRL system which annotates both explicit and implicit roles is expected to annotate the sentence as:

(11) [AGENT The annual visitors [DESTINATION to Nepal]] have also *brought* [THEME another problem] – litter.

Here you can see another example, 12, in which the implicit role is not realized in the near vicinity of the predicate as in 10, but in the previous sentence.

(12) "This is not talk for a police-constable."
"[COGNIZER I] *know*, sir, I know" [3].

The COGNIZER (the person who knows the content) and the CONTENT (the object of the cognizer's awareness) are the required roles for the predicate *know* to fully express its meaning. While the COGNIZER is present as the subject, the CONTENT is realized in a wider context, one sentence earlier. Implicit roles can be realized even farther. For more details regarding the distance between the role filler and the predicate, cf. Chapter 5.

As the examples show, implicit semantic roles play the same role as the semantic roles with regards to the predicate, except that there is no/weak syntactic relationship between them and the predicates.

1.4 Implicit Semantic Role Labeling

Identification of implicit semantic roles can provide a lot of information that would be beneficial for NLP applications dealing with text understanding, such as information extraction, summarization, question answering, and textual entailment. In other words, Implicit Semantic Role labeling (ISRL) can be viewed as an enrichment of traditional SRL which extends the task from sentence-level to context-level to provide richer semantic information.

[3]Taken from Arthur Conan Doyle's "The Adventure of Wisteria Lodge" and part of the SemEval-10 Task-10 corpus (Ruppenhofer et al., 2010)

This task is not a new phenomenon and has been studied by Palmer et al. (1986) as a special case of anaphora and coreference resolution (CR). Similarly, Whittemore et al. (1991) treated the recognition of unexpressed roles as a special case of CR. Later, Burchardt et al. (2005) proposed that implicit semantic roles might be determined using the observed coreference patterns in a large corpus of text. However, they did not implement and evaluate their approach.

In 2010, a shared task attracted the attention of the research community to Implicit Semantic Role Labeling again. Ruppenhofer et al. (2010) defined a task in SemEval2010, called *Linking Events and Their Participants in Discourse* which addressed identification of correct fillers for the implicit semantic roles. Chen et al. (2010) and Tonelli and Delmonte (2010) participated in the task and obtained an F-score of 0.02 and 0.01, respectively. Since then many studies have tried to improve the results, e.g. Tonelli and Delmonte (2011), Ruppenhofer et al. (2011), Silberer and Frank (2012), Gorinski et al. (2013), Laparra and Rigau (2012), Laparra and Rigau (2013), but they have improved the results only to some extent, with the highest F-score of 0.19.

These studies have found that ISRL is a hard task in NLP (e.g. Roth and Frank (2013)) which involves many challenges. The main challenges of the task are briefly explained here:

- ISRL involves two challenging sub-tasks: identifying which roles are locally unrealized but can be inferred from the context, and finding the correct fillers for these roles in the discourse context.

  While it may seem that the second sub-task is the main focus of the whole task, the first sub-task is not less important. An evidence to this assertion is the number of irretrievable implicit roles in the annotated data set by Ruppenhofer et al. (2010) which is even more than that of recoverable ones, 335 vs. 245 (Laparra and Rigau, 2012) (for more details cf. Chapter 5). One reason to this is that some implicit roles can be *inferred* from the context, but are not directly mentioned. Another reason is that less important semantic roles from the viewpoint of the speaker/author remain completely unrealized at the surface level. For example, while all motion events require a PATH, this semantic role is rarely mentioned, either directly or indirectly.

- Most of the powerful syntactic features which are fundamental in traditional SRL are unavailable across sentence boundaries (cf. Chapter 3).

- The poor performance of the task participants (Chen et al. (2010) and Tonelli and Delmonte (2010)) has proved the inherent difficulty of the task.
- The fully annotated data set annotated with implicit semantic roles by SemEval2010 task organizers is a very small text with around 400 sentences. Prior studies have mentioned the small size of the data set as the main obstacle to effective computational modeling (e.g. Roth and Frank (2013); Chen et al. (2010); Laparra and Rigau (2013) and Silberer and Frank (2012)) and some of them have attempted to solve the data sparseness problem by weakly supervised training (Gorinski et al., 2013), and heuristic annotation of data (Silberer and Frank, 2012). These methods could improve the results, but their performance is still poor, with the highest F1-Score of 0.19. Also, another study (Gerber and Chai, 2012) achieved much higher performance of 0.50 F1-Score on the same task using a considerable amount of annotation instances for 10 nominal predicates and showed that the amount of training data plays a substantial role in ISRL.

In this dissertation, we focus on the last challenge mentioned above, *data sparseness*, because annotating more instances for implicit semantic role labeling task is very expensive, due to various reasons:

1) The task requires full-text annotation, which is time-consuming and implies each annotator to read and understand the whole discourse.

2) Consistency/reliability is harder than for overt roles because the task involves the complexity of a traditional SRL with that of coreference annotation (Gerber and Chai, 2012). The annotators may annotate different mentions of the same entity. So, to precisely evaluate the inter-annotator agreement, the annotation of coreference chains is also required.

3) Implicit roles are rarer than overt roles. For example, in the SemEval2010 training set 2726 overt roles are annotated, while this number for resolvable implicit roles is only 245 (Laparra and Rigau, 2012), which is around 8% of the overt roles. Therefore, to annotate a reasonable number of implicit roles to train a model, one needs to annotate large texts.

1.5 This Dissertation

Within this thesis, we address data sparseness problem of ISRL. For this purpose, we follow two approaches. In our first study, we focus on annotating more data in a simple

and cheap way. So, we utilize crowdsourcing as a time- and cost-effective way to have data annotated by non-experts.

Since potentially there are a large number of predicates in each text and annotating all of them with all semantic roles involves a few number of annotations per predicate, we follow a domain-specific approach and simplify the task by focusing on a subset of semantic roles for a few predicates. This approach is similar to Gerber and Chai (2012) who focused on 10 predicates from the financial topic and Kordjamshidi et al. (2010) who introduced a task called *Spatial Role Labeling* concentrating on annotating spatial roles answering WHAT/WHO/WHERE questions about the semantic structure of the given sentences.

In our crowdsourcing study, we focus on motion domain and pick a few predicates carrying the notion of movement, including *arrive*, *reach*, *pass*, etc. We present the predicates along with their previous context to the non-experts to annotate both explicit and implicit semantic roles.

The result shows that crowdsourcing can be an effective and cheap method to increase the amount of data annotated with implicit semantic roles and accordingly can help to solve the problem of data sparseness in ISRL. However, this method requires appropriate task design to ensure the reliability of the annotations. In addition, the resulting annotations are domain-specific which makes it difficult to scale up the task to arbitrary vocabularies.

In our second approach, we combine some existing corpora to address the data sparsity issue. However, the existing corpora belong to different genres and it is well known that the performance of NLP models reduces when applied across domains. This holds for traditional SRL (Carreras and Màrquez, 2005) and is likely to extend to ISRL as well. Therefore, to overcome the problem of domain difference, we propose applying a domain adaptation technique to bridge the differences between the corpora and ensure that reasonable generalizations can be learned.

In our experiments, we combined Gerber and Chai (2012) data set and the SemEval2010 corpus by applying feature augmentation (Daume III, 2007), an effective domain adaptation technique. The outcome indicated that we can profit from increasing the training data with data points from another genre and significantly improve the performance. The improvement however are more affected by the variety of the data than the amount of the additional data set.

To summarize our contributions, in this thesis, we investigate data sparseness issue which has been mentioned as the bottleneck of developing computational models for ISRL and propose two methods two solve this issue:

1) We collect annotations by non-experts using crowdsourcing and show that even without defining precise definitions for semantic roles, we can have a large text annotated reliably by non-expert people.

2) We apply domain adaptation as an effective technique to combine out of domain and in-domain data for ISRL and improve the performance of the system. This approach solves the problem of lack of enough data in a simple way without needing any additional annotation effort. We use this approach to combine two corpora and evaluate the system on both corpora and observe significant improvements in both.

1.6 Structure of this Dissertation

This thesis is structured as follows:

Part I, **Introduction**, is composed of three chapters. The first chapter presents the introduction. Chapter 2 introduces the two large corpora annotated with semantic roles, FrameNet (Baker et al., 1998) and PropBank (Palmer et al., 2005), and how they have annotated (Implicit) Semantic Roles. Chapter 3 provides a literature review on semantic roles, semantic role labeling and implicit semantic role labeling.

Part II, **Domain-focused Annotation of Implicit Semantic Roles**, consists of one chapter. Chapter 4 first discusses automatic identification of motion events using FrameNet (Baker et al., 1998) and WordNet (Miller, 1995) and then presents our crowdsourcing experiment to annotate implicit semantic roles of some motion events in a text.

Part III, **Domain Adaptation Technique in ISRL**, includes one chapter in which we discuss how we combine existing corpora and improve the performance of our ISRL system and evaluate scalability of the proposed method.

Part IV,**Summary and Conclusion**, includes only one chapter, Chapter 6, which concludes the thesis and suggests some ideas for future work.

Chapter 2

Corpora Annotated with Semantic Roles

2.1 Introduction

In this chapter, we introduce the linguistic foundation of SRL and the two major role-annotated corpora which are developed based on the proposed linguistic theories. These corpora, FrameNet (Baker et al., 1998) which is based on *Frame Semantics Theory* and PropBank (Palmer et al., 2005) which is a more syntax-oriented, semantic role annotated corpus, have made the data-driven modeling of SRL and ISRL possible. In this chapter, we present detailed information about these corpora.

2.2 Linguistic Foundation

2.2.1 Semantic Roles

Semantic roles, also called *thematic relations*, characterize the existing semantic relationship between a predicate and its components. As the presented examples in Chapter 1 showed, these relationships are usually indicated by labels such as AGENT, THEME, EXPERIENCER, etc.

Although many studies and theories have attempted to present an exhaustive inventory of semantic roles, no consensus has been reached so far by the linguistics community about the precise list and definition of these roles. A few thematic roles, including

AGENT, PATIENT, THEME, SOURCE and DESTINATION were introduced by Gruber (1965) but the leading theories on development of such roles were introduced by Fillmore (1968), Jackendoff (1972) and Dowty (1991). The main points and results of these studies are described in the following sections.

2.2.2 Case Grammar

In modern generative grammar, Fillmore (1968) introduced *case grammar* whose focus was on presenting a semantic grammar. Prior studies on *case* had considered various semantic relationships between nouns and other components of the sentence in the same way as studying semantic function of inflectional affixes on nouns. In other words, in these studies *cases* did not exist in the deep structure of the sentence, but they were seen as morphophonemic realization of syntactic relations in the surface structure (Fillmore, 1968, p. 5). By introducing *case grammar*, Fillmore (1968) presented a new theory in which syntactic components such as subject and object were removed from the deep structure and case relationships formed its basic components.

In this theory, he attempts to find some "deep-structure cases" which can be realized by different syntactic surfaces in various languages and can represent the semantic function of arguments of the predicates. He introduced the notion *case* as the base component of the grammar of every language and defined it as "a set of universal, presumably innate, concepts which identify certain types of judgments human beings are capable of making about the events that are going on around them" and proposed six semantic cases including (Fillmore, 1968, pp. 46-47):

- AGENTIVE (A): the case of the typically animate perceived instigator of the action identified by the verb.

(13) [AGENT John] broke the window.

- INSTRUMENTAL (I): the case of the inanimate force or object causally involved in the action or state identified by the verb.

(14) [INSTRUMENT A hammer] broke the window.

- DATIVE (D): the case of the animate being affected by the state or action identified by the verb.

(15) John gave the books to [DATIVE my brother].

- FACTITIVE (F): 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.

(16) John ruined [FACTITIVE the table].

- LOCATIVE (L): the case which identifies the location or spatial orientation of the state or action identified by the verb.

(17) It is windy in [LOCATIVE Chicago].

- OBJECTIVE (O): 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.

(18) [OBJECTIVE The door] opened.

(Fillmore, 1968, p. 46) notes that the proposed list is not exhaustive and additional cases may be added.

The aim of the case grammar theory is to explain that none of the defined cases corresponds to a specific surface structure relationship, such as subject or object. Therefore, *John* in (19) and (20), *the key* in (21) and (22), and Chicago in (23) and (24) are similar in terms of their cases which are AGENT, INSTRUMENT and LOCATIVE, respectively.

(19) John opened the door.

(20) The door was opened by John.

(21) John used the key to open the door.

(22) John opened the door with the key.

(23) Chicago is windy.

(24) It is windy in Chicago.

Another evidence presented by Fillmore (1968) in support of defining deep cases can be exemplified by comparing (25), (26) and (27).

(25) John broke the window.

(26) A hammer broke the window.

(27) *John and a hammer broke the window.

Case grammar argues that only noun phrases representing the same case can be conjoined. Therefore, *John* in (25) which fills the AGENT case and *A hammer* in (26) which fills the INSTRUMENT case can not be conjoined. In other words, (27) shows that an attempt to assign two different deep cases, AGENT and INSTRUMENT, to the subject fails.

Cases in case grammar theory, or semantic roles in general, are considered to have the following main features:

- Cases (or semantic roles) can be listed as a fixed set.
- Each syntactic arguments of the predicate is assigned with a case (or semantic role).
- No argument of the predicate can be assigned more than one case (or semantic role).

The first main characteristic of cases, i.e. presenting a list of semantic roles, was the focus of many studies after the case grammar theory. These studies attempted to present more precise definitions and a universal set of roles. Huddleston (1970), for example, questioned the distinction between the INSTRUMENTAL and AGENTIVE thematic roles in Fillmore's theory and stated that according to Fillmore's definitions, *the key* and *the wind* in (28) and (29) have both INSTRUMENTAL case while comparing their semantics determines that (28) assumes presence of an agentive participant, while (29) does not. Thus, he classified *the wind* in (29) under a new category of thematic roles, called FORCE (Huddleston, 1970, p. 504).

(28) The key opened the door.

(29) The wind opened the door.

In another study, Starosta (1988) suggested a different set of semantic roles composed of PATIENT, AGENT, EXPERIENCER, LOCUS, CORRESPONDENT, and MEANS. Later attempts of Fillmore to complete the list of semantic roles firstly caused adding some more roles to the list, such as SOURACE, GOAL, LOCATION and EXPERIENCER (cf. Table 2.1), but in his later studies he found that a small list of deep cases is not enough to characterize all arguments of all predicates and introduced *Frame Semantics* theory which motivated the development of the FrameNet resource (cf. Section 2.3).

TABLE 2.1: Most common semantic roles and their prototypical definitions

Semantic role	**Definition**
AGENT	animate instigator of the action
THEME	entity which undergoes movement
PATIENT	the affected entity
EXPERIENCER	the entity receiving emotional feeling
INSTRUMENT	the entity used to carry out an action
FORCE	the entity performing an action mindlessly
SOURCE	where the action originates
GOAL	the place to which the action is directed
LOCATION	where the action occurs

2.2.3 Frame Semantics

The term *Frame Semantics* (Fillmore, 1982) refers to a theory which is based on the premise that "word meanings are relativized to scenes"(Fillmore, 1977). This theory describes the meaning of lexical items in terms of *prototypical scenes* and explains that words can be categorized under classes which are motivated by evoking *scenes* or *situations*. Such motivating scenes are called *frames* (Fillmore, 1982).

As Fillmore and Atkins (1992) state, in *Frame Semantics*

"A word's meaning can be understood only with reference to a structured background of experience, beliefs, or practices, constituting a kind of conceptual prerequisite for understanding the meaning. Speakers can be said to know the meaning of the word only by first understanding the background frames that motivate the concept that the word encodes. Within such an approach, words or word senses are not related to each other directly, word to word, but only by way of their links to common background frames and indications of the manner in which their meanings highlight particular elements of such frames." (Fillmore and Atkins, 1992, pp. 76-77)

2.2.3.1 Frame

The notion *frame* is defined as "any system of concepts related in such a way that to understand any one of them you have to understand the whole structure in which it fits." (Fillmore, 1982, p. 111)

Frame in Frame Semantics can be exemplified by a "commercial event". There are a large number of English verbs which are semantically related to commercial transaction and evoke a commercial *scene*. Such a *scene* includes the following elements: a buyer, a

seller, goods and money. "buy", "sell", "pay", "spend" and "cost" are examples of verbs which evoke a commercial situation. The difference between them is that "buy" focuses on the *buyer* and *goods*, "sell" focuses on *seller* and *goods*, "pay" focuses on *seller*, *buyer* and *money*, and so on. The main idea is that whoever who knows the meaning of one of these verbs, knows the details of a commercial scene and also knows the meaning of the other verbs (Fillmore, 1982, pp. 116-117). Therefore, based on Frame Semantics theory, all these verbs must be classified under the same frame.

Another example which highlights the notion of frame is the difference between the words "land" and "ground". While both these words refer to the dry surface of the earth, they are different regarding the distinction they make. "Land" is distinct from the "sea", whereas "ground" is distinct from the "air". Therefore, a bird that lives on the land means that it does not spend its life in water and a bird which lives on the ground is described as a bird that does not fly. As can be seen, while "land" and "ground" are not much different in what they are, their difference can be seen in how they settle that thing in a larger frame (Fillmore, 1982, p. 121). Therefore, to illustrate the similarities between such words and distinguishing them from other words, the frame semantics theory proposes to define some larger structures, called *frames*.

2.2.3.1.1 Prototype As mentioned, situations in frame semantics are *prototypically* described by frames, that is, not all the conditions which define the prototype are required to be present in order to properly understand and use the word. For example, understanding the meaning of the word *breakfast* requires understanding the culture that people eat three meals a day at almost fixed times, one of which is eaten in the morning, after a period of sleep and includes a special menu. However, even when one of the above conditions is absent, there is no change in using the word by native speakers. The word *breakfast* defines a category which can be used in many different background situations. (Fillmore, 1982, p. 119)

Defining words using frame and prototype notions is a useful approach to avoid the problem of defining boundary conditions for linguistic categories and distinguishes frame semantics theory from checklist theories which require a precise checklist of conditions which must be satisfied in order for the word to be used appropriately (Petruck, 1996). This approach led to development of FrameNet which categorizes the words under different frames and define a set of semantic roles for each frame (cf. Section 2.3).

2.2.4 Jackendoff's Theory

Another dominant theory on semantic roles was introduced by Jackendoff (1990) who extended the inventory of semantic roles which were previously developed by Gruber (1965). Gruber (1965) studied the semantics of verbs of motion, possession and position and presented a list of thematic roles involved in such events: THEME, SOURCE, GOAL and AGENT. Jackendoff (1990) proposed considerable modifications to this list based on his theoretical framework, called *conceptual semantics*.

He believed that the language is organized in three different levels of structure - phonological, syntactic and semantic/conceptual- each of which has its own primitives, combination principles, and sub-components. For example, surface and deep structures are sub-components of the syntax level (Jackendoff, 1987).

According to this theory, the meaning of linguistic expressions are represented by a conceptual structure whose building blocks are *conceptual constituents* which belong to one of the following six main ontological categories: *thing*, *event*, *state*, *place*, *path*, or *property*. Each syntactic constituent in the sentence corresponds to a conceptual constituent in the meaning of the sentence. For example, in (30), *John* and *the house* can be mapped to *thing* constituents, *toward the house* to a *path* constituent and the whole sentence correspond to an *event* constituent (Jackendoff, 1990, p. 22)

(30) John ran toward the house.

Each conceptual constituent is composed of a *function-argument* structure. The function determines the conceptual constraints on the arguments of the function and also the relation between the arguments. It is noticeable that the arguments must correspond to the conceptual constituents. For example, as illustrated in (31), the *place* conceptual constituent can be expanded into a *place function* with an argument of type *thing*. In this case, the argument is a spatial point which enables the *place function* to define a region. (32) and (33) show *event* and *state* conceptual constituents and their function-argument structures. In these cases, the functions take two arguments. (Jackendoff, 1987).

(31) Place $\rightarrow$ [$_{\text{Place}}$ Place-function (thing)]

(32) Event $\rightarrow$ [$_{\text{Stay}}$ Event-function (thing, path)]

(33) State $\rightarrow$ [$_{\text{Be}}$ State-function (thing, place)]

Some concrete examples of such formation rules can be found in (34), (35) and (36).

(34) Under the table

(35) Bill stayed in the kitchen.

(36) The dog is in the park.

In (34), *the table* serves as a reference object and *under* determines a *place-function* which defines a specific region, *under the table*. In (35), *stay* represents an *event-function* which expresses stasis of something over a time span. In this case, *Bill* is the *thing* stayed still and *in the kitchen* is the *place* in which *Bill* is located. In (36), *is* acts as the *state-function* which determines the location of an object. In this case, the *thing* is the *dog* and *in the park* denotes the *place* of *the dog* (Jackendoff, 1987).

In Jackendoff's theory, thematic roles are assumed to be structural relations within conceptual structures. For example, THEME thematic role which was defined as "the object in motion or being located" in Gruber (1965) can be defined as the first argument of *go*, *stay*, *be* and *orient* event-functions of conceptual structures. SOURCE and GOAL can be defined as the argument of *from* and *to* path-functions, respectively and AGENT can be defined as the first argument of *cause* event-function.

The advantage of Jackendoff's theory in defining thematic roles compared to Gruber (1965) and Fillmore (1968) is that it defines semantic roles as correspondence to positions in the defined semantic structures instead of presenting them as a list of labels. In addition, in contrast to the definition of thematic roles in Fillmore (1968) and Gruber (1965), the relationships between different roles are clear from the conceptual structures. Also, this theory avoids defining a default thematic role for syntactic constituent (as done in Fillmore (1968)) (Wagner, 2004, p. 60).

2.2.5 Dowty's Theory

Dowty (1991) motivates his theory by referring to prior studies whose attempts to provide a universal set of semantic roles have yielded no consensus. He mentions the main problem lack of a consensus about "what semantic roles are". Therefore, instead of defining discrete categories as semantic roles, he proposes to view semantic roles as prototypical concepts. He states that semantic roles are not "discrete categories", but "cluster concepts" and assumes that different arguments can have "different degrees of membership" in the role types. Then, he postulates only two roles: proto-agent and

proto-patient and presents a list of properties which entities belonging to these role types should have. The list is as follows:

Contributing properties for the proto-agent (Dowty, 1991, p. 572)

a. Volitional involvement in the event or state
b. Sentience (and/or perception)
c. Causing an event or change of state in another participant
d. Movement (relative to the position of another participant
e. (Exists independently of the event named by the verb)

Contributing properties for the proto-patient (Dowty, 1991, p. 572)

a. Undergoes change of state
b. Incremental theme
c. Causally affected by another participant
d. Stationary relevant to movement of another participant
e. (Does not exists independently of the event, or not at all)

According to the list above, an entity assigned to the proto-agent role should volitionally involve in the event or state, whereas an entity in the proto-patient role category should involuntarily undergo a change of state. However, constituents are assigned to these roles depending on how many of these contributing properties they have. An NP which meets all or most of the criteria for either proto-agent or proto-patient is a *good* example to be assigned to the relevant category. But it is still possible that it meets only one of the criteria of one of the role types and be excluded from the other class. For example in

(37) John *builds* a house.

John meets all the criteria to be assigned to the proto-agent role and *a house* meets all the criteria to be a proto-patient. In contrast, in

(38) John *is* disappointed (Dowty, 1991, p. 573).

John meets only the second criteria (sentience and/or perception) of the proto-agent role type and is assigned to it. It means that Dowty's roles are essentially prototypes

(similar to FrameNet frames) and it lets the theory to solve the problem of formulating precise definitions for traditional semantic roles and determining the boundaries between different semantic roles. This theory is the basis of the argument annotation in PropBank.

2.3 FrameNet

Frame semantics was proposed to account for phenomena in syntax (e.g. Lambrecht (1984)) and morphology (e.g. Petruck and Boas (2003)), but it was most useful in semantics, more specifically computational lexicography. The first operationalization was in terms of a study by Fillmore and Atkins (1992) who did a large-scale study on "risk" lexeme in English as the first attempt to describe a *frame-based* dictionary. A frame-based lexicon is distinct from traditional print dictionaries in that in such a dictionary "word senses, relationships between the senses of polysemous words and relationships between senses of semantically related words are linked using frames" (Fillmore and Atkins, 1992, p. 75).

After this study showed the feasibility to apply frame semantics for free text occurrences, the FrameNet project was initiated. The FrameNet project has produced frame-semantic descriptions for thousands of English lexical items and has annotated example sentences from contemporary English corpora, The British National Corpus and the LDC North American Newswire Corpora manually (Baker et al., 1998).

As of 01.03.2016, according to https://framenet.icsi.berkeley.edu/fndrupal/current_status, FrameNet lexicon database contains more than 13,000 lexical units (cf. Section 2.3.2) which belong to more than 1200 frames. The whole corpus includes more than 170,000 annotated sentences.

2.3.1 Frame

The key concept of FrameNet annotation is *semantic frame* which was introduced in frame semantics theory (cf. Section 2.2.3.1). For example, *ARREST*, *SLEEP*, and *TELLING* are examples of frames which can be evoked by words such as *apprehend*, *hibernate* and *notify*. These examples are fairly specific frames which refer to specific situations. However, frames range from highly abstract to very specific. For example, *MOTION* frame is an abstract frame which is evoked by more than 20 lexical units,

while the frame *PRANK* is a very specific one which includes only two lexical units: *practical joke.n* and *prank.n*.

2.3.2 Lexical Unit

Each sense of a word is called a Lexical Unit (LU) and typically different senses of polysemous words belong to different frames (Ruppenhofer et al., 2006, p. 5). For example, in

(39) He will *break* your arm.

(40) I *broke* the car windows with a stone to try to reach Mr Pickering [1].

the two *break* verbs belong to different frames: *CAUSE_HARM* and *CAUSE_FRAGMENT*.

Each frame is evoked by lexical units which linguistically express the situation of the frame. For example, *ARRIVING* frame is evoked by *appear*, *approach*, *arrival*, *arrive*, *come*, etc. These frame evoking lexical units are also called *target*s and can be verbs, nouns or adjectives.

2.3.3 Frame Element

Frame Elements (FEs) are participants associated with the frame (Baker et al., 1998) which are shared among lexical units of the same frame. For example, in *ARRIVING* frame, THEME (the object that moves) and GOAL (the place where the object ends up its movement) are the required frame elements by all lexical units evoking the frame.

In the simplest case, the frame-evoking lexical unit is a verb and the FEs are its syntactic dependents. (41) describes an *arriving* situation in which the lexical unit *arrived* evokes the *ARRIVING* frame and the two frame elements, THEME and GOAL are realized by *they* and *in this country*.

(41) [THEME They] *arrived* [GOAL in this country].

Defining a frame in FrameNet includes describing the situation based on the relations between the frame elements, defining the frame elements and naming the lexical units that evoke the frame. Table 2.2 shows an example of a FrameNet frame.

[1] All examples in the FrameNet section are taken/adapted from FrameNet, unless otherwise mentioned.

TABLE 2.2: An example of a FrameNet Frame (taken from FrameNet)

Frame Name	**Giving**
Frame Definition	A DONOR transfers a THEME from a DONOR to a RECIPIENT. This frame includes only actions that are initiated by the DONOR (the one that starts out owning the THEME). Sentences (even metaphorical ones) must meet the following entailments: the DONOR first has possession of the THEME. Following the transfer the DONOR no longer has the THEME and the RECIPIENT does.
Frame Elements	DONOR: The person that begins in possession of the THEME and causes it to be in the possession of the RECIPIENT. RECIPIENT: The entity that ends up in possession of the *theme*. THEME: The object that changes ownership.
Lexical Units	advance.v, bequeath.v, charity.n, contribute.v, contribution.n, donate.v, donation.n, donor.n, endow.v, fob off.v, foist.v, gift.n, gift.v, give out.v, give.v, hand in.v, hand out.v, hand over.v, hand.v, leave.v, pass out.v, pass.v, treat.v, volunteer.v, will.v

(42) is an exemplar sentence from FrameNet for the frame *GIVING* which has annotated all participants of the events.

(42) Katy and Jamie got ready very quickly and [DONOR Mum] *gave* [RECIPIENT each of them] [THEME two wee spoons.]

According to the definitions in Table 2.2, *Mum* who is the giver is annotated as the DONOR, the *two wee spoons* which are the transferred objects are the THEMEs and *each of them* which refers to Katy and Jamie, the recipients of the spoons, are the RECIPIENTs.

2.3.3.1 Core and Non-core Frame Elements

Frame elements are classified, among other things, in terms of how central they are to the given frame. This classification distinguishes three levels of frame elements: core, peripheral and extra-thematic frame elements.

A *core* frame element is a conceptually necessary participant of a frame, which makes the frame unique and distinct from other frames (Ruppenhofer et al., 2006, p. 19). For example, in the *TOPIC* frame, COMMUNICATOR, TEXT, and TOPIC are all core

frame elements, because an event like *discussing* necessarily includes these participants. One cannot imagine an act of discussing which does not involve any TOPIC or COMMUNICATOR. In other words, core FEs are, by default, assumed by frame semantics to be realized. However, this is not always the case. The next section, section 2.3.3.2, discusses the situation in which a core frame element is missing in more details. There is also another class of frame elements, called *incorporated frame elements*, which seems to be missing and is discussed in section 2.3.3.3.

A *peripheral* frame element is a frame element which is not necessary conceptually and does not uniquely characterize the frame. Peripheral frame elements are usually repeated in different frames and mark notions such as TIME, PLACE, MANNER, MEANS, etc (Ruppenhofer et al., 2006, p. 20). It is remarkable however that peripheral frame elements of a frame can be core frame elements of another frame. For example, PATH which is a peripheral FE in *ARRIVING* frame, is a core FE in *MOTION* frame. In (43), core and peripheral frame elements of a sentence from FrameNet are shown. In this sentence, COMMUNICATOR, TOPIC and TEXT are core frame elements and DEGREE is a peripheral one.

(43) [COMMUNICATOR I] shall *discuss* [TOPIC these effects] [DEGREE further] [TEXT in section 2.3.

Extra-thematic frame elements have a considerably different interpretation compared with *core* and *peripheral* frame elements. They have no direct relationship with the situation identified by the frame they are listed in, but they usually evoke a larger frame which embeds the frame in which they appear (Ruppenhofer et al., 2006, p. 89).

For example, in

(44) I *rode* to school [COTHEME with her] all the time.

the word *rode* evokes the *RIDE_VEHICLES* frame, whose COTHEME frame element is an extra-thematic one, because it implies a *motion* event and therefore evokes the *MOTION* frame whose THEME frame element is realized by the COTHEME of *RIDE_VEHICLES* frame, and whose other FEs are co-identified with FEs of *RIDE_VEHICLES*. Therefore, not only *I*, but also *her* is described as moving to school (Ruppenhofer et al., 2006, p. 89).

Peripheral and extra-thematic frame elements are called *non-core* frame elements.

2.3.3.2 Null Instantiation (NI)

Sometimes FEs that are conceptually necessary, i.e. core frame elements, remain unexpressed as a lexical or phrasal material and are called *Null Instantiations* (Ruppenhofer et al., 2006, p. 24). In such cases, FrameNet indicates their absence and classify them based on their omissibility conditions which are discussed in the following paragraphs.

