

Knowledge-based Supervision for Domain-adaptive Semantic Role Labeling



Vom Fachbereich Informatik
der Technischen Universität Darmstadt
genehmigte

Dissertation

zur Erlangung des akademischen Grades Dr. rer. nat.

vorgelegt von
Silvana Hartmann, Dipl. Ling.
geboren in Darmstadt

Tag der Einreichung: 3. August 2016
Tag der Disputation: 30. September 2016

Referenten: Prof. Dr. Iryna Gurevych, Darmstadt
Prof. Martha Palmer, Ph.D., Boulder
Prof. Dr. Simone Paolo Ponzetto, Mannheim

Darmstädter Dissertationen 2017
D17

Please cite this document as

URN: [urn:nbn:de:tuda-tuprints-67700](https://nbn-resolving.org/urn:nbn:de:tuda-tuprints-67700)

URL: <http://tuprints.ulb.tu-darmstadt.de/id/eprint/6770>

This document is provided by tuprints,
E-Publishing-Service of the TU Darmstadt

<http://tuprints.ulb.tu-darmstadt.de>

tuprints@ulb.tu-darmstadt.de



This work is published under the following Creative Commons license:
Attribution – Non Commercial – No Derivative Works 4.0 International

<https://creativecommons.org/licenses/by-nc-nd/4.0/>

Abstract

Semantic role labeling (SRL) is a method for the semantic analysis of texts that adds a level of semantic abstraction on top of syntactic analysis, for instance adding semantic role labels like *Agent* on top of syntactic functions like *Subject*. SRL has been shown to benefit various natural language processing applications such as question answering, information extraction, and summarization.

Automatic SRL systems are typically based on a predefined model of semantic predicate argument structure incorporated in lexical knowledge bases like PropBank or FrameNet. They are trained using supervised or semi-supervised machine learning methods using training data labeled with predicate (word sense) and role labels. Even state-of-the-art systems based on deep learning still rely on a labeled training set. However, despite the success in an experimental setting, the real-world application of SRL methods is still prohibited by severe coverage problems (lexicon coverage problem) and lack of domain-relevant training data for training supervised systems (domain adaptation problem). These issues apply to English, but are even more severe for other languages, for which only small resources exist.

The goal of this thesis is to develop knowledge-based methods to improve lexicon coverage and training data coverage for SRL. We use linked lexical knowledge bases to extend the lexicon coverage and as a basis for automatic training data generation across languages and domains. Links between lexical resources have already been previously used to address this problem, but the linkings have not been explored and applied at a large scale and the resulting generated training data only contained predicate (word sense) labels, but no role labels. To create predicate and role labels, corpus-based methods have been used. These rely on the existence of labeled training data as sources for label transfer to unlabeled corpora. For certain languages, like German or Spanish, several lexical knowledge bases, but only small amounts of labeled training data exist. For such languages, knowledge-based methods promise greater improvements.

In our experiments, we target FrameNet, a lexical-semantic resource with a strong focus on semantic abstraction and generalization, but the methods developed in this thesis can be extended to other models of predicate argument structure, like VerbNet and PropBank. This

work presents the first knowledge-based method for training data generation that covers both predicate and role labels. The main contributions of this work are the following:

- First, we lay the foundations for the improvement of FrameNet coverage through the development of standardized models of interoperability for semantic knowledge bases like FrameNet, specifically modeling FrameNet in the standard-compliant lexicon model UBY-LMF, and through modeling links between lexical knowledge bases on the level of predicate argument structure, i.e., the level of frames and roles.
- Second, we develop a novel method for creating a FrameNet in any language based on the automatic alignment of FrameNet and Wiktionary. The method is evaluated on the example of German, effectively building a larger FrameNet knowledge base to improve FrameNet semantic role labeling for English and German.
- Third, we present the application of DistantSRL, a novel knowledge-based distant supervision method for the creation of frame- and role-labeled training data, which we evaluate for English and German. We find that the resulting training data are of high quality and complementary to the manually labeled data, i.e., the FrameNet fulltext corpus and SALSA.
- Fourth, we assess the domain generalization capabilities of open-source FrameNet SRL and analyze how DistantSRL can contribute to the adaptation of FrameNet semantic role labeling to new domains, including the domain of user-generated discourse. Therefore, we create a new, substantially-sized test dataset based on English texts from the community question-and-answer forum *Yahoo! Answers*. We find that domain adaptation is required for the predicate labeling step of FrameNet SRL. Our experiments include a comparison of DistantSRL to alternative methods for training data generation for English, e.g., paraphrasing and monolingual annotation projection. We find that our automatically generated training data contribute to best results on a range of out-of-domain test sets.

Finally, we summarize the main findings of our work and discuss open research questions that result from our work. Results show that our method creates high-quality training data that complement the available FrameNet data for English and German. However, our automatically generated data is noisy and may require methods for training SRL systems that better deal with the large and noisy generated training data, for instance by using additional domain adaptation methods.

We point out possible directions and recommendations for future work, e.g., the automatic linking of semantic knowledge bases on the level of predicate argument structure to improve the coverage of our automatically labeled data, and the further integration of FrameNet-like resources with large semantic knowledge bases like Wikidata with the goal to extend the ontology coverage of FrameNet.

Zusammenfassung

Die automatische Annotation semantischer Rollen (*Semantic Role Labeling*, kurz *SRL*) ist eine Methode der automatischen Textanalyse, die auf der syntaktischen Analyse aufbaut und syntaktische Argumente um Annotationen ihrer *semantischen* Funktion ergänzt.

Die syntaktische Funktion *Subjekt* erhält so beispielsweise die semantische Funktion, oder semantische Rolle, *Agent*. Frühere Arbeiten zeigen, dass Semantic Role Labeling eingesetzt werden kann um verschiedene Anwendungen, die semantische Informationen voraussetzen, zu verbessern. Beispiele sind das automatische Beantworten von Fragen (*Question answering*), die Informationsextraktion (*Information extraction*) oder die automatische Textzusammenfassung (*Summarization*).

Systeme für die automatische Rollen-Annotation nutzen üblicherweise ein theoretisches Modell semantischer Prädikat-Argument-Struktur, das in lexikalischen Wissensbasen wie PropBank oder FrameNet implementiert ist. Diese Modelle weisen semantischen Prädikaten, zumeist Verben, eine Lesartenannotation (*Word Sense*) zu, und annotieren (oft abhängig von der Lesart) syntaktische Argumente der Prädikate mit semantischen Rollen.

Überwachte oder teilüberwachte Verfahren des Maschinellen Lernens werden auf entsprechend annotierten Trainingsdaten angewendet, um automatische Systeme zur Annotation der Prädikat-Argument-Strukturen zu trainieren. Auch Systeme, die dem neuesten Stand der Forschung entsprechend *Deep Learning* einsetzen, benötigen annotierte Trainingsdaten. Diese üblicherweise von Experten manuell annotierten Datensätze zu produzieren ist sehr aufwändig. Die mangelnde Abdeckung der Vielfalt natürlicher Sprache durch die Trainingskorpora (mangelnde Lexikonabdeckung) ist ein Grund dafür, dass Systeme für die automatische Annotation semantischer Rollen zwar in Laborexperimenten erfolgreich sind, in praktischen Anwendungen jedoch noch nicht umfassend eingesetzt werden können. Ein weiterer Grund ist der Mangel an Trainingsdaten für verschiedene Textarten oder Genres, auch Domänen genannt, denn trainierte Systeme müssen auf neue Genres, für die sie eingesetzt werden sollen, angepasst werden (Domänenadaption). Diese beiden Probleme bestehen für das Englische, sind jedoch noch stärker ausgeprägt für andere Sprachen,

für die es nur wenige, kleine Ressourcen mit semantischen Rollen, also lexikalische Wissensbasen und annotierte Korpora, gibt.

Das Forschungsziel dieser Arbeit ist die Entwicklung wissensbasierter Methoden, mit denen die Lexikonabdeckung und Abdeckung mit Trainingsdaten für die automatische Annotation semantischer Rollen verbessert werden kann, sowohl für neue Sprachen als auch für neue Genres. Die Verlinkung lexikalischer Wissensbasen auf der Ebene von *Word Sense* und semantischer Prädikat-Argument-Struktur dient als Grundlage für die automatische Generierung von Trainingsdaten mit Lesarten und semantischen Rollen für verschiedene Sprachen und Genres. Verlinkte lexikalische Wissensbasen wurden bereits in früheren Arbeiten mit dem Ziel eingesetzt, die Lexikonabdeckung zu erhöhen. Diese Arbeiten nutzen jedoch nicht die umfassende Verlinkung *mehrerer* Wissensbasen. Zudem erstellen sie zwar neue Lesartenannotationen, nicht jedoch vollständige Prädikat-Argument-Strukturen, die eine schwierigere Aufgabe darstellen. In dieser Arbeit stellen wir die erste Methode vor, die wissensbasiert Trainingsdaten mit vollständigen Prädikat-Argument-Strukturen annotiert.

Eine Alternative zu wissensbasierten Methoden sind korpusbasierte Methoden. Diese übertragen Annotationen von vorhandenen annotierten Trainingsdaten auf noch unannotierte Korpora und setzen daher die Existenz annotierter Korpora mit guter Abdeckung bereits voraus. Für einige Sprachen, wie Deutsch oder Spanisch, stehen zwar umfangreiche lexikalische Wissensbasen zur Verfügung, jedoch kaum annotierte Korpora. Für solche Sprachen bieten wissensbasierte Methoden daher essentielle Vorteile.

Im Zentrum dieser Arbeit steht die Wissensbasis FrameNet und ihr Modell der Prädikat-Argument-Struktur. FrameNet implementiert ein feinkörniges semantisches Modell mit einem hohen Grad semantischer Abstraktion, das es für Anwendungen im Bereich Sprachverstehen interessant macht. Die Methoden, die in dieser Arbeit vorgestellt werden, können jedoch auch auf weitere Modelle semantischer Prädikat-Argument-Struktur angewendet werden, beispielsweise VerbNet oder PropBank. Die wichtigsten Forschungsbeiträge dieser Arbeit sind:

1. Erstellung der notwendigen Voraussetzungen für die Verbesserung der Lexikonabdeckung: Wir präsentieren ein standardkonformes Modell lexikalischer Interoperabilität für Wissensbasen, die wie FrameNet semantische Prädikat-Argument-Struktur modellieren. Konkret wird ein Modell von FrameNet in UBY-LMF entwickelt. Zudem ergänzen wir das UBY-LMF Modell um eine Modellierung von Links zwischen Wissensbasen auf der Ebene der Prädikat-Argument-Struktur.
2. FrameNet für neue Sprachen: Wir präsentieren eine neue Methode für die Übersetzung von FrameNet in beliebige Sprachen aus Wiktionary. Die Methode basiert auf einer automatischen Verlinkung von FrameNet und Wiktionary und nutzt Wiktionary als Interlingua. Wir wenden die Methode exemplarisch für das Deutsche an und erstellen eine neue Wissensbasis mit FrameNet-Informationen für das Deutsche.

3. Wissensbasierte Trainingsdatengenerierung für FrameNet Semantic Role Labeling: Wir verwenden DistantSRL, eine neue Methode zur wissensbasierten Trainingsdatengenerierung für FrameNet-Prädikate und -Rollen, um große Mengen neuer Trainingsdaten für englisches und deutsches FrameNet-SRL zu generieren. Unsere Experimentelle Auswertung zeigt, dass die automatisch erstellten Trainingsdaten hohe Qualität aufweisen und bereits vorhandene Trainingsdaten ergänzen.
4. Experimente zur Domänenadaption: Wir untersuchen die Eigenschaften eines open-source Systems für FrameNet Semantic Role Labeling im Kontext der Domänenadaption für das Englische. Bisher werden FrameNet-basierte Systeme, auch mangels entsprechender Testdatensätze, ausschließlich domänenintern evaluiert. Aus diesem Grund entwickeln wir einen neuen Testdatensatz basierend auf nutzergenerierten Texten aus dem Frage-und-Antwort Forum *Yahoo! Answers*. Anhand dieses Datensatzes, sowie weiterer Datensätze, stellen wir fest, dass das Problem der Domänenadaption auf der Ebene der Lesartenannotation auftritt, die der Rollen-Annotation vorangeht. Wir vergleichen in unseren Experimenten DistantSRL mit korpusbasierten Methoden zur Trainingsdatengenerierung und deren Einfluss auf die Domänen-Generalisierung. Im Vergleich mit korpusbasierten Methoden zur Annotationsprojektion und Methoden, die Paraphrasen verwenden, erzielt DistantSRL die besten Ergebnisse auf domänenfremden Testdatensätzen der englischen Sprache.

Die Ergebnisse unserer Experimente zeigen, dass unsere Methode der wissensbasierten Trainingsdatengenerierung zur Entwicklung von SRL Systemen für neue Sprachen und neue Domänen beitragen kann. Weitere Forschungsarbeit ist jedoch erforderlich, um die vorhandenen Trainingsmethoden besser auf bestimmte Eigenschaften der wissensbasiert generierten Daten anzupassen, d.h. das Training auf großen Datenmengen automatisch erstellter Annotationen zu verbessern.

Die Arbeit schließt mit einer Zusammenfassung der Ergebnisse und einem Ausblick auf offene Fragen und weitere Forschungsideen, die sich aus unserer Arbeit ergeben, beispielsweise die automatische Verlinkung auf der Ebene der Prädikat-Argument-Struktur, sowie weitere Integration von FrameNet mit großen semantischen Wissensbasen wie Wikidata mit dem Ziel, die Abdeckung von Weltwissen in FrameNet weiter zu vergrößern.

Acknowledgments

Many people contributed directly or indirectly to the creation of this thesis, and I gratefully acknowledge their contributions. First and foremost, I would like to thank my advisor Iryna Gurevych for giving me the opportunity to pursue my doctoral studies at UKP Lab and explore the topic of semantic role labeling. I am grateful for her patience and support during my time at UKP Lab, and for her scientific guidance that substantially shaped this work. I would also like to thank my committee members, Martha Palmer and Simone Paolo Ponzetto, for finding the time to review my dissertation. This work has been funded by the German Research Foundation (DFG) through the projects QAEL (grant № GU 798/3-1), InCoRe (grant № GU 798/9-1), and AIPHES (grant № GRK 1994/1).

I would like to thank the co-authors of my publications for the fruitful collaboration and discussions, in particular those with Judith Eckle-Kohler and Ilia Kuznetsov on semantic role labeling, and with Christian Meyer and the UBY team on lexical resources. I am also grateful to Martha Palmer for her helpful feedback and encouragement, and for being my host when I visited CU Boulder on a research grant by the Fulbright Commission.

I am glad to have been part of a vibrant research lab, and I am grateful to all my colleagues for inspiring discussions and moral support during lunch and coffee breaks, especially Torsten Zesch, György Szarvas, Emily Jamison, Nicolai Erbs, and Richard Eckhart de Castilho. I wholeheartedly thank Christian Meyer, Emily Jamison, Ilia Kuznetsov, Judith Eckle-Kohler, Lisa Beinborn, and Tristan Miller for their thoughtful reviews of my thesis. I also thank Orin Hargraves for sharing his experience on annotation studies, Anette Frank for insights on SALSA, and the student assistants and annotators who contributed to my work. Moreover, I would like to thank the scientific community that has shared datasets and software, and answered many questions, including, but not limited to Collin Baker, Hagen Fürstenau, Mihai Surdeanu, Meghana Kshirsagar, Nancy Ide, and Nathan Schneider.

I am grateful to my parents who encouraged me to follow my interests and pursue an academic career, and to my friends and family for providing support and much-needed distraction on the journey leading to this thesis. My deepest gratitude goes to Dorian. Thank you for accompanying me on this ride, I couldn't have done it without you.

Contents

1	Introduction	1
1.1	Semantic Role Labeling	2
1.2	Knowledge Bases for Semantic Role Labeling	5
1.3	Training Data Generation for Semantic Role Labeling	7
1.4	Domain Adaptation of Semantic Role Labeling	9
1.5	Research Questions	11
1.6	Approach	12
1.7	Contributions and Findings	13
1.8	Publication Record	15
1.9	Thesis Outline	16
1.10	List of Abbreviations	18
2	Integrating Semantic Knowledge Bases	19
2.1	Motivation: Extending FrameNet	20
2.1.1	FrameNet	21
2.1.2	English Coverage	24
2.1.3	FrameNet for Languages other than English	26
2.2	Lexical-semantic Knowledge Bases	28
2.2.1	VerbNet	28
2.2.2	PropBank	32
2.2.3	SALSA	34
2.2.4	WordNet	36
2.2.5	GermaNet	38
2.2.6	Wiktionary	39
2.3	Linking Semantic Knowledge Bases	42
2.3.1	Sense-level Alignments	44
2.3.2	Predicate Argument Structure Alignments	47
2.4	Extending and Translating FrameNet using Wiktionary as Interlingua	51

2.4.1	Method Overview	52
2.4.2	Related Work	53
2.5	Creating the FrameNet – Wiktionary Alignment	55
2.5.1	Automatic FrameNet – Wiktionary Alignment	55
2.5.2	Resulting Resource FNWKxx	60
2.6	Translating FrameNet to German via Wiktionary	64
2.6.1	Disambiguating German Lexical Entries	65
2.6.2	Resulting Lexical Knowledge Base FNWKde	66
2.6.3	Discussion: a Multilingual FrameNet based on FNWKxx	68
2.7	Standardizing Semantic Knowledge Bases	71
2.7.1	Modeling Semantic Knowledge Bases in UBY-LMF	73
2.7.2	Modeling Resource Links	78
2.8	A Linked Lexical Knowledge Base for FrameNet	80
2.9	Summary of Chapter 2	84
3	Knowledge-based Supervision for Semantic Role Labeling	87
3.1	Knowledge-based Training Data Generation with DistantSRL	88
3.1.1	Formalization	90
3.1.2	Knowledge-based Sense Labeling	91
3.1.3	Knowledge-based Role Labeling	95
3.1.4	Distant Supervision	99
3.2	Semantic Role Labeling using Knowledge Bases	101
3.2.1	Knowledge-based Semantic Role Labeling	102
3.2.2	Semantic Role Labeling using Linked Lexical Knowledge Bases	103
3.3	Generating Training Data for Semantic Role Labeling	107
3.3.1	Generating Training Data for Word Sense Disambiguation	107
3.3.2	Generating Training Data for Semantic Role Labeling	108
3.4	Application of DistantSRL to English	115
3.4.1	Unlabeled Corpora and Gold Standard Data	115
3.4.2	Frame Labeling: Corpus Creation and Experiments	118
3.4.3	Role Labeling: Corpus Creation and Experiments	125
3.5	Application of DistantSRL to German	131
3.5.1	Automatically Generated Training Data and Test Data	131
3.5.2	Frame Labeling Experiments	132
3.5.3	Role Labeling Experiments	133
3.6	Full Semantic Role Labeling with DistantSRL	135
3.6.1	Experiments with SEMAFOR	135
3.6.2	Discussion in Relation to State-of-the-art FrameNet SRL	140
3.7	Summary of Chapter 3	142

4	Domain Adaptation via Training Data Generation	145
4.1	Motivation: Domain Adaptation for SRL	146
4.1.1	Domain Adaptation	146
4.1.2	Domain Adaptation for Semantic Role Labeling	147
4.1.3	Semantic Role Labeling for User-generated Discourse	151
4.2	YAGS – a Gold Standard for User-generated Text	152
4.2.1	FrameNet Frame and Role Annotation	154
4.2.2	Data Selection and Preparation	155
4.2.3	Annotation Task	155
4.2.4	Annotation Study	157
4.2.5	Inter-rater Agreement	159
4.2.6	Gold Standard	160
4.2.7	Comparison to Other Test Datasets	160
4.3	Assessing the Domain Generalization of FrameNet SRL	166
4.3.1	Experimental Setup	166
4.3.2	Experiment Results	167
4.4	Experiments on Training Data Generation for Domain Adaptation	173
4.4.1	Methods of Training Data Generation	173
4.4.2	Experimental Setup	177
4.4.3	Experiment Results	178
4.4.4	Error Analysis	184
4.5	Discussion	188
4.6	Summary of Chapter 4	192
5	Conclusion	195
5.1	Summary and Contributions	196
5.2	Open Issues and Outlook	198
5.2.1	Directions for Further Work	198
5.2.2	Outlook: FrameNet Model Coverage	202
5.3	Closing Remarks	203
A	Appendix A: List of Resources	205
B	Appendix B: Annotation Guidelines of FrameNet Annotation Study	209
	List of Tables	221
	List of Figures	223
	Bibliography	225

CHAPTER 1

Introduction

The semantic analysis of text is an important step on the path towards natural language understanding. It is crucial for analyzing and organizing the large amounts of text that accumulate daily on the World Wide Web.

Semantic role labeling (SRL) is a kind of semantic analysis that builds upon sentence-level syntactic analysis to answer the questions “*Who did what to whom, when, where, how, and why?*” (Palmer et al., 2009), replacing syntactic labels with semantic role labels that abstract from the surface realization of a text to the semantic function of an argument of the sentence, for instance labeling the syntactic *subject* with the role *Agent*.

SRL is increasingly requested for real-life applications and has been successfully evaluated for various applications from question answering (Narayanan and Harabagiu, 2004; Shen and Lapata, 2007) to reading comprehension (Berant et al., 2014) and identifying reason in on-line debates (Hasan and Ng, 2014).

Contemporary semantic role labeling systems are supervised or semi-supervised machine learning systems. Even state-of-the-art systems that use deep learning still rely on a labeled training set.¹

Two problems related to the generalization abilities of the supervised systems impact their use in downstream applications. The first one concerns the coverage of the available training data, e.g., can a system deal with previously unseen training data? The second problem concerns domain adaptation, i.e., how well does a supervised semantic role labeling system generalize to other domains and text types? Both problems extend to and are more severe for languages other than English which often lack the required labeled corpora.

This thesis presents a knowledge-based attempt to answer these questions. It investigates knowledge-based training data generation and its potential to enhance semantic role labeling and domain adaptation for semantic role labeling. This introductory chapter pro-

¹Fully unsupervised approaches receive increasing attention in research, but have not yet progressed to a level in which they are widely used in applications.

vides the motivation for our work on knowledge-based supervision for domain-adaptive semantic role labeling with a particular focus on FrameNet semantic role labeling that is the main subject of study in this thesis.

1.1 Semantic Role Labeling

Semantic role labeling identifies predicates and semantic arguments in text and assigns role labels to the semantic arguments, thus modeling the semantic predicate argument structure on the sentence level. It builds upon the syntactic analysis of sentences: the semantic arguments in semantic role labeling are typically based on syntactic arguments and adjuncts of the predicate – typically a verb or an event noun. Syntactic labels mark the *grammatical function*, e.g., the *subject* or the *direct* and *indirect object* of a sentence. Semantic role labeling adds a layer of abstraction to syntactic analysis by marking the *semantic function* of an argument, for instance whether it performs the role of the *Agent*, the *Patient*, or the *Beneficiary* of an event. This level of abstraction is important for natural language understanding. It however poses a difficult task for automatic analysis, because it solves a complex problem: there is no default mapping between syntactic and semantic functions, not even for the same predicate. The *Agent* role, for example, is not always associated with the grammatical *subject* of a sentence. It could also be represented by a *direct object*, or by a *prepositional object*.

Most automatic semantic role labeling systems are based on a predefined inventory of semantic roles. The most popular and well-known role inventories are FrameNet (Baker et al., 1998), PropBank (Palmer et al., 2005), and VerbNet (Kipper-Schuler, 2005). They are distinguished by the different size and granularity of the role inventories, ranging from a large inventory of fine-grained, predicate specific roles with descriptive labels for FrameNet, over the more compact set of thematic roles for VerbNet, to a set of abstract role labels for PropBank. The degree to which they relate to the syntactic structure of a sentence also varies, with PropBank roles strongly associated with syntactic arguments, and FrameNet abstracting more strongly. FrameNet also puts a stronger focus on predicate labels: it provides a hierarchy of predicate labels called *frames* that group predicates with different lemmas, leading to a setup where role labels are frame-specific, but not predicate-specific. Role labels in VerbNet and PropBank are predicate-specific. As a result, FrameNet provides a more detailed modeling of semantic knowledge than the other semantic role resources and abstracts stronger from the surface text, creating a semantic representation paraphrasing the sentence. Figure 1.1 shows an example sentence from FrameNet for the predicate *buy* that evokes the *Commerce_buy* frame. The arguments of *buy* are labeled with the frame-specific semantic roles *Buyer*, *Seller*, *Goods*, and *Money*. This has advantages for applications relying on detecting paraphrases, for instance when trying to find potential answers for questions in automated question answering: the event described by the sen-

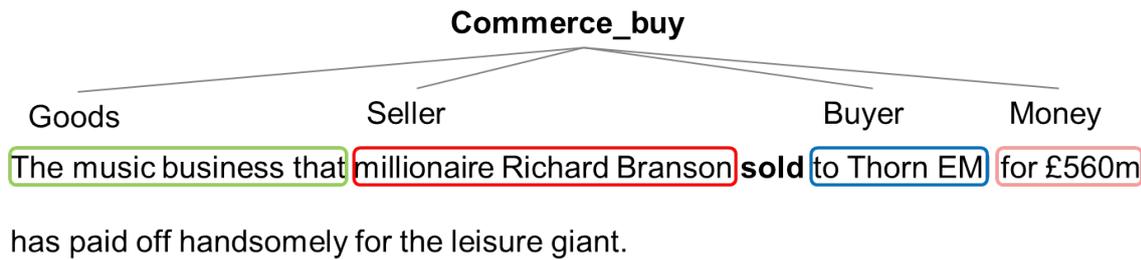


Figure 1.1: Example: FrameNet-annotated sentence from the FrameNet lexical unit examples.

tence in Figure 1.1 can also be paraphrased as “*Thorn EM bought Richard Branson’s music businesses*”. Frame-semantic analysis of the latter sentence also associates the predicate *bought* with the frame *Commerce_buy*, and it labels the participants *Thorn EM*, *Richard Branson*, and *music businesses* with the same roles as shown in the figure. Based on the frame-semantic analysis, the two sentences can be identified as descriptions of the same event.

An additional advantage of FrameNet is that it is largely language-independent. Therefore, lexicology projects in many languages have adopted FrameNet as their preferred model for lexical-semantic representation, and efforts to automatically translate FrameNet to other languages emerged.

The higher granularity of FrameNet roles increases the coverage problems of FrameNet. Therefore, it is likely to benefit from comprehensive resource integration, which motivates the focus on FrameNet in this work.

FrameNet semantic role labeling. Since Gildea and Jurafsky (2002) introduced automatic semantic role labeling, FrameNet semantic role labeling has developed significantly, also fostered by a series of SemEval shared tasks (Baker et al., 2007; Ruppenhofer et al., 2010b). FrameNet semantic role labeling generally follows a two-step approach: first, the predicates are labeled with a frame (word sense) label, then the arguments are identified and labeled with a predicate-specific role label. PropBank semantic role labeling in contrasts is focused on the identification and labeling of arguments. This is possible because the inventory of role labels is a small set of predicate-independent labels.

Increasingly, global optimization and semi-supervised methods are used to create systems that generalize well to unseen data, or can easily be adapted to new languages. For several years, different instances of the SEMAFOR frame-semantic parser showed state-of-the-art performance (Das et al., 2010, 2014). They were recently superseded by systems based on deep learning strategies (FitzGerald et al., 2015; Täckström et al., 2015).

Formalization. In this work, we formalize the task of FrameNet semantic role labeling as follows: semantic role labeling consists of two subsequent analysis steps. The second step relies on the first step. The first step is called *frame identification* – or *FrameId*:

$$\text{FrameId}(s, t_i) = f_i \quad (1.1)$$

It receives a predicate target t_i and the surrounding sentence s as input and returns the frame label f_i for the predicate target t_i . There can be $i = 1 \dots n$ predicate targets t_i per sentence s . Some definitions of FrameNet semantic role labeling also include the identification of the predicate target as a preliminary step, but we follow recent work that uses predefined targets for evaluation (Das et al., 2014).

The second step is called *role labeling*, or *RoleId*. It can be split into two subtasks, *argument identification* (*ArgId*) and *role classification* (*RoleC*), that are often solved jointly. If formalized separately, *ArgId* receives the output of the *FrameId* step f_i as input together with the predicate target t_i and the sentence s . It returns a set of argument spans $\{a_{i0}, \dots, a_{ik}\}$ for $k = 0 \dots m$ arguments per sentence s . Each a_{ij} correspond to a list of tokens associated with the argument span:

$$\text{ArgId}(s, t_i, f_i) = \{a_{i0}, \dots, a_{ik}\} \quad (1.2)$$

These argument spans are the input for the subtask of role classification, which optionally assigns a role label r_{ij} to each a_{ij} , resulting in a list of pairs of argument span and role label:

$$\text{RoleC}(s, t_i, f_i, \{a_{i0}, \dots, a_{ik}\}) = \{(a_{i0}, r_{i0}), \dots, (a_{ik}, r_{ik})\} \quad (1.3)$$

If these two steps are performed in a joint fashion, a set of pairs $\{(a_{i0}, r_{i0}), \dots, (a_{ik}, r_{ik})\}$ is returned which pairs the argument spans with their role labels:

$$\text{RoleId}(s, t_i, f_i) = \{(a_{i0}, r_{i0}), \dots, (a_{ik}, r_{ik})\} \quad (1.4)$$

where a_{ij} is the j th argument span associated with t_i , and r_{ij} is the role label associated with a_{ij} , and (a_{ij}, r_{ij}) is a pair of argument span and role label associated with t_i .

Full frame-semantic role labeling combines these two steps:

$$\text{SRL}(s, t_i) = \text{RoleId}(s, t_i, \text{FrameId}(s, t_i)) \quad (1.5)$$

To illustrate this formalization, we refer to the example sentence in Figure 1.1. The verb *sold* in the sentence corresponds to the predicate t_1 . The frame label f_1 associated with this predicate is *Commerce_buy*. The example contains $k = 4$ semantic arguments $\{a_{11}, \dots, a_{14}\}$. The argument spans are marked by boxes in Figure 1.1. Counting the arguments in the sentence from left to right, the argument span a_{12} , associated with the phrase “*millionaire Richard Branson*”, receives the role label $r_{12} = \text{Seller}$.

1.2 Knowledge Bases for Semantic Role Labeling

Most of the current semantic role labeling systems are associated with a lexical knowledge base (LKB) like FrameNet, also called lexical-semantic resource. These knowledge bases come in different degrees of elaboration: the PropBank corpus was specifically created to provide training data for semantic role labeling systems and its associated knowledge base is often neglected or dismissed as an annotation help for the PropBank corpus, whereas VerbNet and FrameNet are considered highly elaborate knowledge bases motivated by specific linguistic theories. VerbNet is based on Levin's verb classes (Levin, 1993), which have been significantly extended by additional verb classes and revised during the continued development of VerbNet. The motivating theory of FrameNet is frame semantics (Fillmore, 1976), which aims to describe common situations and their participants, modeled by frames and roles respectively. All of these resources have in common that they incorporate some notion of word sense. VerbNet and FrameNet moreover group word senses, VerbNet in verb classes that gather syntactically and semantically related verbs; FrameNet groups word senses, called *lexical units*, by frames: a frame gathers words of various parts-of-speech that are associated with (or *evoke*) the frame. FrameNet additionally spans a hierarchy of relations between frames, for instance inheritance or temporal relations. Like PropBank, FrameNet distinguishes between the lexical knowledge base and a corpus of annotated sentences that is used for the training of SRL systems, the FrameNet fulltext corpus.

Coverage problems. Also in common are the coverage problems of the lexical knowledge bases. They do apply to a different degree for the different resources, e.g., there is a larger set of labeled data available for PropBank than for VerbNet and FrameNet.

The coverage problems are caused on the one hand by insufficient amounts of training data for the given labeling problem, and on the other hand by insufficient coverage of the lexical knowledge base. For FrameNet, for example, Palmer and Sporleder (2010) analyzed the coverage problems in detail. They identified three types of coverage problems: *example coverage*, when a certain frame or role has not been seen in the training data, *lexicon coverage*, when a certain word sense is not included in FrameNet, and *ontology coverage*, when certain real-world concepts are missing from FrameNet. *Ontology coverage* is also referred to as *model coverage*.

Despite these issues, lexical-semantic knowledge bases incorporate a lot of valuable knowledge and theory compiled by expert linguists. In order to use these valuable knowledge sources efficiently, methods to enhance them have been proposed. These include standardization of lexical knowledge bases to improve access to the various types of information contained in different lexical resources, and interlinking them to increase their coverage and combine the complementary information types contained. These approaches address the problems of *lexicon coverage* and *example coverage*. Linkings between lexical

knowledge bases have been proposed on the sense and role level. They link, for instance, a FrameNet sense to the corresponding sense in WordNet, or a FrameNet role to the corresponding VerbNet role. Such linkings between two or more lexical knowledge bases result in a linked lexical knowledge base, short LLKB.

Linking lexical knowledge bases. Early work in semantic role labeling already used semi-automatically created links between FrameNet, VerbNet, and WordNet to assuage coverage problems and improve a rule-based semantic role labeling system (Shi and Mihalcea, 2005); others used word sense disambiguation to create linkings to WordNet (Burchardt et al., 2005), or used implicit linkings to WordNet and exploited the FrameNet frame hierarchy (Johansson and Nugues, 2007b). SemLink (Palmer, 2009; Bonial et al., 2013) is an ongoing effort that provides a manually curated linking between FrameNet, VerbNet and PropBank on the sense and role level. Recent work also attempts to improve semantic role labeling using available external resources, typically in the form of corpora, e.g., by adding information from PropBank (Kshirsagar et al., 2015; FitzGerald et al., 2015), but does not exploit the linked knowledge bases.

Knowledge-based methods in natural language processing (NLP) received additional leverage in the recent years when a) collaboratively created knowledge bases such as Wikipedia and Wiktionary emerged as resources for NLP, and b) advanced methods to automatically link lexical knowledge bases on the word sense level were developed (Ruiz-Casado et al., 2005; Navigli and Ponzetto, 2010). The latter are also called methods for *sense alignment*. Previous work in linking resources typically focused on specific combinations of knowledge bases, e.g., linking FrameNet to WordNet or Wikipedia (Laparra and Rigau, 2009; Tonelli and Giuliano, 2009), mostly used for translating FrameNet to other languages via multilingual wordnets. As a larger linking effort, Gurevych et al. (2012a) created UBY as a lexical database that links the major expert-built and collaboratively created knowledge bases. There has been no comprehensive evaluation that exploits linkings of several knowledge bases in parallel for FrameNet semantic role labeling.

Standardizing lexical knowledge bases. Linking lexical knowledge bases on the sense or role level does not suffice for an efficient use of the various information types contained therein. The lexical information needs to be accessible in an efficient manner. To solve this problem, Ecker-Kohler et al. (2012) developed the lexicon model UBY-LMF as a standardized format for modeling the major lexical knowledge bases and the information types provided by them. Gurevych et al. (2012a) provided the methods to convert the major lexical knowledge bases to this standardized format, representing them in a single large MySQL database, and created a Java API to access them programmatically. The present author's contributions to the standardized model of FrameNet in UBY-LMF, which include the modeling of seman-

tic lexica in UBY and the corresponding conversion routines, are presented in Chapter 2 of this work.

1.3 Training Data Generation for Semantic Role Labeling

Automatic training data generation is another method to approach the coverage problem of frame-semantic resources. It is motivated by the lack of annotated training data for FrameNet, i.e., the problem of *example coverage*. This is already a problem for the well-resourced English language,² but it is particularly severe for other languages for which annotated corpora are lacking and often cannot be created manually at a large scale because of prohibitive cost.