Null instantiation in FrameNet originates from the Fillmore's null complementation theory (Fillmore, 1986) which states that the interpretation process of a sentence is not finished until all obligatory complements of a predicate are located. Therefore, if an obligatory complement of a predicate is missing, one must look for the explanation of the omission. For example, some grammatical constructions allow omission of some of the participants of the event: imperative constructions commonly lack a subject (Fillmore, 1986) (for more details about related literature cf. Chapter 3).

Following this approach, FrameNet identifies three categories of null instantiations: Definite Null Instantiation (DNI), Indefinite Null Instantiation (INI) and Constructional Null Instantiation (CNI):

A) Definite Null Instantiations:

Definite null instantiations (DNIs) (or anaphoric null instantiation) are those missing frame elements which can be understood from the discourse context (Ruppenhofer et al., 2006, p. 24). For example, the GOAL of *come* predicate in (45) is a DNI which is not overtly expressed but can be understood from the context. The reader can recognize *this lonely, silent house* in the previous sentence as the filler of the GOAL frame element.

(45) "Well, sir, it is this lonely, silent house and the queer thing in the kitchen. Then when you tapped at the window I thought [$_{\text{THEME}}$ it] had *come* again." [GOAL_{DNI}][2]

B) Indefinite Null Instantiations:

Indefinite Null Instantiations (INIs)(also called existential null instantiations) are usually the missing frame elements of typically transitive verbs such as eat, bake, drink, etc., which are used intransitively (Ruppenhofer et al., 2006, pp. 24-25).

In such cases, the semantic type of the INI can be understood from the text, but no specific discourse referent can be retrieved. For example, with *eat* the missing frame

[2]The example is adapted from the training set of SemEval2010-Task 10 (Ruppenhofer et al., 2010).

element is understood to be a meal, and with *drink* it is likely to be beverage. (46) indicates an *eating* event which requires INGESTOR (the eater) and INGESTIBLE (the thing eaten). The INGESTOR is overtly realized, whereas the INGESTIBLE remains unrealized as an INI.

(46) [INGESTOR We] *ate* [PLACE on the boat]. [INGESTIBLE$_{\text{INI}}$].

C) Constructional Null Instantiations:

Constructionally unrealized elements (CNIs) (also called structurally omitted elements) are those missing elements whose omission is due to a grammatical construction which lets a missing argument. Omitted subjects of imperative sentences and missing agents in passive sentences are examples of constructionally licenced null instantiations (Ruppenhofer et al., 2006, pp. 25-26). (47) and (48) are instances of the above grammatical structures with CNIs.

(47) includes an imperative sentence in which the THEME (the moving object) is missing and (48) involves an *eating* predicate whose INGESTOR is missing due to the passive structure. (47) has also another missing argument, GOAL which is a DNI.

(47) Finally, check the airspeed carefully and *approach* [MANNER with an adequate amount of height and speed]. [THEME$_{\text{CNI}}$][GOAL$_{\text{DNI}}$]

(48) [Ingestible Nuts] can be *eaten* in moderation. [Ingestor$_{\text{CNI}}$]

2.3.3.3 Incorporated Frame Element

Some frames contain lexical units which **incorporate** information about one of their frame elements in their definition, i.e. the identity of one of the frame elements is understood from the meaning of the verb. Some of *PLACING* and *BODY MOVEMENT* verbs are such cases (Ruppenhofer et al., 2006, p. 26). For example in

(49) [THEME The wine] is *bottled* [MANNER under pressure].

bottled is a *PLACING* predicate which requires a GOAL. However, the GOAL, which is understood to be *in the bottle*, is not mentioned explicitly but is implied by the predicate itself.

As an example of *BODY MOVEMENT* predicates, in

(50) [AGENT Cranston] *blinked* [PURPOSE to hide his tears].

blinked has its BODY_PART frame element as an incorporated role.

Nevertheless, it should be noticed that it is still possible to specify the incorporated frame elements further. Therefore, (49) and (50) can be re-written as (51) and (52):

(51) [THEME The wine] is *bottled* [GOAL in red bottles] [MANNER under pressure].

(52) [AGENT Cranston] *blinked* [BODY_PART his eyes] [PURPOSE to hide his tears].

2.3.4 Locality

FrameNet annotates roles realized by direct dependents of the target predicate as overt semantic roles and the roles that are not realized in the same sentence as the predicate as null instantiations. However, there are some roles which are realized in the same sentence as the target predicate but farther away and not as a syntactic argument of the predicate. In this case, FrameNet uses its definition for the locality concept to annotate the frame element as *overtly realized* or *null instantiation*.

Normally, FrameNet annotates all frame elements which are realized by constituents inside the maximal phrase headed by the target word. However, there are two exceptions in which syntactically non-local constituents are annotated as overtly realized elements: raising and control structures. In such cases, one of the arguments of the higher raising/control predicate is also interpreted as an argument of the target, even though the maximal phrase headed by the target word does not dominate the shared argument (Ruppenhofer et al., 2006). (53) and (54) are instances of raising and control structures.

(53) [AVENGER John] *seems* to have *avenged* [INJURY the death of his brother] [PUNISHMENT by luring Smithers into a trap].

(54) [AVENGER They] are *hoping* to *get even* [OFFENDER with Smithers] [INJURY for the insult][3].

In these examples, *John* and *they* are arguments of *seems* and *hoping* predicate which are shared with *avenged* and *get even* predicates, respectively. All the annotated semantic roles in these examples are in the *local* vicinity of the target predicate and therefore considered overt semantic roles.

[3]Examples are adapted from Ruppenhofer et al. (2006).

2.4 PropBank

While FrameNet is primarily a lexicographical project, development of PropBank, another corpus for semantic role annotation, has followed a more practical goal: providing predicate argument annotation for Penn Treebank (Marcus et al., 1994) corpus (Ellsworth et al., 2004). The Penn Treebank is an English newswire text which has been provided with manual annotations of syntactic structures (Marcus et al., 1994). PropBank annotates the Wall Street Journal section of the Penn Treebank II, the second release of the Penn Treebank with an updated bracketing style, which is designed to allow the extraction of simple predicate-argument structures. PropBank focuses on the argument structure of verbs and annotates all verbal predicates with their semantic roles, that is, constituents which are predicates' semantic arguments and also adjuncts. It includes more than 112,000 semantic roles annotated for 3,256 distinct verbs (Gerber and Chai, 2012). PropBank is also distinct from FrameNet in terms of the text it annotates. While FrameNet provides annotations for individual sentences and includes a few running text annotations, the texts annotated in PropBank are only running texts.

2.4.1 Predicates in PropBank

PropBank makes only coarse word sense distinctions, that is, senses are kept apart only if they require different number of arguments (Palmer et al., 2005). It names each verb along with its set of roles as a *frameset*.

2.4.2 Arguments in PropBank

PropBank calls the semantic roles *Arguments* and follows a different approach from FrameNet in annotating them. FrameNet is based on *frame semantics* theory (cf. Section 2.2.3) in which specific labels are assigned to the participants of an event based on the relation between them and the given event. In contrast, PropBank follows Dowty's proto-agent\proto-patient theory (cf. Section 2.2.5). PropBank determines two main types of semantic roles, numbered arguments and adjuncts which are described in the following sections.

2.4.2.1 Numbered Arguments

Core semantic roles in Propbank are labeled as numbered arguments (ARG0–ARG4). While ARG0 and ARG1 correspond to Dowty's porto-agent and porto-patient roles (cf. Section 2.2.5) respectively, the argument types ARGn for n≥ 2 are generalizable across verbs to a much lower degree than Frame Elements of a Frame in FrameNet.

For example, while the definition of ARG3 for the prdicate *arrive* is START_POINT, it is defined as BENEFACTIVE for predicate *choose*. Table 2.3 illustrates the equivalent of the numbered arguments of PropBank in Dowty's theory and traditional set of semantic roles.

TABLE 2.3: Numbered arguments in PropBank

Argument Tag	**Equivalent**
A0	Proto-agent
A1	Proto-patient
A2	Beneficiary, instrument, attribute, end state
A3	Start point, beneficiary, instrument, attribute
A4	End point

(55) and (56) present annotation of some sentences from PropBank and the definition of their arguments.

(55) [ARG1 Imports] have *gone* down [ARG2 33%].

ARG1: Entity in motion/goer

ARG2: Extent

(56) [ARG0 John] *left* [ARG1 Mary] [ARG2 alone].

ARG0: Entity Leaving

ARG1: Place, person or thing left

ARG2: Attribute of ARG1

Imports and *Mary*, fillers of ARG1, can be both classified under proto-patient role type of Dowty's theory as they undergo changes (cf. Section 2.2.5) and *John* in the second example can be categorized as proto-agent due to its agentive role. In contrast, *33%* and *alone* fill ARG2 role whose definition depends on the predicate in question.

2.4.2.2 Adjuncts

In addition to the verb-specific numbered arguments, PropBank defines several more general roles which can apply to any verb. They are adjunct-like arguments, such as direction, manner and time, which are shown as ARGMs and can be compared to non-core frame elements in FrameNet (cf. Section 2.3.3.1). The numbered arguments are conceptually necessary for the predicate to present its meaning, while adjuncts provide additional unnecessary information about the event. ARGM_MNR determines *how* the action/event happens; ARGM_TMP characterizes the temporal placement of the event; and ARGM_MOD tags an embedding modal verb if there is one.

Table 2.4 provides the complete list of adjuncts in PropBank.

TABLE 2.4: Adjuncts in PropBank

Argument Tag	**Argument Tag**
LOC: Location	CAU: Cause
EXT: Extent	TMP: Time
DIS: Discourse Connective	PNC: Purpose
ADV: General Purpose	MNR: Manner
NEG: Negation Marker	Dir: Direction
Mod: Modal Verb	

(57) and (58) present full annotation for some sentences from PropBank, including both numbered and adjunctive arguments.

(57) [ARG1 Boeing's plans] for the 767 *went* [ARGM_MNR without a hitch].

ARG1: Entity in motion/goer

(58) [ARG0 Argentine negotiator Carlos Carballo] [ARGM_MOD will] *meet* [ARG1 with banks] [ARGM_TMP this week].

ARG0: Meeter

ARG1: Person/ Entity/ Object being met

2.4.2.3 Null Instantiations

PropBank does not provide any distinction between different types of uninstantiated arguments like in FrameNet (cf. Section 2.3.3.2) and it also does not determine if the

predicate has a null instantiated argument. For example PropBank does not present any clue that the ARG0 in (59) is missing.

(59) [$_{\text{ARG1}}$ John] was *hit*.

2.4.3 Locality

PropBank is built on a specific syntactic structure, Penn Treebank. Among the syntactic structures, Penn Treebank annotates movements of the constituents by traces and their antecendents. For example, (60) presents the trace annotation of a passive sentence in Penn Treebank. *"*"* in (60) represents a passive trace and the number beside the "*" determines its reference in the sentence (*John*).

(60) John-1 was *hit* [*-1] by Mary.

In such cases, PropBank annotates the trace as the role filler and presents the relation between the trace and its reference as a coreference chain. (61) presents the annotation of PropBank for a moved constituent.

(61) John was *hit* [$_{\text{ARG1}}$ *-1] [$_{\text{ARG0}}$ by Mary].

PropBank's approach in annotating such cases is similar to FrameNet which annotates structures like raising. The difference is that when a movement occurrs in the sentence which leaves a trace, PropBank annotates the trace rather than the NP, (in (59) *John*), as the argument of the predicate. Missing elements which leave a trace in the parse tree are called *null element*s in Penn Treebank. It is noticeable however that PropBank annotates them as overt roles and the term *null element* must not be mistaken with null instantiation in FrameNet.

Other examples of moved constituents with a trace are fronted or dislocated arguments:

(62) There-1 I *put* the book [*T*-1].

(63) There-1 [$_{\text{ARG0}}$ I] *put* [$_{\text{ARG1}}$ the book] [$_{\text{ARG2}}$ [*-1] [4].

*[*T*]* in (62) represents the Penn Treebank annotation for a movement trace and (63) illustrates that PropBank annotates the trace as the role filler rather than the fronted NP.

[4]The examples are taken from Bonial et al. (2010)

2.5 NomBank

A complementary project to PropBank is NomBank (Meyers et al., 2004) which addresses the annotation of nominal predicates and their arguments. This corpus presents semantic annotation of arguments that co-occur with nouns in the PTB sections annotated by PropBank. Therefore, it complements PropBank which focuses on annotation of verbal predicates. The NPs whose arguments are annotated in NomBank, markable NPs, satisfy one of the following conditions (Meyers, 2007, p. 7):

"I. *NP* must contain at least one (unincorporated) argument.

II. The head of NP must be of a *propositional* type (representing an event, state, etc.) and NP must contain at leas one *proposition-modifying* adjunct.

III. The head of N takes an argument which occurs in structures such as support verbs and transparent nouns (cf. Section 2.5.2)."

To better understand the first condition, consider (64) and (65):

(64) A [$_{\text{ARG1}}$ math] [$_{\text{ARG0\_Incorporated}}$ *teacher*]

(65) A teacher

The word *teacher* in (64) is considered a markable NP, because it has one argument, ARG1, in addition to its incorporated argument, ARG0. But *teacher* in (65) is not treated as a markable NP because the word itself is the only argument (Meyers, 2007, p. 8).

Propositional NP, as stated in the second condition, means an NP representing an event, relation or state (Meyers, 2007, p. 8).

2.5.1 Arguments

Similar to PropBank, NomBank arguments are classified as numbered arguments and adjuncts.

To determine arguments of a nominal predicate, all non-heads inside the noun phrase are considered as potential arguments: possessives, pre-nominal, modifiers, adjectives, PPs, post-nominal clauses. However, two points here are remarkable: firstly, not all non-heads inside the NP necessarily fill a semantic role and secondly, the realization of

the arguments is not bounded to the NP boundaries and in some cases role fillers can be realized out of the NP (cf. Section 2.5.2).

To determine the argument status of a non-head inside a markable NP that fills a role, the relation between these non-heads and the head noun must be considered. For example, while *math* in *the math teacher* fills an argument of the *teacher* predicate, *tall* in *the tall teacher* is not an argument. The reason is that *tall* can co-occur with many head nouns and their co-occurrence depends on the compatibility of the meaning of *tall* and the head noun, while the meaning of *the math teacher* is derived from the argument-taking properties of *teacher*, i.e., teachers teach subjects (Meyers, 2007, pp. 7-8).

2.5.1.1 Adjunctive Arguments

Annotation of adjunctive arguments in NomBank is exactly the same as PropBank. (66) presents full annotation of numbered and adjunctive arguments of a nominal predicate, *appearance*, in NomBank.

(66) [ARG1 His] [ARG2 uncombed] *appearance* [ARGM_MNR among these buttoned-up chaps]

2.5.1.2 Incorporated Arguments

NomBank considers annotation of incorporated roles in representation of argument structure. (67) and (68) indicate examples of incorporated roles in NomBank.

(67) When the [*ARG2 price*] [ARG1 of plastics] took off in 1987, [ARG0 Quantum Chemical Corp.] went along for the ride.

(68) The [*ARG2 funds]* should help ease a cash bind at HealthVest. [5]

As can be seen, *price* and *funds* are incorporated *ARG2* of themselves as predicates.

2.5.1.3 Null Instantiations

NomBank follows the same approach as PropBank in annotation of null instantiations, that is, it only annotates local semantic roles and does not provide any information

[5]Examples are taken from Gerber and Chai's dataset (Gerber and Chai, 2010)

about the missing roles, e.g. if there is an implicit role, or what is the type of the missing role.

2.5.2 Locality

As mentioned earlier, arguments of a nominal predicate may be realized outside the NP. (69) exemplifies an instance of such arguments.

(69) [ARG0 John] took a *walk*. [6]

In this example, *John* fills the ARG0 role of the *walk* predicate.

There are two main constructions which let the arguments of a markable NP be realized outside the NP: support verbs and transparent nouns (e.g., partitive/share constructions) (Meyers, 2007, p. 63).

"A support verb is a verb V that takes an argument-bearing NP A as one of its arguments, and at least one other argument B, such that A also takes B as an argument"(Meyers, 2007, p. 63). Due to their argument-sharing capability, these support verbs are comparable to raising and equi (control) verbs (Meyers, 2007, p. 63). Common support verbs are give, have, get, bring, carry, do, obtain, need, make, take and undergo (Meyers, 2007, p. 68).

(70) [ARG0 It] might [Support take] *action* [ARG1 to cure the default]. [7]

In (70), *take* is a support verb which takes *action to cure the default* as its argument. Therefore, *it* which is its argument can be shared with *action* predicate and is therefore annotated as its ARG0.

Transparent nouns, the other group which cause the arguments of a markable NP to be realized outside the NP, "take a special argument B such that the whole noun phrase represents either a multiple of B, a fraction of B, a part of B, or any other possible quantification over an amount of B. B is assigned the role ARG1 on analogy of nouns that are simultaneously nominalizations on the one hand and *partitive* (variety, cascade), *share* (slice, share) or *group* nouns (assembly, band) on the other" (Meyers, 2007, p. 42).

[6] The example is adapted from (Meyers, 2007, p. 8)
[7] The example is adapted from (Meyers, 2007, p. 65)

Almost all (transparent) partitives (*partitive* nouns or quantifiers in partitive constructions) and *share* nouns share one of their argument with the head noun, just like what support verbs do (Meyers, 2007, p. 72).

For example, in (71), *share of* is a *share* noun which lets its argument, *his* be shared with the nominal predicate *accomplishment*.

(71) [ARG0 his] [Support share of] *accomplishments* [8]

As the examples show, in the two introduced constructions, support verbs and transparent nouns, the arguments realized outside the NP are annotated as local semantic roles which is similar to FrameNet approach in annotating such constructions.

2.6 Comparing FrameNet, PropBank and NomBank

While FrameNet and PropBank are similar in terms of their goal which is providing semantic annotation, they have the following dissimilarities:

- FrameNet classifies verbs under frames and define semantic roles per frame while PropBank lacks such a classification.

- PropBank tries to be more syntax-oriented, in particular with its higher numbered arguments, while FrameNet attempts to be more semantics-oriented.

- PropBank annotates all sentences of Wall Street Journal section of Penn Treebank, regardless of their semantic or syntactic complexity. In contrast, FrameNet is a lexicography whose main annotation part is composed of exemplar sentences (though it has some full annotation texts as well). As a result, PropBank has been somewhat more successful for training statistical systems.

- PropBank provides full annotation per sentence, whereas FrameNet exemplar sentences are annotated only with arguments of one predicate. That is, in a sentence annotated by PropBank, all predicates are annotated with their arguments. In contrast, FrameNet provides exemplar sentences per predicates which evoke a frame. Therefore, in each case, only the predicate in question is annotated with its semantic roles and other predicates in the sentence, if any, are remained unannotated.

[8] The example adapted from (Meyers, 2007, p. 73)

- PropBank annotates only verbs, whereas FrameNet provides annotations for verbs, nouns and adjectives.

- All the differences between FrameNet and PropBank hold also for FrameNet and NomBank comparison, as NomBank follows exactly the same annotation approach as PropBank, except that NomBank focuses on annotation of nominal predicates, while PropBank annotates verbal predicates.

2.6.1 Summary

In this chapter, we introduced the linguistic foundation of SRL and the proposed theories about semantic roles. We also reviewed the two major role-annotated corpora which are developed based on the proposed linguistic theories: FrameNet (Baker et al., 1998) which has been built based on frame semantics theory and PropBank (Palmer et al., 2005) which is a more syntax-oriented corpus. We also introduced the annotation of implicit semantic roles, Null Instantiations, in these corpora and finally compared them regarding their similarities and differences.

Chapter 3

Related Work

This chapter reviews the related work and is composed of three parts. The first part discusses studies on automatic SRL, the second part reviews automatic identification of implicit semantic roles, which is a newer and broader version of semantic role labeling, and the last part presents some related work on domain adaptation, the technique that we use in our experiments in Chapter 5.

3.1 Semantic Role Labeling

Since NLP researchers have already created accurate syntactic analysis tools, such as POS taggers and NP chunkers, a natural next step is to develop systems which predict semantic entities and relations between them. Therefore, SRL can be understood as part of the traditional NLP analysis pipeline (morphology, syntax, and then semantics), i.e. a deeper structural analysis of the sentence with a focus on predicate-argument relations. Correct identification of semantic arguments of predicates (SRL) and the relation between them and their predicates is an important step for many NLP task, such as textual entailment (de Salvo Braz et al., 2006), text summarization (Melli et al., 2005), question answering (Narayanan and Harabagiu, 2004), and information extraction (Yakushiji et al., 2005).

The advances in performing automatic SRL have been made feasible by the availability of the large manually annotated corpora such as FrameNet and PropBank (cf. Chapter 2). In this section, we review prior studies focusing on automatic SRL using these corpora.

3.1.1 FrameNet-based SRL

Performing automatic semantic role labeling using FrameNet annotation framework is composed of four steps:

- **Target identification:** Identifying the target words in text
- **Frame identification:** Selecting the correct frame for the target word
- **Role recognition:** Identifying whether a constituent, or sub-string, in a sentence fills a semantic role for a given target word
- **Role classification:** Assigning the correct role label to the constituent or substring

For simplicity, many studies take the first two steps for granted and focus on recognition and classification of the semantic roles. These two steps can also be merged and performed as a single step. In the following subsections, some of the prior studies on SRL are reviewed.

3.1.1.1 Gildea and Jurafsky (2002)

The first supervised learning approach to SRL was proposed by Gildea and Jurafsky (2002). They developed a semantic role labeling system relying on FrameNet corpus. For simplicity, they assumed that the target predicate and their frames are already known and focused on the third and fourth steps.

Their system relied on lexical and syntactic information. To obtain the syntactic information, they automatically generated the parse trees using Collin's parser (Collins, 1997) because no manual syntactic annotation is available for FrameNet sentences. In the next step, to train a supervised system, they extracted various features for each constituent. The features they used are the fundamental features which have been used in many later studies:

- **Path:** The path in the parse tree from the candidate constituent to the target predicate. For example, in the parse tree of Figure 3.1, the path from the target word *ate* to the frame element *He* is described as VB ↑ VP ↑ S ↓ NP.
- **Phrase type:** The syntactic category of the constituent, e.g. NP, VP, S, etc.
- **Position:** Whether the candidate constituent precedes or follows the target word.
- **Voice:** The voice of the target predicate, active vs. passive

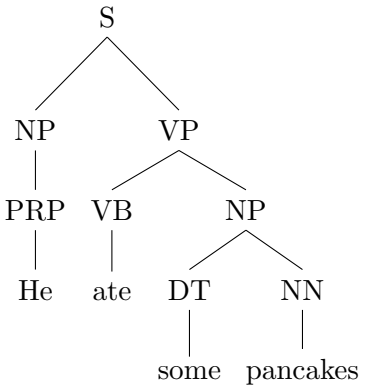

FIGURE 3.1: Parse tree example

- **Head word:** The syntactic head of the constituent
- **Predicate:** The target word lemma
- **Sub-categorization:** The phrase structure rule that expands the target word's parent node in the parse tree. For example, in Figure 3.1, the sub-categorization rule of the target word, *ate*, is VP $\rightarrow$ VB-NP.

After extracting the features, they developed two probabilistic models, the first one determined if a constituent fills a role and the second one labeled the constituents with semantic roles. In this framework, the identification of the constituents filling a role and their classification were performed in a pipeline architecture.

The results showed that when the boundaries of the frame elements are manually determined, the system assigns semantic roles with an accuracy of 82% and when it does identification of frame elements and semantic role classification simultaneously, it achieves precision and recall of 65% and 61%, respectively.

This work is considered as the foundation of current SRL systems and provides a general SRL system architecture.

3.1.1.2 Feature Engineering

Since the pioneering work by Gildea and Jurafsky (2002) interests in SRL rapidly spread and many studies addressed this problem, some of which were developed due to the introduced shared tasks, like SENSEVAL-3 (Litkowski, 2004) and SemEval2007 (Baker et al., 2007). These subsequent systems were built based on Gildea and Jurafsky's feature set, and added some more features. The features used in the early studies, such

as Surdeanu et al. (2003) and Pradhan et al. (2004) (cf. Table 3.1) are the baseline features which have been repeated in many later studies.

Later studies added some new features. For example, Croce et al. (2011) added more sophisticated features using convolution kernels to represent the predicate-argument structures and the similarity between the dependency structures more accurately. Some approaches exploited additional semantic features. Che et al. (2010), for instance, added features based on annotated word senses and replaced features based on word *lemma* and POS with word sense information. Also, some studies extended the features to beyond sentence-level, i.e. contextual features (Roth and Lapata, 2015).

3.1.1.3 Machine learning techniques

Apart from feature engineering, studies have tried to improve the task by using different approaches. For example, Fleischman et al. (2003) utilized a more efficient machine learning technique compared to Gildea and Jurafsky (2002), maximum entropy classifier, and could improve the results on frame element classification.

Johansson and Nugues (2007), the best system in SemEval2007, worked based on dependency parse trees instead of constituent-based ones. Their SVM-based classifier was trained using features including those extracted from the dependency parse tree such as the set of dependencies of the target word, and path through the dependency tree from the target word to the candidate constituent.

Das et al. (2010b) developed a supervised system, called SEMAFOR, which used probabilistic log-linear models. This system was later changed to a graph-based semi-supervised system and was improved in different ways (cf. Das and Smith (2011), Das et al. (2012), Das and Smith (2012) and Das et al. (2014)).

Most recently, Hermann et al. (2014) focused on frame identification using distributed representations of predicates and their syntactic context. Their system used word embeddings as the input and identified the semantic frames. This system, with an F-score of 69.91 on frame and argument identification performance jointly and an F-score of 88.93 on frame identification sub-task, has outperformed Das et al. (2014) which used log linear model and therefore set a new state of the art on SRL task. Another similar approach was followed by Foland and Martin (2015), Zhou and Xu (2015) and Foland and Martin (2015) who exploited neural networks to reduce the use of hand-crafted features. These studies also showed improvements. But due to the difference in the data sets, a direct comparison with other mentioned studies is not possible.

TABLE 3.1: Features used in Surdeanu et al. (2003) and Pradhan et al. (2004) SRL systems

Feature	**Definition**
Content word	selecting an informative word from the constituent which is different from the head word
POS of head word	POS of the head word
POS of content word	POS of the content word
Named Entity class of content word	the class of the Named Entity containing the content word
Named Entity flags	a set of binary features determining which types of NEs (e.g. location, organization, etc) exist in the constituent
Phrasal verb collocation features	including two features identifying the frequency of collocation of the verb and any/predominant preposition or particle
Verb clustering	verb cluster based on the co-probabilistic co-occurrence model of Hofmann and Puzicha (1998)
Partial path	the path from the candidate constituent to the lowest common ancestor of the predicate and the constituent
Verb sense	sense information extracted from PropBank
Head word of prepositional phrase	replacing the head word of a PP with the head word of the first noun phrase in it
First and last words/-POSs in constituent	a set of four features representing the first and last words and their POS in the constituent
Ordinal constituent position	concatenation of the constituent type and its ordinal position from the predicate
Constituent tree distance	a finer representation of the position feature
Constituent relative features	a set of nine features representing the phrase type, head word and head word POS of the parent, and left and right siblings of the candidate constituent
Temporal cue words	a binary feature determining if a temporal cue is present
Dynamic class context	a set of feature representing the hypotheses at most previous two nodes belonging to the same tree as the node being classified

3.1.2 PropBank-based SRL

In addition to models' differences, SRL systems are different regarding the training set and test set resources. While some systems are built based on FrameNet, as stated in previous section, PropBank is also widely used due to its close connection with Penn

Treebank. Penn Treebank provides manually annotated syntactic information and also full text annotation rather than exemplar sentences like FrameNet. These features make PropBank more appealing for SRL when applying machine learning methods.

3.1.2.1 Shared tasks

CoNLL-2004 (Carreras and Màrquez, 2004) and CoNLL-2005 (Carreras and Màrquez, 2005) were the first shared tasks on PropBank-based SRL. The former aimed at development of SRL systems based on partial parsing information, while the latter focused on increasing the amount of semantic and syntactic information. The results in CoNLL-2005 showed a great improvement compared to CoNLL-2004 which indicates the availability of full parsing information as an effective factor.

CoNLL-2008 (Surdeanu et al., 2008) was another shared task which explored the possibility of performing syntactic parsing and semantic role labeling jointly. Nevertheless, only a few participants performed the tasks jointly. The rest followed a pipeline architecture. The results on this task showed that system performances are remarkably lower when tested on a domain other than the training set (Surdeanu et al., 2008).

3.1.2.2 Feature engineering

Similar to FrameNet-bases systems, many PropBank-based systems developed for the shared tasks and also many other studies focused on supervised learning and used features introduced by Gildea and Jurafsky (2002), Surdeanu et al. (2003) and Pradhan et al. (2004). Many of the feature engineering attempts for FrameNet-based SRL, such as utilizing selectional preferences, were also carried out on PropBank (e.g. Zapirain et al. (2013)) and obtained significant improvements.

3.1.2.3 Learning approaches

Learning algorithms of PropBank-based SRL systems are, analogous to FrameNet systems, different. Maximum Entropy (ME) (Baldewein et al. (2004) and Lim et al. (2004)), memory-based learning (MBL) (van den Bosch et al. (2004) and Kouchnir (2004)), and support vector machines (SVM) (Hacioglu et al. (2004) and Park et al. (2004)) are among the exploited learning algorithms. Integer linear programming (ILP) has also

been used to enforce structural and linguistic constraints, such as prohibiting a predicate from have more than one core role of each type (A0-A5) (e.g. Punyakanok et al. (2004) and Punyakanok et al. (2008)).

Despite the existence of the large manually annotated corpora, many studies in recent years have applied unsupervised methods to induce semantic roles, e.g. Titov and Klementiev (2012a), Lang and Lapata (2014) and Woodsend and Lapata (2015). The main reason of this trend of work is that supervised models are domain-specific, and their performance drops dramatically when applied to a new domain (Pradhan et al., 2008). These systems try to develop open-domain SRL systems and have shown substantial improvements. For example, Woodsend and Lapata (2015) introduced an approach for learning distributed representations for predicates and their arguments. In their approach, the argument embeddings are learned from the contexts involving the predicate and its neighboring arguments, and the predicate embeddings are learned from the argument contexts. Their system (with an F-score between 0.81 and 0.89 in different settings) showed improved performance over previous unsupervised semantic role labeling approaches.

3.1.3 Comparing SRL systems

Systems developed using different training sets and frameworks are not directly comparable, however, comparing the FrameNet-based and PropBank-based systems shows that PropBank-based SRL systems perform better. The main reason is the difference between the granularity of the defined roles. FrameNet includes very fine-grained semantic roles while PropBank has coarse-grained numerical labels. Previous SRL research has demonstrated granularity as an important factor in performance of SRL systems (Bonial et al., 2014). For example, Yi et al. (2007) and Loper et al. (2007) demonstrated that since the labels in VerbNet are more generalizable across verbs than PropBank tags, they are easier for SRL systems to learn. In addition, PropBank frame files and rolesets are specifically determined based on the usage of the predicates in natural textual data while FrameNet provides conceptual frames and makes them semantically rich by defining all potential frame elements that can be realized in the specified event (Bonial et al., 2014). This feature makes PropBank-based systems more successful than FrameNet-based ones. An overview of some of the developed SRL systems can be found in Table 3.2. Though the results are not directly comparable due to the differenes between the training sets, test sets and the input information (e.g. syntactic information), the better performance of PropBank-based systems over FrameNet-bases systems can be seen.