Similar to the previous work in linking lexical knowledge bases introduced above, some efforts at training data generation aim at extending the FrameNet lexicon and creating labeled training data at the same time. Other approaches only focus on creating FrameNet-labeled training data for predicates already present in the FrameNet lexicon. In this work, we distinguish between *knowledge-based* approaches to training data generation and *corpus-based* approaches.

Knowledge-based training data generation. Knowledge-based approaches to training data generation have been used for the generation of frame labels via linked lexical knowledge bases, in particular for languages other than English (Tonelli and Pianta, 2009a; Larraza and Rigau, 2009). They only rely on information provided by the linked knowledge base and do not use labeled corpus information. Since the knowledge bases are linked on the sense level, but not on the level of predicate argument structures or roles, they typically focus on frame labels and do not provide role-labeled training data.

Corpus-based training data generation. Corpus-based approaches have been popular for the generation of both frame and role labels on previously unlabeled texts. They rely on existing annotated corpora as a basis for the transfer of labels to unlabeled text in the same language or in a different target language, and do not use semantic information from linked lexical knowledge bases. There are two types of corpus-based approaches: approaches to annotation projection, which have been explored in monolingual and cross-lingual varieties, and monolingual approaches based on paraphrasing the annotated sentences in existing role-labeled corpora.

Annotation projection. Annotation projection uses alignments of labeled sentences to unlabeled candidate sentences on the token level or on the level of syntactic dependen-

²FitzGerald et al. (2015), for example, report that the performance of their state-of-the-art system on FrameNet test data suffers from the small training set available.

cies. For monolingual projection, the candidate sentences are identified on the basis of a matching predicate lemma. If the candidate sentences can be aligned successfully, the role labels are transferred to the aligned target constituents (Fürstenau and Lapata, 2012). For cross-lingual projection, parallel corpora are automatically labeled with an existing semantic role labeling tool, typically for English. Then, the created role labels are transferred to the aligned tokens (Padó and Lapata, 2009).

Fürstenau and Lapata (2012) also explore a variant of monolingual annotation projection that attempts to expand the FrameNet lexicon by projecting to lemmas not yet in the FrameNet lexicon. This opens up a prohibitively large search space. To make the method viable, they use a method for frame acquisition based on lexical similarities to filter the set of potential frames for a new lemma.

Exner et al. (2015) explore a variant of cross-lingual projection that borrows from distant supervision methods (Mintz et al., 2009) in relying on named-entity matching to establish the alignment: they use loosely parallel texts – Wikipedia articles in English and Swedish – and align sentences based on matching named-entities in the source and target language. The source language is labeled automatically with PropBank roles and redundancy is used to filter noisy labels. They evaluate this approach for PropBank labels.

Paraphrasing-based approaches to training data generation. Monolingual projection approaches have in common that they build on a low-resource scenario, starting out with very few labeled seed sentences. Paraphrasing-based approaches rely on a larger corpus. They attempt to variegate existing labeled sentences with the goal to create a broader range of phenomena in the training corpus. Woodsend and Lapata (2014) use synchronous grammars derived from comparable corpora based on Wikipedia and bitext from the Paraphrase Database – a large database of paraphrases generated from bilingual parallel texts (Ganitkevitch et al., 2013) – to extract paraphrase rules for PropBank-labeled data. Using these rules, they multiply the number of training instances in PropBank by 24 and improve on the state of the art for PropBank semantic role labeling.

Pavlick et al. (2015a) use the Paraphrase Database to extend the FrameNet lexicon. They replace predicates (lexical units) in the FrameNet fulltext corpus with previously unseen predicates suggested by the rewrite rules extracted from the Paraphrase Database and rely on manual postprocessing via crowdsourcing to filter out the noise. The number of predicates in the resulting resource, FrameNet+, is three times as high as the number in the original FrameNet lexicon. The coverage extension, however, does not extend to role labels: FrameNet+ only provides additional frame-labeled training data based on the FrameNet fulltext corpus.

Each of the aforementioned approaches to training data generation has advantages and disadvantages. Corpus-based approaches do not use the wealth of information encoded in linked lexical knowledge bases. Monolingual annotation projection approaches create

training data that are very similar to the source data and do not promise to generalize well to other domains. The paraphrasing-based approach for FrameNet only expands the FrameNet lexicon, but does not add additional training sentences. Previous approaches to knowledge-based training data generation focused on frame labels, but did not create additional role labels. The present work fills this gap and successfully employs linked lexical knowledge bases for the generation of frame- and role-labeled training data in Chapter 3.

1.4 Domain Adaptation of Semantic Role Labeling

Supervised machine learning systems perform worse on data with different underlying label distributions, e.g., different domains. This has also been confirmed for semantic role labeling systems (Pradhan et al., 2007b). In order to avoid the expensive labeling of large amounts of target domain data, methods for domain adaptation of supervised machine learning systems have been developed that aim to create systems that generalize to a specific target domain, or to variable domains. They range from supervised domain adaptation, which combines labeled source domain and target domain training data, via semi-supervised domain adaptation (labeled source-domain data, labeled and unlabeled target domain data), to unsupervised domain adaptation (labeled source domain, unlabeled target domain data), and finally to blind domain adaptation that attempts to create domain independent – also called open-domain – systems.

Domain adaptation for semantic role labeling has been mostly evaluated for PropBank-style semantic role labeling. The CoNLL shared tasks on semantic role labeling (Carreras and Màrquez, 2005; Surdeanu et al., 2008; Hajič et al., 2009) provide the appropriate test bed: typically, newspaper text from the Wall Street Journal Corpus is contrasted to the target domain of fiction texts from the Brown Corpus. Conventional SRL systems do not focus on the aspect of domain generalization and therefore show a large drop in performance when applied to the out-of-domain test sets. Recently, Yang et al. (2015b) used unsupervised domain adaptation based on deep belief networks to create a PropBank semantic role labeling system with improved out-of-domain performance.

For FrameNet, due to its comparatively smaller training dataset, but larger set of frame and role labels, similar, if not stronger, domain adaptation problems are expected. Both subtasks of FrameNet semantic role labeling, frame identification and role labeling, may require domain adaptation.

However, domain adaptation for FrameNet semantic role labeling has so far only been evaluated sparsely. Johansson and Nugues (2008b) evaluated the impact of different parsers on FrameNet semantic role labeling, using the Nuclear Threats Initiative (NTI) data as an out-of-domain training set, observing low domain generalization abilities of their supervised system. Croce et al. (2010) aim to create an open-domain FrameNet semantic role labeling system by integrating a distributional model into their semantic role labeling sys-

The image shows a screenshot of the Yahoo! Answers website. At the top left is the 'YAHOO! ANSWERS' logo. To the right is a search bar with 'Search Answers' and 'Search Web' buttons. On the left side, there is a navigation menu with categories like 'Answers Home', 'All Categories', and 'Cars & Transportation'. The main content area features a question: 'Who now own chrysler i know daimler sold them but i want to know who is the company that bought them.?' with a 'Follow' button and '5 answers'. Below the question is an 'Answers' section. The first answer is marked as the 'Best Answer' by Cerberus, with the text 'The Oft-Suspended RP McMurphy · 8 years ago' and 0 thumbs up/down. The second answer is by Littleball, with the text 'Daimler sold 80% of Chrysler to Cerberus Capital Management. Daimler still owns 20% of Chrysler.' and a source cited as 'Wikipedia'. It has 1 thumbs up and 0 thumbs down. A 'Relevance' dropdown menu is visible on the right side of the answers section.

Figure 1.2: Example of user-generated text: question and answers from Yahoo! Answers.

tem that generalizes lexicalized features for argument classification to previously unseen arguments and thus contributes to a system with similar performance on test data from the source domain and the target domain NTL.

Distributional methods that may help to generalize to unseen data have also been adopted by state-of-the-art systems, including dense word representations obtained via deep learning (Hermann et al., 2014; FitzGerald et al., 2015). These systems, however, have not been evaluated on out-of-domain test data. Nowadays, standard FrameNet semantic role labeling evaluation makes use of an in-domain test set, more specifically a split of the FrameNet 1.5 fulltext corpus that is randomly sampled and contains texts from the same sources in the training, development, and test portions. Out-of-domain evaluation is lacking, as are appropriate datasets that enable this kind of evaluation.

Discourse type adaptation of semantic role labeling. Domain adaptation is typically focused on different domains of professionally edited text, e.g., newspaper text versus fiction, or newspaper text versus biomedical texts (Biber and Conrad, 2009). A subtype of domain adaptation concerns the difference between discourse types, for instance between edited texts and unedited, colloquial texts, that can be found in large quantities on the web, called user-generated discourse. The most popular variants of user-generated discourse are Twitter tweets and forum posts.

Figure 1.2 shows an example of a user-generated question and two answers from the question-and-answer forum Yahoo! Answers.³ It contains two predicates evoking the *Commerce_buy* frame: *sold* and *bought*. The text shows some typical properties of user-generated text, for instance careless consideration of spelling (*own* instead of *owns*), capitalization (*i, chrysler* instead of *I, Chrysler*), and punctuation conventions, which makes it difficult to segment and analyze automatically.

This now-ubiquitous discourse type of user-generated text has sparked research on natural language processing for user-generated discourse, for instance adapting natural language processing tools to data from the Twitter service (Xu et al., 2015; Han and Baldwin, 2011). There is, however, almost no related work that evaluates semantic role labeling for user-generated discourse. Liu et al. (2010) developed a PropBank semantic role labeling system for Twitter data in the news domain, and Søgaard et al. (2015) recently created a small FrameNet-labeled dataset based on Twitter tweets in the context of their experiments in knowledge extraction. They, however, do not present a detailed evaluation of the domain generalization capabilities of FrameNet semantic role labeling.

We present such an evaluation in Chapter 4 and discuss the potential of various methods of training data generation to the domain adaptation of FrameNet. Therefore, we create a new, substantial test dataset based on user-generated questions and answers from the community question-and-answer forum Yahoo! Answers, complementing the available Twitter dataset. Surdeanu et al. (2011) showed that PropBank semantic role labeling can benefit automated community question answering, and we expect that FrameNet semantic role labeling, with its higher level of semantic abstraction, could further benefit automated question answering once sufficient semantic role labeling performance can be reached on this domain.

1.5 Research Questions

In the previous sections, we introduced the coverage problems of FrameNet semantic role labeling that hamper its use in real-life applications for English and other languages, and we also addressed the lack of research on domain adaptation of FrameNet SRL.

The goal of this thesis is to develop methods that deal with problems of lexicon coverage and domain adaptation for semantic role labeling from a knowledge-based perspective. It aims to assuage coverage issues by integrating existing lexical knowledge bases comprehensively into a linked lexical knowledge base – in previous work, only small-scale integration

³Accessed at October 16 2015 from <https://answers.yahoo.com/question/index?qid=20080212084227AAxdp15>.

efforts of selective resources have been reported. On this background, the main research question of this thesis is:

- **Can the comprehensive integration of lexical knowledge bases benefit semantic role labeling in the context of domain adaptation and adaptation to other languages?**

This question raises several secondary research questions that are addressed in the course of this thesis:

1. How can the semantic information in lexical-semantic knowledge bases like FrameNet be standardized such that their integration can be modeled effectively?
2. Can the comprehensive integration of expert-built and user-generated lexical knowledge bases alleviate FrameNet coverage issues for English and other languages?
3. How can linked lexical knowledge bases contribute to enhance semantic role labeling systems, for instance via automatic training data generation?
4. What are the domain generalization capabilities of open-source FrameNet semantic role labeling, and what are the needs and requirements for domain adaptation?
5. How can linked lexical knowledge bases support domain adaptation for semantic role labeling?

The long-term objective of this research is making semantic role labeling viable for real-world applications on user-generated web data, for instance automatic question answering. The result is an evaluation of knowledge-based supervision for domain-adaptive semantic role labeling as outlined in the next section.

1.6 Approach

Figure 1.3 summarizes the approach to the research questions that was taken in this thesis. It can be roughly segmented into two parts. The first part, on the left-hand side of the figure, describes work on enhancing lexical knowledge bases by their standardization and integration into a linked lexical knowledge base, answering research questions 1 and 2.

The second part, on the right-hand side of the figure, focuses on the use of the integrated knowledge base for semantic role labeling. We explore a novel distant supervision approach for the automatic generation of large-scale role-labeled training data in various domains that is inspired by related work in relation extraction (Mintz et al., 2009). The lower box on the right-hand side represents the analysis of the proposed method in an in-domain semantic role labeling setup for English and German, answering research question 3. It

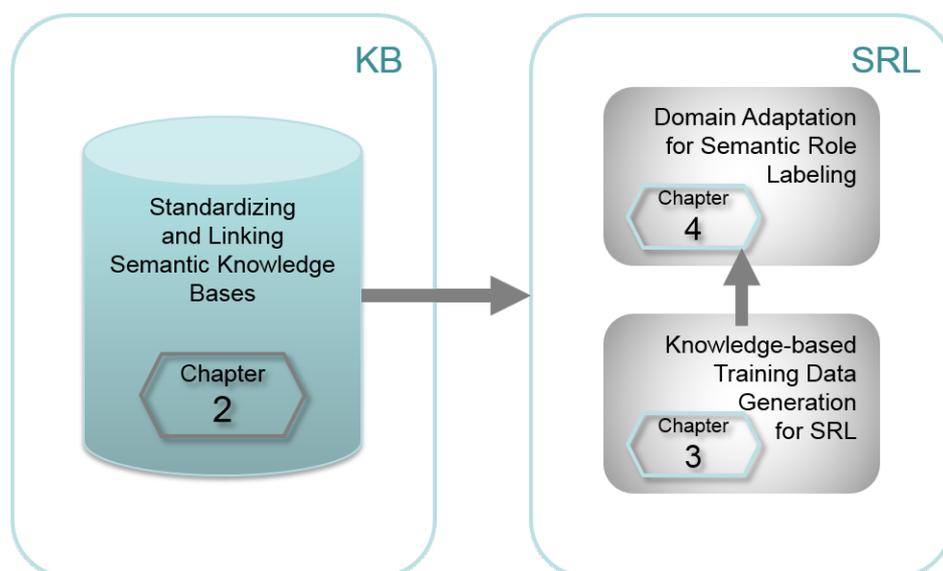


Figure 1.3: Thesis overview diagram.

provides a detailed evaluation for the tasks of frame identification and role classification. Additionally, training an open-source semantic role labeling system with the English data leads to a semi-supervised semantic role labeling system for English. The upper box on the right-hand side represents the application of the distant supervision approach to domain adaptation, which is evaluated for English. This includes a comparison to other approaches of training data generation, answering research questions 4 and 5.

1.7 Contributions and Findings

In the course of this thesis, we show that the large-scale integration of lexical knowledge bases on the sense and semantic role level can benefit semantic role labeling for English and other languages. The main contributions and findings of this thesis can be summarized as follows:

1. The development of standardized models of interoperability of semantic knowledge bases like FrameNet, specifically modeling FrameNet in UBY-LMF and modeling links on the frame and role level (see Chapter 2).
2. A novel method for creating a FrameNet in any language based on the automatic alignment of FrameNet and Wiktionary. The method is evaluated on the example of German, effectively building a larger FrameNet knowledge base to improve FrameNet semantic role labeling for English and German (see Chapter 2).

3. The study of the knowledge-based distant supervision method DistantSRL for the creation of frame- and role-labeled training data for English and German. The resulting training data are of high quality and complementary to the manually labeled data, i.e., the FrameNet fulltext corpus and SALSA (see Chapter 3).
4. A detailed analysis of the domain generalization capabilities of contemporary open-source FrameNet semantic role labeling for English; it shows that domain adaptation is required for the task of frame identification (see chapter Chapter 4).
5. The application of the introduced knowledge-based methods for training data generation in the context of adapting FrameNet SRL to the domain of user-generated discourse in community question answering and their comparison to alternative methods for training data generation for English, e.g., FrameNet+ and monolingual annotation projection (see Chapter 4). We find that our automatically generated training datasets have the potential to improve the domain generalization capabilities of FrameNet semantic role labeling, but that the training of semantic role labeling systems needs to be adapted to better deal with large and noisily labeled training data.

Along with the methodological contributions and studies listed above, a number of lexical resources were created and published for research purposes:

- UBY_{FN} , a lexical knowledge base in UBY-LMF format containing WordNet, FrameNet, VerbNet, PropBank, the English Wiktionary, the German Wiktionary, GermaNet, and SALSA, and links between them from SemLink.
- FNWKde, a lexical knowledge base in UBY format containing FrameNet, the English and German Wiktionary, and automatically created sense-level links between FrameNet and the English and German Wiktionary.
- several large-scale corpora automatically labeled with FrameNet frames and roles using DistantSRL and monolingual annotation projection.
- a manually annotated gold standard dataset of 2,789 matching and non-matching pairs of word senses between FrameNet and the English Wiktionary, with annotation guidelines.
- YAGS, a large manually annotated gold standard dataset of 3,091 FrameNet frame and 6,081 role annotations on the user-generated texts from the Yahoo! Answers Manners dataset, with annotation guidelines.

The software developed for the thesis was partially integrated into and contributed to the DKPro UBY and DKPro Core open-source software repositories. A detailed list of resources and download links are contained in Appendix A.

1.8 Publication Record

Large parts of the research presented in this thesis have been published previously in peer-reviewed conference proceedings or journals. The central publications are:

- FrameNet on the Way to Babel: Creating a Bilingual FrameNet Using Wiktionary as Interlingual Connection ([Hartmann and Gurevych, 2013b](#)). This work was presented at ACL 2013 in Sofia, Bulgaria and is described in Chapter 2.
- Generating Training Data for Semantic Role Labeling based on Label Transfer from Linked Lexical Resources ([Hartmann et al., 2016](#)). This article was published by the Transactions of the Association for Computational Linguistics. It describes joint work with Dr. Judith ECKLE-KOHLER who provided the methods for the knowledge-based role label transfer on sense-labeled data. The improved approach to sense labeling and the experimental evaluation of the method presented in this paper are contributions of the author. The content of this paper is described in Chapter 3.
- Out-of-domain FrameNet Semantic Role Labeling ([Hartmann et al., 2017a](#)). This paper describes joint work with Ilia Kuznetsov and Teresa Botschen (née Martin) and was presented at EACL 2017 in Valencia, Spain. The creation of YAGS, the FrameNet-labeled gold standard based on user-generated question-and-answer data, and the out-of-domain evaluation of SEMAFOR in this paper are contributions of the present author and reported in this thesis. They are described in detail in Chapter 4.