TABLE 3.2: Some of traditional SRL systems based on PropBank and FrameNet corpora

System	Test Set	Framework	Task	F-score
Hacioglu et al. (2004)	SenseEval-3	FrameNet	FE identification and classification	0.72
Johansson and Nugues (2007)	SemEval2007	FrameNet	full task	0.49
Das et al. (2010b)	SemEval2007	FrameNet	full task	0.50
Das and Smith (2011)	part of FrameNet	FrameNet	frame and FE identification and classification	0.69
Foland and Martin (2015)	CoNLL2009	FrameNet	argument identification and classification	0.86
Koomen et al. (2005)	CoNLL2005 (WSJ)	PropBank	full task	0.79
Johansson and Nugues (2008)	CoNLL2008 (WSJ)	PropBank	full task	0.86
Titov and Klementiev (2012a)	CoNLL2008	PropBank	argument identification and classification	0.84
Lang and Lapata (2014)	CoNLL2008	PropBank	argument classification	0.79
Woodsend and Lapata (2015)	CoNLL2008	PropBank	full task	0.89

3.1.4 Implicit Semantic Role Labeling (ISRL)

While development of large manually annotated corpora and introduction of shared tasks on SRL motivated development of a large number of different SRL systems, providing a data set of locally unrealized semantic roles (cf. Chapter 2) and their identification remained unfocused until 2010. This section presents early and recent studies on Implicit Semantic Role Labeling.

3.1.4.1 Early studies on ISRL

Palmer et al. (1986)'s study can be mentioned as the first study considering computational treatment of implicit roles. The aim of this study was to make the implicit information of the context explicit. It focuses on the interaction between syntax, semantics and pragmatics to identify implicit information in text. This study distinguishes between two types of implicit entities, missing syntactic arguments and missing semantic

roles. In the first step, it determines if the missing argument is obligatory or optional and in the second step, if the implicit argument is obligatory, it tries to link it to its reference in the wider discourse. Palmer et al. (1986) manually developed a knowledge base for entities in a specific domain, maintenance reports, and provided hand-coded syntactic and semantic rules to identify implicit roles in discourse. For example, in

(72) Disk drive (was) down (at) 11/16-2305.
Spindle motor is bad.

motor is a dependent entity which is part of another entity, i.e. it presupposes the existence of a system which includes a spindle motor. Considering the discourse context, the entity it presupposes is not a computer system, but the *disk drive* which is mentioned in the previous sentence.

With a similar approach, Whittemore et al. (1991) tried to find the implicit semantic roles by applying semantic constraints on arguments. They treat semantic arguments as a type of anaphor and when a role is not filled by a local syntactic argument, they call it a "not-yet-instantiated slot". This study shows that the information from different levels of processing, syntax, semantics and discourse, can be brought together to build an event representation incrementally.

Both Palmer et al. (1986) and Whittemore et al. (1991) studies rely on some hand-coded processing rules on small texts from specific domains which makes it difficult to generalize the approach to larger data sets. After development of large semantic role annotated corpora more studies were conducted which led to computational modeling of ISRL. For example, Fillmore and Baker (2001) analyzed a small text and noted that not all the required frame elements of a predicate are necessarily available in the same sentence. However, they did not propose any computational model to detect or recognize such arguments.

In another study, Burchardt et al. (2005) suggested that extracting frame-to-frame relations in a text can result in a partially connected predicate argument structure which can form a coherent discourse interpretation. They showed that the existing lexico-semantic and contextual relations in a text can induce further semantic relations between frames and roles, e.g. a filled frame element of a predicate can be linked to a missing frame element of another frame, provided that there is enough supportive evidence. They, however, never implemented and evaluated their proposed method.

3.1.4.2 SemEval2010

SemEval-2010-task 10, called *Linking Events and their Participants in Discourse*, was the first shared task which addressed identification of implicit roles (also called Null Instantiations; cf. Chapter 2) and provided a data set annotated with implicit roles. This task attracted the attention of many studies to ISRL (Ruppenhofer et al., 2010).

The provided data set of the task is selected from novel genre (in contrast to data sets of normal SRL which have mainly been selected from newswire texts), parts of Arthur Conan Doyle's fiction works and the annotations are originally made using FrameNet-style but can be mapped semi-automatically to PropBank annotations.

Two systems participated in this task. Tonelli and Delmonte (2010) developed a knowledge bases system, called VENSES++, which applied a rule-based anaphora resolution procedure and exploited different resolution strategies for verbal and nominal predicates.

For verbal predicates, they considered the Predicate Argument Structure (PAS) of the current sentence and looked for comparable PASs in previous sentences. Comparable means that the predicates are semantically related according to WordNet (Miller, 1995). If such a PAS is found, it checks whether the two predicates share at least one argument. If yes, it finds the best head word available in the PAS by semantic matching between the candidate filler and the frame element label.

For nominal predicates, the system made use of a common sense reasoning module that was builds based on ConceptNet (Liu and Singh, 2004). This system achieved an F-score of 0.01 which is far from satisfatory. Their analysis showed that their system is much poorer in identification of INIs, as the system does not provide any specific strategy for INI detection. In addition, the gold standard data includes many un-resolved DNIs, while their system searches for an antecedent in case of detecting a DNI. Also, they argue that the very poor result of the system illustrates ISRL as a challenging task.

Chen et al. (2010) was the other participant of the task. Their system, SEMAFOR, extended an existing semantic role labeler by extending the search window to other sentences. The feature set of the traditional SRL system which focused on the current system was also modified to make the system practical in searching the wider window. For this purpose, the syntactic path feature was replaced by features derived from FrameNet. For each candidate span, two features were computed. The first one indicated whether the head word of the candidate filler is used as filler of the same role in at least one of the exemplars of FrameNet. The second one determined the maximum distributional similarity between the candidate filler head word and any word heading a filler of that

role in FrameNet exemplars (Chen et al., 2010). Although this system obtained the best performance in the task, its performance was still very low, with an F-score of 0.02. They mentioned data sparseness the biggest obstacle in ISRL and claimed that without increasing the number of NIs, it is unlikely to to develop an effective supervised learner. Because there are more than 20,000 training examples of overt roles, while this number for null instantiations is only 600 cases, among which 2/3 are without any referent in the context Chen et al. (2010).

3.1.5 Approaches to ISRL

Since SemEval-2010 shared task, many ISRL systems were developed which could improve the performance on the task significantly. These systems can be categorized under two groups: 1) studies focusing on linguistic knowledge and 2) studies trying to solve the problem of data sparseness.

3.1.5.1 Linguistic-motivated approaches

Many studies on ISRL have used exclusively the available training set by the SemEval2010 task organizers and have tried to improve the performance using linguistic knowledge and feature engineering. For example, Tonelli and Delmonte (2011) tuned their previous system (Tonelli and Delmonte, 2010) and developed a system relying on linguistic knowledge. They showed that an algorithm that reflects the linguistic motivations behind identification of null instantiation in the FrameNet paradigm can improve the results significantly. For example, they used the *exclude* and *require* relations between frames. The former means that some frame elements can not occur together and the latter determines that if a specific frame element is present, then another specific role must also be overtly realized. In the antecedent binding sub-task, their system relied on the frequency of the observed heads of frame elements in the training set and assigned a relevance score to each candidate antecedent. This system could improve the results on SemEval2010 task to an F-score of 0.08.

Ruppenhofer et al. (2011) was another study with linguistically motivated strategies. Their system firstly determines which roles are implicit using interdependencies between frame elements in FrameNet (similar to Tonelli and Delmonte (2011)). Then, in the second step, they evaluate whether the omission is licensed by a grammatical construction or not to determine if the missing role is an INI or DNI. If no relevant construction could support the omission, they consider the omission lexically licensed

and follow to the next step, antecendent finding, using the semantic type information provided in FrameNet. In this step, they extract the semantic type of the missing frame element in the relevant frame in FrameNet and then try to select an active coreference chain as the correct filler. *Active coreference chain* means a coreference chain with at least one mention in the considered context window, whose elements have at least one common semantic type with the missing frame element. They showed that a more sophisticated linguistically motivated model can lead to noticeable better performance and an error reduction of 14% compared to Chen et al. (2010).

Laparra and Rigau (2012) exploited the information of explicit role annotations to identify implicit roles. At the first step, their system collects the most common frame element patterns of the corresponding frame from the training data. Then, the system defines the missing core frame elements of the most common pattern as DNIs of the predicate in question. To better understand this step, consider (73).

(73) Apparently the *tenants* had brought little or nothing with them.

In (73), the word *tenants* evokes the frame *RESIDENCE* which requires three core frame elements: RESIDENT, CO_RESIDENT and LOCATION, among which RESIDENT is filled by *the tenant* and the other frame elements are implicit. To identify which of the missing frame elements must be assumed as DNI, they search for the most common patterns of frame elements for *residence* predicate which include RESIDENT frame element. The obtained result is presented in Table 3.3[1].

TABLE 3.3: Most common frame element patterns of the *residence* predicate which contain RESIDENT frame element

Frame element pattern	**Frequency**
RESIDENT LOCATION	384
RESIDENT CO_RESIDENT LOCATION	34
RESIDENT CO_RESIDENT	14
RESIDENT	13
RESIDENT LOCATION MANNER	1
RESIDENT LOCATION TIME	1

As can be seen, the most common pattern includes two frame elements: RESIDENT and LOCATION. Considering that the RESIDENT frame element is already filled by the phrase *the tenants*, LOCATION is the only missing frame element whose referent must be searched in the context.

[1]The table is taken from Laparra and Rigau (2012)

In the next step, to select the correct filler for the missing frame element, the model uses probability distributions of semantic types of frame elements to select the most probable missing implicit arguments.

The results demonstrated that this method, with an F-score of 0.19, outperforms all prior studies evaluated on the SemEval-2010 dataset.

Laparra and Rigau (2013) presented another ISRL study with a linguistic approach. They explored alternative linguistic and semantic strategies to study implicit argument resolution as a special case of coreference resolution. They evaluated features of traditional coreference and anaphora models to solve the problem of ISRL and showed that these theories and models can be successfully applied and can obtain an F-Score of 0.16.

3.1.5.2 Data-based approaches

Since the small number of implicit roles in the SemEval2010 training (only 245 recoverable implicit roles in the FrameNet annotation framework) is a main bottleneck, some studies have tried to use extra data sets. For example, Silberer and Frank (2012) created an artificial data set. For this purpose, they selected a data set annotated with both semantic roles and coreference chains and extracted coreference chains with anaphoric pronouns which filled a semantic role of a predicate. They removed the pronoun's role label and annotated its closest antecedent in its chain with its semantic role. For example, in

(74) Riady spoke in his 21-story office building on the outskirts of Jakarta. [...] The timing of [SPEAKER his] *statement* is important.

(75) Riady spoke in [SPEAKER his] 21-story office building on the outskirts of Jakarta. [...] The timing of *statement* is important.[2]

the pronoun *his* which fills the role of SPEAKER is removed from (74) and the role SPEAKER is assigned to its previous mention in the coreference chain in (75).

They approached the problem as an anaphora resolution task and exploited three sets of features: CR-oriented features, SRL-oriented features and features relating to both phenomena. They found that the first class of features yielded better performance compared to the SRL features, which means that ISRL is closer to the CR task than SRL. Their system using the artificially created data set reached an F-score of 0.10.

[2]The example is taken from Silberer and Frank (2012).

Moor et al. (2013) addressed providing a higher number of predicate-specific annotations for implicit roles using OntoNotes (Weischedel et al., 2011) as the underlying corpus. Their annotation included a total of 630 implicit roles for five verbal predicates. They evaluated efficiency of using their corpus in ISRL by training Silberer and Frank (2012)'s system on their created data set and showed that using a moderate amount of implicit roles can improve the results by around 2 points in F-score.

In another study, Roth and Frank (2013) followed a different approach. They exploited aligned predicate argument structures in comparable texts. In this approach, when a pair of predicates are comparable in two different texts and have different number of arguments, the missing arguments of one of the predicates can be induced using the arguments of the other one in the pair. They created a data set with 701 implicit roles and used it as an additional data set to train an ISRL system. Although their created data set was one third of the size of Silberer and Frank (2012)'s data set, it outperformed Silberer and Frank (2012)'s system and obtained an F-score of 0.12.

Gorinski et al. (2013) proposed a weakly supervised approach to resolve DNI resolution. They built four different resolvers for the task: the semantic type-based resolver, the vector resolver, the participant-based resolver and the string-based resolver. The first one looks at FrameNet to find the semantic types that FrameNet specifies for the missing frame element. The vector resolver focuses on determining the similarity between the mentions of a coreference chain, which is a candidate filler of a missing role, and the known fillers of that role in FrameNet. The participant-based resolver focuses on co-occurring roles and string-based resolver is based on the assumption that a constituent which has filled a role before is expected to fill it again. This system obtained F-scores of 0.15 and 0.12 for chapter 13 and 14 of the test set, respectively which is comparable to the developed supervised systems by Silberer and Frank (2012) and Roth and Frank (2013).

Gerber and Chai (2010) did another study which addressed ISRL with a different framework. While all previous studies had worked with FrameNet annotation framework and had utilized the training set of SemEval2010 task, Gerber and Chai (2010) adopted PropBank annotation framework and restricted the task to annotation of implicit roles for nominal predicates. For this purpose, they annotated a large number of instances of implicit roles for 10 nominal predicates of NomBank data set. Then, they extracted different syntactic, lexical, and discourse features for candidate fillers. To evaluate their system, they trained a logistic regression model over 816 annotated predicate instances which included 650 implicit roles and tested it on a test set of 437 predicate instances associated with 246 locally unrealized roles. The F-score of their system was 0.42.

Schenk and Chiarcos (2016) proposed an unsupervised approach using pre-trained word embeddings which required an SRL system and a large number of unannotated instances. They used the PropBank annotation framework and the NomBank test section for testing to compare their results with Gerber and Chai (2010) and Laparra and Rigau (2013). Though their approach had the advantage of being simple and allowing to induce implicit roles for arbitrary predicates, it performed around 10 points lower than Laparra and Rigau (2013) and around 6 points lower than Gerber and Chai (2010).

In this dissertation, we address the data sparseness problem of the task and try to improve the performance of a baseline system by increasing the number of training instances. However, due to the expensive and time consuming task of implicit semantic role annotation, we try to design some experiments to do the annotation in a cheap and easy way or use the existing corpora and avoid much effort on annotating new instances.

In our first approach to annotate instances without much effort, we follow a similar way as Gerber and Chai (2010) and focus on a small number of predicates (for more details cf. Chapter 5) and in our second approach to use the existing corpora, we utilize the Gerber and Chai (2010) data set. Due to the differences between the Gerber and Chai (2010) data set and the SemEval2010 benchmark training set, however, we benefit from a domain adaptation approach to reduce the effect of domain dissimilarities. In the following section, we briefly introduce some studies on domain adaptation techniques which can be helpful in understanding our experiments in Chapter 5.

3.2 Domain Adaptation

Domain difference is a main problem of many NLP tasks because one of the basic assumption in development of supervised models is the uniformity of training and test set. In practice, however, this assumption does not hold in many cases. Therefore, many studies have focused on tackling this problem by developing methods to easily port models trained on one domain to applications in other domains with minimum error rate.

Domain adaptation is a well known approach which lets us apply a model developed for a domain (source domain) to another domain (target domain) when there is not enough training data on the target domain and it is difficult to collect data. This method helps to overcome the problem of differences between the writing styles and vocabularies in the two domains which may cause considerable drop in the performance of supervised models.

3.2.1 Domain definition

The notion "domain" in computational linguistics has been widely used in previous work to refer to differences in topics, writing styles, genre, register, the categories such as *general fiction* vs. *romance and love story*, etc (Banerjee, 2013, p. 5). Thus, there is no specific definition of domain. In our experiments, we define domain based on the text genre, e.g., newswire text and novel text are considered as different domains.

3.2.2 Domain adaptation approaches

Similar to machine learning techniques, domain adaptation techniques are also classified under three classes: supervised (e.g. Daume III and Marcu (2006)), unsupervised (e.g. McClosky et al. (2006)), and semi-supervised (e.g. Daumé III et al. (2010)) domain adaptation.

The difference between these approaches is the type of the available data. In supervised domain adaptation, a large annotated data set of the source domain is available, while the amount of annotated data in the target domain is limited. Unsupervised domain adaptation is used when there is a large amount of unannotated data from the target domain available and semi-supervised approach uses both unlabeled and labeled data from the target domain, usually the size of the unlabeled data in this approach is much larger than the annotated data. In the following sections, we explain some of the most commonly used methods to implement these approaches.

3.2.2.1 Semi-supervised domain adaptation

Semi-supervised domain adaptation refers to the utilization of both labeled and unlabeled data (e.g. Daumé III et al. (2010)). However, nowadays there has been a shift in the terminology and these methods are called unsupervised domain adaptation, similar to co-training and self-training which use both labeled and unlabeled data.

In our experiments, we use supervised domain adaptation, i.e. using only data points which have already been annotated (the details of the experiments can be found in Chapter 5).

3.2.2.2 Supervised domain adaptation

One of the well-known techniques to adapt a model to be used in a different domain was developed by Daume III (2007). In this method, called *feature augmentation*, they make three versions of each feature in the original problem: a general version, a source-specific and a target-specific version. The augmented source data contains the general and source-specific versions and the augmented target data contains the general and the target-specific versions. If this approach is applied to D domains, the feature space is segmented into D+1 subspaces: general domain, domain 1, domain 2, ..., domain D, which can be expressed as

$$\text{h(f,e)= } <h_g, h_1, ..., h_D> \tag{3.1}$$

where h_g and h_i denote the feature vectors of the general domain and the domain-specific spaces, respectively. All features are deployed to the general space, but each domain space includes only features that match that domain (Imamura and Sumita, 2016). That is,

$$h(g) = \Phi(f, e)$$
$$h(i) = \begin{cases} \Phi(f, e) & if\ domain(f) = i \\ \emptyset & Otherwise \end{cases}$$

To understand why this method helps, let us consider a POS tagging task. Assume that the source domain is from the newswire genre and the target domain includes review of computer hardware. In this setting, a word like *the* is a *determiner* in both cases, while the word *monitor* can have different POS tags: in the newswire text it is probably a verb, and in the hardware review data it is more likely to be a noun. In this case, assuming that in the original model there are two features determining whether the word is *the* and whether the word is *monitor*, in the domain-adapted model there are six features (h_1 and h_2 for the general domain, h_3 and h_4 for the source domain and h_5 and h_6 for the target domain). The algorithm lets the feature augmented model to set the *determiner* weight vector to something like $\langle 1,0,0,0,0,0 \rangle$ which places more weight on

the common version of *the* which shows that *the* is commonly a *determiner* regardless of its domain. The weight vector of *noun* in this model would be like ⟨0,0,0,0,0,1⟩ which determines that *monitor* is most likely to be a noun in the target domain. In a similar way, the model can create a weight vector like ⟨0,0,0,1,0,0⟩ for *verb* to determine that the word *monitor* is most likely a verb in the source domain data.

Daume III (2007) applied this technique to some NLP tasks, such as POS tagging, named entity recognition, and shallow parsing and observed improvements over the results. This method was also applied to SRL in biomedical text by Dahlmeier and Ng (2010) and indicated improvements in argument identification.

3.2.2.3 Unsupervised domain adaptation

An effective method of bootstrapping is self-training using the available unlabeled data. In this method, the baseline model trained with the existing labeled data set firstly labels the unlabeled data. Then, the newly labeled data is treated as correctly annotated and and combined with the actual labeled data to train a new model. This process can be iterated over various sets of unannotated data (McClosky et al., 2006). McClosky et al. (2006) applied this method to parsing and could improve the results.

Another well-known bootstrapping method is co-training (Blum and Mitchell, 1998) in which two or more models are trained on the same data set, each one on one *view* of the data. For example, in classification of the web pages, the pages can be classified based on the text of their hyperlinks or the content of the web page. To apply co-training, the features are split to two sets and two classifiers are trained on the available labeled data set using each feature set. Then, the unlabeled data is labeled using each of the learners and the result is added to the pool of labeled examples from which both classifiers are trained. Blum and Mitchell (1998) applied this technique to classification of web pages and showed that iterative retraining on the pseudo-labeled data set provides improvement in the performance of both learners. This method has also been effectively used in other NLP tasks like parsing (Steedman et al., 2003). This method however, has two main assumptions: features must be dividable into distinct groups (views) and each group must contain sufficient information to perform the labeling. Nigam and Ghani (2000) showed that while arbitrary features splits can be used to apply co-training technique, it is most effective when the features can be separated as disjoint sets naturally.

3.2.3 Summary

In this chapter, we introduced the linguistic background of semantic roles and reviewed prior studies on semantic role labeling and implicit semantic role labeling and introduced domain adaptation approaches.

By reviewing prior studies and their results on SRL and ISRL we demonstrates that ISRL can be seen as a more difficult task whose computational modeling is much harder. One reason is the fact that ISRL systems can not rely on syntactic information which is a main part of any SRL system.

In addition, the overview of ISRL systems shows that the approaches toward ISRL can be categorized under two classes. One group of studies attempt to use more linguistic information and another group attempts to increase the amount of annotated data to solve the problem of data sparseness. Almost all prior studies on ISRL have claimed data sparseness as the main problem of performing ISRL effectively (e.g Chen et al. (2010); Silberer and Frank (2012); Laparra and Rigau (2013); and Roth and Frank (2013)). The studies by Moor et al. (2013) and Gerber and Chai (2010) which manually annotated large corpora with implicit roles provide the strong evidence that the amount of annotated data plays an important role. The performance of their system, F-scores of 0.36 and 0.42, respectively, is around/more than double the performance of the best ISRL system which has exclusively used the training set of the task.

In this thesis, we approach the ISRL task from a similar perspective to Gerber and Chai (2010) and Moor et al. (2013), trying to solve the problem of data sparseness. Thus, we use an expressive but conservative feature set and focus on increasing the amount of annotated data, but at the same time, reducing the manual effort or at least the need for expert annotators. In the next chapters we propose two methods for this purpose. In the first one, we evaluate crowdsourcing as a cheap way to annotate more data instances and assess if the data annotated by non-experts can be helpful in ISRL. In the second approach, we consider combination of the available ISRL data sets and avoid annotating new data sets. We then discuss that even data sets from different genres can be effectively combined to enhance the system.

Part II

Domain-focused Annotation of Implicit Semantic Roles

Chapter 4

Annotation of Implicit Semantic Roles using Crowdsourcing

Aspects of the work in this chapter have been published in Feizabadi and Padó (2012) and Feizabadi and Padó (2014).

4.1 Introduction

As mentioned in previous chapters, the availability of large annotated corpora, such as FrameNet (Baker et al., 1998), PropBank (Palmer et al., 2005) and TreeBank (Marcus et al., 1994) has inspired development of many systems for semantics applications.

However, lack of enough annotated data has been mentioned as an important bottleneck of ISRL which makes it difficult to develop high-performance ISRL systems. In this chapter, we assess crowdsourcing as a possible solution to this problem and show that provided that the task is defined suitably, reliable annotations can be obtained, even without providing definitions of semantic roles.

4.2 Crowdsourcing for text annotation

Since manual annotation of large texts is time and cost consuming, crowdsourcing was introduced as a solution to this problem to annotate the data in a time and cost effective way with the help of a large number of non-experts. This approach helps to avoid

training expert annotators, often in person, and to ensure that they do the task with a high quality.

In crowdsourcing, an annotation task can be assigned to hundreds or even thousands of computer-literate workers and get the results back quickly. To define a crowdsourcing task however one faces two problems. The first one is that not all tasks can be provided to non-experts. The tasks which require professional information are normally not suitable for crowdsourcing. Semantic roles annotation is such a task which requires knowing the definition of the semantic roles. Therefore, it requires precise design for the task to be able to have the semantic roles annotated by non-expert. The second problem is the quality of the annotations. The quality of the obtained annotations varies. Some annotators provide low quality labels due to misunderstanding the task while others may do the task reliably. The standard solution to the problem of noisy annotations is assigning the same task to a number of different annotators, hoping that a consensus can be reached from the majority of the obtained annotations (Welinder and Perona, 2010).

To implement this mechanism, different platforms have been developed. The most important and common ones are Amazon Mechanical Turk(AMT) (https://www.mturk.com/) and CrowdFlower (https://www.crowdflower.com/). These platforms follow the same goal, but provide different tools to the task definers. Also, another difference is that AMT is officially available only to the workers who live in the USA while CrowdFlower is open to the workers from all over the world. Nevertheless, more workers are available in AMT than CrowdFlower. Therefore, most studies use the former to do their annotation tasks. For example, Callison-Burch (2009), Mellebeek et al. (2010) and Heilman and Smith (2010b) have effectively used crowdsourcing via AMT for machine translation evaluation, sentiment analysis and student answer rating tasks.

A task which is closely related to implicit semantic role annotation and has already been carried out by crowdsourcing is frame semantic role annotation. This task is composed of two sub-tasks of FrameNet-based SRL: identification of the frame evoked by the given predicate and labeling its semantic roles in the sentence (cf. Chapter 3).

For example, the target predicate *gain.v* can belong to two different frames: *GETTING* or *CHANGE_POSITION_ON_A_SCALE* frame. So, when the workers are presented with the sentence " *You will have to* ***gain*** *their support, if change is to be brought about.*", they should select the appropriate frame from a multiple choice list including *GETTING*, *CHANGE_POSITION_ON_A_SCALE* and *None of the above*.

For this purpose, Hong and Baker (2011) defined a task in which they removed the frame names and used some hand-crafted synonyms instead to make it more comprehensible to the workers. Because, for instance, non-experts can not easily understand what lexical units must be categorized under *BODY_MARK* and *INSTANCE* frames. The experiment was performed on a number of predicates and showed that the results can reach a level of accuracy which makes the data appropriate for other NLP tasks.

In another study, Fossati et al. (2013) attempted to perform full FrameNet annotation using crowdsourcing, by adopting FrameNet definition of a frame as a description of a type of event and its participants. In other words, frames are distinguishable from each other based on their involved participants and it sounds more cognitively plausible to perform frame identification by firstly annotating its frame elements, and not the other way round. Thus, they tried to elicit full frame annotation in a single step and in a bottom-up manner, using simplified definitions for frame elements. To evaluate the effect of FE definition simplification, they performed the experiments with three different types of FE definitions: the original definitions presented by FrameNet, a manually simplified version, and an automatically simplified one using a tool provided by Heilman and Smith (2010a) which does some syntactic simplification for complex sentences. In the manual simplification part, a linguistic expert simplified the frame element definitions based on the following rules: replacing the frame elements by their semantic type when its associated semantic type is a common concept (e.g. Location); simplifying the syntax of the sentence; trying to make homogeneous definitions, e.g. letting all definitions to start with "This element describes..."; Replacing technical concept (e.g ARTIFACT) with common words (e.g. object). Using these rules and simplifying the task and at the same time doing the task in a single step, let them obtain a high accuracy of 0.80 which is promising for doing further annotation using crowdsourcing.

In our experiment to perform implicit semantic role annotation using crowdsourcing, we focus on a specific domain, motion domain, to simplify the task and avoid frame annotation. Therefore, after identifying the motion events, we select a text with a large number of such events, "Around the World in Eighty Days" by Jules Verne, to evaluate annotation of location roles. In the following sections, we discuss more details about domain selection, identification of motion events, and finally the crowdsourcing experiment and the results.

4.3 Domain selection

In our study, we focus on annotation of a small number of predicates and their semantic roles because annotating all semantic roles of all predicates in a text is complex (in terms of how to define the semantic role in a simple and precise manner) and time consuming (to prepare the definitions of semantic roles).

We restrict our study to a number of *motion* predicates. The intrinsic reason for this decision is that they all have very similar semantic role sets according to FrameNet which simplifies the task and let us deal with only one frame and the extrinsic reason is that determining location information plays an important role in many NLP tasks, such as information Extraction (Leidner et al., 2003), Question Answering (Greenwood, 2004) and the analysis of narratives (Howald, 2015).

Figure 4.1 shows the general steps of our crowdsourcing experiment: given a running text, we (a) define a set of motion verbs and disambiguate the verb instances in the text and assume an instance to be a motion verb if it is included in the obtained set of motion verbs; then we (b) annotate the implicit and explicit roles using crowdsourcing, and finally we (c) use the annotated text to train a model for automatic identification of implicit semantic roles.

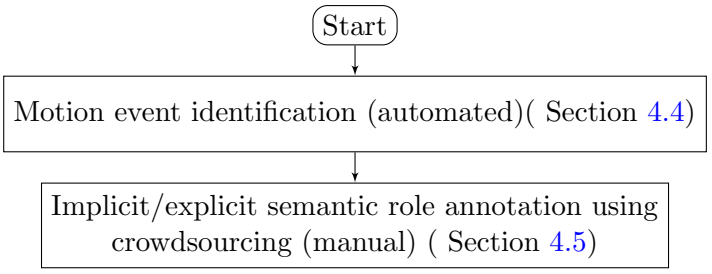

FIGURE 4.1: Using crowdsourcing for ISRL

To perform the first step, we firstly need to consider word sense disambiguation. Because many verbs have motion as well as non-motion senses. An example of such verbs is *cross* which represents a motion event in (76), while in (77) and(78) its non-motion senses are used.

(76) The ship *crossed* the ocean.

(77) Peter *crossed* himself.

(78) John was *crossed* by the conman.

TABLE 4.1: Motion frame in FrameNet

Frame: MOTION		
Frame Elements	AREA	Emily *moved* restlessly **around the room**.
	DIRECTION	She shied, she reared, *went* **backwards**.
	DISTANCE	The twig *floated* atop the water **for about 100 yards**.
	GOAL	The car *moved* **into the slow lane**.
	PATH	Jo *moved* **past dad** into the hall.
	SOURCE	The policeman *moved* **away from the door**.
FEEs	abandon.v, desert.v, depart.v, departure.n, emerge.v, emigrate.v, emigration.n, escape.v, escape.n, leave.v, quit.v, retreat.v, retreat.n, split.v, withdraw.v, withdrawal.n	

In addition, the motion/position domain is used for metaphorical mappings (Lakoff and Johnson, 2008). Examples of such metaphorical usage can be found in (79) and (80).

(79) Colin *moves* towards the Labour position.

(80) Colin *is* on the Tory side.

In these examples, *moves* and *is* which are typically seen as motion and position events, do not express any physical movement or position. Therefore, the open question here is if literal senses should be distinguished from metaphorical ones. Our observation shows that the semantic role analysis of metaphorical usages should be very close to the literal usage. That is, *towards the Labour position* in (79) and *on the Tory side* in (80) can be considered as the metaphorical DESTINATION and PLACE semantic roles of the predicates. Since we are primarily interested in predicate-argument structure, we decided to keep these readings of the motion predicates in and distinguish literal/metaphorical motion on one side from non-motion senses on the other side. Some other studies which have considered metaphorical senses are (Ellsworth et al., 2004) and (Burchardt et al., 2006) which have shown that FrameNet frames are generally reusable for metaphorical senses.