The author’s contributions to the following publications are also reported in this thesis:

- UBY – A Large-Scale Unified Lexical-Semantic Resource Based on LMF ([Gurevych et al., 2013](#)). This paper was presented at EACL 2012 in Avignon, France.
- UBY-LMF – A Uniform Model for Standardizing Heterogeneous Lexical-Semantic Resources in ISO-LMF ([Eckle-Kohler et al., 2012](#)). This paper was presented at LREC 2012, Istanbul, Turkey.
- UBY-LMF - Exploring the Boundaries of Language-Independent Lexicon Models ([Eckle-Kohler et al., 2013](#)). This publication is a chapter in the book “LMF Lexical Markup Framework” published by ISTE–HERMES–Wiley in 2013. It contains a detailed description of the model of FrameNet in UBY-LMF.

The main contributions of the present author to these three papers are the model of FrameNet in UBY-LMF and the corresponding conversion routines. They are presented in Chapter 2.

During the work on the thesis, the author also contributed to a number of publications that are not reported in this thesis, but are placed in related research areas, namely the acquisition of multiwords for lexical knowledge bases and ontologies (Hartmann et al., 2012; Hartmann and Gurevych, 2013a), which provides another example on how the lexicon coverage of FrameNet could be improved. Other publications present further work on integrating lexical knowledge bases in UBY and making them accessible to researchers in linguistics and natural language processing (Gurevych et al., 2012b, 2013; Chiarcos et al., 2012a,b). Another publication (Mújdricza-Maydt et al., 2016) provides the basis for a better integration of SALSA, the frame-semantic resource for German, with other lexical resources in the future: it presents the reannotation of parts of the SALSA corpus with GermaNet senses and VerbNet-style roles to facilitate the comparative evaluation of semantic role resources and their linking on the sense and role level for German. It also provides the foundations for the work reported in Hartmann et al. (2017b). In this study, we create a small German corpus with parallel FrameNet-style, VerbNet-style, and PropBank-style annotations and use it as a benchmarking dataset for the experiment-based comparison of the three semantic role labeling frameworks.

1.9 Thesis Outline

The main part of this thesis follows the order of the research contributions that build upon another as visualized in Figure 1.3.

Chapter 2 introduces work on enhancing linked lexical knowledge bases by standardizing and linking them; it builds the foundation for the following chapters. We study two aspects of how standardization and integration of lexical knowledge bases can improve coverage issues associated with expert-built semantic knowledge bases. The first aspect is the extension of lexical knowledge bases by standardizing their format and linking them on the sense and role levels. The second aspect is the translation of lexical knowledge bases to other languages based on their linking to multilingual resources. The chapter presents a novel approach using Wiktionary as an interlingual index for the creation of FrameNet resources in any language and its application to German. It results in UBY_{FN} , a linked lexical knowledge base for English and German centered around FrameNet.

Chapter 3 describes a new method for knowledge-based training data generation for FrameNet frame identification and semantic role labeling. The method uses knowledge-based distant supervision from UBY_{FN} , the LLKB created in Chapter 2, for the automatic labeling of large corpora with FrameNet word senses, i.e., frames, and roles. The method requires large corpora, because it labels the input data only sparsely.

Experiments on the tasks of frame identification and role classification on diverse test sets show that the method creates large-scale training data of high quality that complement the FrameNet fulltext corpus, indicating that they can be used to support domain gener-

alization of FrameNet semantic role labeling systems. An evaluation for the English and German languages shows that the method also generalizes to different languages. Integrating the labeled data in an open-source semantic role labeling system results in a semi-supervised system for semantic role labeling for English. Experimental evaluation of this system further proves that our automatically generated training set has the potential to improve full FrameNet semantic role labeling, but also shows that semantic role labeling systems trained on these data need to be adapted to better deal with large amounts of noisily labeled training data.

Chapter 4 evaluates the approach to training data generation introduced in the previous chapter and the resulting semi-supervised semantic role labeling system for English in the context of domain adaptation to user-generated discourse. To this end, we introduce a new manually labeled test dataset of FrameNet role labels based on user-generated data from the Yahoo! Answers question-and-answer forum. We use this dataset to assess the domain generalization capabilities of an open-source FrameNet semantic role labeling system, finding that the main bottleneck for domain adaptation of FrameNet semantic role labeling is the frame identification step. We therefore study the potential of our automatically generated training data to support domain adaptation of frame identification. In these experiments, the knowledge-based approach to training data generation is also compared to other approaches for training data generation, e.g., monolingual annotation projection and FrameNet+. The results further confirm our results from Chapter 3: the automatically generated training set has the potential to improve frame identification across domains, but systems making full use of its potential need to be equipped to deal with large amounts of training data and may need to use additional domain adaptation methods.

The thesis concludes in Chapter 5 with a summary of the work presented and a detailed discussion of open issues, for instance how to efficiently use the large amounts of automatically labeled training data for FrameNet semantic role labeling, and future work, for instance improving FrameNet coverage not only for the lexicon and training data, but also improving the model coverage.

1.10 List of Abbreviations

In this section, we introduce a number of abbreviations that we use in the course of this work.

- **SRL** abbreviates *semantic role labeling* as introduced in Section 1.1.
- **LKB** abbreviates *lexical knowledge base*, a term referring to lexical resources such as FrameNet, WordNet, Wikipedia, and Wiktionary.
- **LLKB** abbreviates *linked lexical knowledge base* and refers to LKBs linked on the level word senses and/or semantic roles.
- **WSD** abbreviates *word sense disambiguation*, the assignment of a word sense label according to a sense inventory like WordNet or FrameNet to a given target word in context.
- **VSD** abbreviates *verb sense disambiguation*, word sense disambiguation for verbs.
- **POS** abbreviates the term part-of-speech.
- **UGD** abbreviates the term user-generated discourse.
- **Acc** stands for accuracy, an evaluation metric that reports how many system predictions are correct with respect to a gold standard annotation.
- **P** stands for precision, an evaluation metric that reports how many of the instances labeled by the systems are correctly labeled with respect to a gold standard annotation.
- **R** stands for recall, an evaluation metric that reports how many of the available test instances are labeled correctly by the system.
- **F₁** is the F₁-score, the harmonic mean between between precision P and recall R.

CHAPTER 2

Integrating Semantic Knowledge Bases

This chapter presents our contributions to the integration of lexical-semantic knowledge bases through their linking and standardization. It builds the foundation for our work on knowledge-based semantic role labeling in the following chapters.

The work presented in this chapter is motivated by the deficiencies of existing semantic knowledge bases like FrameNet: as a lexical resource created manually by experts, FrameNet is not complete. Even for English, gaps in the lexicon coverage and a lack of training data are detrimental to the quality of automatic semantic role labeling. For other languages, semantic role resources are even smaller than FrameNet, or non-existent.

Links between lexical resources have already been previously used to address this problem, but they have not been explored and applied at a large scale, for instance by exploiting links on the sense level and on the level of semantic predicate argument structure, and by exploiting links to several resources at the same time. To perform such an exploration, we embed FrameNet in a network of linked lexical knowledge bases, first by integrating existing linkings that were created either manually or automatically, and second by automatically creating new links to further lexical knowledge bases.

We address two of the research questions outlined in Chapter 1. The first research question is: can the lexicon coverage of semantic knowledge bases like FrameNet be improved for English and other languages by linking them comprehensively on the sense level and on the semantic role level? To answer this question, we integrate existing sense and predicate argument structure linkings into a large linked lexical knowledge base, create a new automatic sense alignment of FrameNet to Wiktionary, and analyze the resulting linked lexical knowledge base in detail. Automatic sense alignment is an intricate task, due to the different sense granularity in the different lexical knowledge bases, and due to the different representational models they provide for the same, or similar, semantic information types.

We also address the aspect of FrameNet coverage for other languages, namely the transfer of lexical knowledge bases for English to other languages lacking such resources. This

task is even more difficult than coverage extension for English, because it requires automatic methods to provide a connection across languages, for instance cross-lingual similarity; automatic methods may encounter different conceptualizations of word senses in different languages.

In this work, we propose a new answer to the question on how to extend the coverage increase to other languages, i.e., can we induce semantic role resources by linking lexical knowledge bases cross-lingually? The solution suggested in this chapter is a simple, but powerful approach to construct a FrameNet lexicon in other languages using Wiktionary as an interlingual representation. The approach is applied to and evaluated on the German language, resulting in a German FrameNet lexicon. We present a detailed evaluation of the induced German FrameNet and a discussion of Wiktionary as an interlingual connection for the cross-language transfer of lexical-semantic resources to various languages.

The second research question we address concerns the efficient access to the various information types contained in the linked lexical knowledge bases. To efficiently use the various types of information encoded in the linked lexical knowledge bases, they need to be represented in a standardized format and provided with a common API. To achieve this goal, the different linguistic information types, and the different terminology used to represent these types in the various lexical knowledge bases need to be unified, which requires diligent analysis of the structure of the considered lexical knowledge bases and their information types. We aim to provide a comprehensive model that conforms to the metamodel implemented in the ISO standard LMF (Francopoulo et al., 2006), and represents all information types in the major lexical knowledge bases; we specifically present our contributions to creating such a model for a) semantic lexicons like FrameNet, and b) for links on the level of semantic predicate argument structure.

In this chapter, we first introduce FrameNet and motivate our work in enhancing and translating FrameNet via resource linking. We then introduce relevant lexical knowledge bases, discuss earlier efforts in linking them, and finally present our own work on automatically aligning FrameNet to Wiktionary on the sense level, and using the resulting alignment to create a German FrameNet lexicon. We then describe our contributions to standardization and modeling of linked semantic knowledge bases, which is a prerequisite for their efficient use in NLP. The chapter closes with a detailed description of the resulting linked lexical knowledge base UBY_{FN} .

2.1 Motivation: Extending FrameNet

In this section, we motivate the work on linking and integrating lexical-semantic knowledge bases presented in this chapter. The main driving factors are the coverage gaps of FrameNet and the lack of FrameNet-style resources in other languages. Therefore, this sec-

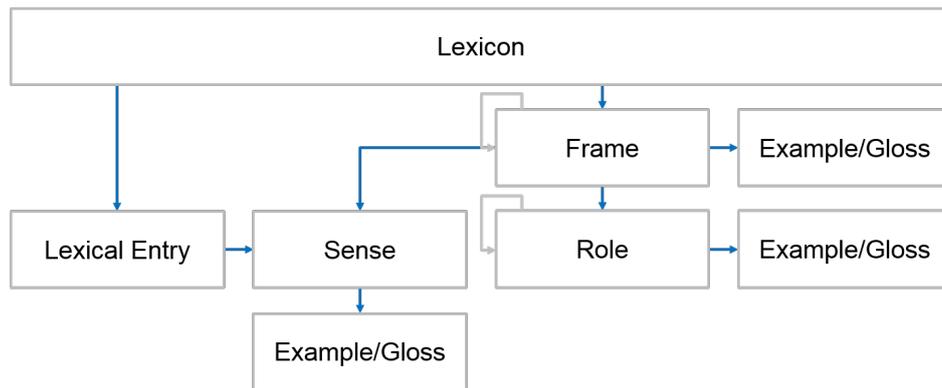


Figure 2.1: FrameNet lexicon structure.

tion first introduces FrameNet in detail, and then discusses coverage gaps and problems associated with the creation of FrameNet-like resources in other languages.

2.1.1 FrameNet

FrameNet (Baker et al., 1998; Fillmore and Baker, 2010) is an expert-built lexical-semantic resource incorporating the theory of frame semantics (Fillmore, 1976). It groups word senses in *frames* that represent particular prototypical situations, including predicates for both events and states: the verb *complete* and the noun *completion* belong to the frame *Activity_finish*.

The participants of these situations, typically realized as syntactic arguments, fill the semantic roles of the frame called *frame elements* in FrameNet, for instance the role of the *Agent* performing an activity, or the role of the *Activity* itself. Roles are frame-specific, leading to a large inventory of roles. FrameNet distinguishes between obligatory roles that are crucial for the understanding of the frame and optional roles, that are often realized as adjuncts in example sentences. These are called *core roles* and *non-core* roles respectively. Examples for core roles of the frame *Activity* are *Agent* and *Activity*, examples for non-core roles are *Place*, *Duration*, and *Manner*.

FrameNet lexicon structure. Figure 2.1 shows the schematic structure of the FrameNet lexicon. As shown by the two blue arrows leaving the *Lexicon* box in Figure 2.1, there are two entry points to the FrameNet lexicon. One is the lexical entry that is defined by lemma and part-of-speech, represented by the box labeled *Lexical Entry*, the other is the list of frames in FrameNet, represented by the box labeled *Frame*. A lexical entry can have one or several word senses, called *lexical units* in FrameNet, that are represented by the box labeled *Sense* in the figure. These are defined by lemma, part-of-speech, and frame label, which is why they are linked to the *Frame* box. Each frame provides a set of roles, called *frame*

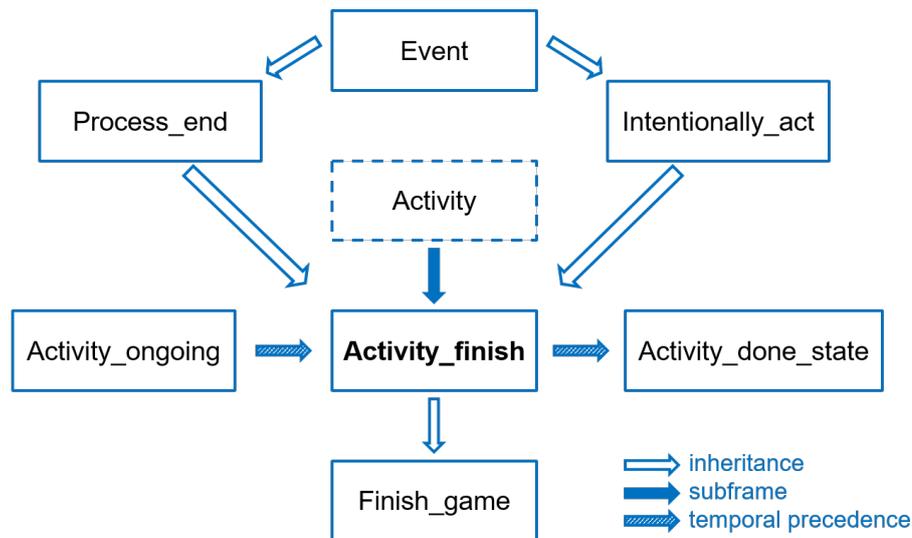


Figure 2.2: Example: FrameNet frame hierarchy around *Activity_finish*.

elements in FrameNet that are represented by the box labeled *Role*.⁴ Senses, frames, and roles are equipped with a definition gloss and example sentences in the FrameNet lexicon, as shown by the box labeled *Example/Gloss*. To simplify the lexicon schema, there is only a single box for examples and/or definition glosses. FrameNet also defines relations between frames and roles that are represented by the grey arrows in Figure 2.1 and will be introduced in the next paragraph.

Frame relations. FrameNet frames are connected by a hierarchy of frame relations, the frame *Activity_finish*, for instance, is a subframe of the abstract frame *Activity*, inherits from the frames *Intentionally_act* and *Process_end*, temporally precedes the frame *Activity_done_state* and is preceded by the frame *Activity_ongoing*. These relations are visualized in Figure 2.2. The inheritance relation describes the *is-a* relation present in many ontologies. A frame that inherits from another frame, just as *Activity_finish* inherits from *Process_end*, inherits its semantic properties, for instance its set of roles. The subframe relation is used to segment events into smaller parts, an *Activity* contains several subframes that describe the beginning, running, and the end of the *Activity*. Some of these subframes are shown in Figure 2.2: *Activity_ongoing*, *Activity_finish*, and *Activity_done_state* are subframes of *Activity*. The subframes follow a temporal order, therefore they are linked by a temporal precedence relation. The dotted lines in Figure 2.2 mark *Activity* as an *abstract*

⁴Note that we use similar diagrams to characterize the structure of other lexical knowledge bases later in this chapter and contrast them to FrameNet. To ensure comparable terminology between different lexical knowledge bases, we diverge from FrameNet terminology when describing certain aspects of the lexicon structure, for instance using *sense* instead of *lexical unit* or *role* instead of *frame element*.

Lemma:	complete
Part-of-speech:	verb
Sense id:	11352
Sense definition:	COD: finish making or doing
<hr/>	
Frame:	Activity_finish
Frame definition:	An Agent finishes an Activity, which can no longer logically continue. This frame is a subframe of Activity.
Core roles:	Activity, Agent
Role definition Activity:	This Fe identified the Activity that the Agent has finished
Role definition Agent:	This FE identifies the Agent who has finished an Activity
Non-core roles:	Circumstances, Containing_event, Degree, Depictive, Explanation, Manner, Mean, Place, Result, Subevent, Time
Role definition Circumstances:	...
...	
<hr/>	
Example sentences:	<ol style="list-style-type: none"> 1 [The 17-year-old]_{Agent} has [already]_{Time} COMPLETED [one course of 10 lessons]_{Activity}. 2 [Neil]_{Agent} has [recently]_{Time} COMPLETED [an 'A' level course in Art]_{Activity} [at his local college]_{Place} and is an enthusiastic painter. 3 In addition, [the project]_{Activity} was COMPLETED [within budget]_{Circumstances}.
...	

Figure 2.3: Example: FrameNet lexicon entry for the verb *complete*.

frame that, unlike other frames, is not directly associated with word senses in FrameNet, but provides a set of roles that derived frames inherit. *Activity_finish* for instance inherits the roles *Agent* and *Activity* from the abstract frame *Activity*.