To identify motion events, we firstly referred to the *MOTION* frame in FrameNet which covers a large number of motion events. This frame requires six core location roles which are shown with examples in Table 4.1

Among the above roles, DISTANCE is rarely realized among the roles of motion predicates. Therefore, we discard it from our analysis and define motion events as the events that require the other five roles.

In addition to verbs of motion (including self motion and caused motion), verbs of orientation, and verbs of position are also events in which locations play a central role in their semantics, as can be argued on the basis of decomposition (Jackendoff, 1990) or of corpus evidence (Baker et al., 1998). Based on these observations, in this dissertation, we focus on these three classes of verbs which require a location: motion, orientation, and position verbs (henceforth called motion verbs for the sake of simplicity). (81) and (82) represent examples of position and orientation verbs which both belong to *CHANGE_POSTURE* frame.

(81) She *leaned* [$_{DIRECTION}$ back] [$_{GOAL}$ against the seat].

(82) He *sat* [$_{GOAL}$ beside her], close enough for her to feel the heat from his body.

To be able to evaluate the efficiency of our proposed methods for automatic identification of motion events, we annotated the first three chapters of "Around the World in Eighty Days" novel. We decided on a fiction text since we can expect to find more natural narrative structures in it compared to a newswire text. In addition, due to the topic of the selected novel which is a trip round the world, it includes a large number of motion events.

The annotation was performed by two annotators with excellent proficiency in English and graduate-level linguistic background on raw text, with no linguistic analysis other than sentence segmentation. To perform the annotation, the MMAX2 Annotation Tool (Müller and Strube, 2006) was used, which is a graphical user interface for creating, browsing, visualizing and querying linguistic annotations on multiple levels.

Among the first three chapters of the novel which included 4591 words, 114 instances of motion events belonging to 71 different lemmas, were found and annotated. The inter-annotator agreement on identification of motion events can be found in Table 4.2. As the results show, the agreement is very high and reliable.

TABLE 4.2: Motion events annotation reliability

Agreement on motion verb (exact match)	87%
Inter-annotator agreement on motion verb (Cohen's Kappa)	0.83

An analysis of the results showed that the disagreement occurrences can be categorized under the following groups:

1) Phrasal verbs: some phrasal predicates were not annotated by both annotators. For example, *make the tour* in *I will make the tour of the world in eighty days or less* and

get away in *And also why the thief can get away more easily* were annotated only by one annotator.

2) Predicates which imply a movement, but are not directly categorized under *MOTION* frame or its similar frames (e.g. *CAUSE_MOTION*) in FrameNet. *Give* and *hand* which belong to the *GIVING* frame are examples of such verbs. Another example of such frames is the *PLACING* frame which includes verbs like *place*.

3) Predicates implying a position: The *RESIDENCE* frame includes verbs like *live* and *inhabit* which are not annotated by both annotators.

4) Metaphorical usage: Though both annotators have read the same guideline and have annotated metaphorical senses, some instances of metaphorical usages are disagreed among them. For example, *overspread inthe smile overspread his features* and *fell* in *the discussion fell during the rubber* were annotated by just one annotator.

4.4 Mining lexical resources for motion events

For automatic identification of motion events, we followed a knowledge-based approach, using two standard CL resources, WordNet and FrameNet, which cover a large number of words, and compare how well they capture motion events.

At the pre-processing step, we performed WSD for the annotated text to distinguish the sense of the motion predicate as discussed before. For this purpose, we used UKB, a state-of-the-art unsupervised graph-based word sense disambiguation tool (Agirre and Soroa, 2009) with a broad coverage and outperformance compared to other unsupervised WSD tools and performed WSD using WordNet 3.0. Since UKB requires part-of-speech information, we also performed part-of-speech tagging using Stanford CoreNLP (Manning et al., 2014). Our analysis showed that Stanford CoreNLP recognized 97% of the gold standard motion verbs as verbs, which leads to an upper bound of 97% recall for any model building based on the Stanford CoreNLP output. The missing cases were phrasal verbs which the tagger could not handle correctly.

We also annotated the motion verbs with their WordNet synsets manually to evaluate UKB efficiency in performing WSD on the presented text. When compared against the gold standard, we found an exact match accuracy of 50%, which is comparable to the verb results reported by Agirre and Soroa (2009) on the SensEval all-words data sets. However, if we evaluate just the coarse-grained decision motion verbs vs. non-motion verbs, the accuracy improves to 75%.

One of the reasons of disagreement was the phrasal verbs, like *go on*, *take a step* and *make the tour*. In such cases, UKB annotates the words separately, which causes the wrong annotation for the motion events.

In addition, some words like *go* have many motion senses which are just slightly different from each other and it sometimes cause annotation agreement when considering coarse-grained classification of motion verbs vs. non-motion verbs, despite a disagreement when considering WSD. For example, *go* in

(83) But in the present instance things had not *gone* so smoothly.

the UKB annotated the predicate with the first sense of *go* in WordNet, i.e. *travel, go, move, locomote (change location; move, travel, or proceed, also metaphorically)*, while the gold standard annotation assigns the sense *go, proceed, move (follow a procedure or take a course)* to it.

Another example of such instances is the predicate *pass*. For example, in

(84) He took it up, scrutinised it, *passed* it to his neighbour, he to the next man, and so on.

the UKB assigns the first sense of the predicate to it, i.e. *pass (go across or through)*, while the gold standard assigns the fifth sense in WordNet, i.e. *pass, hand, reach, pass on, turn over, give (place into the hands or custody of*, to the *pass* predicate. As such cases show, though the senses picked by UKB are not correct, they still refer to the motion senses of the predicate.

These examples give rise to the question of whether WSD is good enough to serve for the selection of motion verbs in our application. To explore this question, in our analysis we compare three strategies. The first strategy is no disambiguation at all (NoWSD). It classifies a verb instance as a motion verb if any sense of the lemma is in the WordNet-derived list of motion verbs. The second strategy, WSD, classifies an instance as a motion verb if its UKB-assigned synset is in the list of motion verbs. The third strategy, PredomSense, leverages the observation that WSD has a hard time beating the predominant sense heuristic (McCarthy et al., 2004) which assigns the predominant, or first, sense to all instances. Here, this strategy means that we treat all verbs as motion verbs whose first sense, according to WordNet, is in the list of motion verbs.

4.4.1 Mining motion events using WordNet

In the first approach, we used the troponymy relation between the verbal predicates in WordNet which structures the verbs as a hierarchy. For example, *jog* and *climb* are troponyms of *run*. In this approach, we use this hierarchical organization to define the motion domain (motion, orientation, and position verbs) as a set of sub-trees in WordNet. Therefore, the main challenge is to identify the sub-trees in a way that they cover as much as possible of the motion domain while avoiding over-generation. We selected two nodes in WordNet as the root nodes: "move, locomote, travel, go"(synset ID 01818343-V), which covers the motion domain, and "to be (occupy a certain position or area; be somewhere)" (synset ID 02629830-V), which covers the orientation and position domains. Figure 4.2 and Figure 4.3 show a small part of WordNet including these synsets. These synsets have many hyponyms but do not belong to any synset, i.e. these synsets are top nodes in the WordNet tree.

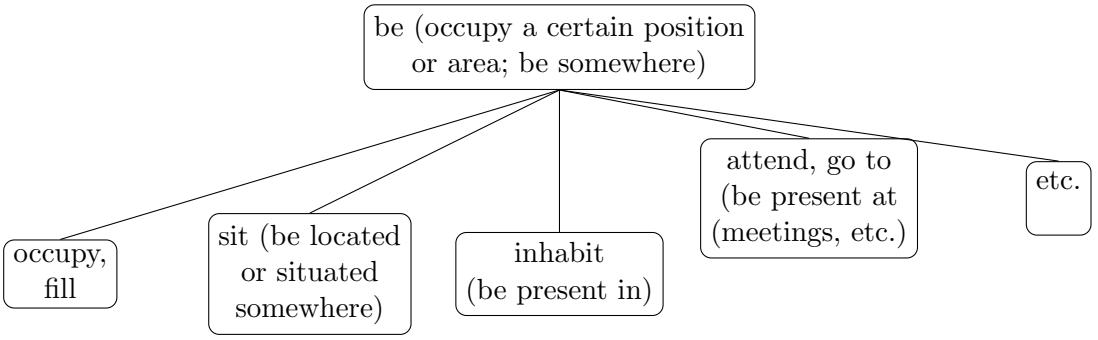

FIGURE 4.2: WordNet hierarchy of the "be(occupy a certain position)" synset

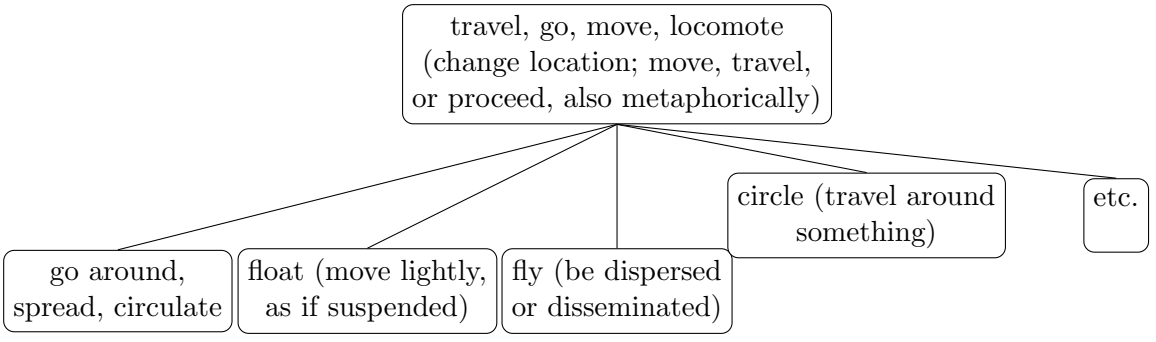

FIGURE 4.3: WordNet hierarchy of the "move,go,location" synset

To evaluate the efficiency of the WordNet approach, we consider three different strategies as mentioned before: NoWSD, WSD and Predominant Sense.

4.4.1.1 Evaluation of WN-based method in motion event identification

The results of the WN-based method are shown in Table 4.3.

TABLE 4.3: Motion verb recognition results with WordNet

	Precision	**Recall**	**F-score**
NoWSD	0.17	0.56	0.26
WSD	1.0	0.14	0.25
PredomSense	0.58	0.21	0.31

Using WordNet to identify motion events notably suffers from low F-score. Even in the NoWSD condition, many motion verbs are not included in the WordNet-derived list of motion verbs, the F-score being only 26%. The reason appears to be that a number of motion verbs are scattered in WordNet outside our two chosen subtrees. Examples include *put*, a hyponym of the synset *put into a certain place or abstract location)*, and *stand*, a hyponym of the synset *to be (have the quality of being)*. However, these motion verbs are not easy to assign to complete subtrees that can also be designated as motion subtrees.

Not surprisingly, NoWSD has by far the lowest precision of the three conditions, since many instances of verbs that have motion senses but also other senses are mistagged as motion verbs. The precision improves dramatically for the WSD condition, from 0.17 to 1.0. However, the recall takes a further major hit down to 0.14, and thus the resulting F-Score is not much different from the NoWSD condition. In the PredomSense condition, precision decreases compared to WSD condition, due to the predicates with both motion and non-motion senses, and the decline in recall compared to NoWSD is not quite as pronounced. Consequently, the PredomSense condition shows the overall best F-Score for WordNet-based models. That is, it seems currently preferable to employ a predominant sense heuristic over performing full-fledged word sense disambiguation. The best WordNet-based result, 0.31 F-score. An example of an instance that is wrongly classified as a motion verb even in the PredomSense condition in both WN and FN-based methods is the verb *go*, whose first Word-Net sense concerns motion, but has many other non-motion senses, as well. For example, in (85) and (86) the verb *go* belong to the *be or continue to be in a certain condition* and *pass, fare, or elapse; of a certain state of affairs or action* senses.

(85) The children went hungry that day.

(86) How is it going?

4.4.2 Mining motion events using FrameNet

Figure 4.4 shows the steps in our second approach, that is using FrameNet, as a basis for making the decision about motion events.

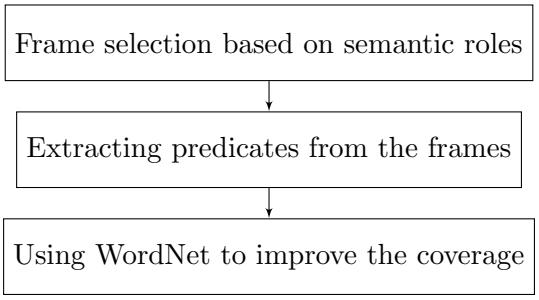

FIGURE 4.4: Using FrameNet to extract motion events

In this approach, we characterized motion predicates through a set of motion frames which were recognized based on their location semantic roles (cf. Table 4.1). We started firstly with the *MOTION* frame as the central frame and then selected all the frames which included core semantic roles as the *MOTION* frame (cf. Table 4.4). The verbal lexical units listed in these frames were considered motion events. However, the obtained list was far from complete, because only 67% of the manually annotated predicates (i.e. 47 different lemmas out of the total of 71 lemmas) in "Around the World in Eighty Days" were listed in the FrameNet list of motion events.

The missing predicates can be categorized under two groups: (a) verbs which are not listed in FrameNet, although the correct frame is available (e.g., *take a walk* which is supposed to belong to the same frame as *walk*, that is *SELF_MOTION*) and (b) verbs belonging to frames which are missing in our list of motion events. For example, the frame *CAUSE_TO_MOVE_IN_PLACE* lists several motion verbs but was not mentioned in our FrameNet list because according to FrameNet, the locational roles in these frames are not among the core roles.

To address the above missing coverage, we mapped the FrameNet lexical units to WordNet synsets using the mapping developed by Shi and Mihalcea (2005) to determine the matching synsets and lexical units. Assuming that hyponyms of a lexical unit evoke the same frame as the lexical unit, the coverage of the motion predicates was improved by including all hyponyms of the obtained synsets. The resulting set comprised 2838 synstes from WordNet.

TABLE 4.4: List of Motion frames in FrameNet

Adorning	Arriving
Cause_Fluidic_Motion	Cause_Motion
Cotheme	Departing
Emanating	Emitting
Emptying	Event
Filling	Fluidic_Motion
Giving	Import_Export
Light_Movement	Mass_Motion
Motion	Motion_Directional
Motion_Noise	Operate_Vehicle
Path_Shape	Placing
Posture	Precipitation
Quitting	Receiving
Removing	Residence
Ride_Vehicle	Self_Motion
Sending	Smuggling
Taking	Travel

4.4.2.1 Evaluation of FN-based method in motion events identification

TABLE 4.5: Motion verb recognition results with FrameNet

	Precision	**Recall**	**F-score**
NoWSD	0.31	0.73	0.43
WSD	0.51	0.44	0.47
PredomSense	0.59	0.58	0.58
WN-based approach best result (PredomSense)	0.58	0.21	0.31

Similar to WN-based approach, we evaluate the results in three settings: NoWSD, WSD and PredomSense. NoWSD evaluation in Table 4.5 shows that without disambiguation, 73% of the gold standard motion verbs in novel text are detected, but the precision is only 31%. The 27% false negatives are either cases of missing frames, or of missing lexical units in FrameNet and WordNet (phrasal verbs).

The low precision of 31% is due to ambiguous words with non-motion senses. Using WSD substantially improves precision (by 28%), but hurts recall, with an improvement in F-Score. Finally, the predominant sense heuristic outperforms WSD in both precision and recall. It cannot rival NoWSD in recall, but the higher precision yields net gains of 11% and 15% in F-Score compared to WSD and NoWSD conditions, the overall best results.

These numbers show that when the first sense of a verb is a motion sense, the heuristic assumption that instances of this verb belong to the motion domain outperforms the

UKB-provided disambiguation. It should, however, be noticed that UKB tries to solve a more difficult task, namely fine-grained sense assignment.

The heuristic is nevertheless far from perfect: Among the false positives, there are high-frequency high-ambiguity verbs like *take*, but we also find that many motion verbs specifically have a concrete motion sense but also a more abstract non-motion sense, often in the mental or cognitive domain. For example, the *lean* and *follow* predicates in (87) and (88) are used with their motion sense and (89) and (90) include their more abstract senses:

(87) She *leaned* over the banister

(88) Please *follow* the guide through the museum.

(89) We can *lean* on this man.

(90) A terrible tsunami *followed* the earthquake.

Comparing the results with the WordNet-based results shows that FrameNet-based approach works better in all settings. One main reason of worse results for WordNet-based approach is limiting the selected main nodes of motion predicates in WordNet to just two nodes, which causes losing many motion instances. In terms of precision however both approaches are comparable.

4.4.3 Summary

In this section, we followed a knowledge-based approach and performed an experiment to compare WordNet and FrameNet for the recognition of the motion domain (motion, orientation, and position verbs) and found the combination of FrameNet and WordNet a useful tool to define the motion domain. In this approach, we firstly extracted motion frames based on their core elements and then extended the list of motion events by hyponymy relationship of WordNet.

We also realized that the processing can proceed when just WordNet is available, but unsurprisingly with lower results. Comparing different word sense disambiguation schemes, the unsupervised WSD system UKB could not beat the simple "predominant sense heuristic".

One reason of the low performance of WordNet-based method is limiting our experiment to two subtrees in the WordNet verb hierarchy which leads to a very high precision but a

low recall. Since WordNet was not designed specifically with the motion domain in mind, many motion verbs are scattered throughout the verb hierarchy, and their distribution is difficult to describe succinctly.

Though our experiment showed using the combination of FrameNet and WordNet a useful approach to identify motion events with an F-score of around 60%, one avenue of future research is the refinement of the characterization of motion verbs in WordNet and FrameNet. As we have observed, our proposed list of motion frames misses some frames which should be added to the list of frames. Also, our current definition of motion events in WordNet is limited to two subtrees which results in a low recall. The experiments can be repeated with the evaluation of precision/recall trade-off that arises from adding more frames/synsets to the list of motion verbs. Furthermore, the version of the UKB tool used in our experiments was 0.1.6 which was an up to date version at the time of the experiment. Since newer versions of this tool have been released since then, one can repeat the experiments with the latest version.

4.5 Annotating Implicit Semantic Roles using Crowdsourcing

4.5.1 Introduction

After automatic identification of motion events, we follow our approach to solve the problem of data sparseness in ISRL by annotating implicit semantic roles using crowdsourcing (cf. Figure 4.1). Annotating implicit semantic roles using crowdsourcing is a more challenging task compared to semantic role annotation. While it has the challenges of explicit semantic role annotation task, such as Fossati et al. (2013)'s observation that presenting linguistic definitions to non-experts can negatively affect the results and the definitions must be simplified as much as possible, it has its own challenges, as well. While in semantic role annotation a single sentence is generally enough for the annotators to determine the role of different text spans or to determine the text spans which fill a given role, in implicit semantic role annotation it depends on the setting. In ISRL, the fillers of a role can occur anywhere in the text, that is, maybe many sentences before or even some sentences after the predicate. This makes the task more time-consuming. In addition, there can be many predicates and fillers between the predicate and the role filler which can distract or confuse the annotators. The following sections show how we try to find a solution to these challenges.

4.5.2 Designing a crowdsourcing experiment

Crowdsourcing paradigms for text annotation can be categorized under three classes: mechanised labour, which is based on financially rewarding the workers; games with a purpose, where the task is presented as a game; and humanitarian work, relying on goodwill (Sabou et al., 2014). A wide range of NLP problems, such as word similarity, textual entailment, event annotation and word sense disambiguation, have been performed using mechanised labour (Snow et al., 2008). The process of crowdsourcing can be broken down into four main stages: project description, data preparation, project execution and results evaluation (Bontcheva et al., 2014).

In the first stage, the NLP problem must be decomposed into a set of simple crowdsourcing tasks, which can be comprehensible for and performed by non-experts with minimal training and guidelines. There are a couple of common task types which have been used to perform many NLP problems. For example, a *selection task* is a common task type in which workers are presented with some information and required to select from a list of possible answers. Word sense disambiguation, sentiment analysis, and entity disambiguation are among the NLP tasks which cab be implemented using this approach. Another common task type is *sequence marking* where workers highlight parts of the text as the answer. This approach can be used in performing tasks such as named entity labeling, timex extraction, and actor identification. Using these common task types as templates, and also keeping the task simple are important principles which improves the efficiency of the experiment (Sabou et al., 2014).

In the second stage, data preparation, user interfaces are designed, the data is collected and prepared. In this stage, it might be needed to do some filtering to remove objectionable content.

The third stage, project execution, is the main phase of each crowdsourcing process and consists of three types of tasks: performing the task by the workers, managing the workers and doing the quality control. In the first task, it is important to attract a large number of workers to do the task. In the second one, some filtering may be required to determine which workers, e.g. with what level of knowledge, are allowed to do the task, and in the third task, the results must be controlled against spamming and cheating by the workers.

In the last step, the challenge lies in evaluating multiple workers' inputs and assessing their quality. Determining the workers' agreement is an important task in this step

and contributor aggregation primarily relies on majority voting or average computation based algorithms (Sabou et al., 2014).

4.5.2.1 Designing an ISRL crowdsourcing experiment

In our design of the ISRL crowdsourcing experiment, we chose the mechanised labour approach similar to many other NLP problems, as rewarding the contributors influences the time-completion of the task and the quality of the gathered data (Sabou et al., 2014). To determine how much to pay the workers per each task, we considered previous studies which had mostly offered 0.01-0.05$ per task (e.g. Finin et al. (2010), Lawson et al. (2010), and Mellebeek et al. (2010)), and paid 0.15$ per each task which included the annotation of (a maximum of) four semantic roles, i.e. around 0.04$ per each semantic role.

In the data preparation step, we prepared parts of the novel "Around the World in Eighty Days" to present to the workers (cf. Section 4.3). To simplify the task, we focused on a small number of motion predicates as operationalized by the lists from Section 4.4. These predicates could be clearly recognized as motion events by non-experts and did not include any motion verb like *blinking* whose recognition can be difficult for non-experts. As another step towards simplification, we shorten the context presented as the search window for the implicit roles to a three-sentence window (i.e. the sentence including the predicate and two prior sentences) and avoid presenting long texts to the workers, though this filtering can lead to losing some implicit roles. Previous studies have shown that around 90% of implicit roles can be realized in a window of three sentences (Gerber and Chai, 2010).

To keep the experiment's web interface user-friendly and simple, we provided the workers with a detailed description of the task and some examples through an external link. We followed one of the common types of experiment design, sequence marking. We did not consider the selection task pattern as a good solution in our case, due to the large number of candidates for each implicit role in the presented text and the results of prior studies which have shown that ideally the workers must not be asked to choose from a list of more than 10 items (Sabou et al., 2014).

The implementation of the experiment was carried out using Amazon Mechanical Turk because it allows only workers from the US to participate and we assumed it to be a more reliable platform to do the annotation for an English text compared to other platforms which are available for all people from all over the world. Then, we defined our HITs

(Human Intelligent Task) in Amazon Mechanical Turk. A HIT is a single task which must be done by a worker. In our case, each HIT included a text with a predicate in bold whose semantic roles were required to be determined in the text.

To determine how many workers to hire for each HIT, considering the trade-off between the cost and the quality of the annotation, we decided for five workers. Previous studies have evaluated hiring different numbers of workers, from 1 to 50, and the results have shown that increasing the number of workers to 5 causes a dramatic decrease in the error rate (Carvalho et al., 2016).

Figure 4.5 shows the workflow of our experiment. As can be seen, the HITs are firstly annotated by the turkers using one of the two different setups (we firstly had the annotation performed using the marking task and then switched to the gap filling approach. For more details cf. Section 4.5.2.2 and Section 4.5.2.3) and the result is evaluated regarding the agreement among the annotators. Then, a canonicalized version of the annotation is created as the final version of the crowdsourcing annotation (cf. Section 4.5.2.4) and is compared to the expert annotation to evaluate the overall results of the crowdsourcing and evaluating if the implicit semantic role annotation can be reliably performed using crowdsourcing. The following sections describe different parts of the workflow in more details.

4.5.2.2 Marking setup

As discussed earlier, the sequence setup (which is called marking setup in our experiments) is one of the common types of crowdsourcing designs. In this setup, in each HIT, we present a text including a target predicate to the the annotators and mark the target predicate in boldface. The annotators are asked to read the text and answer four questions about "the event in bold". Figure 4.6 shows a screenshot of the web interface presented to the turkers.

The asked questions correspond to the PLACE, SOURCE, GOAL and PATH semantic roles of the motion events. We used these questions to follow Fossati et al. (2013) in avoiding presenting linguistic definition of the roles to the turkers. This decision was made based on two reasons: (1) The FrameNet definitions of location roles require defining some other roles, as well. For example, in *MOTION* frame, the GOAL semantic role is defined as "The GOAL is the location the THEME ends up in." which calls for the definition of THEME. (2) Not all the selected predicates for annotation belong to the same frame. Due to the differences between the definitions of the semantic roles for

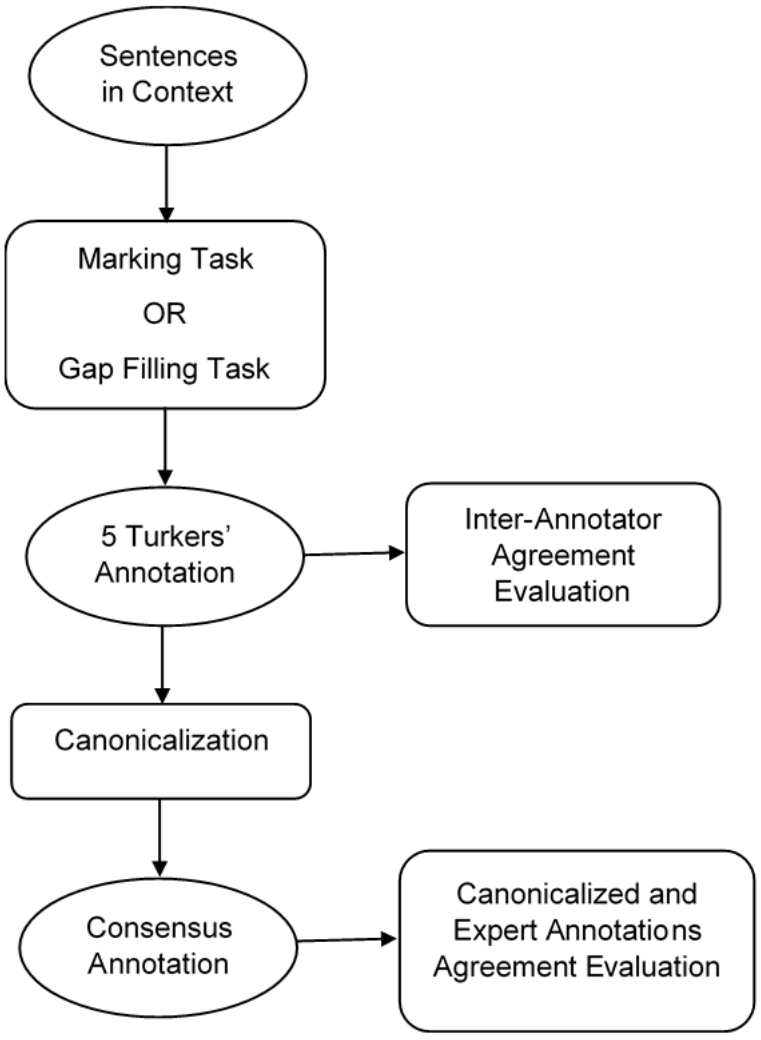

FIGURE 4.5: Workflow of the Crowdsourcing experiment for ISRL

different FrameNet frames, it is not possible to select a definition from the FrameNet which is common between all motion predicates. For example, the GOAL semantic role for the *PATH_SHAPE* frame is defined as "Any expression which tells where the fictive mover travelling along a ROAD would end up." which is different from its definition for *MOTION* frame. Therefore, simplifying the role definitions in the way done makes them usable for all motion predicates, while using FrameNet definition requires to present turkers with different definitions per predicates of each frame.

To answer the questions, the turkers could mark a text span (shown in a non-editable field below the question) or click on the button "no answer found". The goals of this setup were (a) to minimize annotation effort by marking the text instead of typing, and (b) to make the task as layman-compatible as possible.

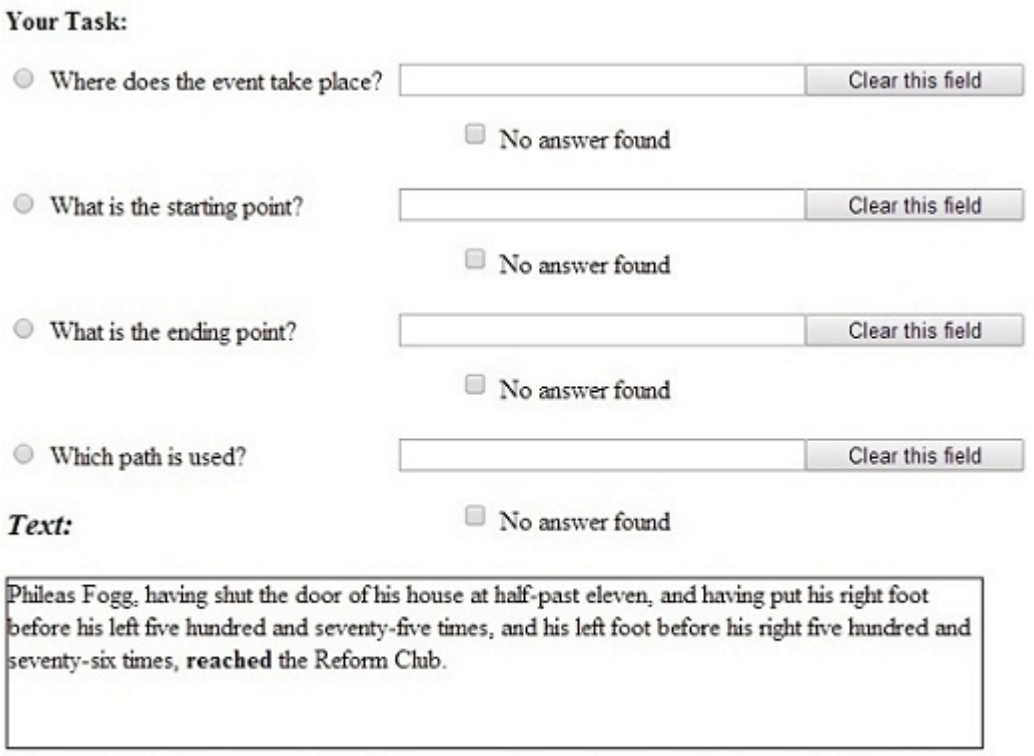

FIGURE 4.6: Web interface of the marking setup

The presented task did not limit the turkers to annotate just the implicit roles. In other words, the turkers were asked to answer the given questions, regardless of where the answers can be found in the text. In this way, both explicit and implicit roles were annotated and whether an annotated text was an implicit or explicit role was later determined by comparing with the expert annotation.

After annotating some instances, we computed raw inter-annotator agreement (IAA) in two conditions (average pairwise exact match and word-based overlap) for the first 49 instances. (1) and (2) present the equations we used to calculate the inter annotator agreement.

(1) Exact Match: $\sum_{i=1}^{n} \frac{\sum_{j=1}^{\frac{k(k-1)}{2}} m_j}{\frac{k(k-1)}{2}}$

(2) Word-based Overlap Match: $\sum_{u=1}^{n} \frac{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \frac{P_{(i,j)}}{max(s_i, s_j)}}{\frac{k(k-1)}{2}}$

in which

k represents the number of turkers per each annotation (i.e. 5 in our case)

m_j represents the result of the exact match pairwise comparison (0 or 1)

n represents the number of all HITs

$P_(i, j)$ represents the number of common words in the pairwise comparison of annotations for a HIT

s_i represents the length of the annotation i

and

s_j represents the length of the annotation j

The results can be seen in Table 4.6. Since the the same HIT was not necessarily done by the same people, it was not possible to calculate kappa measure for the inter annotator agreement.

TABLE 4.6: Raw inter-annotator agreement in the "marking" task

	Source	Goal	Path	Place
Exact Match	0.35	0.44	0.48	0.24
Overlap	0.35	0.46	0.52	0.27

As can be seen, the overall IAA is 37.7% in Exact Match setting and 40.0% in Overlap setting. Due to the low IAA, we did not continue with this approach. The low result even in the overlap condition indicates that the problems can not be due mainly to minor differences in the marked spans. Thus, we performed an analysis and realized that the main reason was that annotators were often confused by the presence of multiple predicates in the paragraph such that many marked roles pertaining not to the boldfaced target predicate but to other predicates. For example, in

(91) Leaving Bombay, it passes through Salcette, *crossing* to the continent opposite Tannah, goes over the chain of the Western Ghauts, runs thence north-east as far as Burhampoor, skirts the nearly independent territory of Bundelcund, ascends to Allahabad, turns thence eastwardly, meeting the Ganges at Benares, then departs from the river a little, and, descending south-eastward by Burdivan and the French town of Chandernagor, has its terminus at Calcutta.

the annotators were expected to annotated *the continent opposite Tannah* as the GOAL of *crossing* predicate, but some annotators marked *Calcutta*, the final destination of the chain of the motion events described.

As an another example, in

(92) At five minutes before eight, Passepartout, hatless, shoeless, and having in the squabble lost his package of shirts and shoes, rushed breathlessly into the station. Fix, who had followed Mr. Fogg to the station, and saw that he was really going to leave Bombay, was there, upon the platform.He had resolved to follow the supposed robber to Calcutta, and farther, if necessary. Passepartout did not observe the detective, who stood in an obscure corner; but Fix heard him relate his adventures in a few words to Mr. Fogg. "I hope that this will not happen again," said Phileas Fogg coldly, as he got into the train. Poor Passepartout, quite crestfallen, *followed* his master without a word.

the annotators were expected to annotate *the train* as the GOAL of the *followed* predicate, but *Calcutta*, the more general destination, was wrongly annotated.

Also, in

(93) Passepartout jumped off the box and followed his master, who, after paying the cabman, was about to enter the station, when a poor beggar-woman, with a child in her arms, her naked feet smeared with mud, her head covered with a wretched bonnet, from which hung a tattered feather, and her shoulders shrouded in a ragged shawl, approached, and mournfully asked for alms. Mr. Fogg took out the twenty guineas he had just won at whist, and handed them to the beggar, saying, "Here, my good woman. I'm glad that I met you;" and passed on. Passepartout had a moist sensation about the eyes; his master's action touched his susceptible heart.Two first-class tickets for Paris having been speedily purchased, Mr. Fogg was *crossing* the station to the train, when he perceived his five friends of the Reform.

Paris was wrongly annotated as the GOAL of the *crossing* predicate, whose expected GOAL was *the train*.

This problem can be called global understanding, that is, the annotators answer the given questions based on their global understanding of the whole given paragraphs or sentences instead of focusing on one individual predicate which is the boldfaced target predicate (for more detailed discussion on what other information affects the inference process, cf. Graesser and Singer (1994).

Since the essence of the implicit semantic role annotation requires presenting prior context to the annotators, the problem can not be solved by shortening the text. Because

it may cause losing the correct fillers. In the following section, we discuss our solution to overcome this problem.

4.5.2.3 Gap filling setup

To solve the problem of the marking task which did not constrain the interpretation of the turkers sufficiently, we moved to a gap filling setup to anchor the turkers' attention better to a single predicate rather than the complete set of predicates present in the given text. In this setup, the annotators were asked to complete the sentence by filling in the blanks in two sentences. The web interface of this setup can be seen in Figure 4.7.

As the picture shows, the first sentence corresponds to annotation of the SOURCE, GOAL, and PATH roles and the second one of the PLACE role. The rationale is that the presence of the predicate in the sentence focuses the turkers' attention on the specific predicate and its actual roles. The annotators were asked to fill in the gaps with the words from the given context, i.e. avoid paraphrasing or leaving them empty in the case of unrealized roles.

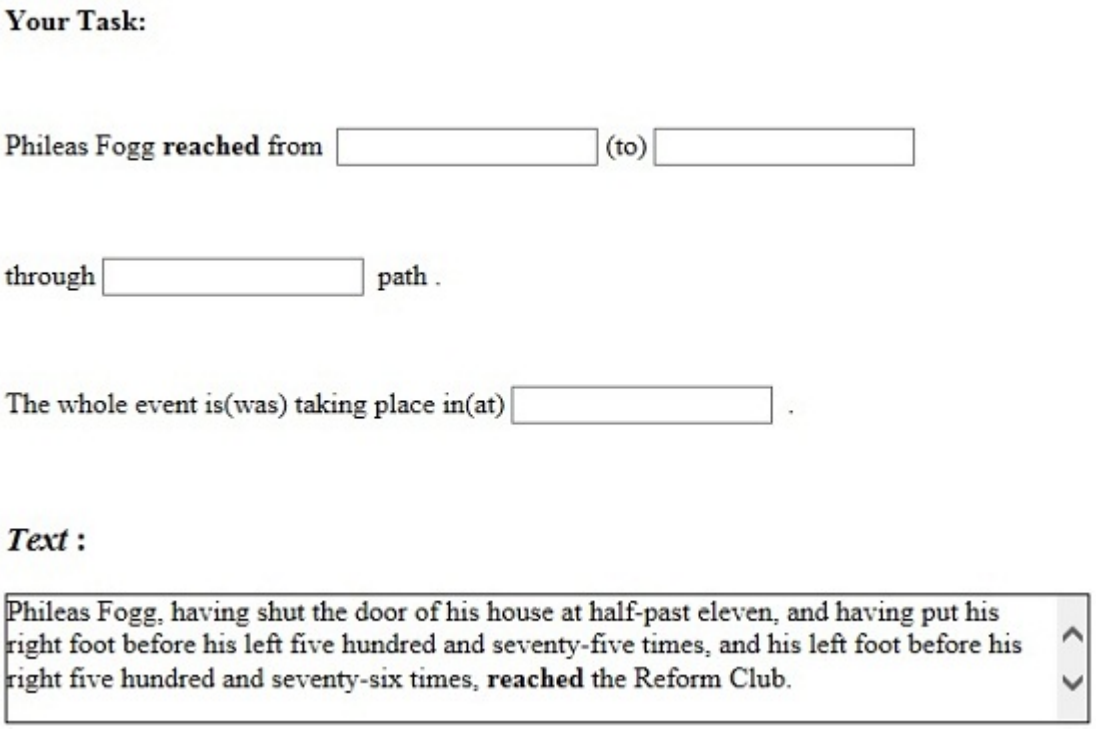

FIGURE 4.7: Web interface of the gap filling setup

The analysis of the annotations showed that by paying 0.15$ per each task annotated by each worker, the whole annotation, which included 384 predicates, cost 288$. The whole annotation was carried out in around 60 hours, that is, an hourly wage of 4.88$. On average each HIT was annotated in 1 minute and 48 seconds, which means 27 seconds per each role.

Table 4.7 shows the pairwise IAA for this setup, which is again calculated as the agreement between each pair of annotators. The results show that with the new approach which requires to make a meaningful sentence using the semantic roles, the IAA improves considerably in both Exact Match (by 11%) and Overlap (by 15%) conditions in the base case (for Source). The overall IAAs in this setup are 44.5% and 50.0% in the Exact Match and Overlap conditions, respectively, which mean 6.8% and 10.0% improvements. Overall, the numbers are still fairly low. However, they are the IAA numbers among turkers and not the agreement among a "canonical" version of the turkers' annotation and an ideal gold standard (cf. Section 4.5.2.4). In addition, a data analysis showed that in this setup, many of the disagreements are cases where annotators disagree on the exact range of the string to fill into the gap which can also be seen by higher numbers for the overlap match compared to exact match in Table 4.7

TABLE 4.7: Raw inter-annotator agreement in the "gap filling" task

	Source	Goal	Path	Place
Exact Match	0.46	0.46	0.56	0.30
Overlap	0.50	0.54	0.58	0.38

For example, in

(94) Skillful detectives have been *sent* to all the principal ports of America and the Continent, and he will be a clever fellow if he slips through their fingers.

turkers annotated different spans, including *all the principal ports of America*, *ports*, and *America and the Continent* as the correct span for the GOAL role of the *sent* predicate, while experts would annotate *all the principal ports of America and the Continent*.

The lowest IAA is found for the PLACE role due to its vague nature compared to the other roles which made it more difficult for annotators to tag consistently. For example, in

(95) Phileas Fogg, having shut the door of [$_{SOURCE}$ his house] at half-past eleven, and having put his right foot before his left five hundred and seventy-five times, and his left foot before his right five hundred and seventy-six times, *reached* [$_{GOAL}$ the Reform Club].

the PLACE could be, *the city*, *London*, *England*, etc. We also found that for some predicates such as *arrive* and *reach* many turkers attempted to resolve the ambiguity by (erroneously) annotating the same text as both GOAL and PLACE, which runs counter to the FrameNet guidelines.

These observations show that to get more reliable results, one must firstly try to solve the problem of specificity of the vague roles like PLACE, because many locations mentioned in a text are related hierarchically.

4.5.2.4 Canonicalization

To compare the turkers' annotations and the expert annotation, we needed to compute a "canonical" annotation that combines the five turker's annotations. Creating the canonical version mainly relies on majority voting or average computation based algorithms (Sabou et al., 2014). However, we found it necessary to be more flexible because manual analysis of a few instances showed that cases of agreement between two turkers' annotations with non-empty overlap could be accepted as non-local roles. That is, turkers frequently miss non-local roles, but if two out of five annotate an overlapping span with the same role, this is reasonable evidence. For example, two turkers have annotated *the Indian Ocean* as the PATH of the *reach* predicate in (96) and the other three annotators have missed it. In such cases, we assume the annotated text as the canonicalized version.

(96) At six p.m. the Mongolia slowly moved out of the roadstead, and was soon once more on the Indian Ocean. She had a hundred and sixty-eight hours in which to *reached* Bombay.

Regarding the role's span, we used the consensus span if it existed, and the maximal (union) span otherwise, given that some turkers filled the gaps just with head words and not complete constituents. For example, in (96), if the annotations of the turkers are *the Indian Ocean* and *Indian Ocean*, we pick the longest span as the canonicalized version.

To test the quality of the canonical annotation, an expert annotator annotated the instances that were presented to the turkers. We considered the result to be an expert annotation approximating a gold standard and used it to judge the quality of the canonical turker annotations. The comparison between the canonical annotation and the expert annotation with regard to the roles and realization status are shown in Table 4.8 and Table 4.9.

TABLE 4.8: Raw agreement between canonical crowdsourcing annotation and expert annotation by role

	Source	Goal	Path	Place
Exact Match	0.72	0.67	0.82	0.50
Overlap	0.72	0.69	0.82	0.54

TABLE 4.9: Raw agreement between canonical crowdsourcing annotation and expert annotation by realization status

	Local	Non-local	Unrealized
Exact Match	0.66	0.66	0.69
Overlap	0.69	0.70	0.69

As the Table 4.8 shows, raw agreement between the canonical annotation and the expert annotation is considerably higher than the inter-annotator agreement. Again, we see that the higher numbers belong to PATH role which is the most specific one and the lowest number can be seen for PLACE role which has the vaguest nature. The disagreement cases are, similar to the inter annotator case, due to the differences between the exact span annotated. For example, in 97, the SOURCE annotated by the expert is *before the railway station*, while the canonicalized annotation includes only *the railway station*.

(97) The cab stopped before the railway station at twenty minutes past eight. Passepartout jumped off the box and followed his master.

To evaluate whether the annotations are affected by the realization status of the roles, we assessed the results in terms of realization status as well, cf. Table 4.9, and the results indicated that non-locally realized roles are annotated as reliably as locally realized ones.

The results show that except for the ill-defined PLACE role, our reliability is comparable to Fossati et al. (2013) who obtained an accuracy of 0.79. Given the more difficult nature of the task due to presence of a longer context and the more difficult decision of identifying locally-unrealized roles, we consider the obtained results promising.

This experiment resulted in annotation of 394 instances of motion predicates with 666 semantic roles, composed of 251 implicit roles and 415 explicit ones. However, after

the conversion to PropBank style, only 32 instances of the motion predicates remained which included 40 implicit and 28 explicit roles.

4.5.3 Summary

This section presented a study on crowdsourcing the annotation of non-local semantic roles in discourse context. We designed two setups (marking setup and gap filling setup) to annotate the implicit semantic roles of a small number of motion predicates using the power of crowdsourcing. We found that gap filling is the more reliable choice since the repetition of the predicate helps focusing the turkers' attention on the roles at hand rather than understanding of the global text. Thus, the semantic role-based crowdsourcing approach of Fossati et al. (2013) appears to be generalizable to the area of non-locally realized roles, provided that the task is defined suitably. Our results also support Fossati et al. (2013)'s observation that reliable annotations can be obtained without providing definitions of semantic roles. However, we also found large differences among semantic roles. Some (like PATH) can be annotated reliably and should be usable to train or improve SRL systems. PLACE, in contrast, was annotated least consistently. Part of the problem is presumably that PLACE is defined relatively vaguely in FrameNet as "the general area in which a specific motion (with SOURCE, PATH, or GOAL) occurs", that is, a kind of location specification that can potentially be filled by more than description in a whole text, due to the meronymy relations that hold among locations: If an event happened at Phileas Fogg's house, it also happened, in London, in the UK, etc.

4.5.4 Assessment of limitations

Though Table 4.8 show reliable results, we did not go on with further studies using the obtained results. There were two main types of limitations for performing further experiments. The first one was the scalability problem which is briefly explained here:

- We simplified our crowdsourcing task by choosing a specific domain and selecting predicates belonging to a specific frame, *MOTION* frame, which all need the same set of roles. However, this is not generally the case and FrameNet roles are not easily generalizable to other frames beyond the sets of frames related by frame-to-frame relations (cf. Ruppenhofer et al. (2006)). Therefore, the crowdsourced data was not suitable for training a model to identify implicit roles of other frames. On the other hand, it was also not easy to apply the same approach for gaining more annotations for other frames. If one needs to get annotations for different predicates, he has to define a mapping from

roles to slots/questions in our crowdsourcing experiment. For example, to get the annotations for the GOAL role for the *MOTION* frame, a question like "what is the ending point of the movement?" must be defined. In this case, the mapping must be defined for each frame (not for each verb) but it is still hard to do as it would ultimately mean manually defining slots/questions for hundreds, if not thousands, of roles.

- Many of the best systems working on SRL have shown that PropBank style annotation leads to better results compared to FrameNet style (cf. Chapter 3). But converting the data to PropBank framework resulted in a small number of implicit roles as many roles where converted to adjuncts and therefore it resulted in a very incompletely annotated corpus, seen from the perspective of PropBank roles.

- One possible solution to avoid the conversion problem may be to annotate the PropBank roles instead of FrameNet roles in the procedures. But annotating PropBank roles with the procedures proposed in this chapter was also not possible because PropBank roles are much more abstract than FrameNet roles and have different meanings for different predicates. Therefore, it was harder to formulate them as questions/slots, unless we defined a mapping onto questions/slots for each predicate- a task an order of magnitude larger than the corresponding definition for FrameNet.

In addition to the scalability problem, there were some technical issues which limited the possibility of training a supervised model on the crowdsourced "around the world in 80 days" corpus:

- During the conversion of roles to PropBank annotation, roles like PATH which had the highest agreement among the annotators were removed due to being categorized under adjunct arguments, and the remaining instances suffered from a low agreement and were somewhat questionable as gold standard annotations.

- The overt roles of the non-target predicates in the crowdsourced data set, i.e. the predicates which were not annotated by the annotators, must be annotated automatically. Thus, features extracted using this information could be inaccurate.

- ISRL models profit strongly from discourse level information such as conference (cf. Chapter 5). Since gold standard coreference chains were not available for this corpus, this would have made features relying on this information unusable.

4.6 Summary

In this chapter, we evaluated crowdsourcing as a potential approach to obtain annotated data for ISRL task in a cheap way. Our findings showed that by simplifying the task and presenting an appropriate task design to the crowd annotators, we could obtain fairly reliable annotations which are well-agreed by the annotators. However, due to the mentioned limitations, this approach could not scale up to larger corpora which motivated our decision to adopt a different approach in solving the problem of data sparseness in ISRL. For this purpose, we utilized an additional data set which had already been annotated with semantic roles and combined it with the SemEval2010 training set to train a model. Due to the different genres of the data set, we used feature augmentation as a domain adaptation technique to overcome the domain difference problem. More details about this approach and the results are presented in the next chapter.

Part III

Domain Adaptation Technique in ISRL

Chapter 5

Domain Adaptation in ISRL

Aspects of the work in this chapter have been published in Feizabadi and Padó (2015).

5.1 Introduction

Since the annotated data using the crowdsourcing experiment was not considered scalable enough to train a model for other corpora, and considering that there are multiple annotated corpora for implicit semantic roles, we experiment with the use of domain adaptation to combine them to address the data sparsity problem of ISRL. In this approach, we exploit a domain adaptation technique (cf. Chapter 3) to combine the available large corpora which have already been reliably annotated by linguistics expert.

Then, we evaluate the performance of our model on both source and target domain and also evaluate the scalability of the approach by combining more than two corpora.

5.2 Existing corpora for ISRL

We use three main corpora in our experiments. The first one is the SemEval2010 task 10 (Ruppenhofer et al., 2010) data set which is the benchmark on ISRL and has been used by many previous studies, the second one is the Gerber and Chai (2010) corpus which is a large annotated corpora in newswire domain focusing on annotating instances of 10 nominal predicates, and the third one is the Moor et al. (2013) corpus which provides annotation of implicit semantic roles for five verbal predicates in OntoNotes corpus. By selecting these corpora, we can evaluate the effect of combining different corpora on

annotating implicit roles in different domains for predicates with different POSs. In the following sections, you can find more details about these corpora, their similarities and dissimilarities.

5.2.1 The SemEval 2010 corpus

SemEval2010 corpus, henceforth called SemEval, is the data set provided by the organizers of the SemEval2010 shared task on ISRL (task 10), called *Linking Events and Their Participants in Discourse* (Ruppenhofer et al., 2010). The data set is from Arthur Conan Doyle's fiction works. They selected fiction genre rather than news because they assumed that fiction texts generally include more recoverable null instantiations. They also assumed that these texts have a more linear structure compared to newswire texts which normally present different levels of details about each fact by considering it from different perspectives. The data set which presents full-text annotation is composed of three sections: one as the training set and two sections as the test sets. The training set is taken from *The Adventure of Wisteria Lodge* story which has two parts. The annotated part is called *The Tiger of San Pedro* and is composed of 7917 words. The test set is composed of the last two chapters of *The Hound of the Baskervilles* story, which is a different book and can cause dissimilarities between the training set and the test set. These chapters include a total of 9083 words.

5.2.1.1 FrameNet vs. PropBank annotation

The SemEval corpus is originally presented in FrameNet style which can be converted to PropBank framework automatically using the FrameNet-PropBank mapping provided by the task organizers. The verbal predicates are converted to PropBank annotation using SemLink [1]. For nominal predicates there is no direct mapping between FrameNet and NomBank predicates. Therefore, the task organizers obtained the mapping indirectly. In this approach, they considered the fact that PropBank verbs and eventive nouns in NomBank both have a mapping to VerbNet classes which are covered by SemLink. In this way, they could convert the nominal predicates of FrameNet to NomBank style with the help of SemLink, VerbNet, PropBank and Nombank. In this manner, however, only eventive nouns could be mapped and other types of predicates in FrameNet, like adjectives, which did not have any counterpart in PropBank were missing.

[1] http://verbs.colorado.edu/semlink/

Table 5.1 and Table 5.2 show the number of implicit roles in the data set in both annotation frameworks.

TABLE 5.1: Number of implicit and explicit semantic roles in the SEMEVAL corpus in FrameNet framework (adapted from Laparra and Rigau (2012))

	Implicit Roles	Resolvable Implicit Roles	Explicit Semantic Roles
Training set	325	245	2726
Test set	349	259	3233

TABLE 5.2: Number of implicit and explicit semantic roles in the SEMEVAL corpus in PropBank framework

	Implicit Roles	Resolvable Implicit Roles	Explicit Semantic Roles
Training set	122	122	1091
Test set	133	133	1440
Total	255	255	2531

As the Table 5.2 illustrates, the PropBank annotation of the data set includes only recoverable implicit roles. Comparing Table 5.1 and Table 5.2 shows that the number of explicit and implicit roles in PropBank framework is almost half of FrameNet. There are two main reasons for it: Firstly, as mentioned, the predicates in FrameNet can be verbs, nouns and adjectives, while PropBank (and its companion project, NomBank) cover only verbal and nominal predicates. Thus, during the automatic conversion, some of the predicates and their corresponding semantic roles are lost. For instance, in

(98) A [INCORPORATED_ATTRIBUTE *cold*] and melancholy walk [CIRCUMSTANCES of a couple of miles] brought us to a high wooden gate, which opened into a gloomy avenue of chestnuts [2].

cold is a target predicate belonging to *AMIBIENT_TEMPERATURE* frame which requires six core roles, ATTRIBUTE, DEGREE, PLACE, TIME, WEATHER TEMPERATURE and can have a non-core role, CIRCUMSTANCES. Among these roles, ATTRIBUTE is incorporated in the lexical unit (cf. Chapter 2), CIRCUMSTANCES is realized by the phrase *of a couple of miles*, PLACE and TIME are lexically licensed definite null instantiations and DEGREE is lexically licensed indefinite null instantiation which are all left unrealized.

As the example shows, annotation of *cold* adjective predicate causes annotation of two additional semantic roles which are missing in the PropBank version.

[2] Adapted from SemEval2010 training set.

Secondly, the PropBank annotation of the data set includes only the core roles, while the FrameNet annotation covers both core and non-core roles, e.g. CIRCUMSTANCES in (98).

As the goal of this dissertation was to address the data sparseness problem, we decided to work with the PropBank style annotation in which the number of argument classes are fewer compared to FrameNet, though it caused losing around half of the instances. Predicates from different frames have different sets of roles in FrameNet and it causes a sparse data set for implementing a classification task. In contrast, PropBank which includes a fixed number of numbered-argument reduces the number of classses to only five classes (A0-A4) which simplifies the task.

5.2.1.2 Annotation analysis

Table 5.3 presents details about the number of explicit and implicit roles in SemEval data set.

Table 5.3: Number of implicit and explicit roles per PropBank semantic roles in SemEval

Role	# of explicit roles	# of implicit roles
A0	568	69
A1	699	110
A2	148	45
A3	11	7
A4	14	24
Overall	1440	255

As the table shows, the realization status is different for different roles, while A0 and A1 are more probable to be realized explicitly, A3 and A4 are more likely to remain locally unrealized, and A2 can be placed between these two groups. At the same time, A0 and A1 are much more frequent and therefore are still the majority classes among the implicit roles.

Beside the number of implicit roles and their distribution among different roles, the distance between the target predicate and the filler of the implicit role is another important point which can affect the difficulty of the task, because the automatic systems must always consider the trade-off between the search window and the upper bound of recall. When the search window is extended, the upper bound recall, i.e. the number of recoverable implicit role fillers, will also increase. On the other hand, the total number

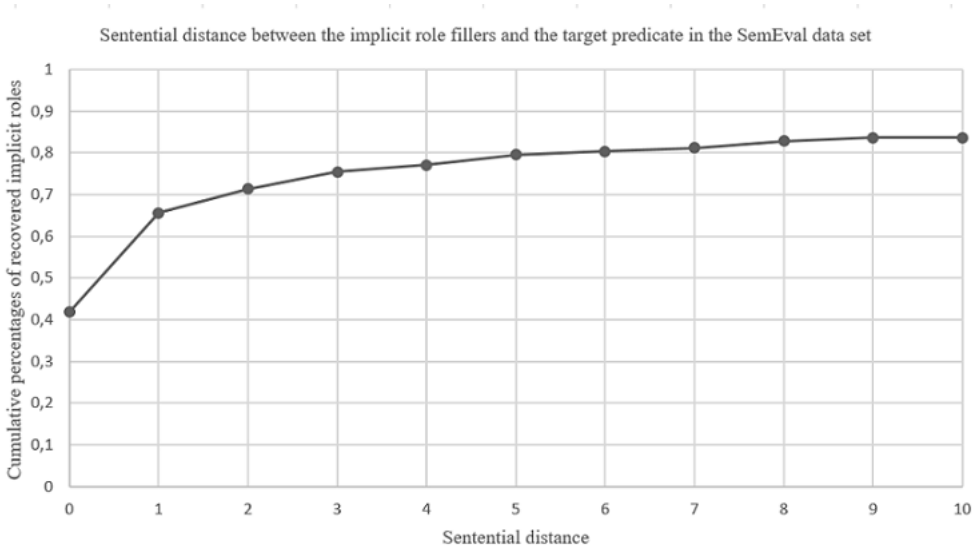

FIGURE 5.1: Sentential distance between the implicit role fillers and the target predicate in SEMEVAL training set

of candidate fillers for a role increases as well, which leads to a more complicated classification task. Figure 5.1 depicts the percentage of the recoverable implicit role fillers in a window of 10 sentences, i.e. current sentence and nine sentences of the pre-context.

Around 42% of the implicit role fillers are realized in the same sentence as the target predicate. A wider window of 3 sentences (current sentence and two proceeding sentences) can cover around 75% of the implicit role realizations. Around 16% of the role fillers are not covered even in a window of 10 sentences. It means that extending the window size from 3 to 10 can increase the number of recoverable implicit roles only by 9%, i.e. around one percent per each extra sentence. Considering the task complexity which results from extending the window size and including tens or hundreds of candidate filler per each role, we find a window of three sentences an ideal choice for our experiments.

5.2.1.3 Data annotation

The data formats for the FrameNet and PropBank style annotation are TIGER/SALSA XML (Erk and Pado, 2004) and a modified version of CoNLL format, respectively. The data is annotated with POS tags and also parse trees which are produced by Shalmaneser (Erk and Pado, 2006) using the Collins parser (Collins, 1997). The coreference chains are also manually annotated in the training set. In the test set, the manual coreference chains are only used for evaluation purposes and are not used for making predictions.

Since SemEval2010 task on ISRL was the first task which provided a full-text annotation corpus for ISRL, it has been used as the main corpus of many follow-up studies on ISRL. Therefore, we selected this corpus for our experiments to be able to compare our results with others'.

5.2.2 Gerber and Chai's corpus

Gerber and Chai (2010) developed another corpus, henceforth called GerberChai, whose focus was on annotation of implicit roles for nominal predicates. In this corpus Penn TreeBank, the textual basis of NomBank, is annotated with implicit roles.

5.2.2.1 Predicate selection

Due to existence of a large number of predicates in Penn TreeBank, Gerber and Chai (2010) limited their study to a selected group of predicates and considered the following criteria for this purpose (Gerber and Chai, 2012):

- The selected predicates must be derived from a verb. These predicates are also called eventive predicates. NomBank includes some non-eventive predicates, such as % symbol.

- The selected predicates must have an unambiguous role set, i.e. given the arguments supplied by NomBank, one can determine the noun's role set to determine which roles are missing.

- The selected predicates should have a high frequency in the Penn TreeBank corpus. The frequency of the predicates are calculated by considering morphological normalization, i.e., counting *bids* and *bid* as the same predicate.

- The selected predicates should include a large number of implicit roles in NomBank. To estimate the number of implicit roles before annotating the text, Gerber and Chai (2010) calculated the average number of roles expressed by the nominal predicate in NomBank, N_p, and the average number of roles expressed by its counterpart predicate in PropBank,V_p, and assumed N_p - N_p as an indication of the number of implicit roles which can be recoverable from the context. The motivation for this hypothesis can be explained by some examples:

(99) * John LOANED (the money to Mary).
* John INVESTED (his money).

(100) John's LOAN was not repaid.
John's INVESTMENT was huge [3].

In (99) and (100), removal of the arguments in parentheses in (99) causes ungrammatical structures, while the nominal predicates of the events can exist without those arguments and still be grammatical. It is worthwhile to mention, however, that the examples in (100) are meaningful only if a referent to the missing argument exists in the context and this is exactly the type of predicates which should be considered for implicit role annotation.

The predicates were selected using the first and second criteria and then ranked according to the third and fourth ones. Then, the top 10 predicates were selected for annotation. The list of the selected predicates includes: *bid*, *sale*, *loan*, *cost*, *plan*, *investor*, *price*, *loss*, *investment*, *fund*.

5.2.2.2 Data annotation

As the explicit roles in NomBank are annotated in PropBank paradigm, annotation of the implicit roles follows the same approach, i.e. the implicit roles can be any of A0-A4 (cf. Chapter 2). The annotation procedure was as follows:

- From the document d, all non-proper singular and non-proper plural nouns belonging to the list presented in Section 5.2.2.1 were selected.

- Due to the unambiguity of the selected nominal predicates, the implicit roles are determined by considering the explicit roles presented by NomBank.

- For each implicit role, the current sentence and all preceding sentences are searched for a suitable filler.

- If possible, the textual bounds of the implicit role filler are matched to the textual bounds of an argument presented by either PropBank or NomBank.

This approach led to annotating a total of 1172 implicit roles in the whole data set which is composed of 5702 sentences including a total of 140536 words. Since there was no pre-defined training set and test set split in this data set, during our experiments, we performed 3-fold cross validation by splitting the data set based on the number of documents. That is, we selected one third of the data set as the test set and the

[3] Examples in this section are adapted from Gerber and Chai (2012).

remaining two third as the training set. By this approach, we obtained 869 vs. 303, 772 vs. 400 and 703 vs. 469 as the proportion of implicit roles in the training set vs. test set.

Compared to SemEval, GerberChai is around 4.6 times larger in terms of the number of annotated implicit roles while the whole size of the data set is around 8.2 times the size of SemEval. Since GerberChai has focused on annotating implicit roles of a pre-selected list of nominal predicates, the relatively larger size of the data set is reasonable.

5.2.2.3 Annotation Analysis

Table 5.4 presents the number of explicit and implicit roles in GerberChai data set per each role.