Examples for relations between roles are *excludes* and *requires*. The *Damaging* frame has two core roles, *Agent* and *Cause*, that exclude each other: either the *Damaging* is performed by an *Agent*, typically a person, or brought about by a *Cause*, for instance a thunderstorm. The *requires* relation occurs for the roles that describe the participants of a conversation in the *Discussion* frame, *Interlocutor_1* and *Interlocutor_2*. In the example sentence “*She discussed the menu with the chef*”, “*She*” is labeled with the role *Interlocutor_1* and “*with the chef*” is labeled with the role *Interlocutor_2*. The *Interlocutor_2* cannot be omitted. An alternative phrasing would be to group the several participants of a discussion as in the sentence “*They discussed the menu*”. In this example, “*They*” receives the *Interlocutors* role that excludes both *Interlocutor_1* and *Interlocutor_2*.

FrameNet lexical entry. To further illustrate the FrameNet lexicon, Figure 2.3 shows some of the information associated with the verb *complete* in FrameNet. The verb *complete* has a single word sense in FrameNet that is characterized by the frame *Activity_finish* and receives the unique sense id label 11352. Figure 2.3 also shows the frame definition of the frame *Activity_finish* and an abbreviated list its roles. The frame *Activity_finish* has two core roles and twelve optional non-core roles. The FrameNet lexical entry for the example

sense also contains 24 example sentences annotated with frame and role labels. Three of them are shown in Figure 2.3.

FrameNet release 1.5. The creation of FrameNet is a major lexicographic effort: corpus evidence was collected for the most frequent verbs in the British National Corpus and aggregated by syntactic behavior and word sense to identify the concepts modeled as frames. The FrameNet release 1.5 (Ruppenhofer et al., 2010a) contains 1,019 frames, 9,633 frame-specific roles, and 11,942 word senses defined as a combination of lemma, part-of-speech, and frame, also called *lexical unit* in FrameNet. Many of those word senses are equipped with role-annotated sense examples that are based on corpus evidence from the British National Corpus, leading to overall 154,485 example sentences like the sentences shown in Figure 2.3. Additionally, running text has been annotated with FrameNet frames and roles and is available as the FrameNet fulltext annotations corpus containing 5,946 sentences and 23,944 annotated frames.⁵ The corpus texts annotated with FrameNet frames and roles have been used to train automatic semantic role labeling systems, for instance the SEMAFOR system (Das et al., 2014).

2.1.2 English Coverage

As an expert-built resource, FrameNet is growing slowly. As a result, there are coverage gaps on the ontology level, the lexicon level, and also on the level of annotated corpus instances (Palmer and Sporleder, 2010). On the ontology level, certain frames may be missing from the frame hierarchy. FrameNet, for instance, contains a frame for *Intentional_deception*, but does not contain a more specialized frame to describe a doping scenario in sports competitions. Additionally, the frame hierarchy may be more fine-grained for certain topic areas than for others.

On the lexicon level, certain word senses – or lexical units – are missing in FrameNet, even though there is a frame that represents their meaning. The word senses associated with the frame *Cause_to_make_progress* include the verb *improve*, as in “*We improve FrameNet*”, but they do not contain the synonymous verb *enhance*.

To get an estimate on the coverage of the FrameNet lexicon, we analyze how many unique words defined by their lemma and POS (types) and instances of these words (tokens) the FrameNet lexicon covers in various corpora. We use the word frequency lists from the British National Corpus (BNC)⁶ and the written part of the American National Corpus (ANC),⁷ and the web-based corpus ukWAC (Baroni et al., 2009). Table 2.1 shows the results

⁵Earlier versions of FrameNet do not distinguish between example sentences and fulltext annotations. In September 2015, FrameNet release 1.6 has been published. Since we used FrameNet 1.5 in the present work, all reported details concern FrameNet release 1.5.

⁶We use Adam Kilgarriff’s BNC frequency lists: <https://www.kilgarriff.co.uk/bnc-readme.html>.

⁷<http://www.anc.org/data/anc-second-release/frequency-data/>

selected POS	BNC		ANC-written		ukWAC 1-4	
	type	token	type	token	type	token
FrameNet						
adjective, noun, verb	55.44	80.58	3.41	36.41	0.31	29.56
adjective	47.69	68.81	2.21	51.86	0.42	42.83
noun	49.48	70.51	2.45	6.30	0.18	23.32
verb	77.44	95.48	34.03	60.15	3.24	75.67
WordNet						
adjective, noun, verb	98.99	99.48	17.45	61.70	2.56	56.01
adjective	97.33	98.93	13.60	84.43	4.03	83.94
noun	99.33	99.58	16.48	20.54	2.06	46.41
verb	99.61	99.55	70.88	79.58	9.07	86.82

Table 2.1: FrameNet lexicon coverage of several corpora in percent.

of this analysis and provides numbers for WordNet for comparison. In total, the FrameNet lexicon covers 55.4% of the open-class words, e.g., nouns, verbs, adjectives, in the BNC, accounting for 80.58% of the corpus instances. The coverage of the BNC is high, very high for verbs, since the BNC was used as a basis for the development of FrameNet. For other corpora, the picture changes: the FrameNet lexicon only covers 3.41% of the open-class types and 36.41% of the open-class tokens of the ANC. For the web-based corpus ukWAC (Baroni et al., 2009), the FrameNet lexicon only covers 0.31% of the open-class words and 29.56% of the corpus instances for open-class words. The type and token coverage is higher for verbs in all corpora. In the web-based ukWAC corpus, the token coverage is high, even despite the low type coverage. The low noun coverage in ukWAC is caused to a large degree by the automatic preprocessing in ukWAC that lists various kinds of proper names and URLs as nouns. The ANC contains a lot of biomedical terminology, which leads to the low noun coverage observed for the ANC.

The comparison to WordNet shows that the WordNet coverage of the considered corpora is in general higher. The difference is, however, smaller for verb tokens. This shows that FrameNet coverage for verbs is already fairly well, but can be improved in comparison to a larger LKB like WordNet.

This coverage analysis can only provide information on the lexicon coverage on the level of lexical entries, i.e., lemma and part-of-speech entries, it cannot determine, whether the word senses in the corpora are present in the FrameNet lexicon. To estimate the sense coverage would require a large sense-labeled corpus labeled with FrameNet senses and word senses not in FrameNet.

On the level of corpus instances, or example coverage, there are senses in FrameNet that are not equipped with a role-annotated example sentence, nor with an annotated example in the FrameNet fulltext corpus. This has direct impact on training automatic semantic role labeling systems based on FrameNet and their applicability in downstream applications.

Palmer and Sporleder (2010) evaluate these coverage gaps for FrameNet release 1.3, training the Shalmaneser semantic role labeler on data from the FrameNet fulltext annotations and testing on the SemEval 2007 shared task on Frame Semantic Structure Extraction (Baker et al., 2007). They find that more than 75% of the errors in frame identification are caused by the lack of information on the evaluated target word or the frame (lexicon coverage and model coverage), more than 13% of the errors result from misclassifications where the correct frame label for a word does not appear in the training data (example coverage). Only 9% of the errors are based on misclassifications of instances for which the gold label was seen in the training data. As a result of this study, they recommend to investigate methods for FrameNet semantic role labeling that add capabilities of dealing with previously unseen data.

Ruppenhofer et al. (2010c) suggest to automatically merge certain related frames and their associated word senses in order to reduce the sense-granularity in FrameNet and make FrameNet semantic role labeling more robust. They find that a coarser-grained FrameNet increases the semantic role labeling performance in a cross-validation setup for FrameNet release 1.3. This does not cover the coverage problems on the ontology level, but can treat the problems of lexicon coverage and example coverage to a certain extent: by merging frames, the sets of predicates for two frames are merged, leading to a larger number of predicates and a larger number of training instances for the new frame.

Another method to alleviate the coverage problems on the lexicon level is linking FrameNet to other lexical knowledge bases. We will introduce this method in detail in Section 2.3 below. Now, we discuss FrameNet coverage for languages other than English.

2.1.3 FrameNet for Languages other than English

The FrameNet frame hierarchy can be considered largely language-independent. From a theoretical perspective, Boas (2005) confirms that FrameNet frames are generally appropriate for modeling situations across languages. Padó and Erk (2005) also consider frames as mostly language-independent. They suggest to use frame-semantic analyses of sentence translation pairs to investigate similarities and differences in how different languages express similar meaning.

Boas (2005) however also reports that, while many frames are largely language-independent, other frames receive culture-specific or language-specific interpretations, for example calendars or holidays. Also, fine-grained sense and frame distinctions may be more relevant in one language than in another language. Such granularity differences also led to the addition of proto-frames in the German SALSA (Rehbein et al., 2012), a German Frame-

name	URL
FrameNet	https://framenet.icsi.berkeley.edu
FrameNet Brasil	http://www.ufjf.br/framenetbr/
Chinese FrameNet	http://sccfn.sxu.edu.cn/portal-en/home.aspx/
Japanese FrameNet	http://jfn.st.hc.keio.ac.jp/
Danish FrameNet	http://framenet.dk/
French FrameNet	https://sites.google.com/site/anrasfalda/
Spanish FrameNet	http://spanishfn.org
Swedish FrameNet	http://spraakbanken.gu.se/eng/swefn
German FrameNet	http://www.laits.utexas.edu/gframenet/
SALSA	http://www.coli.uni-saarland.de/projects/salsa/

Table 2.2: Overview on frame-semantic resources for languages other than English.

Net resource. SALSA aims to re-use frames from FrameNet release 1.3, but introduces new rudimentary frames with only few core roles for a German lexical unit that requires a finer sense distinction than available in FrameNet. These new frames are called proto-frames, because they present a preparatory step for the definition of a new frame. They are predicate-specific, i.e., they do not connect several senses like regular FrameNet frames, and do not contain a meaningful frame label, but do provide a brief definition of the frame and its core roles. An example for a proto-frame will be shown in Figure 2.11 in Section 2.2.3 that introduces SALSA in detail.

Both, the general applicability and the necessity of considering language-specific aspects led to the creation of FrameNet resources in several other languages, for instance, FrameNet Brasil, Chinese FrameNet, Japanese FrameNet, Danish FrameNet, French FrameNet, Spanish FrameNet, the Swedish FrameNet, German FrameNet, and the German SALSA (Burchardt et al., 2006). They have in common that they use the semantic model introduced by FrameNet, but adapt it to language-specific requirements. Table 2.2 shows an overview of these international FrameNet initiatives.

The increasing number of initiatives to create FrameNet resources for other languages together with the increasing use of FrameNet in NLP applications indicate the need for FrameNet resources in multiple languages. This need also inspired work on generating FrameNet-like resources for languages other than English, either automatically or semi-automatically, including our own work. Some of these approaches use sense-level linkings to other lexical knowledge bases, which requires knowledge on their structure and the types of lexical-semantic information they contain. Therefore, we introduce these lexical knowledge bases in the next section, before introducing related work and our own approach in creating FrameNet resources in various languages in the subsequent sections.

2.2 Lexical-semantic Knowledge Bases

Aside from FrameNet, there are several large expert-built lexical knowledge bases for English and German. WordNet and GermaNet focus on modeling word sense information and relations between senses. The most popular lexical resources with predicate argument structure models are FrameNet, VerbNet, and PropBank for English, and SALSA for German. These resources have been built to embody different linguistic and semantic theories and thus contain complementary types of linguistic information. They also show a large overlap in linguistic information, but represent similar information types in different ways. This motivates our work on the standardization of lexical knowledge bases introduced in Section 2.7 below.

The expert-built lexical knowledge bases have in common that they are mostly created manually, which entails large cost and large efforts in creation. On the upside, they contain fine-grained semantic information in machine-readable formats, thus providing valuable information for natural language processing applications like word sense disambiguation and semantic role labeling. If the information in different resources can be accessed efficiently in a uniform way, the larger coverage in larger lexical knowledge bases like WordNet can, for instance, be used to increase the coverage of smaller lexical knowledge bases like FrameNet.

Another important type of resource are collaboratively created resources that emerged with the Web 2.0 and were soon discovered as valuable resources for NLP ([Gabrilovich and Markovitch, 2007](#)), for example Wikipedia and Wiktionary. They promise to solve the coverage issues associated with expert-built resources by accumulating the efforts of crowds of volunteer contributors. At the same time, they pose new questions for electronic dictionaries and lexical knowledge bases, for instance how to ensure a high standard of quality, e.g., by identifying faulty entries resulting from vandalism in Wikipedia ([Adler et al., 2011](#)). [Meyer \(2013\)](#) discusses vandalism in Wiktionary.

This section introduces English and German resources that are central to this work, starting with those resources that, like FrameNet, provide a model of semantic predicate argument structure. We introduce their lexicon structure, the information types therein, and discuss similarities and differences between the different lexical knowledge bases.

2.2.1 VerbNet

VerbNet groups verb senses in a hierarchy of classes that go back to Levin's classification ([Levin, 1993](#)), but have been extended substantially during the development of VerbNet ([Kipper-Schuler, 2005](#); [Kipper et al., 2006](#); [Kipper-Schuler et al., 2008](#)). The classification groups verbs based on their syntactic and semantic properties. The verb semantics are modeled in two ways: first, there is an event semantics representation following Montague semantics that describes the verb semantics as a boolean combination of a base inventory

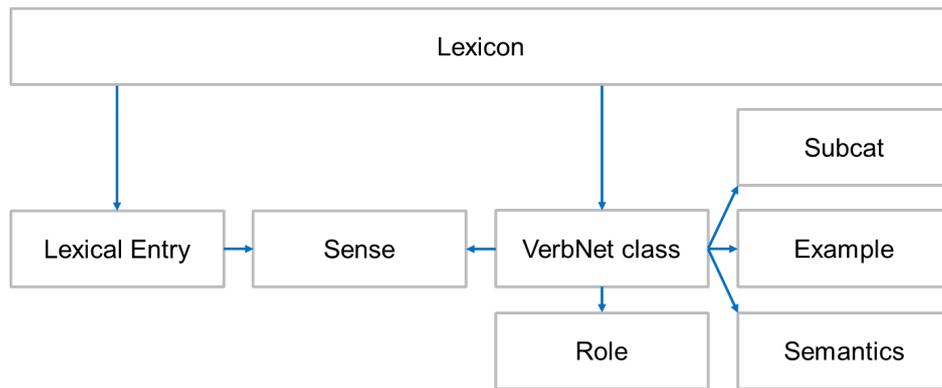


Figure 2.4: VerbNet lexicon structure.

of semantic predicates. An example is the semantic representation for *complete-55.2* (also shown in Figure 2.5 below): *complete* in the sense of *finish making or doing* is represented as $END(E, Theme) CAUSE(Agent, E)$. This can be paraphrased as follows: the *Agent* causes an event *E* that ends the action referred to by the *Theme*, for instance *smoking* in the example “*She quit smoking*”. These semantic representations can be used to infer relations like antonymy and entailment between verb senses, and describe the semantic aspect of verbs (Kipper-Schuler, 2005). Second, there is a coarse-grained set of up to 35 semantic roles based on thematic roles like *Agent*, *Patient*, and *Theme*. Additionally, the dictionary lists the corresponding semantic roles for each *syntactic* argument for the different syntactic frames in a VerbNet class. Thus, VerbNet provides an explicit model of the syntax-semantics interface that is missing from FrameNet.

VerbNet lexicon structure. The overall structure of VerbNet is very similar to the structure of FrameNet, as shown in Figure 2.4. The central information types *Lexicon Entry*, *Sense*, and *Role* are also present, and the box labeled *VerbNet class* matches the *Frame* box in Figure 2.1. The differences are in the additional representation of syntactic subcategorization in the *Subcat* box, and in the fact that the examples and semantic representations in the *Example* and *Semantics* boxes attach to the VerbNet class that subsumes several similar senses. There are no sense-specific definition glosses and examples like in FrameNet.

VerbNet lexical entry. The instantiation of the VerbNet lexicon structure for the verb *complete* is shown in Figure 2.5. The example illustrates the model of the syntax-semantics interface incorporated in VerbNet: for two different syntactic subcategorization frames, it shows an example that instantiates the frame, lists the corresponding semantic roles and the Montague semantics representation that also uses the role labels. In contrast to FrameNet, there is no sense-specific example and gloss, and from the list of other class members *accomplish*, *achieve*, *discontinue*, *quit* we can infer that the semantics of the VerbNet class

Lemma:	complete		
Part-of-speech:	verb		
Sense id:	complete-55.2		
<hr/>			
VerbNet class:	55.2		
Class members:	accomplish, achieve, complete , discontinue, quit		
Roles:	Agent [+animate +organization], Theme		
Syntax-semantics interface			
Subcat frame 1:	NP	V	NP
Example:	Wilma	completed	the assignment
Roles:	Agent	V	Theme
Semantics:	END(E, Theme) CAUSE(Agent, E)		
Subcat frame 2:	NP	V	S_ING
Example:	She	quit	smoking
Syntax:	Agent	V	Theme
Semantics:	END(E, Theme) CAUSE(Agent, E)		

Figure 2.5: Example: VerbNet lexicon entry for the verb *complete*.

55.2 are more coarse-grained than those of a FrameNet frame: in FrameNet, *accomplish* and *achieve* belong to the *Achievement* frame, while *discontinue* and *quit* belong to the *Activity_stop* frame.

VerbNet model of predicate argument structure. Unlike the FrameNet semantic roles, the inventory and the definition of the semantic roles in VerbNet is independent of the verb sense. The semantic roles have been supplemented with information on their selectional preferences, describing typical properties of role fillers for a specific role. For the *Agent* role, these include properties like *being animate* or *being an organization*. As can be seen in Figure 2.4, VerbNet does not provide annotated example sentences for each word sense, but prototypic examples for each class. The VerbNet classes group word senses with similar syntactic and semantic properties and thus, like FrameNet frames, group senses into sets. The distinctions that are made are, however, differ from those made in FrameNet, as a result of the stronger focus on syntactic similarities in VerbNet. This frequently leads to *n-to-m* mappings between FrameNet frames and VerbNet classes where $n \geq 1$ and $m \geq 1$.

An example of such a mapping from SemLink (Bonial et al., 2013) is shown in Figure 2.6. Starting from the VerbNet class 55.2 on the left-hand side of the figure, there are links to four different FrameNet frames shown in the center. The connections to the right-most column in the figure show that the frames are in turn linked to four additional VerbNet classes, which again link to new FrameNet frames like *Firing*. In the example, 5 frames map to 5 VerbNet classes, which in turn map to other frames in FrameNet, as indicated by the arrows without targets in Figure 2.6. On the level of frames and VerbNet classes, there are no closed groups or pairs that build a 1-to-1 mapping.

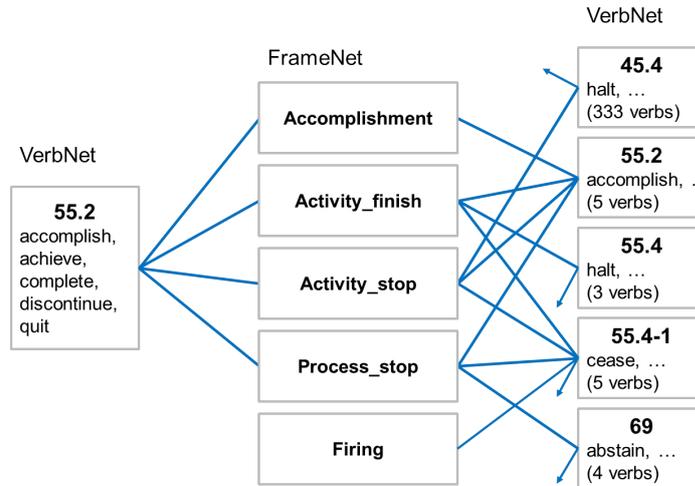


Figure 2.6: Example: FrameNet frame and VerbNet class linkings: n -to- m mappings.

Currently such mismatches between FrameNet frames and VerbNet classes are frequent. Thus, creating a mapping between VerbNet classes and FrameNet frames is not trivial. Note that VerbNet is continuously revised to make it more semantically coherent, which should make it easier to create a coherent mapping between VerbNet and FrameNet in the future.

VerbNet semantic role labeling. The applicability of the VerbNet model to other languages has been confirmed in several studies. There are VerbNet resources for Arabic (Mousser, 2010), Basque (Salaberri et al., 2014), Catalan and Spanish (Taulé et al., 2010), French (Falk et al., 2012), German (Mújdricza-Maydt et al., 2016), and Urdu (Hautli-Janisz et al., 2015). In contrast to FrameNet, VerbNet does not provide a role-labeled corpus (only the SemLink corpus (Bonial et al., 2013) provides VerbNet class and role labels). This may be one reason why VerbNet attracted less interest as a resource for automatic semantic role labeling compared to FrameNet and PropBank, which were both endorsed by shared tasks. Nevertheless, VerbNet has been suggested as an appropriate model for semantic role labeling, providing an appropriate degree of semantic abstraction compared to PropBank roles and a smaller, more coarse-grained role inventory than FrameNet (Merlo and van der Plas, 2009). Yi et al. (2007) and Loper et al. (2007) suggest that VerbNet roles generalize better across verbs than PropBank roles and should therefore be easier to learn for semantic role labeling systems. Silberer and Frank (2012) observe stronger generalization capabilities of VerbNet roles compared to FrameNet roles for the task of binding non-local roles, i.e., roles realized in a different sentence.

Despite the positive assessment of VerbNet labels for SRL, only few instances of VerbNet semantic role labeling systems can be found in the literature (Swier and Stevenson, 2005; Zafirain et al., 2008). Zafirain et al. (2008) find that VerbNet SRL is improved by per-

FrameNet

[Daimler]_{Seller} **sold**_{Financial_Transaction} [the Chrysler Group]_{Goods} [to Cerberus]_{Buyer} [for \$7.4 billion]_{Money}.

PropBank

[Daimler]_{A0} **sold**_{sell.01} [the Chrysler Group]_{A1} [to Cerberus]_{A2} [for \$7.4 billion]_{A3}.

VerbNet

[Daimler]_{Agent} **sold**_{13.1-1} [the Chrysler Group]_{Patient} [to Cerberus]_{Recipient} [for \$7.4 billion]_{Asset}.

Figure 2.7: Example: sentence annotated with FrameNet, VerbNet, and PropBank labels.

forming VerbNet class disambiguation prior to role labeling, similar to FrameNet semantic role labeling, and report lower cross-domain generalization for VerbNet compared to PropBank. Their results indicate that VerbNet SRL performance can compete with PropBank SRL. However, there are no recent VerbNet SRL systems that use state-of-the-art techniques.

2.2.2 PropBank

PropBank (Palmer et al., 2005) is a lexical resource that consists of a lexicon in the form of so-called frame files and a corpus based on the Penn Treebank annotated with the PropBank set of semantic roles. Note that the concept of *frame* used for the frame files is different from the concept of a FrameNet frame. PropBank frame files contain the rolesets for a verb, combination of roles that occur for the different verb senses for a verb lemma.

PropBank semantic roles do not have descriptive labels, obligatory arguments are for instance labeled *A0* to *A5*, and – unlike FrameNet roles – they mostly have a predicate-specific semantic interpretation that closely follows the syntactic behavior of the predicate. An exception are the roles *A0* and *A1*. Their definition follows Dowty’s theory of proto-roles (Dowty, 1986): *A0* corresponds to role fillers that display properties of a prototypical Agent, *A1* corresponds to role fillers that display properties of a prototypical Patient or Theme. Thus, *A0* and *A1* capture syntactic alternations of verbs. *A2* to *A5* receive verb-specific interpretations. The remaining arguments, for instance those represented by adjuncts in a sentence, receive labels that refer to their function and are valid across verbs, for instance *ARGM-LOC* for locations. In total, PropBank provides 25 distinct role labels, six numbered roles, the label *Arg-A* for secondary agents, and 18 optional *ARGM*-roles (Bonial et al., 2010).

Comparison between FrameNet, VerbNet, and PropBank annotations. The example in Figure 2.7 shows an example sentence labeled with predicate and role labels from FrameNet, PropBank, and VerbNet. It illustrates the above-mentioned syntactic orientation of

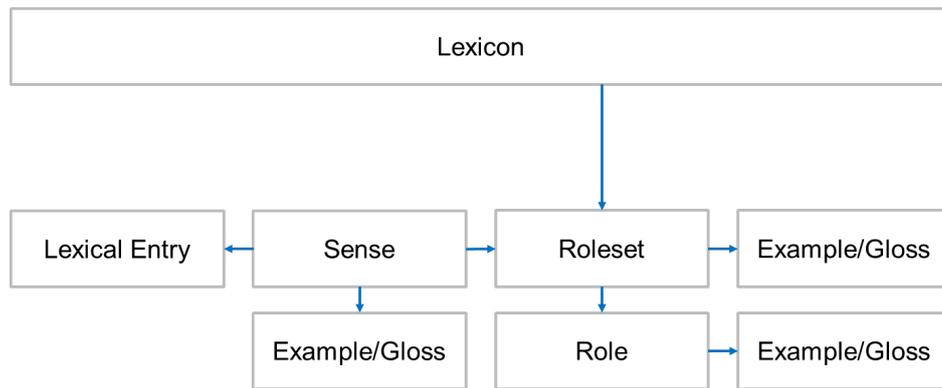


Figure 2.8: PropBank lexicon structure.

PropBank roles: the subject of the sentence receives the PropBank role label *A0*, and the direct object the role label *A1*. While VerbNet role labels are more meaningful than PropBank labels, they are more generic than the precise FrameNet role labels, as shown by the comparison between the corresponding roles *Agent* for VerbNet and *Seller* for FrameNet.

PropBank lexicon structure. The structure of the PropBank lexicon as modeled by the frame files is visualized in Figure 2.8. An instantiation of the lexicon structure for the verb *complete* is shown in Figure 2.9. The PropBank lexicon contains descriptions of predicates per lemma represented by the *Roleset* box. For each semantic predicate, there is a predicate-specific set of roles together with brief definition glosses for the predicate and the roles. The predicate provides a list of lexemes defined by lemma and part-of-speech, defining several senses and lexical entries associated with this predicate, see also *Alternative Lexemes* in Figure 2.9. Example sentences attached to the predicate level, i.e., the *Roleset* box, illustrate the usage of the predicate for PropBank annotators.

The model of verb semantics in PropBank closely follows syntactic distinctions. The PropBank lexicon model does not contain any groupings of word senses to create a more abstract generalization like VerbNet classes or semantic frames in FrameNet. PropBank started out with only verb senses, but has been extended to include also nouns, adjectives, and multiword predicates (Bonial et al., 2014).

PropBank semantic role labeling. Because it provides large annotated corpora, PropBank has become a popular role inventory for semantic role labeling, as shown by a series of CoNLL shared tasks (Carreras and Màrquez, 2005; Surdeanu et al., 2008; Hajič et al., 2009). A main difference to FrameNet semantic role labeling is that PropBank semantic role labeling typically performs role labeling directly, without preceding predicate sense labeling. PropBank resources have also been developed for other languages than English, for instance Catalan, Chinese, Czech, German, Japanese, and Spanish (Hajič et al., 2009).

Lemma:	complete
Sense id:	complete.01
Sense definition:	bring to an end
<hr/>	
Alternative Lexemes:	complete.v(erb), completion.n(noun), complete.(adj)
FrameNet Frame:	complete.v: Activity_finish, Process_completed, completion.n: Activity_finish, Activity_stop
VerbNet class:	55.2
Roles:	A0, A1
Role definition A0:	finisher
Role definition A1:	task, action coming to an end
VerbNet roles	A0: Agent, A1: theme
Example sentences:	<ol style="list-style-type: none"> 1 [The thrift holding company]_{A0} expects to complete [the transaction]_{A1} [by year.end]_{Argm-TMP}. 2 [The \$2.5 billion Byron I plant near Rockford IL]_{A1} was completed [in 1985]_{Argm-TMP}. 3 Apogee to advance cash to ensure the completion of [the project]_{A1} <p>...</p>

Figure 2.9: Example: PropBank lexicon entry for the verb *complete*.

2.2.3 SALSALSA

SALSALSA (Burchardt et al., 2006; Rehbein et al., 2012) is a German FrameNet-annotated corpus. The goal of the SALSALSA project was to annotate the syntactically annotated German TIGER corpus (Brants et al., 2004) with FrameNet frames and roles. The annotation is based on a lexical sample of lemmas from different frequency bands.

The first release (Burchardt et al., 2006) mainly covered verbs. It contains 20,380 verb instances for 493 verb lemmas, and 348 noun instances for 15 nominal lemmas. A second release, SALSALSA 2, extended the number of nouns to 15,871 for 155 lemmas, mostly verb nominalizations and relational nouns (Rehbein et al., 2012).

Differences to FrameNet. SALSALSA annotations are based on the FrameNet frame hierarchy from FrameNet releases 1.2 and 1.3, but SALSALSA added new word senses, so-called *proto-frames*, that are used when there is no FrameNet frame that covers the word sense of an annotation target. Those proto-frames are predicate-specific and do not generalize over several predicates like FrameNet frames. Proto-frames are equipped with brief definition glosses for the frame and typically few core roles. They do not contain a meaningful frame label, and were not postprocessed to create full-fledged frames. Figure 2.11 shows an example of the proto-frame *aufhoeren2-salsa*. Additionally, a small number of FrameNet frames has been adapted to accommodate new semantic arguments that are relevant for the German frame instances, see Burchardt et al. (2006). The resulting corpus covers 1,349 verb senses and 477 nominal senses, and contains 36,251 annotated frame instances for more than 1,000 different frames. The raw number of frame instances is comparable to the num-

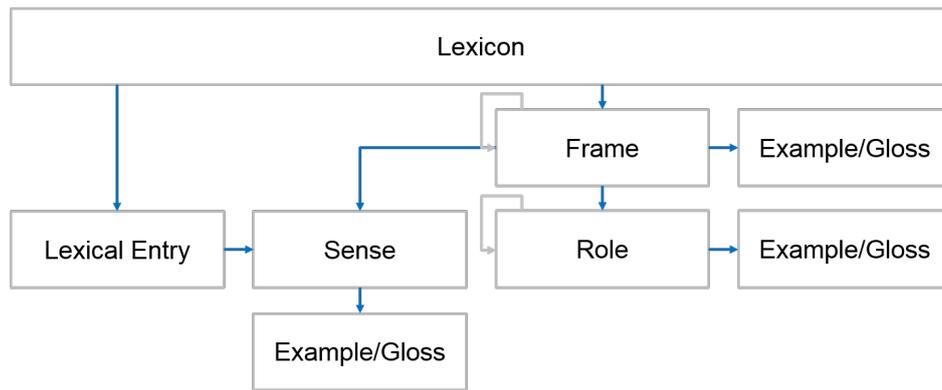


Figure 2.10: SALSAs lexicon structure.

bers for FrameNet, the lexicon coverage is however lower, because SALSAs contains fewer lemmas and predicate senses.

Incidental to the differences in construction to FrameNet, the SALSAs resource diverges from FrameNet and the FrameNet corpus in several ways: SALSAs allowed the annotator to assign several frames per target, thus modeling underspecification for ambiguous targets and annotating literal and metaphorical meaning of a target at the same time. The same applies to role labels and argument spans. Because only the most frequent lemmas in the TIGER corpus are annotated and thereby marked as lexical unit for a frame, the SALSAs lexicon is less complete than the FrameNet lexicon: the number of senses per frame is smaller than for FrameNet. Moreover, SALSAs only covers those word senses for a lemma that are attested in the TIGER corpus.

SALSAs corpus and lexicon. The SALSAs release corpus (Rehbein et al., 2012) includes the annotated corpus files, and a frame file that describes the frames.⁸ The frame file is less detailed than the FrameNet lexicon, but includes descriptions of the used (proto-)frames and roles, the frame extensions, and some German examples – likely used as reference examples for the annotators. From the corpus files and frame files, the lexicon structure shown in Figure 2.10 can be inferred from SALSAs. It is identical to the lexicon structure shown for FrameNet in Figure 2.1.

There is no equivalent of *complete* in SALSAs, therefore we use the verb *aufhören* (*cease doing something*) as a German example verb. An exemplary SALSAs lexicon entry for the verb *aufhören* with proto-frame *aufhoeren2-salsa* is shown in Figure 2.11. The proto-frame has two core roles and brief definition glosses for the frame and role labels, and there are several labeled sentences in the SALSAs corpus.

⁸The SALSAs corpus can be obtained via <http://www.coli.uni-saarland.de/projects/salsa/corpus/>.

Lemma:	aufhören
Part-of-speech:	verb
Sense id:	1409
Sense definition:	-
<hr/>	
Frame:	aufhoeren2-salsa
Frame definition:	entspricht Event_stop (= Zusammenlegung von Activity_stop und Process_stop). Beispiele: Wann werden die Kämpfe aufhören? Wann werden sie aufhören zu kämpfen..
Core roles:	Agent, Event
Role definition Agent:	Agens
Role definition Event:	The Event that stos or is stopped by the Agent.
Example sentences:	<p>1 [Die kontraproduktive Debatte]_{Event} muss AUFHÖREN, damit wir den Blick frei bekommen - für die wirklichen Schwächen, zum Beispiel bei den technischen Innovationen und der wirtschaftlichen Steuerung...</p> <p>2 [Frankreich]_{Agent} hat in Wirklichkeit nicht AUFGEHÖRT, dem Regime dem Regime militärisch, logistisch und wirtschaftlich zu helfen]_{Event}.</p> <p>...</p>

Figure 2.11: Example: SALSA lexicon entry for the verb *aufhören* (*cease doing something*) with frame *aufhoeren2-salsa*.

Erk and Padó (2006) published Shalmaneser, a SALSA semantic role labeling system for German, but there is no recent, state-of-the-art semantic role labeling system for SALSA. Likely due to the low coverage of SALSA, it only contains 493 verb lemmas, and because PropBank semantic role labeling was fostered by the CoNLL shared tasks, the development of German semantic role labeling systems focused on PropBank.

2.2.4 WordNet

WordNet (Fellbaum, 1998) was the first large-scale lexical database. Its motivation stems from psycholinguistics: grouping word senses with the same meaning into sets of cognitive concepts called *synsets*. WordNet models word senses for nouns, verbs, and adjectives in a hierarchy of synsets, and includes relations that hold between synsets and word senses. The most important relation is the *is-a* relation that links the synsets in a network of hypernyms and hyponyms, others are antonymy for opposite meanings, meronymy for part-whole relations, or entailment relations between verbs.

WordNet lexicon structure. Figure 2.12 illustrates the lexicon structure of WordNet. Instead of semantic verb classes or frames, the synset, represented by the *Synset* box, is the central class that groups frames. The grey arrows attached to the *Sense* and *Synset* boxes represent the sense- and synset relations that are essential to WordNet. In addition to its positioning in the relational network, the meaning of a synset is represented in WordNet

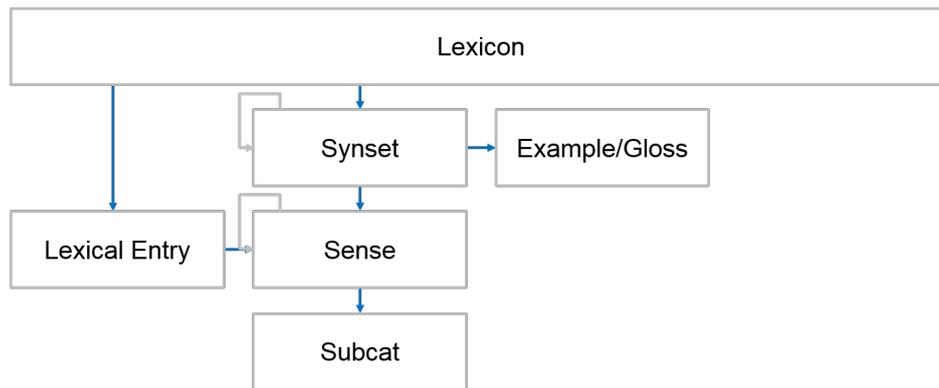


Figure 2.12: WordNet lexicon structure.