Table 5.4: Number of implicit and explicit semantic roles in GerberChai data set

Role	# of explicit roles	# of implicit roles
A0	453	484
A1	641	290
A2	330	291
A3	112	104
A4	10	3
Overall	1546	1172

As the table shows, the selected predicates have frequent implicit roles, such that the number of implicit instances of A0 are even higher than the explicit ones which reflects their choice of nominalization phenomenon. In general, however, the number of explicit roles are higher than the implicit ones. Comparing the results with the implicit roles in SemEeval, one can find that the implicit roles in nominal predicates are much more frequent. While implicit roles in SemEeval make up 15% of the realized roles, they include 43% of the realized roles for the selected nominal predicates in GerberChai.

Figure 5.2 shows the sentential distance between the implicit role fillers and the target predicate in GerberChai data set. More than 60% of the fillers are covered by the same sentence as the target predicate and a window of three sentences covers around 90% of all implicit role antecedents in discourse. If the window is extended to 10 sentences, 98% of the fillers are covered.

Comparing these numbers with SemEeval corpus reveals that the implicit roles in the newswire text are recoverable in a much closer distance compared to the implicit roles in the novel text. In addition, the closer distance between the predicate and its implicit

role filler can be related to the POS of the predicate, i.e. nominalizations have their implicit roles in a closer distance compared to predicates of all types (noun and verb).

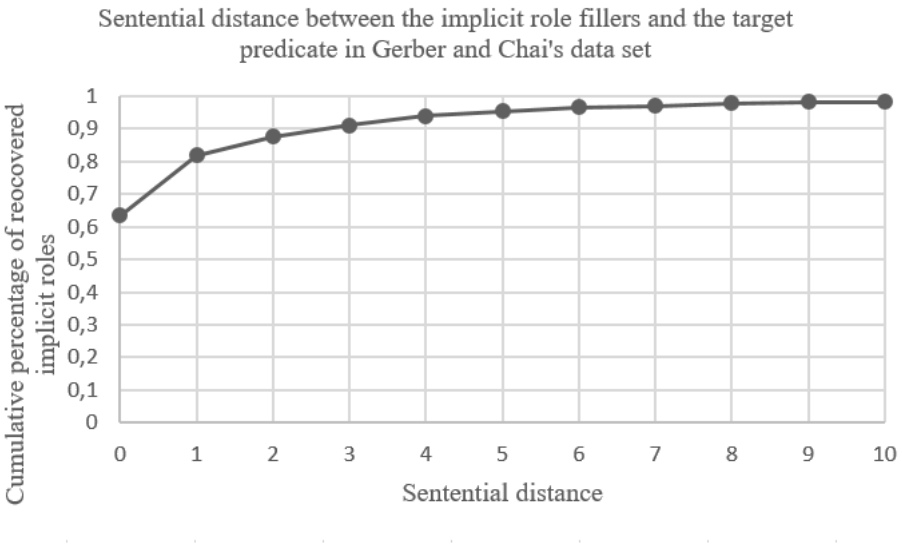

FIGURE 5.2: Sentential distance between the implicit role fillers and the target predicate in GERBERCHAI data set

5.2.3 Moor et al's corpus

Following a similar approach to Gerber and Chai (2010), Moor et al. (2013) developed another corpus for a small number of verbal predicates to complement GERBERCHAI data set which was developed for nominal predicates. They targeted five high-frequency verbs in OntoNotes corpus (Weischedel et al., 2011). They selected OntoNotes as the underlying corpus, because it provides semantic role labels in PropBank annotation style and they could focus on annotation of implicit roles. They annotated a corpus composed of 145136 words which included 215 instances of implicit roles for the selected list of predicates. Similar to GERBERCHAI, they also did not present any pre-defined split for the training/test set.

5.2.3.1 Predicate selection

To select the target predicates, the following criteria were considered (Moor et al., 2013):

- The target predicates can not be light verbs, as they typically involve difficult sense disambiguation.

- The target predicates (and their senses) must be covered in VerbNet, PropBank and FrameNet, according to the Unified Verb Index. This criterion makes the corpus usable in later experiments using any of the VerbNet, FrameNet or PropBank paradigm.

- The target predicates must have considerable number of recoverable implicit roles considering the identification of DNIs in the FrameNet annotation data set. In case little or no such cases were found for the candidate predicate, the predicate was removed from the list.

The final list of target predicates includes *bring*, *give*, *leave*, *pay*, and *put*.

5.2.3.2 Annotation Analysis

Table 5.5 presents the number of explicit and implicit roles in the whole Moor data set.

TABLE 5.5: Number of implicit and explicit semantic roles in Moor data set

Role	# of explicit roles	# of implicit roles
A0	387	74
A1	470	26
A2	255	76
A3	72	27
A4	0	12
Overall	1184	215

As the table shows, the number of recoverable implicit roles for the verbal target predicates is 15% of all realized roles which is closer to what can be seen in SEMEVAL corpus rather than GERBERCHAI.

Figure 5.3 shows the recoverable implicit roles in the context. As shown, the current sentence and its proceeding one include the majority of the implicit roles (70%) and a window of 10 sentences can cover 90% of the instances.

These numbers show that around 82% and 90% of the implicit roles are recoverable from a window of three and 10 sentences, respectively, which shows a closer behaviour to GERBERCHAI than SEMEVAL. Since the texts of this corpus are also taken from the newswire text, the distance pattern existing between the predicate and its implicit roles can be assigned to the this feature, and we can hypothesize that novel texts are more probable to have farther connections between the predicates and their missing pieces of information.

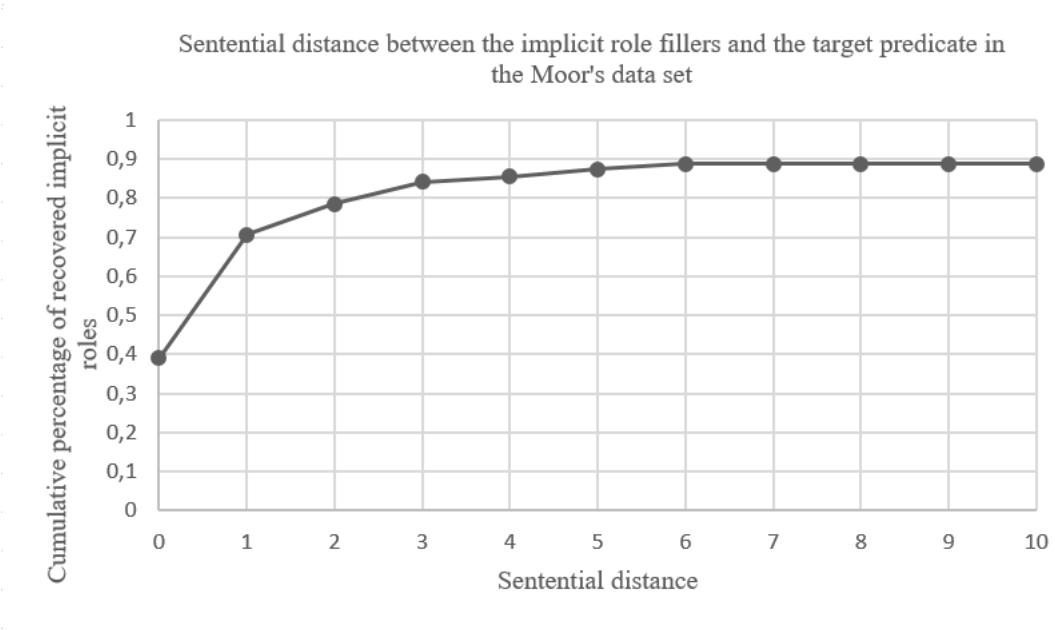

FIGURE 5.3: Sentential distance between the implicit role fillers and the target predicate in Moor data set

5.2.4 Comparing data sets

Comparing the introduced data sets shows that in all cases the number of implicit roles is lower than the number of explicit roles. The difference between the explicit and implicit roles, however, is not alike in all data sets and we can find the following differences between the data sets:

- While implicit roles comprise 15% of the realized roles in SEMEVAL, which includes both verbal and nominal predicates, they also form 15% of the roles in Moor data set, which includes five selected verbal predicates, and 43% in GERBERCHAI which covers 10 selected nominal predicates. Since SEMEVAL was a full-text annotation while GERBERCHAI and Moor were selective anntoations and GERBERCHAI was specifically selected for implicit roles, a direct comparison between these data sets is not possible. However, comparing GERBERCHAI and Moor data set which both tried to select predicates with frequent implicit roles indicates that nominal predicates are more likely to have recoverable implicit roles (43% vs. 15%). This finding is in accordance with Dowty (1989)'s hypothesis that verbs follow an ordered-argument method, i.e. they have a fixed number of arguments and a predicate with fewer than its determined number of arguments does not have any well-defined interpretation while the association between event nouns and their arguments can be better explained by the Neo-Davidsonian representational framework which is motivated by the missing arguments of nouns. This approach

assumes only one argument for the predicate, which is the event, and allows all other arguments optional to present a better analysis of "semantically optional" arguments. This approach behaves nouns and verbs similarly, i.e., missing arguments of verbs can also be explained by Neo-Davidsonian, though nominal predicates' arguments are shown to be more optional than verbs'. The Neo-Davidsonian approach however does not say anything about the regularities of the missing arguments which are discussed in the following.

- Regarding the realization of different roles, one can see that nominal predicates (as in GerberChai) have an A0 (agent role) that is recoverable from the context rather than explicitly mentioned, while verbal predicates (as in Moor data set) have a strong preference on overtly expressed A0. In contrast, verbal predicates remain the A4 role always locally unrealized, whereas the nominal predicates in GerberChai prefer it to be overtly expressed. Nevertheless, due to the low number of instances for A4 role in both data sets, the comparison is not necessarily extensible to all nominal or verbal predicates.

In terms of realization of A1, A2 and A3, the data sets behave differently. While SemEval is closer to Moor data than GerberChai in realization of A2 (in both SemEval and Moor data set, 23% of the A2 roles are implicit roles), its difference from Moor data set and GerberChai is almost the same for A1 and A3.

- Comparing the distance of the non-locally realized roles shows that SemEval data set has the lowest coverage of implicit roles in a window of 10 sentences (83%) and GerberChai has the most coverage (98%). GerberChai also has the most covered implicit roles in a window of one sentence (63%). SemEval and Moor are similar in this regard, with a coverage of around 40%. As the window is enlarged to more sentences, the trend of different data sets is different; GerberChai shows a smooth increase in the number of realized roles, while the slope of the line in SemEval and Moor data set is steeper and changes from around 40% at the beginning to 71% and 78% in a window of 2 sentences. However, the increase tendency of these two data sets is slightly different for larger windows. Changing the window size from 6 to 10 does not add anything to the list of recoverable roles in Moor data set, but in SemEval every increase adds more instances.

Taking all of these observations together, it could be said that the realization distance of the implicit roles in newswire text (i.e. GerberChai and Moor data set) is less than novel texts (SemEval data set) and the nominal predicates are more likely to have

their fillers in their near vicinity. Table 5.6 summarizes the information of different data sets:

TABLE 5.6: Data sets comparison

Data set	# of tokens	# of predicate tokens	# of different predicates	Predicate POS	Type of annotation	Genre
SEMEVAL	17000	1816	526	nominal and verbal	full-text	novel
GERBERCHAI	140536	28210	3491	nominal	selected-nouns	newswire
Moor data set	145136	17720	1849	verbal	selected-verbs	newswire

5.3 Corpus Combination for ISRL

As mentioned in previous chapters, in our experiment we combine two main existing corpora, SEMEVAL corpus and GERBERCHAI, to evaluate the effectiveness of combining some fragmented ISRL corpora in improving a baseline model, and also to increase the amount of annotated data to address the data sparsity problem. In the easiest case, combination means literally concatenating the corpora. However, the differences between these corpora give rise to some new challenges:

- **Annotation Framework**: The first challenge is the different annotation frameworks of the data sets. SEMEVAL is annotated with FrameNet roles, while GERBERCHAI is annotated in the PropBank paradigm. Though semi-automatic conversion schemes are provided by SemEval task organizers in both directions, we decided to adopt the PropBank paradigm. There are two reasons for this selection: firstly, we believe that, in parallel to results on traditional SRL (cf. Chapter 3), PropBank roles should be generally easier to label than FrameNet roles and this effect should be particularly important when facing sparse data problems, as in our case.

Secondly, GERBERCHAI does not provide any annotation for adjective and adverb predicates and also "ArgM" modifier roles which are part of FrameNet annotation of SEMEVAL data set (cf. Section 5.2.1.1). It means that converting GERBERCHAI to FrameNet version results in a very incomplete annotation. For example, (101) is an example sentence from SEMEVAL. In the FrameNet annotation version, the word *enough* is annotated as a predicate and *hardly* is annotated as the DEGREE role, while GERBERCHAI data set annotates neither adverbs nor non-core roles.

(101) For five days this cruel imprisonment continued, with [DEGREE hardly] *enough* [ITEM] food [ENABLED_SITUATION to hold body and soul together].

- **Predicate POS**: SEMEVAL covers both verbal and nominal predicates, while GERBERCHAI contains only nominal predicates. Given the absence of syntactic features in ISRL, we believe that this is not a huge drawback. We will, however, evaluate the results on a per-POS basis to test this assumption.

- **Differences in text type**: SEMEVAL is based on novels which deal with everyday affairs, while GERBERCHAI consists of newswire text focusing on finance and politics and it is well known that the NLP models perform worse when applied across domains and genres. Since this holds for traditional SRL (Carreras and Màrquez, 2005) and is likely to extend to the ISRL, we will experiment with and without domain adaptation methods to test the prerequisites for leaning reasonable generalizations.

5.3.1 A simple ISRL system

We first developed a simple classification-based ISRL system which was used as a baseline to evaluate domain adaptation in ISRL. Similar to most studies on ISRL in SemEval2010, we focused on the "null instantiation" sub-task (cf. Section 5.4.3) which assumes that the explicit roles are given.

Our system performed two steps. In the first step, we determined the implicit roles which are recoverable from the context and in the second step we determined the antecedents of the missing roles. To do the first step, we followed a similar approach to Laparra and Rigau (2012) and extracted the predominant role set (i.e., most frequently realized role set) of each predicate by counting realization patterns in a large corpus, OntoNotes (Hovy et al., 2006). We assumed that all missing roles from the predominant role set are retrievable from the context and must be considered in the second step, that is, if a role occurs frequently when realized, then it is also likely to be retrievable if unrealized. This assumption is warranted because the non-realization is often due to discourse considerations while the conceptual presence/absence of a role is a lexical property of the predicate which is shown by the core vs. non-core frame elements in FrameNet (cf. Chapter 2). Table 5.7 shows an example of this step for predicate "leave". As the table shows, the most frequent role set of the predicate includes A0 and A1 which are described as "entity leaving" and "place, person or thing left" in PropBank definitions, respectively. These roles are the ones which are searched for in the second step of the system.

TABLE 5.7: Role set frequencies for the predicate "leave" in OntoNotes

Role set	Frequency
A0, A1	308
A0, A1, A2	211
A0	110
A1, A2	51
A1	38
A0, A2	12

Though the role set frequencies of the "leave" predicate are high and reliable to determine recoverable semantic roles, low frequencies of some other predicates cause unreliable estimates. For example, "victim" with an overall frequency of 6 has only two role sets, A0 and A1 with a frequency of 1 and A1 with a frequency of 5.

Another bottleneck of this approach is that it neglects the frequency differences between the predominant role set and the other role sets. For example, as Table 5.8 shows, the most frequent role set of "arrest" is A1 with a frequency of 38, while the second most frequent role set, A0 and A1, has a frequency of 35. Though the difference between these role sets is just three instances, it makes the system not to search for unrealized instances of A0.

TABLE 5.8: Role set frequencies for the predicate "arrest" in OntoNotes

Role set	Frequency
A1	38
A0, A1	35
A1, A2	10
A0, A1, A2	2
A0	1

For some frequent predicates, there is a similar problem, with the difference that in these cases the difference between the role sets is significant, however, the number of instances in each role set is fairly high and ignoring them may lead to losing instances of recoverable semantic roles. For example, for the predicate "go", the top role set consists of just A1 which was seen 204 times. But the second top role set, A1 and A4, has also a high frequency of 157 which constitutes 31% of the instances.

Due to the above limitations, this approach led to a recall upper bound of 68.42% on the SEMEVAL test sets. We nevertheless adopted this approach because preliminary experiments with the obvious alternative (looking for all roles) led to worse results (somewhat higher recall, but much lower precision). An informed selection procedure for the appropriate role set of the current predicate instance is a question of future work.

In the second step, we followed a similar approach to Das et al. (2010a) and performed a binary classification. The items to be classified were triples ⟨target predicate, implicit role, candidate realization⟩.,i.e, for a given predicate and a given implicit semantic role, we determined if the given candidate could be a filler or not. The set of candidate fillers was defined as all syntactic constituents from the target predicate's sentence and the two prior sentences which do not fill an explicit role for the target predicate and do not include the target predicate. Selecting a three-sentence-window (current window plus two prior sentences) was performed based on the observation that more than 70% of the implicit role fillers are realized in this window (cf. Section 5.2). To perform the classification, we employed a Naive Bayes classifier[4] that can deal relatively well with sparse data and used 10 features, shown in Table 5.9, which attempt to capture relevant syntacto-semantic and discourse features.

TABLE 5.9: Feature Set (above: syntacto-semantic features; below: discourse features)

Name	Description
Expected role	Set of roles required by the target predicates (based on PropBank and NomBank). This feature serves as a delexicalized target representation
Semantic Type	Semantic type of the candidate realization's head word (WordNet supersenses) or, if pronoun, of the next content word in the coreference chain
Word Frequency	Lemma frequency of the candidate filler's head word
POS	Part of Speech of candidate realization's head word
Constituent type	The constituent type of the candidate filler, e.g. NP, PP, VP, etc.
Distance	Distance between candidate realization and target predicate (in sentences)
Salience	Whether the candidate realization's head word is included in a non-singleton coreference chain
Previous Role	Whether the candidate realization has overtly realized any semantic role in the data set
Same Role	Whether the candidate realization has realized the implicit role as an overt role in the data set
Role Percentage	The percentage with which the candidate realization has realized the implicit role

As can be seen in the table, the *expected role* feature is composed of a set of binary features which determine which roles are required by the target predicate according to PropBank or NomBank. For example, according to PropBank, the required roles for *go.01* and *come.01* are A1 A2, A3 and A4. In the first one, A1 means *entity in motion/goer*, A2 means *extent*, A3 means *start point*, and A4 means *end point, end*

[4]We experimented with other classifiers, such as SVM, but no better result was obtained.

state of A1. In the second one, A1 means *entity in motion/comer* and the rest of the arguments have a similar definition to the predicate *go*. Therefore, the set of *expected roles* for these two target predicates would be 01111 which corresponds to A0-A4. With these features we avoid using the predicate lemma feature which causes a sparse feature space due to low frequencies of many predicates and at the same time capture the similarities between the predicates. In the mentioned cases for example, including the lemmas as a feature causes two different values for the feature, though both predicates have a similar behaviour in terms of their semantic roles.

We should mention that the roles represented by these features do not necessarily correspond to the predominant role set of the predicate (as obtained in the first step). Because in many cases, only one or two roles are remaining in the predominant role set which are not a good representative of the predicate. For example, for the predicate *go*, the predomiant role set includes only A1 while the predicate requires A1-A4. In this manner, the expected role feature provides information that is complementary to the contribution of the first processing step.

The next feature in the table, *semantic type*, is determined using WordNet supersenses which include a list of 44 classes (cf. appendix A). The remaining syntacto-lexical features, i.e. *word frequency*, *POS* and *constituent type* show the frequency of the head word in the document, the part of speech of the head word and the type of the constituent (e.g. NP, VP, etc.). Among the discourse features, *distance* shows the sentential distance between the target predicate and the candidate filler, *salience* determines if the candidate filler occurs in a non-singleton coreference chain, *previous role* determines if the candidate filler has already filled an overt semantic role, *same role* shows if the filler has already filled the given role as an overt role and the last feature, *role percentage* shows in what percentage of its realization, the filler has filled the current missing role.

5.3.2 Domain Adaptation

In our application, SemEval and GerberChai can be understood as two domains. Therefore, to profit from combining the two corpora, we adopt simple but effective feature augmentation (cf. Section 3.2.2.2) as a supervised domain adaptation technique to combine the data sets. In this manner, the model balances global and domain-specific trends against each other. As an example, the distance feature (cf. Table 5.9 for the definition and Section 5.2 for the analysis of the corpora on this feature) is likely to change across domains.

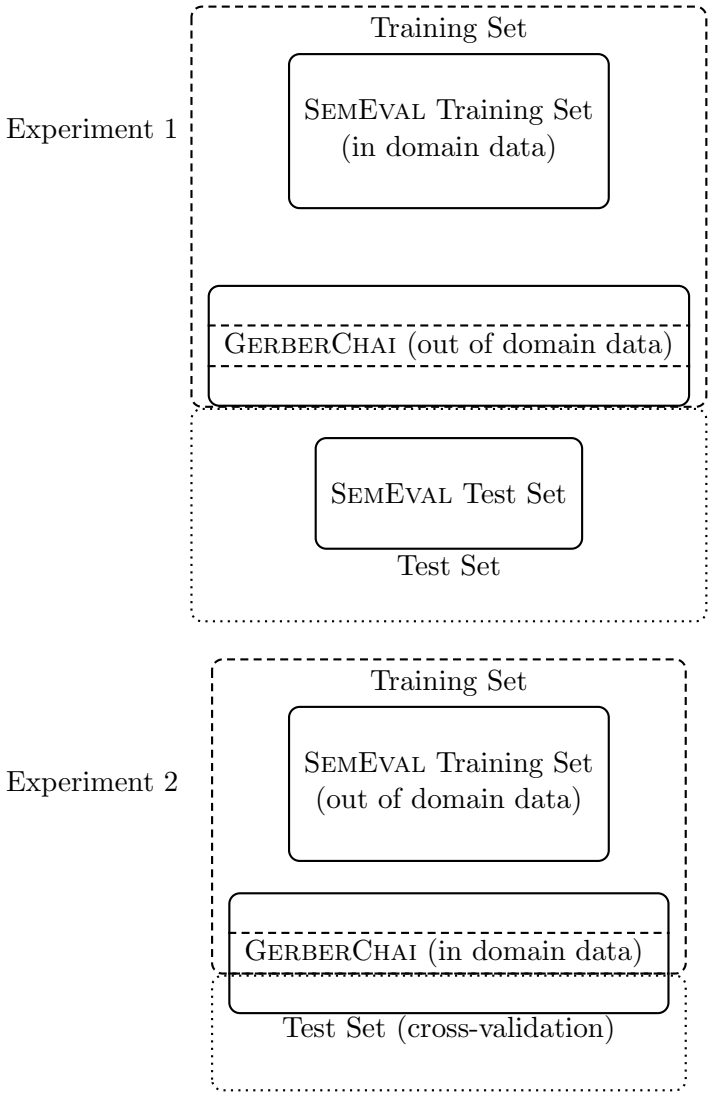

Figure 5.4: Experimental setup

5.4 ISRL for the SemEval corpus

To evaluate the domain adapted system, we performed two main experiments, in the first one we extended the SemEval corpus using the out of domain data from GerberChai and evaluated the results on SemEval. In the second experiment, we swaped the setup, extending the GerberChai with SemEval and evaluating on GerberChai (cf. Figure 5.4). As shown in Figure 5.4, in the second experiment, we split the GerberChai data set and used one third of it as the test set, because there was no pre-defined train/test set for this corpus (cf. Section 5.5).

5.4.1 Experimental setup

We compared four experimental scenarios (cf. Table 5.10): (1) The standard "in-domain" setup which used only SemEval as the training data, as assumed by most studies on the data set. (2) A pure "out-of-domain" setup where we used only GerberChai to train the model. (3) A simple "concatenation" setup where we trained on the union of GerberChai and the SemEval training corpus. (4) The "feature augmentation" setting where the model was trained on the combined corpus with applying the domain adaptation method proposed by Daume III (2007).

5.4.2 Pre-processing

SemEval was pre-parsed with the Collins parser (Collins, 1997). So, we parsed GerberChai with the same parser, ignoring the Penn Treebank gold trees, to have the same level of correctness in terms of parse trees in both corpora. Since all data sets were manually annotated with semantic roles, no overt SRL was necessary. Coreference information, which we required for "Salience", "Previous Role", "Same Role" and "Role Percentage" features was computed with the Stanford CoreNLP tool (Manning et al., 2014). This tool has been evaluated on CoNLL data set (Pradhan et al., 2012) and has obtained an average F-score of 0.63% in identification of coreference chains (Clark and Manning, 2015). At the time of evaluation, this F-score number was higher than the state-of-the-art systems (Clark and Manning, 2015).

5.4.3 Experiment 1: Evaluation on SemEval corpus

The evaluation was carried out using precision, recall and F1-score computed according to the SemEval task guideline. In SemEval task, the task participants could participate in two sub-tasks: full task or null instantiations only. The first one requires identification of both explicit and implicit semantic roles and the second one assumes the explicit role are already given and the focus is on annotation of implicit roles. In both sub-tasks, however, identification of the target predicates and whether the implicit role is recoverable must be carried out by the participants. We evaluated our system against the null-instantiation task.

Since implicit roles can have more than one referent in the context, that is different mentions of the same entity, the coreference chains are also considered during the evaluation. Therefore, if a system links an implicit role with any of the mentions of the correct filler,

it is considered a correct annotation. In addition, as it may be sometimes difficult to determine the exact extent of the correct filler of the implicit role, the evaluation metrics of a correct annotation is the head word of the filler; if the head word of the correct filler is annotated, it is considered a true positive.

5.4.3.1 Evaluation of our model

TABLE 5.10: ISRL evaluation on SEMEVAL test set (PropBank annotation)

Training set	Precision	Recall	F1-score
SEMEVAL training set (in-domain)	0.10	0.20	0.13
GERBERCHAI (out-of-domain)	0.12	0.08	0.10
SEMEVAL training set + GERBERCHAI	0.11	0.19	0.14
SEMEVAL training set + GERBERCHAI, feature augmentation	**0.13**	**0.30**	**0.18**
Laparra and Rigau (2013)	0.12	0.16	0.14

Comparing our results to the state of the art systems was not easily possible, as all prior systems have used the FrameNet annotation framework. Gratefully, Laparra and Rigau shared their prediction files of their system in 2013 (Laparra and Rigau, 2013) with us and we converted their results to the PropBank format using the FrameNet-PropBank mapping provided by the task organizers. Table 5.10 presents the results of the four different settings and Laparra and Rigau (2013) system.

Our baseline system which is trained only on the in-domain data achieves a performance of 0.13 F-score, comparable to Laparra and Rigau (2013)'s 0.14 F-score, but with a different precision-recall tradeoff: our system has P=0.1, R=0.2, while Laparra and Rigau (2013) have a more balanced P=0.12, R=0.16. One reason of this difference is that our system classifies (role,span) labels and it lets multiple constituents to be annotated as the fillers of the same role. The average number of constituents per implicit role in our system was 4.2, while Laparra and Rigau (2013) annotated only one constituent per each given implicit role.

As expected, the pure out-of-domain training does not perform well. Simple data concatenation improves the results slightly, because the original data is still there and the additional data set has helped to gain a little bit of improvement but not a lot. One reason could be that the data is highly biased such that adding more instances from another domain can not help significantly. The feature augmentation method however improves the performance substantially. In this case, there is a major improvement in recall (+10 percentage points) and a smaller improvement in precision compared to the in-domain setting, with a final F1-score of 0.18. The differences between the feature

augmentation model and the baseline were tested for significance with bootstrap resampling (Efron and Tibshirani, 1994) and it was found to be highly significant ($p < 0.01$). We see an improvement of 5% F-Score, despite the differences between the corpora, when feature augmentation is used. Notably, we achieve a high recall, despite the upper bound imposed by the first step (implicit role determination). Our system, like many other ISRL systems traded off recall against precision by restricting the search space. We restricted our search space to the current and two preceding sentences. In addition, we had another filtering step which confined the search space to the predominant role set (cf. Section 5.3.1). The upper bound in recall on the SemEval test set that can be achieved in this setting is 60.1%. More details about the upper bound limitation by the filtering steps are presented in Table 5.11.

Table 5.11: Upper bounds for recall on SemEval test set

Percentage of total implicit roles	Predominant role set	Three-sentence window	Combination of filters
100%	68.42%	81.20%	60.15%

5.4.3.2 Results analysis

To evaluate the effect of each feature in each model, we performed feature ablation for the baseline system (in-domain) and the feature-augmented model. The results showed that discourse features such as Previous Role (cf. Table 5.9) were among the most important features in the feature-augmented model, while they are almost useless in the baseline model. This indicates that discourse-level features particularly profit from the inclusion of out-of-domain data. Because the lexical and semantic features used in the model (e.g. Expected Roles and Semantic Type) do not change a lot across domains and adding more instances does not help in this regard, while discourse features (e.g. Same Role and Previous Role) are more probable to suffer from sparseness and therefore can profit from more instances.

Analysis by amount of out-of-domain data:

Since GerberChai was about ten times larger than the SemEval training set (in terms of the number of implicit roles), we wondered whether the out-of-domain data has overwhelmed the in-domain data. Thus, keeping the SemEval test set for the evaluation, we combined SemEval training set with subsets of GerberChai in increments of 5% of the total number of predicates. The results are shown in Table 5.12.

TABLE 5.12: Results on SEMEVAL test set, training on SEMEVAL training set plus varying amounts of data from GERBERCHAI

% of GERBERCHAI	Precision	Recall	F-score
0	0.10	0.20	0.13
5	0.13	0.29	0.17
10	**0.14**	**0.31**	**0.19**
15	0.13	0.31	0.18
20	0.13	0.30	0.18
100	0.13	0.30	0.18

The results show that almost the complete benefit of the GERBERCHAI data is obtained by adding only 5% of the data, and we can achieve the optimal result by adding 10%. This result is marginally higher than when adding the complete GERBERCHAI but the difference is not significant. The outcome indicates that, in contrast to the proposal by Moor et al. (2013), we do not require many annotations for each predicate, but the results are best when the in-domain and out-of-domain corpora have the right amount of variety. We believe that this result can be understood from an ensemble model perspective. According to Jiang (2008), mixture models, of which Daume III (2007), as a simplified version of Daume III and Marcu (2006) is an example, form a type of ensemble model which consists of three models: general model, source domain and target domain model. Since the source and the target domain are of different distributions, the base model constructed by source domain does not have sufficient performance on the target domain. However, by combining these base models as an ensemble model, we can expect that the final model performs well on the target domain, since diversity among the members of the base models is expected to be advantageous in ensemble learning.

Analysis by Predicate POS:

Since GERBERCHAI contained only nominal predicates, we hypothesized that its inclusion improves results in SEMEVAL specifically for nominal predicates. To test this hypothesis, we evaluated verbal and nominal predicates separately. The results are presented in Table 5.13.

Even though the benefit is somewhat smaller for verbs, there is still a substantial improvement on both types of predicate: +4.1% F1-score for verbs and +5.9% F1-score for nouns. In contrast, studies on traditional SRL have indicated only small (but consistent) improvements for extending training sets with instances of targets with different parts-of-speech (Li et al., 2009).

TABLE 5.13: Evaluation of results on the SEMEVAL test set, by target part of speech

Training set	Verbal			Nominal		
	Precision	Recall	F1-score	Precision	Recall	F1-score
SEMEVAL (in-domain)	0.11	0.20	0.14	0.10	0.21	0.14
GERBERCHAI (out-of-domain)	0.09	0.12	0.10	0.07	0.11	0.09
SEMEVAL + GERBERCHAI	0.11	0.18	0.13	0.11	0.21	0.14
SEMEVAL + GERBERCHAI, feature augmentation	0.13	**0.30**	**0.18**	**0.14**	**0.32**	**0.20**
Laparra and Rigau (2013)	**0.15**	0.20	0.17	0.09	0.11	0.09

One reason could be that ISRL can rely less on syntactic features and must make predictions mostly based on semantic and discourse features, which are more comparable across target parts of speech. Examples below make the point clearer.

(102) SEMEVAL: The wagonette was *paid off* and ordered to return to Coombe Tracy forthwith, while [$_{A0}$ we] started to walk to Merripit House.

(103) GERBERCHAI: His seven-bedroom cedar and brick house outside of Johnstown is up for *sale* to pay for [$_{A0}$ his] lawyers.

(102) and (103) compare a verbal and a nominal predicate with an implicit A0 in the two corpora. In both cases, the correct filler occurs in the same sentence as the predicate, but outside the syntactic domain. While the role realizations are quite different structurally (subject vs. possessive), they are similar in the semantic and discourse levels: both are pronouns referring to agent-like entities and are realized in the immediately following discourse.

In Laparra and Rigau (2013), however, the results on nominal predicates is not as good as verbal predicates. Considering that they consider only a combination of the semantic type and the POS of the candidate filler's head word to identify the correct filler, we can conclude that these features are more effective in identifying verbal predicates' implicit roles and to determine antecedents of nominal predicates' implicit roles the model profits from other features more.

Analysis by Role:

We also evaluated the results by individual semantic roles to assess the effect of the additional data set on identification of each semantic role. The results are provided in Table 5.14.

Our evaluation concentrates on A0 through A2, since A3 and A4 are so infrequent in SEMEVAL that evaluation results are not reliable (cf. Table 5.3). The overall best results

TABLE 5.14: Evaluation of implicit SRL (PropBank roles) on the SEMEVAL test set, by role results

Training set	A0			A1			A2		
	P	R	F1	P	R	F	P	R	F
SEMEVAL train (in-domain)	0.19	0.29	0.23	0.09	0.26	0.13	0.06	0.10	0.07
GERBERCHAI (out-of-domain)	0.19	0.34	0.24	0.03	0.06	0.03	0.00	0.00	0.00
SEMEVAL train + GERBERCHAI, concatenation	0.23	0.34	0.27	0.08	0.22	0.11	0.00	0.00	0.00
SEMEVAL train + GERBERCHAI, feature aug.	**0.24**	**0.42**	**0.31**	**0.11**	**0.37**	**0.17**	0.09	**0.24**	0.13
Laparra and Rigau (2013)	0.21	0.28	0.24	0.10	**0.13**	0.11	0.13	0.19	**0.15**

are seen for A0, followed by A1 and A2. The improvement for combining corpora using feature augmentation correlates with the overall improvement of +8% F1 for A0, +4% for A1, and +6% for A2. Thus, corpus combination seems to benefit all roles. The overall pattern of a major improvement on recall and a minor one on precision is also stable across roles.

A notable observation is the inability of the naive out-of-domain models (the second and third models) to correctly predict any A2 roles. The reason is that for the nominal targets in GERBERCHAI, a large number of A2 roles are incorporated roles, that is, realized by the predicate themselves, while this pattern never occurs in SEMEVAL. Interestingly, the domain adaptation model manages to extract relevant information from GERBERCHAI. Nevertheless, our system can not beat Laparra and Rigau (2013) in terms of A2, though our analysis showed that it has annotated more A2 instances correctly (5 vs. 4 instances). That is, due to the higher number of false positives, the overall performance of our system is not as good as Laparra and Rigau (2013).

Another observation of this analysis is that more frequent roles (A0 and A1) are better predicted than the less frequent one (A2) by all the systems. considering that A2 has a more verb-specific definition, we can conclude that the model requires more instances of A2 to be able to generalize its pattern to cover more instances of the test data.

5.5 Experiment 2: Evaluation on GerberChai corpus

In Experiment 2, we used a combination of GERBERCHAI and the complete SEMEVAL training set for training and evaluated the system on GERBERCHAI. The main question was whether the addition of the much smaller SEMEVAL corpus to GERBERCHAI can improve the performance.

In this experiment, we considered the same four settings as in the first experiment. Since there was no preset train/test split for GerberChai, we split it into three equal-sized parts and the reported numbers are the averages over three cross-validation runs where we always used two thirds for training and one third for testing. The pre-processing steps were performed similar to the first experiment and the evaluation process was also performed as before using the same evaluator of the SemEval task. The only difference with the first experiment was that due to the absence of manually annotated coreference chains in GerberChai data set, only direct matches were considered as true positives.

The upper bound for recall on GerberChai data set (using the same 3-sentence window and predominant role set) was rather low, at 44%. More details about the upper bound filters can be found in Table 5.15. The predominant role set filtration in GerberChai data set loses more roles than SemEval (cf. Table 5.11). This filtering loses all A3 and A4 roles in GerberChai, because most target predicates occur only with one locally realized role, which is mostly A1 or A0. The only predicate with A2 in its predominant role set is *price*, which has its A2 realized as incorporated. The predominant role sets of the predicates can be found in Table 5.16. One possible approach to avoid losing many instances in the predominant role set filteration step in GerberChai is to consider verbal equivalents of the nominal predicates, as well. By applying this approach, we can increase the predominant role set recall upperbound to 59.38%[5]. Nevertheless, it should be noticed that this approach is not applicable to all nominal predicates, but only to the eventive nominal predicates.

Table 5.15: Upper bounds of recall in GerberChai data set

Percentage of total implicit roles	Predominant role set	Three-sentence window	Combination of filtration processes
100%	48.29%	84.89%	44.19%

The results of the experiment are shown in Table 5.17.

The overall patterns are very similar to the first experiment: out-of-domain training works worse than in-domain training, and simple concatenation does not improve over in-domain training. With feature augmentation, however, we see a significant improvement of 8% in precision and recall and 9% in F1-score. The difference is highly significant at $p<0.01$. This confirms the effectiveness of corpus combination, despite the small size of the added SemEval dataset compared to GerberChai.

[5]It should be mentioned that this approach was not used in this dissertation, because this idea was obtained during the final analysis of the experiments at the end of writing the dissertation.

TABLE 5.16: Predominant role set of target predicates in GERBERCHAI data set

Predicate	Predominant role set
Bid	A1
Cost	A1
Fund	A1
Investor	A0
Investment	A0
Loan	A0
Loss	A0
Plan	A0, A1
Price	A1, A2
Sale	A1

TABLE 5.17: ISRL evaluation on GERBERCHAI data set

Training set	Precision	Recall	F1-score
GERBERCHAI training set (in-domain)	0.16	0.10	0.12
SEMEVAL (out-of-domain)	0.11	0.06	0.07
GERBERCHAI + SEMEVAL training set	0.16	0.09	0.11
GERBERCHAI + SEMEVAL training set, feature augmentation	**0.24**	**0.18**	**0.21**
Gerber and Chai (2012)	0.58	0.44	0.50

It is also clear that the results are much worse than those reported in Gerber and Chai (2012). However, the results obtained by Gerber and Chai (2012) are not directly comparable, because their system had focused on some selected nominal predicates from newswire text, while our system was developed for both verbal and nominal predicates from novel genre. In addition, their approach incorporated a large number of detailed linguistic resources (Penn Treebank, Penn Discourse Bank, NomBank, FrameNet) and assumed gold standard information on all levels, while we used, for instance, the Collins parser (Collins, 1997) to parse the sentences and the Stanford CoreNLP tools (Manning et al., 2014) for identifying coreference chains. Because the parse trees of SEMEVAL sentences were provided this way and the gold standard coreference chains of the test set were not provided at all. We therefore see their system as an upper bound rather than a competitor. The results can not also be compared with Laparra and Rigau (2013) since they did not run their system on GERBERCHAI data.

The results of this experiment reinforces the observation that feature augmentation technique acts like an ensemble model with two sub-models for the source and the target domain which performs a weighted averaging and assigns more weight to the target domain model predictions to make the final predictions.

5.5.1 Analysis by role

We did a similar analysis as in the first experiment for the baseline system (in-domain data) and the feature augmented model and evaluated the improvements on individual roles. You can find the results in Table 5.18. We did not perform any other analysis similar to the first experiment because they were not applicable. The analysis by POS was not possible because all the target predicates in GerberChai were nominal predicates and the analysis by amount of additional data set was not applicable because the SemEval was a small corpus consisting of only 122 implicit role instances, i.e. around 1/10 of GerberChai, whose reduction to 10% or 20% (similar to the first experiment) resulted in only 12 or 24 instances.

Table 5.18: Evaluation of ISRL (PropBank roles) on the GerberChai test set, by role results

Training set	A0			A1		
	P	R	F1	P	R	F
GerberChai (in-domain)	0.15	0.10	0.12	0.18	0.23	0.16
SemEval(out-of-domain)	0.10	0.03	0.05	0.05	0.01	0.05
SemEval train + GerberChai	0.15	0.05	0.06	0.16	0.05	0.08
SemEval train + GerberChai, feature aug.	**0.19**	**0.13**	**0.15**	**0.26**	**0.35**	**0.30**

As mentioned earlier, A2-A4 were mostly removed in the pre-processing step. Therefore, we assessed by-role analysis only for A0 and A1: again, we see improvements for both A0 and A1, both regarding precision and recall. We also see that, as expected, the in-domain data performs better than the out-of-domain data. A simple concatenation of the data sets perform slightly better than the out-of-domain data, but outperforms neither the in-domain data nor the feature augmented combination of the corpora. Interestingly, the improvements as well as the performance for A1 exceed those for A0, which is different from SemEval results, where the best results were found for A0.

5.6 Experiment 3: Evaluation of data set size and variety

In the Experiment 1 and 2, we evaluated the effectiveness of adding the additional out-of-domain data and obtained improvements in both experiments, despite the differences between the corpora in terms of target predicates' POS, genre and data set sizes. However, it is still unclear what causes the improvement. Is it simply due to the increased amount of training data, or to the training data becoming more varied? To distinguish between these two hypotheses, Experiment 3 was designed. In this experiment, the total

size of the training set was kept constant and the proportions of the two source corpora, SemEval and GerberChai, were changed in 10% increments, from 100% SemEval (i.e. all data points taken from SemEval) to 100% GerberChai (i.e. all data points taken from GerberChai). The size of the training set was limited by the smaller training set, i.e. SemEval.

We evaluated the feature augmented models on both SemEval and GerberChai test sets. If the improvements seen in Experiment 1 and Experiment 2 are solely due to the larger size of the training sets, it is expected to see the highest performance for the 100% in-domain training set, and decreasing performance with increase of out-of domain data. In contrast, if the variety is the reason of the improvements, it is expected to see the maximum improvement somewhere between the two extremes, at the point where there is enough out-of-domain training data to introduce variety but not enough to overwhelm the in-domain data.

Figure 5.5 and Figure 5.6 show the results on both data sets. On both test sets, we do not see the best result for 100% in-domain data: there is a substantial improvement moving from 100% to 90% in-domain data (from 0.13 to 0.18 F1-score on SemEval and from 0.10 to 0.18 GerberChai). On the SemEval test set, the result for 90%, 80% and 50% are the best results. The F1-score shows minor variation until the 50-50 split and then a mild degradation is seen when the GerberChai training data dominates. This result is consistent with Experiment 1. The changes in precision and recall follow a fairly similar trend, but the recall is more affected with the change of in-domain/out-of-domain composition rather than precision which is almost constant in the whole experiment. This result is also consistent with Experiment 1 where we saw more influence of the additional data set on recall.

On the GerberChai test set, we see a more symmetrical picture, with relatively constant performance for almost all mixtures. We see degradation for the both "pure" (100%) training sets, but still better performance for in-domain than for out-of-domain: F1-score of 0.10 vs. 0.08 when using 100% GerberChai or 100% SemEval. In this setup, the highest F1-score is obtained when 70% or 80% of the data comes from the in-domain data.

Overall, the results are compatible with the second, but not the first hypothesis, that is, the models seem to profit from the combination of different corpora even when this does not involve larger training sets. In other words, it is the *complementarity* of the corpora, rather than the addition of training data, which is responsible for the improvement. This suggests that rather than annotating as many instances as possible, we should

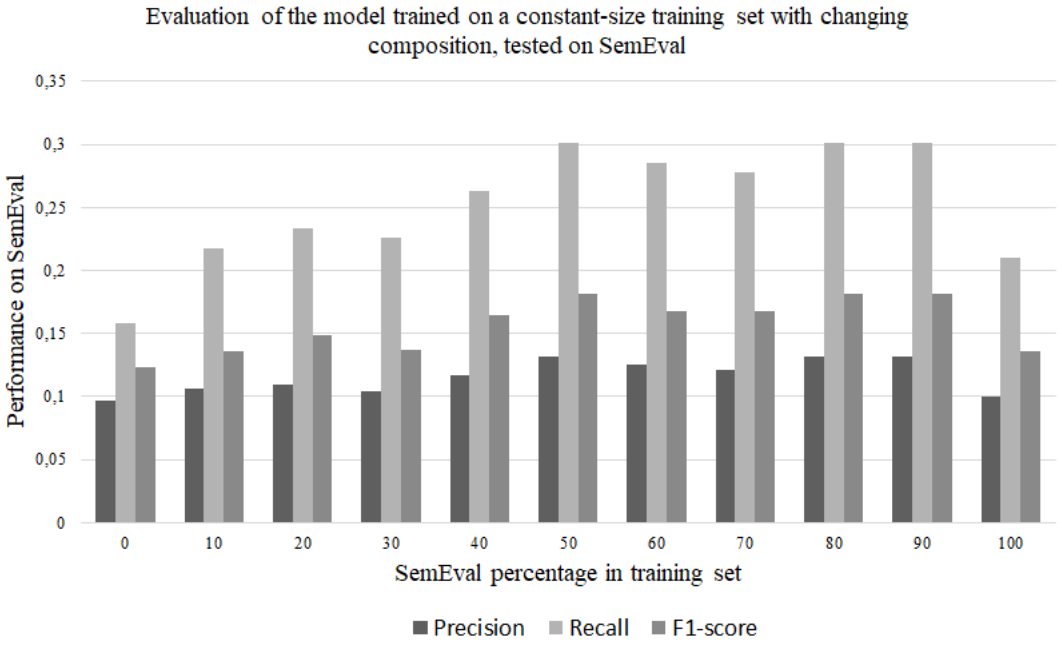

FIGURE 5.5: Evaluation of the model trained on a constant-size training set with changing composition, tested on SEMEVAL data set

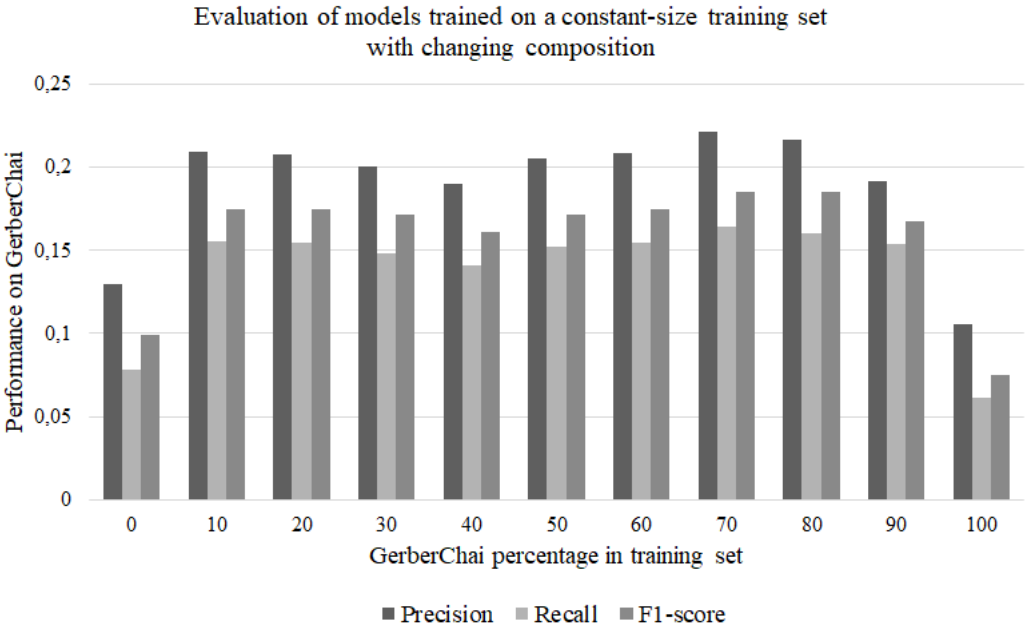

FIGURE 5.6: Evaluation of the model trained on a constant-size training set with changing composition, tested on GERBERCHAI data set

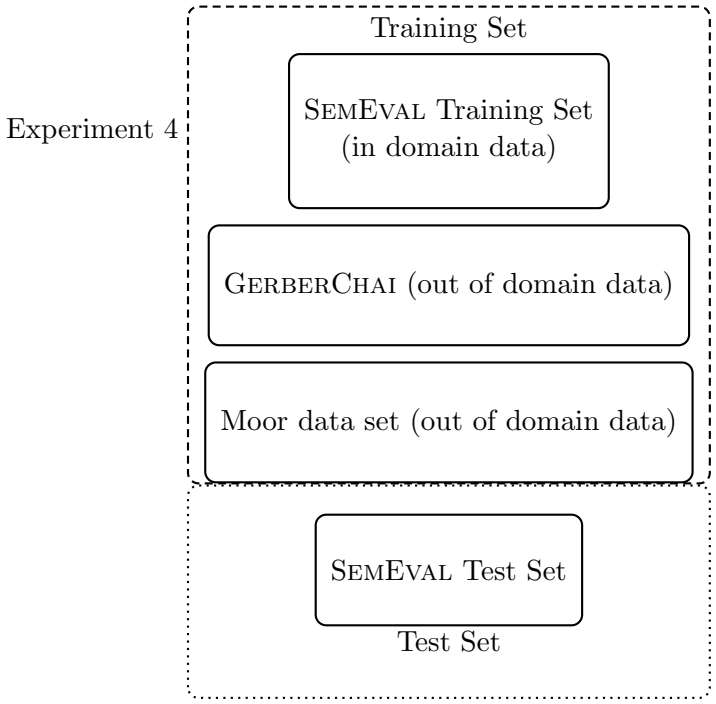

Figure 5.7: Experimental setup

concentrate on annotating informative instances similar to uncertainty sampling in active learning (Lewis and Gale, 1994). The idea of adding instances that are as *varied* as possible can be compared to active learning approaches which notice the diversity of the instances (e.g. Brinker (2003), Zhu et al. (2008) and Yang et al. (2015)), because uncertain data are mostly similar to each other, these studies impose some diversity constraints to make the selected data as diverse as possible.

5.7 Experiment 4: Combining three corpora

In Experiment 1 and 2, we found that adding out-of-domain data can improve the performance on ISRL, when combined with a domain adaptation method (feature augmentation). Experiment 3 collected evidence that this improvement occurs due to the complementarity of the corpora rather than the increase in the size of the training data. In Experiment 4, we evaluate the scalability of the domain adaptation technique in terms of the number of corpora and assess if combining more than two corpora leads to more improvement. This experiment is a follow up experiment to Experiment 3 to evaluate the obtained evidence that the variety of data plays a more important role than the size of the data set. The experimental setup is shown in Figure 5.7.

5.7.1 Evaluation of combining three corpora

In the first part of the experiment, we add the whole Moor data set (cf. Section 5.2.3) to the existing corpora and evaluate the results on the SemEval corpus. To do so, we create four versions of each feature: general version, target version, GerberChai version and Moor version. This experiment can also be seen as an extension of the feature augmentation approach proposed by Daume III (2007) to more than two domains.

To do the experiment, we followed a similar approach to Experiment 1 and used four different settings (cf. Section 5.4.1). Table 5.19 compares the obtained results with other settings.

Table 5.19: ISRL evaluation with combining more than two data sets, tested on SemEval test set

Training set	Precision	Recall	F1-score
SemEval training set (in-domain)	0.10	0.20	0.13
Moor data set (out-of-domain)	0.12	0.26	0.16
SemEval + Moor data set, feature augmentation	**0.13**	**0.30**	**0.18**
SemEval + GerberChai + Moor, feature augmentation	0.12	0.28	0.17
SemEval + GerberChai, feature augmentation	**0.13**	**0.30**	**0.18**

Surprisingly, the pure Moor data set (out-of-domain data) achieves better performance compared to the in-domain data but its combination with SemEval in feature augmentation setting and also its combination with SemEval and GerberChai can improve the results. As can be seen, when only two corpora are combined using feature augmention method, the results show the highest improvements. If all the three corpora are combined, the results still show an improvement but it can not beat the combination of two data sets. This result contradicts the previous results in adding more variety by adding out of domain data but conforms to them in the sense that adding more data does not generally help. We take these results to indicate that when adding out-of-domain data, there is a trade-off between an increase in variety (which is good for the model) and an increase in irrelevant information (which is bad for the model). The best point seems to be mixing a relatively small amount of additional data, which can be achieved by adding a second corpus; adding a third corpus does not lead to further improvements.

To analyze the obtained results, we evaluated feature effectiveness in the out-of-domain model by a feature ablation study (as done for the in-domain model in Section 5.4.3.2). Similar to Section 5.4.3.2, we realized that the discourse features are much more effective in the out-of-domain model than the in-domain trained model. For instance, removing

Previous Role feature (cf. Table 5.9) causes 3% decrease in F1-score in Moor data-trained model and 6% decrease in GerberChai trained model, while it does not change the baseline system.

5.7.2 Comparison with Moor et al. (2013) on the task of implicit role classification

To compare our system with the system developed by Moor et al. (2013), we assumed the same setting as their study, i.e. we assumed that the correct missing roles are given (cf. the second step in Section 5.3.1) and focused on classification of implicit roles. The results can be found in Table 5.20.

Table 5.20: Implicit semantic roles classification, tested on SemEval test set

Training set	Precision	Recall	F1-score
SemEval data set	0.31	0.27	0.29
Moor data set	0.33	0.30	0.31
SemEval + Moor, feature aug. (122 instances each)	0.39	0.37	0.38
SemEval + GerberChai, feature aug. (122 instances each)	0.39	0.37	0.38
SemEval + Moor + GerberChai, feature aug. (122 instances each)	**0.42**	**0.40**	**0.41**
Moor et al. (2013)'s system trained on their annotated corpus + corpus provided by Silberer and Frank (2012)	0.34	0.26	0.30

The results show a considerable improvement in all settings when only the classification of the implicit roles is considered, which shows the significant impact of the recall upper bound imposed by the first step (identification of implicit roles). The improvements for the SemEval trained system and the Moor data trained system compared to the implicit role identification plus classification task are almost similar: +16% and +15% F1-score. That the Moor data set trained system performs better than the in-domain trained system is not surprising as it is in accordance with the first experiment in Section 5.7.1. The results in the first and second line also show that the baseline system is comparable with the system developed by Moor et al. (2013), though with a different proportion of precision and recall. The other settings, however, both outperform Moor et al. (2013) system. Though combining the three corpora did not outperform a combination of two corpora (SemEval and GerberChai) in the previous experiment (doing both implicit role identification and implicit role classification), by skipping the role identification step, it can perform as the best system, with +3% improvement in F1-score compared to the

combination of two corpora. This result makes us reconsider our previous hypothesis (that a concatenation of two corpora is optimal) and clarify that this hypothesis holds in the context of the restricted setup that our experiments considered, i.e. using the predominant role set determination that tends to impose relatively strict upper bounds on recall. This interpretation is further supported by an analysis at the role level that you can find in the following. In a setting where the implicit roles are given, like in the final experiment, it remains to be explored whether the benefit of three corpora carries over to better implicit role determination methods.

Analyzing the annotations showed that the A2 role was more affected than A0 and A1 in the identification step[6], i.e., many A2 roles were not among the roles of predominant role set and were not searched for as an implicit role in the classification step. On the other hand, comparing the annotations by the combination of three corpora in "only classification" setting and "identification and classification" setting showed more improvements in identification of A2 role compared to A0 and A1 which means skipping the first step, lets a higher number of A2 be identified. Since A2 has a more verb-specific meaning (cf. Chapter 2), we assume that it can have a higher variety in different texts. Therefore, when the role identification step is skipped, it profits more than A0 and A1 from adding data sets from different genres which leads to an overall better performance compared to the combination of only two data sets.

5.7.3 Summary

In this chapter, we presented a simple ISRL system and evaluated domain adaptation as an effective technique to profit from increasing the training data by out-of-domain data sets. We assessed the feature augmentation method (Daume III, 2007) and found significant improvements compared with the baseline. It means that feature augmentation can be seen as a cheap but effective method in doing ISRL in which obtaining annotation of more instances by experts is time consuming and expensive.

Our analysis of the results showed that the complementarity of the additional data set plays a more important role than the amount of the data. That is, increasing the amount of the additional data set does not necessarily result in an improved system, but the variety of the data plays the main role. The reason can be explained by the ensemble learning framework. The model trained by feature augmentation can be viewed as an ensemble model which constructs three models, a general model, a first model and a

[6]Since no instances of A3 and A4 were predicted by our developed systems, we did not consider these roles in our analysis.

second model, and makes the final prediction based on these base models. However, due to the differences between the source and the target domain, to make the final predictions, the model assigns higher weights to the target domain examples.

We also evaluated the improvement per POS and found stable improvements across POS, though the additional data set included only nominal predicates, which means that ISRL relies more on semantic features than the syntactic ones. This finding is reasonable, because implicit semantic roles can be realized far from their predicate and have no syntactic relationship with the predicate. We also assessed the improvement per role and realized stable improvements across roles.

To evaluate the scalability of our proposed approach in terms of data set size, we tried adding another corpus to the existing two corpora. The results showed that such a system can not outperform the previous one (with combing two corpora) in a setting where both implicit role identification and implicit role classification must be done by using the predominant role set based method that we used for the first step, though it can still outperform the baseline model. However, if the first step is skipped and it is already known which implicit roles are missing, adding more than two corpora can still be helpful. This improvement, in our case, was due to the fact that many A2 roles were filtered out in the implicit role identification step. However, after removing this step, better results were obtained by combining three corpora, because A2 is defined more verb-dependantly than A0 and A1 and including more varied corpora can affect its identification more than A0 and A1 which can subsequently lead to overall better results.

Part IV

Summary and Conclusion

Chapter 6

Summary and Conclusion

In this chapter, we summarize the contributions of this dissertation and discuss possible future work which can be beneficial for natural language processing.

6.1 Main Contributions

Automatic identification of motion roles. As the first step of our experiments, we followed a knowledge-based approach to determine motion events automatically, using WordNet and FrameNet. In our approach, we firstly extracted motion frames based on their core elements from FrameNet and then extended the list of motion events by hyponymy relationship of WordNet.

Obtaining annotation for implicit semantic roles without expert annotations. In this thesis, we investigated two approaches for obtaining instances of implicit semantic roles and addressed the data sparseness problem of the task. Crowdsourcing was the first approach (cf. Chapter 4) which helped to gain instances of implicit roles in a cheap way from non-experts without providing complicated linguistic definitions of the semantic roles to the annotators. Performing our experiments with two different task setups showed that implicit semantic role annotation requires precise design for the task, because presence of many different predicates and roles in the text can easily distract the annotators from focusing on the target predicate (i.e. the target whose semantic roles are required to be annotated). We showed that in case of a precise design of the experiment, we can obtain data points which are well-agreed among the annotators.

Our findings also showed that the crowdsourcing approach can not be easily scaled up. For annotating FrameNet roles, one needs to define a mapping between roles and

slots/questions for each frame, and for annotating PropBank roles it is even harder as the roles are more abstract and defined differently for each predicate.

Combining corpora from different domains to improve an ISRL system. In the second approach, we utilized two existing corpora which had already been annotated by linguistic experts and combined them to improve a baseline ISRL system. To bridge the dissimilarities between the corpora which were from different domains, we used feature augmentation as a domain adaptation technique (Daume III, 2007). With this study, we showed that the data sparseness problem of ISRL can be addressed by merging the existing data sets, even when they belong to different domains (cf. Chapter 5). This finding can be understood from the ensemble modeling perspective, that is, we can see feature augmented model as an ensemble model which consists of three base models: general, source domain and target domain model. The final model performs better than the individual models on the target domain, because diversity among the members of the base models is advantageous in ensemble learning.

6.2 Possible improvements to the ISRL model

A considerable amount of work is still required to improve ISRL models. Here we propose a number of ways in which the developed models in this thesis could be enhanced:

- ISRL includes two steps: identification of resolvable implicit roles and assigning correct fillers to the identified roles in the first step. In our experiment, we followed Laparra and Rigau (2012)'s approach to perform the first step (cf. Chapter 5). This approach leads to a recall upper bound of around 60%. Our experiments in Section 5.7.2 showed that improving this step can improve the overall results notably. Therefore, one can focus on distinguishing between resolvable implicit roles and non-resolvable ones.

- The finding that complementarity of the extra annotated data set plays a more important role than its size suggests to concentrate more on annotating instances that are as varied as possible, instead of annotating as many instances as possible. This approach recalls the uncertainty sampling in active learning (Lewis and Gale, 1994). Thus, active learning could be considered as a method to increase the amount of helpful instances of implicit roles. This approach has already been used as a domain adaptation technique (e.g. (Rai et al., 2010)), but to the knowledge of the author, no study has focused on ISRL.

- Since there exist not many data sets annotated with implicit semantic roles, one can also try semi-supervised domain adaptation techniques (e.g. (Kumar et al., 2010) which is an extension of the feature augmentation approach used in this dissertation) which do not require labeled data in the target domain.

- Our feature ablation evaluation showed that discourse features play a more important role than syntacto-lexical features. Therefore, one further direction in this field is to focus more on discourse features.

6.3 Future directions

Here are some future directions for research we have found during our study:

Inherent difficulty of ISRL. Our study showed that even with combining multiple corpora and gaining a large number of annotated implicit roles, the results are still not as high as many other NLP tasks. In line with previous studies (e.g. (Roth and Frank, 2013) and (Laparra and Rigau, 2013)), we can say that ISRL is an inherently hard task which requires much more work. Our finding from the crowdsourcing experiment (cf. Chapter 4) that non-experts do not focus on a specific predicate when annotating semantic roles of the given predicate confirms this finding. Our approach in this dissertation could improve the results of a baseline system from 13% F-score to 18% which outperformed state of the art, but still leaves lots of room for improvements.