Lemma:	complete
Part-of-speech:	verb
Sense number, id:	complete#1, complete%2:30:02::
Word class:	verb.change
Sense Relations:	-
<hr/>	
Synset id:	[POS: verb] 484166
Synset definition:	Come or bring to a finish or an end
Synset members:	finish#1
Example sentences:	He finished the dishes; She completed the requirements for her Master's Degree; The fastest runner finished the race in just over 2 hours; others finished in over 4 hours
Synset relations	
-Troponym:	close#17, top#10, top off#1, get through#1, wrap up#2, ...
-Hypernym:	end#2, terminate#1, ...
-Sister term:	close out#3, finish#6, abort#1, culminate#2,...
Derivationally related:	noun completion#2, noun finisher#1, noun finish#7, ...
Subcat frame:	Somebody ---s Somebody ---s VERBing [Applies to complete] Somebody ---s something [Applies to complete] Something ---s something [Applies to complete] They won't complete the story

Figure 2.13: Example: WordNet lexicon entry for the verb *complete*.

by a sense gloss and example sentences. Information on the syntactic subcategorization is disambiguated by sense, and therefore attaches to the *Sense* box in Figure 2.12.

WordNet lexical entry. An example WordNet entry for the verb *complete* is shown in Figure 2.13. It shows a sense of the verb *complete* that shares a synset with a sense of *finish*. The example also shows that this sense of *complete* belongs to the class of change verbs (*verb.change*). This type of label is also known as *semantic field* or *supersense*. The sense number 1 in *complete#1* indicates that this sense is the most frequently labeled sense in the

SemCor reference corpus. There is no sense relation, for instance antonym, for *complete#1*, but there are several synset relations like troponymy and hypernymy. Troponyms are similar words that include information about the *Manner* in which an action is performed, e.g., *completing* something by *closing it*. Hypernyms are terms that are higher in the *is-a* hierarchy, and thus more generic. Thus, *complete#1* is a kind of *end#2*. WordNet also lists sister terms of *complete#1* in the synset hierarchy and terms that are derivationally related to the synset members, including the noun *completion*. Syntactic information is also defined for the whole synset, but can be filtered by the target sense.

WordNet does contain a simple representation of syntactic subcategorization for verbs, but no information on semantic predicate argument structure beyond simple markers of personhood or object properties of verbal arguments that are represented in WordNet's model of verb alternations. An example for such agentive and object markers are shown in the example sentence *somebody completes something* in Figure 2.13. Here, *somebody* stands for a role filled by a person and *something* for a role filled by a non-person, an object or event. A third role label available for WordNet is the label *bodyPart* that is used in constructions such as *somebody's bodyPart moves*.

Even though WordNet does not contain elaborate semantic representations, it is an important resource in the context of FrameNet semantic role labeling: the lexical coverage of WordNet is significantly larger than the coverage of FrameNet: WordNet contains more than 152,000 lexical entries for nouns, verbs, and adjectives compared to 9,702 lexical entries in FrameNet, which means that WordNet multiplies the number of lexical entries in FrameNet by 15. It also contains five times as many verb senses as FrameNet with 25,047 verb senses in WordNet and 4,670 verb senses in FrameNet. Parts of this may, however, be due to the fine sense granularity in WordNet, that also inspired the creation of OntoNotes that clusters WordNet senses (Hovy et al., 2006). Table 2.1 earlier displayed the larger coverage of WordNet in relation to several corpora. Because of its larger coverage, WordNet has been employed before to extend the coverage and improve FrameNet semantic role labeling, for instance by Burchardt et al. (2005), Shi and Mihalcea (2005), and Johansson and Nugues (2007b).

2.2.5 GermaNet

GermaNet (Hamp and Feldweg, 1997; Henrich and Hinrichs, 2010) is the German equivalent to WordNet: it groups word senses for nouns, verbs and adjectives in synsets and provides semantic relations between synsets and word senses, creating a semantic hierarchy of synsets. Its creation followed the model of WordNet with some adaptations, e.g., marking abstract concepts and enforcing cross-classification in the concept hierarchy. The general structure of GermaNet shown in Figure 2.14 is the same as the structure for WordNet shown in Figure 2.12, except for additional sense examples and definitions in GermaNet.

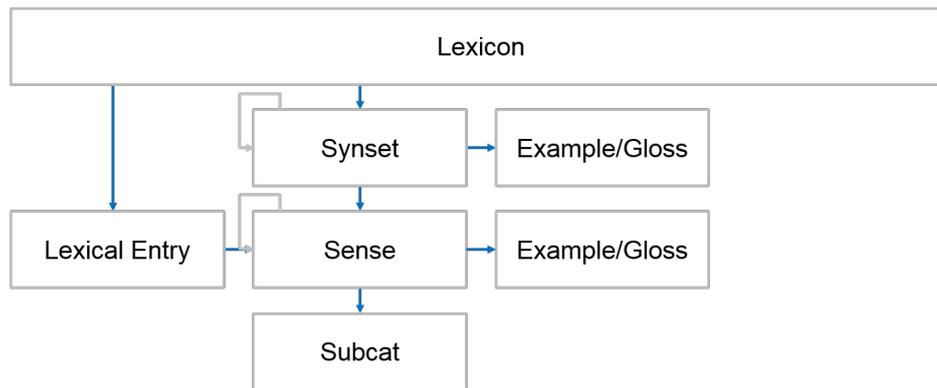


Figure 2.14: GermaNet lexicon structure.

The example in Figure 2.15 shows how this structure is instantiated for the verb *aufhören*. The contained information is very similar to the example for WordNet, but with the addition of two sense definitions.

Similar to WordNet, GermaNet initially did not contain many sense definitions. Additional sense definitions have been acquired later using a manually corrected linking of GermaNet senses to Wiktionary definitions (Henrich et al., 2014). Similar to WordNet, GermaNet does not provide much information on the semantic predicate argument structure. It however does contain information about *temporal*, *locative*, *instrumental*, *comitative*, and *manner* functions of syntactic arguments of verbs. GermaNet is also linked to WordNet on the sense level via the EuroWordNet Interlingual Index (Vossen, 1998). It is the largest expert-built lexical knowledge base for German word senses and semantic relations.

2.2.6 Wiktionary

Wiktionary is a collaboratively created dictionary available in over 500 language editions including dialects and artificial languages. It is continuously extended and revised by a community of volunteer users. The English language edition contains more than 540,000 word senses.⁹

Wiktionary lexicon structure. Wiktionary is organized like a traditional dictionary in lexical entries and word senses. Figure 2.16 illustrates the organizational structure of the English and German Wiktionary. The lexical entry, based on lemma and part-of-speech, is the entry point to the lexicon and contains senses that are defined by a definition gloss and example sentences. The grey arrow attached to the *Sense* box in Figure 2.16 indicates that Wiktionary models semantic relations between senses such as synonymy or antonymy.

⁹as of March 2016, see <http://en.wiktionary.org/wiki/Wiktionary:Statistics>.

Lemma:	aufhören
Part-of-speech:	verb
Sense number, id:	aufhören#3, 73272
Sense definition:	(1) von sich aus stoppen, anhalten (2) mit einer Handlung oder Handlungsweise nicht weitermachen
Word class:	verb.Allgemein
Sense relations:	-

Synset id:	52215
Synset definition:	aufhören etwas zu tun, etwas beenden
Synset members:	-
Example sentences:	(1) Die Studenten hörten mit den Protestaktionen nicht auf. (2) Er hatte aufgehört, sich Hoffnungen zu machen.
Synset relations	
-Hyponym:	stocken
Derivationally related:	-
Subcat frame:	(1) NN PP (2) NN AZ

Figure 2.15: Example: GermaNet lexicon entry for the verb *aufhören* (*cease doing something*).

An important information type in Wiktionary are translations of senses to other languages represented by the box labeled *Translation*.

The example in Figure 2.17 shows the lexical information associated with the verb *complete* in the English Wiktionary, including translations to other languages. For the lexical entries, Wiktionary provides information on etymology, alternative word forms, pronunciation in the form of phonetic transcription and audiofiles, and inflection. For the word senses, definitions and example sentences, as well as other lexical information, such as register (e.g., *colloquial*), and syntactic subcategorization may be available. Senses also provide translations to other languages. These are connected to lexical entries in the respective language editions via hyperlinks. This allows us to use Wiktionary as an interlingual connection between multiple languages.

In the example entry, synonyms are not attached to their corresponding sense, but to the lexical entry. For other senses in Wiktionary, synonyms are marked with the corresponding sense. Similar to the translations, synonyms are linked to the lexical entries of the synonymous words, i.e., *accomplish* and *finish* in Figure 2.17.

The quality of Wiktionary as an electronic dictionary has been confirmed by Meyer and Gurevych (2012b); Meyer (2013) also gives an overview on the usage of Wiktionary in NLP applications such as speech synthesis. They provide a detailed analysis of Wiktionary as a lexicographic resource and as a resource for NLP.

Summary The description of Wiktionary closes our introduction of lexical knowledge bases relevant to this work. The lexical knowledge bases contain different, complementary

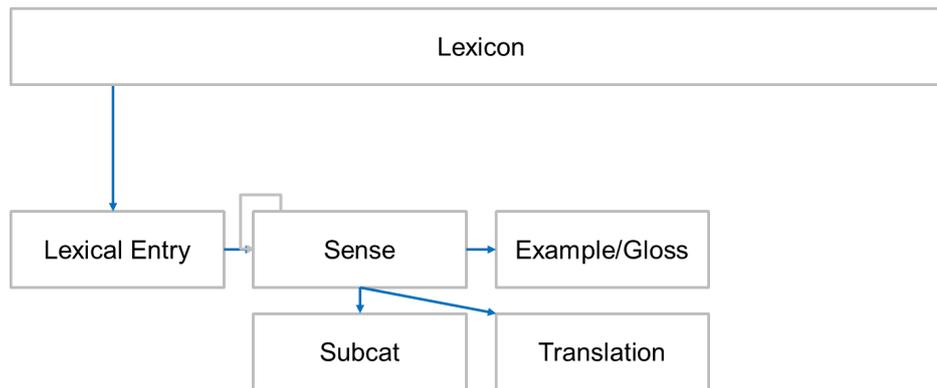


Figure 2.16: Wiktionary lexicon structure.

Lemma:	complete
Part-of-speech:	verb
Etymology:	From middle English compleet („full, complete“), ...
Alternative forms:	compleat (archaic)
Pronunciation:	IPA: /kəmˈpli:t/
Inflection:	3rd-person singular simple present <i>completes</i> , present participle <i>completing</i> , ...
Sense 1:	
Sense definition:	To finish; to make done; to reach the end
Sense example:	He completed the assignment on time
Subcat:	transitive
Translations:	Arabic: intahā, ʾakmala, Catalan: complir, Chinese: wánchéng, Czech: dokončit, ...
Sense 2:	
Sense definition:	To make whole or entire
Sense example:	The last chapter completes the book nicely
Subcat:	transitive
Translations:	Catalan: complir, Dutch: aanvullen, Esperanto: plenigi, Finnish: täydentää, lopettaa, ...
Usage notes:	This is a catenative verb that takes the gerund (-ing)
Synonyms:	accomplish, finish

Figure 2.17: Example: lexicon entry for the verb *complete* in the English Wiktionary.

types of semantic and linguistic information, but also similar information types, which on one hand motivates our work on linking them to increase their coverage, and on the other hand motivates our work on standardizing them to efficiently access the linked information.

The coverage problems introduced in detail for FrameNet also apply to the other lexical knowledge bases, but to different degrees. One way of solving these problems is to link lexical knowledge bases on appropriate semantic levels, such as the level of word sense and semantic predicate argument structure. In the next section, we introduce the concept of linking lexical knowledge bases on these semantic levels. This is a difficult task, be-

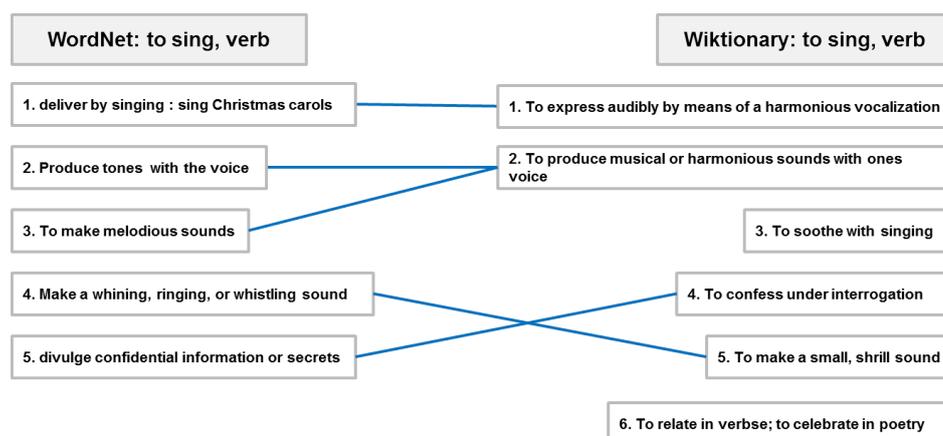


Figure 2.18: Example: word sense alignment for the verb *sing*.

cause of slightly different models of predicate argument structure in the different lexical knowledge bases and different granularity on the sense level and on the level of semantic predicate argument structure. We also present an overview on previous work in linking lexical knowledge bases that was undertaken with the goal of coverage extension of existing lexical knowledge bases, or with the goal to bootstrap lexical knowledge bases in new languages.

2.3 Linking Semantic Knowledge Bases

The various lexical resources introduced above contain complementary information types. FrameNet, for instance, contains information on semantic roles, but does not explicitly model semantic relations like synonymy and antonymy which are represented in WordNet. In order to be able to jointly use those different information types in natural language processing tasks like word sense disambiguation and semantic role labeling, the lexical resources need to be linked on the word sense level. In other words, the sense links provide so-called *semantic interoperability* of lexical resources (Ide and Pustejovsky, 2010). These linkings are also called *sense alignments*. Figure 2.18 shows an example for a sense alignment of the senses for the verb *sing* in WordNet and Wiktionary.

Some lexical resources model highly specialized types of linguistic information, for instance syntactic subcategorization or semantic predicate argument structure. The notion of semantic interoperability can be extended to these information types, e.g., mapping FrameNet frames to VerbNet classes and FrameNet semantic roles to VerbNet semantic roles. The mapping of lexical resources on the level of semantic representations, i.e., semantic predicate argument structure, is particularly relevant to this work. Figure 2.19 illustrates the

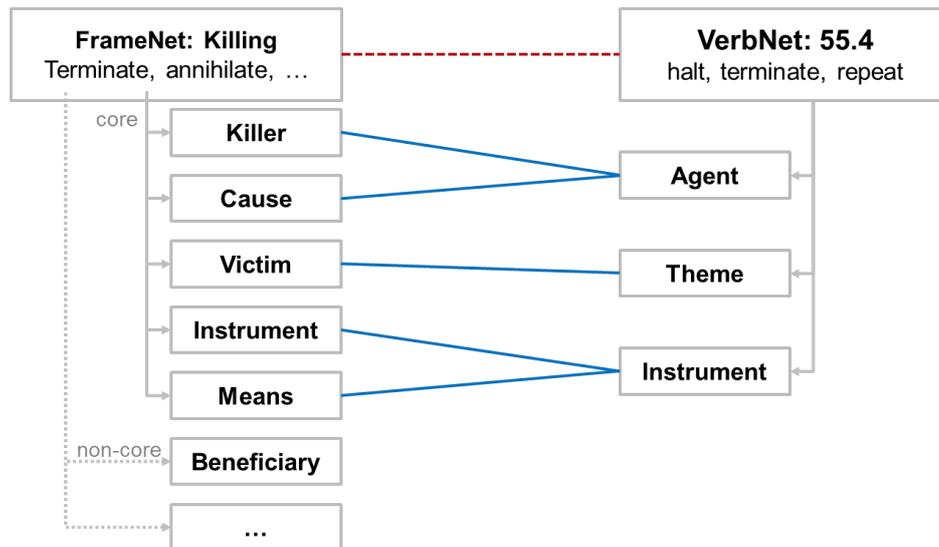


Figure 2.19: Example: predicate argument structure alignment between FrameNet and VerbNet. The red, dotted line represents the predicate-level link, the blue, solid line the argument-level links.

concept of aligning semantic predicate argument structures on the example of the FrameNet frame *Killing* and the VerbNet class *55.4* and their roles.

Manual and automatic alignments. We distinguish several types of alignments based on the way they are created: manually, semi-automatically, and automatically created alignments. Manually created alignments are reliable and have a high precision, but are expensive to obtain and require work by experts. Because of the different granularities of sense inventories and models of predicate argument structure in different lexical knowledge bases, creating these alignments manually is a difficult annotation task that requires work by experts. The example in Figure 2.18 illustrates some of these difficulties: due to the different granularity, one sense in Wiktionary is aligned to several WordNet senses; furthermore, not all senses in Wiktionary are represented in WordNet.

To avoid the manual effort, methods to create sense alignments automatically have been developed. Automatic methods use information encoded in the lexical knowledge bases to link senses, for instance sense definitions and relational structure, and have been shown to reach high precision. Semi-automatic linkings typically perform manual validation of automatically created linkings, as done in the work by [Henrich et al. \(2014\)](#) on linking the German Wiktionary to GermaNet.

Explicit and ad-hoc alignments. We additionally distinguish between two subtypes, ad-hoc and explicit alignments: ad-hoc alignments are created on-the-fly to improve NLP applications – mostly in the context of FrameNet semantic role labeling – only to be discarded

after processing. These are typically created automatically and introduced in detail in Chapter 3 in the context of using resource linkings to enhance semantic role labeling.