Discourse-based ISRL. Due to the inherent difficulty of the task, one possible direction of future work could be viewing the task from a global language understanding view and not considering individual predicates independently. In such a framework, temporal relations between sentences, for instance, can be considered as clues to identify implicit semantic roles. Previous studies have shown the helpfulness of semantic role information in identifying temporal relations between sentences (e.g. (Llorens et al., 2010). Therefore, we suggest to evaluate the inverse relation and evaluating the effectiveness of temporal relations in identifying implicit role. We also suggest to consider event relations in the future work. Considering the studies by Chambers and Jurafsky (2009) and Chambers and Jurafsky (2008) who identified event chains in narrative texts, one can consider that the antecedent of a predicate's implicit role is more probable to occur in the same event chain or in another event chain with a similar/related topic.

Considering domain difference. Our experiments with Moor et al. (2013) data set showed that it outperforms the baseline system developed based on the SemEval

training set, when tested on the SemEval test set. Considering that the SemEval training set and test set are both taken from Arthur Conan Doyle's fiction works, and are considered to be from the same domain, this question arises "what is the definition of in-domain and out-of-domain data?". One possible future avenue could be addressing this question to obtain a better definition of "domain" for ISRL.

Scalability of the proposed approach to other languages. The domain adaptation technique which was used in this dissertation can be applied to bridge the differences between any two corpora in any language. Our baseline system, however, used some features (e.g. Delexicalised Predicate and Semantic Type) which require language resources, such as PropBank and WordNet. In addition, in the implicit role identification, we used OntoNotes as a large enough resource to obtain the predominant role sets. Therefore, a direct implementation of our approach in another language is possible only provided that similar resources exist for the target language.

Deep learning as a new approach. Many of the studies on Semantic Role Labeling and also Implicit Semantic Role Labeling have focused on the large annotated corpora, such as FrameNet and PropBank. In recent years, however, some studies have addressed semantic parsing using deep learning (e.g. Grefenstette et al. (2014), Zhou and Xu (2015) and He et al. (2017)). Considering the promising results and improvements obtained by these studies and data sparsity problem of Implicit Semantic Role Labeling, one possible future avenue could be to try a similar approach for Implicit Semantic Role Labeling.

Implicit Semantic Role Labeling in practice. An important question about ISRL which has not been addressed so far, to our knowledge, is the acceptable/useful quality of an ISRL system. Designing end user tasks which can profit from ISRL is an under-researched question, and one could perform a task-based evaluation that could determine what level of performance we need to make ISRL worthwhile for "practical purposes". Spatial role labeling (Kordjamshidi et al., 2010) might be a promising candidate task for evaluating the usefulness of ISRL systems.

Appendix A

WordNet supersenses

Table A.1: WordNet Supersenses

Name	Description
adj.all	all adjective clusters
adj.pert	relational adjectives (pertainyms)
adv.all	all adverbs
noun.Tops	unique beginner for nouns
noun.act	nouns denoting acts or actions
noun.animal	nouns denoting animals
noun.artifact	nouns denoting man-made objects
noun.attribute	nouns denoting attributes of people and objects
noun.body	nouns denoting body parts
noun.cognition	nouns denoting cognitive processes and contents
noun.communication	nouns denoting communicative processes and contents
noun.event	nouns denoting natural events
noun.feeling	nouns denoting feelings and emotions
noun.food	nouns denoting foods and drinks
noun.group	nouns denoting groupings of people or objects
noun.location	nouns denoting spatial position
noun.motive	nouns denoting goals
noun.object	nouns denoting natural objects (not man-made)
noun.person	nouns denoting people
noun.phenomenon	nouns denoting natural phenomena
noun.plant	nouns denoting plants
noun.possession	nouns denoting possession and transfer of possession
noun.process	nouns denoting natural processes
noun.quantity	nouns denoting quantities and units of measure
noun.relation	nouns denoting relations between people or things or ideas
noun.shape	nouns denoting two and three dimensional shapes
noun.state	nouns denoting stable states of affairs
noun.substance	nouns denoting substances
noun.time	nouns denoting time and temporal relations
verb.body	verbs of grooming, dressing and bodily care
verb.change	verbs of size, temperature change, intensifying, etc.
verb.cognition	verbs of thinking, judging, analyzing, doubting
verb.communication	verbs of telling, asking, ordering, singing
verb.competition	verbs of fighting, athletic activities
verb.consumption	verbs of eating and drinking
verb.contact	verbs of touching, hitting, tying, digging
verb.creation	verbs of sewing, baking, painting, performing
verb.emotion	verbs of feeling
verb.motion	verbs of walking, flying, swimming
verb.perception	verbs of seeing, hearing, feeling
verb.possession	verbs of buying, selling, owning
verb.social	verbs of political and social activities and events
verb.stative	verbs of being, having, spatial relations
verb.weather	verbs of raining, snowing, thawing, thundering
adj.ppl	participial adjectives

Bibliography

Agirre, E. and Soroa, A. (2009). Personalizing pagerank for word sense disambiguation. In *Proceedings of the Annual Meeting of the European Chapter of the Association for Computational Linguistics*, pages 33–41. Association for Computational Linguistics.

Baker, C., Ellsworth, M., and Erk, K. (2007). Semeval'07 task 19: frame semantic structure extraction. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 99–104. Association for Computational Linguistics.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics.

Baldewein, U., Erk, K., Padó, S., and Prescher, D. (2004). Semantic role labeling with chunk sequences. In *Proceedings of CoNLL*, pages 98–101. Association for Computational Linguistics.

Banerjee, P. (2013). *Domain adaptation for statistical machine translation of corporate and user-generated content*. PhD thesis, Citeseer.

Blackburn, P. and Bos, J. (2003). Computational semantics. *Theoria: An International Journal for Theory, History and Foundations of Science*, pages 27–45.

Blum, A. and Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. ACM.

Bonial, C., Babko-Malaya, O., Choi, J. D., Hwang, J., and Palmer, M. (2010). Propbank annotation guidelines. *Center for Computational Language and Education Research, CU-Boulder*.

Bonial, C., Bonn, J., Conger, K., Hwang, J. D., and Palmer, M. (2014). Propbank: Semantics of new predicate types. In *Proceedings of LREC*, pages 3013–3019.

Bontcheva, K., Roberts, I., Derczynski, L., and Rout, D. P. (2014). The gate crowdsourcing plugin: Crowdsourcing annotated corpora made easy. In *Proceedings of the Annual Meeting of the European Chapter of the Association for Computational Linguistics*, pages 97–100.

Brinker, K. (2003). Incorporating diversity in active learning with support vector machines. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 59–66.

Burchardt, A., Erk, K., Frank, A., Kowalski, A., Padó, S., and Pinkal, M. (2006). The salsa corpus: a german corpus resource for lexical semantics. In *Proceedings of LREC*, pages 969–974.

Burchardt, A., Frank, A., and Pinkal, M. (2005). Building text meaning representations from contextually related frames–a case study. *Proceedings of IWCS-6*.

Callison-Burch, C. (2009). Fast, cheap, and creative: evaluating translation quality using amazon's mechanical turk. In *Proceedings EMNLP 2009*, pages 286–295. Association for Computational Linguistics.

Carreras, X. and Màrques, L. (2004). Introduction to the conll-2004 shared task: Semantic role labeling. In *Proceedings of CoNLL*, pages 89–97. Association for Computational Linguistics.

Carreras, X. and Màrquez, L. (2004). Introduction to the conll-2004 shared task: Semantic role labeling. In *Proceedings of CoNLL*. Association for Computational Linguistics.

Carreras, X. and Màrquez, L. (2005). Introduction to the conll-2005 shared task: Semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pages 152–164. Association for Computational Linguistics.

Carvalho, A., Dimitrov, S., and Larson, K. (2016). How many crowdsourced workers should a requester hire? *Annals of Mathematics and Artificial Intelligence*, pages 1–28.

Chambers, N. and Jurafsky, D. (2008). Unsupervised learning of narrative event chains. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, volume 94305, pages 789–797.

Chambers, N. and Jurafsky, D. (2009). Unsupervised learning of narrative schemas and their participants. 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 2-Volume 2, pages 602–610. Association for Computational Linguistics.

Che, W., Liu, T., and Li, Y. (2010). Improving semantic role labeling with word sense. In *Proceedings of NAACL HLT 2010*, pages 246–249. Association for Computational Linguistics.

Chen, D., Schneider, N., Das, D., and Smith, N. A. (2010). Semafor: Frame argument resolution with log-linear models. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 264–267. Association for Computational Linguistics.

Christensen, J., Soderland, S., Etzioni, O., et al. (2010). 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.

Ciaramita, M., Attardi, G., Dell'Orletta, F., and Surdeanu, M. (2008). Desrl: A linear-time semantic role labeling system. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 258–262. Association for Computational Linguistics.

Clark, K. and Manning, C. D. (2015). Entity-centric coreference resolution with model stacking. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1405–1415.

Collins, M. (1997). Three generative, lexicalised models for statistical parsing. In *Proceedings of the Annual Meeting of the European Chapter of the Association for Computational Linguistics*, pages 16–23. Association for Computational Linguistics.

Croce, D., Moschitti, A., and Basili, R. (2011). Structured lexical similarity via convolution kernels on dependency trees. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1034–1046. Association for Computational Linguistics.

Dahlmeier, D. and Ng, H. T. (2010). Domain adaptation for semantic role labeling in the biomedical domain. *Bioinformatics*, 26(8):1098–1104.

Das, D., Chen, D., Martins, A. F., Schneider, N., and Smith, N. A. (2014). Frame-semantic parsing. *Computational Linguistics*, 40(1):9–56.

Das, D., Martins, A. F., and Smith, N. A. (2012). An exact dual decomposition algorithm for shallow semantic parsing with constraints. 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 209–217. Association for Computational Linguistics.

Das, D., Schneider, N., Chen, D., and Smith, N. A. (2010a). Probabilistic frame-semantic parsing. In *Proceedings of NAACL HLT 2010*, pages 948–956. Association for Computational Linguistics.

Das, D., Schneider, N., Chen, D., and Smith, N. A. (2010b). Semafor 1.0: A probabilistic frame-semantic parser. *Language Technologies Institute, School of Computer Science, Carnegie Mellon University.*

Das, D. and Smith, N. A. (2011). Semi-supervised frame-semantic parsing for unknown predicates. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1435–1444. Association for Computational Linguistics.

Das, D. and Smith, N. A. (2012). Graph-based lexicon expansion with sparsity-inducing penalties. In *Proceedings of NAACL HLT 2012*, pages 677–687. Association for Computational Linguistics.

Daume III, H. (2007). Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic. Association for Computational Linguistics.

Daumé III, H., Kumar, A., and Saha, A. (2010). Frustratingly easy semi-supervised domain adaptation. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 53–59. Association for Computational Linguistics.

Daume III, H. and Marcu, D. (2006). Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, pages 101–126.

de Salvo Braz, R., Girju, R., Punyakanok, V., Roth, D., and Sammons, M. (2006). An inference model for semantic entailment in natural language. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 261–286. Springer.

Dowty, D. (1991). Thematic proto-roles and argument selection. *language*, pages 547–619.

Dowty, D. R. (1989). On the semantic content of the notion of 'thematic role'. In *Properties, types and meaning*, pages 69–129. Springer.

Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Ellsworth, M., Erk, K., Kingsbury, P., and Padó, S. (2004). Propbank, salsa, and framenet: How design determines product. In *Proceedings of LREC*, pages 17–23.

Erk, K. and Pado, S. (2004). A powerful and versatile xml format for representing role-semantic annotation. In *Proceedings of LREC*, pages 799–802. Citeseer.

Erk, K. and Pado, S. (2006). Shalmaneser-a flexible toolbox for semantic role assignment. In *Proceedings of LREC*, volume 6.

Feizabadi, P. S. and Padó, S. (2012). Automatic identification of motion verbs in wordnet and framenet. In *Empirical Methods in Natural Language Processing*, pages 70–79.

Feizabadi, P. S. and Padó, S. (2014). Crowdsourcing annotation of non-local semantic roles. In *Proceedings of the Annual Meeting of the European Chapter of the Association for Computational Linguistics*, pages 226–230.

Feizabadi, P. S. and Padó, S. (2015). Combining seemingly incompatible corpora for implicit semantic role labeling. In *Proceedings of *SEM at NAACL HLT 2015*, pages 40–50. Association for Computational Linguistics.

Fillmore, C. (1982). Frame semantics. *Linguistics in the morning calm*, pages 111–137.

Fillmore, C. J. (1968). The case for case. *Universals in Linguistic Theory*, pages 1–88.

Fillmore, C. J. (1977). The case for case reopened. *Syntax and semantics*, 8:59–82.

Fillmore, C. J. (1986). Pragmatically controlled zero anaphora. In *Proceedings of the Twelfth Annual Meeting of the Berkeley Linguistics Society*, pages 95–107.

Fillmore, C. J. and Atkins, B. T. (1992). Toward a frame-based lexicon: The semantics of risk and its neighbors. *Frames, fields, and contrasts: New essays in semantic and lexical organization*, pages 75–102.

Fillmore, C. J. and Baker, C. F. (2001). Frame semantics for text understanding. In *Proceedings of WordNet and Other Lexical Resources Workshop, NAACL HLT 2001*. Association for Computational Linguistics.

Finin, T., Murnane, W., Karandikar, A., Keller, N., Martineau, J., and Dredze, M. (2010). Annotating named entities in twitter data with crowdsourcing. In *Proceedings*

of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pages 80–88. Association for Computational Linguistics.

Fleischman, M., Kwon, N., and Hovy, E. (2003). Maximum entropy models for framenet classification. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 49–56. Association for Computational Linguistics.

Foland, W. and Martin, J. (2015). Dependencybased semantic role labeling using convolutional neural networks. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 279–288.

Fossati, M., Giuliano, C., and Tonelli, S. (2013). Outsourcing framenet to the crowd. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 742–747.

Gerber, M. and Chai, J. Y. (2010). Beyond nombank: A study of implicit arguments for nominal predicates. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1583–1592. Association for Computational Linguistics.

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

Gildea, D. and Jurafsky, D. (2002). Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288.

Gorinski, P., Ruppenhofer, J., and Sporleder, C. (2013). Towards weakly supervised resolution of null instantiations. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers*, pages 119–130.

Graesser, A. and Singer, Murray, . T. T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review,*, 101:371–395.

Greenwood, M. A. (2004). Using pertainyms to improve passage retrieval for questions requesting information about a location. In *Proceedings of the SIGIR Workshop on Information Retrieval for Question Answering (IR4QA)*, pages 17–22.

Grefenstette, E., Blunsom, P., de Freitas, N., and Hermann, K. M. (2014). A deep architecture for semantic parsing. In *Proceedings of the ACL 2014 Workshop on Semantic Parsing*, pages 22–27.

Gruber, J. S. (1965). *Studies in lexical relations*. PhD thesis, Massachusetts Institute of Technology.

Hacioglu, K., Pradhan, S., Ward, W., Martin, J. H., and Jurafsky, D. (2004). Semantic role labeling by tagging syntactic chunks. In *Proceedings of CoNLL*, pages 110–113. Association for Computational Linguistics.

Hajič, J., Ciaramita, M., Johansson, R., Kawahara, D., Martí, M. A., Màrquez, L., Meyers, A., Nivre, J., Padó, S., Štěpánek, J., et al. (2009). The conll-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of CoNLL*, pages 1–18. Association for Computational Linguistics.

He, L., Lee, K., Lewis, M., and Zettlemoyer, L. (2017). Deep semantic role labeling: What works and what's next. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 473–483. Association for Computational Linguistics.

Heilman, M. and Smith, N. A. (2010a). Extracting simplified statements for factual question generation. In *Proceedings of QG2010: The Third Workshop on Question Generation*, pages 11–20.

Heilman, M. and Smith, N. A. (2010b). Rating computer-generated questions with mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 35–40. Association for Computational Linguistics.

Hermann, K. M., Das, D., Weston, J., and Ganchev, K. (2014). Semantic frame identification with distributed word representations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1448–1458.

Hofmann, T. and Puzicha, J. (1998). Statistical models for co-occurrence data.

Hong, J. and Baker, C. F. (2011). How good is the crowd at real wsd? In *Proceedings of the 5th linguistic annotation workshop*, pages 30–37. Association for Computational Linguistics.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006). Ontonotes: the 90% solution. In *Proceedings of NAACL HLT 2006*, pages 57–60. Association for Computational Linguistics.

Howald, B. S. (2015). Linguistic spatial classifications of event domains in narratives of crime. *Journal of Spatial Information Science*, (1):75–93.

Huddleston, R. (1970). Some remarks on case-grammar. *Linguistic Inquiry*, 1(4):501–511.

Imamura, K. and Sumita, E. (2016). Multi-domain adaptation for statistical machine translation based on feature augmentation. In *Proceedings of the 12th Conference of the Association for Machine Translation in the Americas*, pages 79–92.

Jackendoff, R. (1972). *Semantic interpretation in generative grammar*. MIT press Cambridge, MA.

Jackendoff, R. (1987). The status of thematic relations in linguistic theory. *Linguistic inquiry*, pages 369–411.

Jackendoff, R. (1990). *Semantic structures*, volume 18. MIT press.

Jiang, J. (2008). A literature survey on domain adaptation of statistical classifiers. In *Technical report,CS Department at Univ. of Illinois at Urbana-Champaign*.

Johansson, R. and Nugues, P. (2007). Lth: semantic structure extraction using non-projective dependency trees. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 227–230. Association for Computational Linguistics.

Johansson, R. and Nugues, P. (2008). Dependency-based syntactic-semantic analysis with propbank and nombank. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 183–187. Association for Computational Linguistics.

Koomen, P., Punyakanok, V., Roth, D., and Yih, W.-t. (2005). Generalized inference with multiple semantic role labeling systems. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pages 181–184. Association for Computational Linguistics.

Kordjamshidi, P., Van Otterlo, M., Moens, M.-F., and Kordjamshidi, P. (2010). From language towards formal spatial calculi. In *Workshop on Computational Models of Spatial Language Interpretation (CoSLI 2010, at Spatial Cognition 2010)*, pages 17–24. Citeseer.

Kouchnir, B. (2004). A memory-based approach for semantic role labeling. In *Proceeding of CoNLL*, pages 118–121. Association for Computational Linguistics.

Kumar, A., Saha, A., and Daume, H. (2010). Co-regularization based semi-supervised domain adaptation. In *Advances in neural information processing systems*, pages 478–486.

Lakoff, G. and Johnson, M. (2008). *Metaphors we live by*. University of Chicago press.

Lambrecht, K. (1984). Formulaicity, frame semantics, and pragmatics in german binomial expressions. *Language*, pages 753–796.

Lang, J. and Lapata, M. (2010). Unsupervised induction of semantic roles. In *Proceedings of NAACL HLT 2010*, pages 939–947. Association for Computational Linguistics.

Lang, J. and Lapata, M. (2014). Similarity-driven semantic role induction via graph partitioning. *Computational Linguistics*, 40(3):633–669.

Laparra, E. and Rigau, G. (2012). Exploiting explicit annotations and semantic types for implicit argument resolution. In *2012 IEEE Sixth International Conference on Semantic Computing*, pages 75–78. IEEE.

Laparra, E. and Rigau, G. (2013). Sources of evidence for implicit argument resolution. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers*, pages 155–166.

Lawson, N., Eustice, K., Perkowitz, M., and Yetisgen-Yildiz, M. (2010). Annotating large email datasets for named entity recognition with mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 71–79. Association for Computational Linguistics.

Leidner, J. L., Sinclair, G., and Webber, B. (2003). Grounding spatial named entities for information extraction and question answering. In *Proceedings of the NAACL HLT 2003 workshop on Analysis of geographic references-Volume 1*, pages 31–38. Association for Computational Linguistics.

Lewis, D. D. and Gale, W. A. (1994). A sequential algorithm for training text classifiers. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 3–12. Springer-Verlag New York, Inc.

Li, J., Zhou, G., Zhao, H., Zhu, Q., and Qian, P. (2009). Improving nominal srl in chinese language with verbal srl information and automatic predicate recognition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pages 1280–1288. Association for Computational Linguistics.

Lim, J.-H., Hwang, Y.-S., Park, S.-Y., and Rim, H.-C. (2004). Semantic role labeling using maximum entropy model. In *Proceedings of CoNLL*, pages 122–125. Association for Computational Linguistics.

Litkowski, K. (2004). Senseval-3 task: Automatic labeling of semantic roles. In *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, volume 1, pages 141–146.

Liu, H. and Singh, P. (2004). Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Llorens, H., Saquete, E., and Navarro, B. (2010). Tipsem (english and spanish): Evaluating crfs and semantic roles in tempeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 284–291. Association for Computational Linguistics.

Loper, E., Yi, S.-T., and Palmer, M. (2007). Combining lexical resources: mapping between propbank and verbnet. In *Proceedings of the 7th International Workshop on Computational Linguistics, Tilburg, the Netherlands*.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Marcus, M., Kim, G., Marcinkiewicz, M. A., MacIntyre, R., Bies, A., Ferguson, M., Katz, K., and Schasberger, B. (1994). The penn treebank: annotating predicate argument structure. In *Proceedings of the workshop on Human Language Technology*, pages 114–119. Association for Computational Linguistics.

Màrquez, L., Carreras, X., Litkowski, K. C., and Stevenson, S. (2008). Semantic role labeling: an introduction to the special issue. *Computational linguistics*, 34(2):145–159.

McCarthy, D., Koeling, R., Weeds, J., and Carroll, J. (2004). Finding predominant word senses in untagged text. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 279. Association for Computational Linguistics.

McClosky, D., Charniak, E., and Johnson, M. (2006). Effective self-training for parsing. In *Proceedings of NAACL HLT 2006*, pages 152–159. Association for Computational Linguistics.

Mellebeek, B., Benavent, F., Grivolla, J., Codina, J., Costa-Jussa, M. R., and Banchs, R. (2010). Opinion mining of spanish customer comments with non-expert annotations on mechanical turk. In *Proceedings of the NAACL HLT 2010 workshop on Creating*

speech and language data with Amazon's mechanical turk, pages 114–121. Association for Computational Linguistics.

Melli, G., Wang, Y., Liu, Y., Kashani, M. M., Shi, Z., Gu, B., Sarkar, A., and Popowich, F. (2005). Description of squash, the sfu question answering summary handler for the duc-2005 summarization task. *safety*, 1:14345754.

Meyers, A. (2007). Annotation guidelines for nombank-noun argument structure for propbank. *Online Publication: http://nlp. cs. nyu. edu/meyers/nombank/nombank-specs-2007. pdf*, 44.

Meyers, A., Reeves, R., Macleod, C., Szekely, R., Zielinska, V., Young, B., and Grishman, R. (2004). The nombank project: An interim report. In *Proceedings of the NAACL HLT 2004 workshop: Frontiers in corpus annotation*, pages 24–31. Association for Computational Linguistics.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Moor, T., Roth, M., and Frank, A. (2013). Predicate-specific annotations for implicit role binding: Corpus annotation, data analysis and evaluation experiments. In *Proceedings of the 10th International Conference on Computational Semantics*, pages 369–375.

Müller, C. and Strube, M. (2006). Multi-level annotation of linguistic data with mmax2. *Corpus technology and language pedagogy: New resources, new tools, new methods*, 3:197–214.

Narayanan, S. and Harabagiu, S. (2004). Question answering based on semantic structures. In *Proceedings of the 20th international conference on Computational Linguistics*, pages 693–701. Association for Computational Linguistics.

Nigam, K. and Ghani, R. (2000). Analyzing the effectiveness and applicability of co-training. In *Proceedings of the ninth international conference on Information and knowledge management*, pages 86–93. ACM.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

Palmer, M. S., Dahl, D. A., Schiffman, R. J., Hirschman, L., Linebarger, M., and Dowding, J. (1986). Recovering implicit information. In *Proceedings of the 24th annual meeting on Association for Computational Linguistics*, pages 10–19. Association for Computational Linguistics.

Park, K.-M., Hwang, Y.-S., and Rim, H.-C. (2004). Two-phase semantic role labeling based on support vector machines. In *Proceedings of CoNLL*, pages 126–129. Association for Computational Linguistics.

Petruck, M. R. (1996). Frame semantics. *Handbook of pragmatics*, pages 1–13.

Petruck, M. R. and Boas, H. C. (2003). All in a day's week. In *Proceedings of the 17th International Congress of Linguists. CD-ROM. Matfyzpress.*

Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., and Zhang, Y. (2012). Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 1–40. Association for Computational Linguistics.

Pradhan, S. S., Ward, W., Hacioglu, K., Martin, J. H., and Jurafsky, D. (2004). Shallow semantic parsing using support vector machines. In *Proceedings of the NAACL HLT 2004*, pages 233–240. Association for Computational Linguistics.

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

Punyakanok, V., Roth, D., and Yih, W.-t. (2008). The importance of syntactic parsing and inference in semantic role labeling. *Computational Linguistics*, 34(2):257–287.

Punyakanok, V., Roth, D., Yih, W.-t., and Zimak, D. (2004). Semantic role labeling via integer linear programming inference. In *Proceedings of the 20th international conference on Computational Linguistics*, pages 1346–1352. Association for Computational Linguistics.

Rai, P., Saha, A., Daumé III, H., and Venkatasubramanian, S. (2010). Domain adaptation meets active learning. In *Proceedings of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing*, pages 27–32. Association for Computational Linguistics.

Riedel, S. and Meza-Ruiz, I. (2008). Collective semantic role labelling with markov logic. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 193–197. Association for Computational Linguistics.

Roth, M. and Frank, A. (2013). Automatically identifying implicit arguments to improve argument linking and coherence modeling. In *Proceedings of *SEM at NAACL HLT 2013*, pages 306–316. Association for Computational Linguistics.

Roth, M. and Lapata, M. (2015). Context-aware frame-semantic role labeling. *Transactions of the Association for Computational Linguistics*, 3:449–460.

Ruppenhofer, J., Ellsworth, M., Petruck, M. R., Johnson, C. R., and Scheffczyk, J. (2006). Framenet ii: Extended theory and practice.

Ruppenhofer, J., Gorinski, P., and Sporleder, C. (2011). In search of missing arguments: A linguistic approach. In *RANLP*, pages 331–338.

Ruppenhofer, J., Sporleder, C., Morante, R., Baker, C., and Palmer, M. (2010). Semeval2010 task 10: Linking events and their participants in discourse. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 45–50. Association for Computational Linguistics.

Sabou, M., Bontcheva, K., Derczynski, L., and Scharl, A. (2014). Corpus annotation through crowdsourcing: Towards best practice guidelines. In *Proceedings of LREC*, pages 859–866.

Schenk, N. and Chiarcos, C. (2016). Unsupervised learning of prototypical fillers for implicit semantic role labeling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1473–1479.

Shi, L. and Mihalcea, R. (2005). Putting pieces together: Combining framenet, verbnet and wordnet for robust semantic parsing. In *Computational linguistics and intelligent text processing*, pages 100–111. Springer.

Silberer, C. and Frank, A. (2012). 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.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. (2008). Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 254–263. Association for Computational Linguistics.

Srikumar, V. and Roth, D. (2011). 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.

Starosta, S. (1988). *The case for lexicase: an outline of lexicase grammatical theory.* Pinter Pub Limited.

Steedman, M., Osborne, M., Sarkar, A., Clark, S., Hwa, R., Hockenmaier, J., Ruhlen, P., Baker, S., and Crim, J. (2003). Bootstrapping statistical parsers from small datasets. In *Proceedings of the Annual Meeting of the European Chapter of the Association for Computational Linguistics*, pages 331–338. Association for Computational Linguistics.

Surdeanu, M., Harabagiu, S., Williams, J., and Aarseth, P. (2003). Using predicate-argument structures for information extraction. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 8–15. Association for Computational Linguistics.

Surdeanu, M., Johansson, R., Meyers, A., Màrquez, L., and Nivre, J. (2008). The conll-2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 159–177. Association for Computational Linguistics.

Swier, R. S. and Stevenson, S. (2004). Unsupervised semantic role labelling. In *Proceedings of EMNLP 2004*, pages 95–102. Association for Computational Linguistics.

Täckström, O., Ganchev, K., and Das, D. (2015). Efficient inference and structured learning for semantic role labeling. *Transactions of the Association for Computational Linguistics*, 3:29–41.

Titov, I. and Klementiev, A. (2012a). A bayesian approach to unsupervised semantic role induction. In *Proceedings of the Annual Meeting of the European Chapter of the Association for Computational Linguistics*, pages 12–22. Association for Computational Linguistics.

Titov, I. and Klementiev, A. (2012b). Crosslingual induction of semantic roles. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 647–656. Association for Computational Linguistics.

Tonelli, S. and Delmonte, R. (2010). Venses++: Adapting a deep semantic processing system to the identification of null instantiations. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 296–299. Association for Computational Linguistics.

Tonelli, S. and Delmonte, R. (2011). Desperately seeking implicit arguments in text. In *Proceedings of the ACL 2011 workshop on relational models of semantics*, pages 54–62. Association for Computational Linguistics.

van den Bosch, A., Canisius, S., Daelemans, W., Hendrickx, I., and Sang, E. T. K. (2004). Memory-based semantic role labeling: Optimizing features, algorithm, and output. In *Proceedings of CoNLL*, pages 285–292. Association for Computational Linguistics.

Wagner, A. (2004). *Learning thematic role relations for lexical semantic nets.* PhD thesis, Universität Tübingen.

Weischedel, R., Hovy, E., Marcus, M., Palmer, M., Belvin, R., Pradhan, S., Ramshaw, L., and Xue, N. (2011). Ontonotes: A large training corpus for enhanced processing. *Handbook of Natural Language Processing and Machine Translation. Springer.*

Welinder, P. and Perona, P. (2010). Online crowdsourcing: rating annotators and obtaining cost-effective labels. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (ACVHL)*, pages 25–32.

Whittemore, G., Macpherson, M., and Carlson, G. (1991). Event-building through role-filling and anaphora resolution. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, pages 17–24. Association for Computational Linguistics.

Woodsend, K. and Lapata, M. (2015). Distributed representations for unsupervised semantic role labeling. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2482–2491, Lisbon, Portugal. Association for Computational Linguistics.

Yakushiji, A., Miyao, Y., Tateisi, Y., and Tsujii, J. (2005). Biomedical information extraction with predicate-argument structure patterns. In *Proceedings of the first International Symposium on Semantic Mining in Biomedicine (SMBM), Hinxton, Cambridgeshire, UK, April*, pages 93–96.

Yan, S. and Wan, X. (2014). Srrank: leveraging semantic roles for extractive multi-document summarization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(12):2048–2058.

Yang, Y., Ma, Z., Nie, F., Chang, X., and Hauptmann, A. G. (2015). Multi-class active learning by uncertainty sampling with diversity maximization. *International Journal of Computer Vision*, 113(2):113–127.

Yi, S.-T., Loper, E., and Palmer, M. (2007). Can semantic roles generalize across genres? In *Proceedings of the NAACL HLT 2007*, pages 548–555. Association for Computational Linguistics.

Zapirain, B., Agirre, E., Màrquez, L., and Surdeanu, M. (2013). Selectional preferences for semantic role classification. *Computational Linguistics*, 39(3):631–663.

Zhou, J. and Xu, W. (2015). End-to-end learning of semantic role labeling using recurrent neural networks. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1127–1137.

Zhu, J., Wang, H., Yao, T., and Tsou, B. K. (2008). Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 1137–1144. Association for Computational Linguistics.