In this section, we focus on the presentation of explicit alignments of lexical knowledge bases. It provides the background to our own work on automatically linking FrameNet to Wiktionary on the sense level to create FrameNet resources in other languages introduced in Section 2.4 to Section 2.6, and to our work on standardizing models of semantic predicate argument structure in Section 2.7. Explicit alignments are created with the goal to extend lexical knowledge bases. Therefore, they are often published as extensions to the linked LKBs, resulting in what we call a linked lexical knowledge base (LLKB). Manually created alignments are typically also explicit alignments.

Challenges. The main challenge of all sense-level linkings is to unify sense inventories of different granularity and different degrees of lexicon coverage, leading to a) mappings of one sense in resource R to several senses in resource S (1-to- n alignment), or b) senses not being linked to the resource S (1-to-0 alignment). Both kinds of mappings are illustrated in Figure 2.18: Wiktionary sense 2 is mapped to several senses in WordNet, and the two Wiktionary senses 3 and 6 do not have a counterpart in WordNet.

For automatic linkings, the different ways of representation and semantic description of the word senses make it difficult to create a one-size-fits-all algorithms capable of linking different kinds of resources equally well. The same applies to the predicate argument structure linkings, that often build upon sense-level links. Figure 2.19 shows a predicate argument structure alignment between FrameNet and VerbNet that also contains 1-to- n alignments for several VerbNet roles, and 1-to-0 links for the non-core roles in FrameNet. Instances of n -to- m alignments on the predicate level were shown earlier in Figure 2.6.

These challenges make manual and automatic alignment of lexical knowledge bases a difficult task for machines and humans. In the remainder of this section, we present previous efforts in creating explicit linkings, i.e., linkings that are provided as extensions to the linked lexical knowledge bases, of resources on the sense level and on the level of semantic predicate argument structure.

2.3.1 Sense-level Alignments

Starting with the EuroWordNet Interlingual Index (ILI, [Vossen \(1998\)](#)), that provides synset level links between WordNet-like resources in different languages, resource alignments have been created with the goal of creating enhanced resources for various knowledge-based NLP applications, for instance word sense disambiguation and semantic role labeling. The ILI provides a manually created cross-language alignment between wordnets in eight languages. SemLink ([Bonial et al., 2013](#)) links word senses in VerbNet manually to FrameNet and PropBank. Since SemLink also contains predicate argument structure links, we will introduce it in more detail in Section 2.3.2 below.

Automatic alignment of resources gained momentum when large-scale, collaboratively created resources emerged as valuable resources for NLP, starting with the alignment between WordNet and Wikipedia by [Ruiz-Casado et al. \(2005\)](#). A great advantage of automatic alignments compared to manual linkings is that they can be easily recreated on new, extended versions of the linked resources, which is particularly important for continuously updated collaboratively created resources like Wikipedia and Wiktionary. Early approaches to automatic alignments are gloss-based and often rely on word sense disambiguation algorithms, later methods also exploit the graph structure of lexical resources. The next two paragraphs summarize work on automatic gloss-based alignments and automatic graph-based alignments respectively.

Gloss-based alignments. Gloss-based methods are supervised methods that use the similarity between textual representations of word senses to decide on alignments. Typical textual representations are the definition gloss of a word sense or the gloss expanded by additional information such as synonyms. The decision at which degree of similarity a linking between two senses is created can be unsupervised, linking sense *A* to the most similar sense *B*, or they can be supervised, relying on a manually annotated gold standard of positive and negative linkings to determine similarity thresholds for the decision whether to link a pair of senses. This way, [Ruiz-Casado et al. \(2005\)](#) and [Ponzetto and Navigli \(2009\)](#) linked WordNet synsets to Wikipedia entries or categories. Subsequent work improves on the gloss-based approach using advanced similarity measures, e.g., Personalized PageRank on the WordNet graph, for linking WordNet to Wikipedia ([Niemann and Gurevych, 2011](#)). [Meyer and Gurevych \(2011\)](#) adapt this method to a linking between Wiktionary and WordNet, and [Gurevych et al. \(2012a\)](#) expand the same approach by jointly optimizing the cutoff thresholds, creating a cross-lingual linking between WordNet and the German version of the collaboratively created dictionary OmegaWiki ([Matuschek et al., 2013](#)).

There have been several efforts at linking FrameNet and WordNet – mostly with the goal to translate FrameNet to other languages using a third, intermediary resource. [De Cao et al. \(2008\)](#) map FrameNet frames to WordNet synsets based on the embedding of FrameNet lemmas in WordNet. They use MultiWordNet ([Pianta et al., 2002](#)), an English-Italian wordnet, to induce an Italian FrameNet lexicon with 15,000 entries. [Tonelli and Pianta \(2009a\)](#) create a linking between FrameNet and WordNet called *MapNet*. They also aim to translate FrameNet to Italian via MultiWordNet. To create MapNet, [Tonelli and Pianta \(2009a\)](#) align FrameNet senses with WordNet synsets by exploiting the textual similarity of their glosses. They determine alignment candidates based on lemma overlap between the FrameNet predicate and the words in the synset. Their first goal is to increase the lexicon coverage of FrameNet by expanding the existing sets of predicates for a frame with senses from the aligned WordNet synset, Their second goal is the creation of multilingual FrameNet resources, specifically to derive Italian FrameNet predicates via the English-Italian Multi-

WordNet. The similarity measure they use is based on the stem overlap of the candidates' glosses expanded by WordNet domains, the WordNet synset, and the set of senses for a FrameNet frame. In [Tonelli and Pighin \(2009\)](#), they use these features to train a support vector machine classifier to identify valid alignments and report an F_1 -score of 0.66 on a manually annotated gold standard. They report 4,265 new English senses and 6,429 new Italian senses, which were derived via MultiWordNet.

ExtendedWordFramenet ([Laparra and Rigau, 2009, 2010](#)) is also based on the alignment of FrameNet senses to WordNet synsets. Their goal is the multilingual coverage extension of FrameNet, which is achieved by linking WordNet to wordnets in other languages (Spanish, Italian, Basque, and Catalan) in the Multilingual Central Repository. For each language, they add more than 10,000 senses to FrameNet. They rely on a knowledge-based word sense disambiguation algorithm to establish the alignment and report $F_1=0.75$ on a gold standard based on [Tonelli and Pighin \(2009\)](#). [Lopez de Lacalle et al. \(2014\)](#) use the knowledge-based word sense disambiguation methods from [Laparra and Rigau \(2010\)](#) to extend the sense links in their LKB *The Predicate Matrix*. Since they also work on linking lexical knowledge bases on the role level, we will introduce their work in more detail in Section 2.3.2 on predicate argument structure links.

[Henrich et al. \(2014\)](#) use a method based on word overlap in glosses to map German word senses and their definitions in Wiktionary to GermaNet. [Tonelli and Giuliano \(2009\)](#) and [Tonelli et al. \(2013\)](#) align FrameNet senses to Wikipedia entries with the goal to extract word senses and example sentences in Italian, exploiting the inter-language links in Wikipedia. However, Wikipedia only contains few verbal instances that are particularly important for resources like FrameNet. The alignment is restricted to nouns. Therefore, subsequent work on Wikipedia and FrameNet follows a different path and tries to match FrameNet role fillers, typically nouns, to Wikipedia entries in order to enhance the modeling of selectional preferences for FrameNet predicates ([Tonelli et al., 2012](#)).

Graph-based alignments. Graph-based alignment methods exploit the relational structure of lexical knowledge bases, often in combination with gloss-based methods. [Ferrandez et al. \(2010\)](#) add graph-based methods to the gloss-based methods for their alignments between WordNet and FrameNet. In addition to gloss similarity, they use the FrameNet and WordNet graph structure as defined by semantic relations and frame relations to model senses by their relational context and compare these contexts to create a linking.

[Matuschek and Gurevych \(2013\)](#) introduce Dijkstra WSA, an approach that uses monosemous senses to create an initial alignment between resources and then computes the shortest path between two senses in the merged resource to determine whether those senses should be aligned. They use the gloss-based method by [Gurevych et al. \(2012a\)](#) as a backoff. Parameters determined on a gold standard set include the maximal path length of acceptable alignments and similarity thresholds for the gloss-based method. Dijkstra WSA requires a

certain degree of relational structure present for the resources to be aligned successfully. This is not the case for some resources: while FrameNet provides a detailed frame hierarchy, from which some relations between senses can be inferred, sense relations are not encoded explicitly.

[Pilehvar and Navigli \(2014\)](#) address this problem by integrating an approach to ontologization that creates a WordNet-like graph for those resources that do not already possess a rich graph structure. They integrate structural similarities derived from the graph structure and gloss similarities to a combined similarity measure and achieve good results for a supervised setup that tunes decision thresholds for alignments on a training set, and for an unsupervised setup that accepts an alignment if its similarity is higher than the middle point on the similarity scale. With this approach, they achieve state-of-the-art results for the alignment of WordNet with Wikipedia, Wiktionary and the collaboratively created multilingual dictionary OmegaWiki.

Full integration of linked LKBs. Most sense linkings result in a linking of resource pairs, leaving the source resources intact, but there also are approaches that aim at fully integrating several resources into a new resource. Thus, *BabelNet* integrates WordNet and Wikipedia ([Navigli and Ponzetto, 2012](#)), [De Melo and Weikum \(2009\)](#) integrate wordnets in several languages, Wiktionary, OmegaWiki, multilingual thesauri and translation dictionaries into their *Universal WordNet*, and *YAGO* integrates Wikipedia, WordNet, and the GeoNames database with a focus on spatial and temporal information for events ([Hoffart et al., 2013](#)). The advantage of such integrated LLKBs is that all integrated information can be accessed directly. On the downside, knowledge on the source of specific types of information, and information types that have been neglected during the integration, or were added later for collaboratively created knowledge bases cannot be accessed. This speaks in favor of modeling resource integration via a large array of alignments on different levels of information, i.e., on the level of senses, predicates and roles. In the next subsection, we introduce explicit linkings on the level of semantic predicate argument structure.

2.3.2 Predicate Argument Structure Alignments

Most of the alignments introduced so far focus on the sense level. Similar to the correspondences between senses in different lexical resources, the models of predicate argument structure between different resources can also be linked.

Predicate-level links. If two lexical resources explicitly model semantic predicates, the predicate level can be linked. Thus, FrameNet frames can be linked to VerbNet classes. For PropBank, which does not provide a semantic abstraction for its predicates, semantic predicates are synonymous to word senses. Therefore, an alignment of FrameNet frames

to PropBank predicates should be a 1:n alignment, linking a single FrameNet frame to several PropBank predicates. In Figure 2.6, we showed predicate-level alignments between FrameNet frames and VerbNet classes, highlighting n-to-m alignments between them.

Argument-level links. Besides predicates, role labels can be linked across resources to create an alignment on the level of semantic arguments. The correspondence of roles across different models of semantic predicate argument structure was already mentioned in Section 2.2 which introduced the relevant lexical knowledge bases. Thus, for instance, the *Seller* role of the frame *Financial_transaction* can be linked to the role *Agent* for class *get-13.5.1* in VerbNet, or to the agentive *A0* in PropBank, as shown in Figure 2.7. Another example for role-level alignments is shown in the example in Figure 2.19, which displays an alignment between the roles of the *Killing* frame and VerbNet class *55.4*.

Cross-language predicate argument structure links. The most popular lexical resources with predicate argument structure models are FrameNet, VerbNet, and PropBank for English, and SALSA for German. As the model of predicate argument structure in FrameNet is considered largely language-independent, cross-lingual links, for instance between SALSA and FrameNet, are in general possible.

SALSA reuses FrameNet frames and roles. Therefore, establishing a frame- and role-level alignment between FrameNet and SALSA would be straightforward, were it not for an older version of FrameNet used for SALSA frames compared to FrameNet release 1.5. Frame labels have not been stable between FrameNet release 1.3 and 1.5, so not all frames can be mapped directly based on their frame labels.

There are only few resources that provide alignments on the predicate argument structure-level, e.g., Palmer (2009) and Lopez de Lacalle et al. (2014). The next paragraphs describe previous work that provides predicate argument structure links between FrameNet, VerbNet, and PropBank.

SemLink. SemLink (Palmer, 2009; Bonial et al., 2013) is an ongoing effort to link lexical resources on the level of semantic information, including semantic predicate argument structure. It links FrameNet release 1.5, VerbNet release 3.2, PropBank, and the OntoNotes sense groupings, coarse-grained sense representations that were created manually based on clustering WordNet senses to ensure high agreement during annotation (Hovy et al., 2006; Pradhan et al., 2007a). On the role level, SemLink provides two pairs of resource linkings: PropBank roles are aligned to VerbNet roles, and VerbNet roles are aligned to FrameNet roles. Via transitive linkings, a linking of FrameNet to PropBank can be inferred. On the predicate level, VerbNet classes are mapped to FrameNet frames, and predicates from PropBank, called *rolesets*, are mapped to VerbNet classes. Since SemLink includes information on the class members, i.e., lemmas, for the linked VerbNet classes, sense-level alignments

can be inferred from predicate-level alignments and the lemma information. In addition to an explicit representation of these mappings, SemLink includes some corpus text from the Wall Street Journal annotated with senses and roles from PropBank, VerbNet, and FrameNet, as well as OntoNotes senses. The creation of SemLink is mostly a manual effort, but FrameNet corpus annotations were created semi-automatically based on the existing sense- and role-level linkings.

SemLink contains a) 1,716 links between VerbNet classes and FrameNet frames, b) 1,663 links between VerbNet and FrameNet role labels, c) 5,591 links between PropBank predicates and VerbNet class, and d) 12,551 links between predicate-specific PropBank roles and VerbNet roles. These can be used to infer links between FrameNet and PropBank semantic predicates and roles.

SemLink is continuously developed in order to keep it up to date with the developments and changes in the source resources. It is an excellent resource for comparing the differences of the included lexical-semantic resources, for instance identifying gaps in the lexicon coverage for VerbNet compared to PropBank (Bonial et al., 2013), and at the same time it provides data annotated with parallel annotations in the different role schemata that can be used for experimental evaluation. A disadvantage is that the linkings are not complete and contain errors, which are aggravated when using the transitively derived linking from PropBank to FrameNet via VerbNet (Kshirsagar et al., 2015). The sense- and role-level links between VerbNet and FrameNet in SemLink play a crucial role in the automatic generation of training data labeled with sense and role labels introduced in Chapter 3.

Predicate Matrix. The Predicate Matrix (Lopez de Lacalle et al., 2014) is an effort to automatically extend SemLink. The extension is performed on the sense level, and on the semantic level of predicate argument structure. Lopez de Lacalle et al. (2016) extend the work by Lopez de Lacalle et al. (2014). They combine different methods to extend the pairwise linkings between FrameNet, VerbNet, PropBank, and WordNet in SemLink. These methods depend on the properties of the specific pairs of lexical knowledge bases. Lopez de Lacalle et al. (2016) use three main approaches, 1) word sense disambiguation methods to extend sense-level links 2) using existing links on the predicate and role level to fill in missing role-level links, and 3) instance-based methods that use multiple sense or role annotations on the same sentence to infer new sense- or role-level links.

For their first approach, Lopez de Lacalle et al. (2016) use several algorithms for graph-based sense disambiguation to link additional senses from FrameNet and VerbNet to WordNet, backing off to a gloss-based sense alignment method.

Their second approach aims to fill in gaps in SemLink by exploiting information from related predicates and roles that are aligned in SemLink: they use the FrameNet frame and VerbNet class pairs to complete the role-level links for each frame: for every predicate in VerbNet whose class is aligned to a FrameNet frame, they take all those roles that are not

yet aligned to a FrameNet role and perform a most-frequent-label assignment based on the existing role linkings in SemLink and the roles licensed by the FrameNet frame.

The third approach by [Lopez de Lacalle et al. \(2016\)](#), instance-based methods, is used for both sense- and role-level links. For linking WordNet senses to PropBank, they run a WordNet word sense disambiguation system on corpora labeled (manually and automatically) with PropBank roles and aggregate the resulting pairs of sense labels on the same target to create new sense alignments. The same method is used for linking FrameNet to PropBank on the role level, applying FrameNet and PropBank SRL systems to corpora that are manually labeled with the other role schema to establish new links between role instances. [Lopez de Lacalle et al. \(2016\)](#) use filters based on the frequencies of the observed links to increase the quality of the resulting role-level alignment.

A variant of this approach is used to induce new role-level links between VerbNet and FrameNet: exploiting the fixed word order in the English language, [Lopez de Lacalle et al. \(2016\)](#) align annotated examples for a FrameNet frame to the syntactic-semantic patterns associated with the VerbNet class linked to this frame. These patterns have very simple structures like *Agent verb Theme*. The corresponding structure extracted from a FrameNet example sentence has the form *Seller verb Goods*. Aligning those two structures on the token level leads to an alignment of *Agent* to *Seller* and *Theme* to *Goods*.

With these methods [Lopez de Lacalle et al. \(2016\)](#) extend the number of sense alignments between VerbNet and FrameNet to 5,462, between VerbNet and PropBank to 5,462, and between FrameNet and PropBank to 4,163. This is an increase by respectively 47%, 10%, and 61% compared to the numbers in SemLink. They double the number of alignments between VerbNet and FrameNet roles compared to SemLink, leading to 14,259 role linkings, and triple the previously small number of alignments between FrameNet and PropBank roles, leading to 14,194 role-level links.

[Lopez de Lacalle et al. \(2016\)](#) evaluate their automatic methods intrinsically using SemLink as the gold standard, reporting high precision between 0.76 and 0.88 for methods based on word sense disambiguation and gap-filling, and slightly lower precision between 0.64 and 0.76 for the instance-based methods that suffer from errors in the automatic processing involved. Since most of their methods use information from SemLink, they report averages of the evaluation scores when evaluating each sense or role link in a leave-one-out fashion. This evaluation is biased towards role pairs seen in the linkings and thus is not a good predictor for the quality of previously unseen role mappings. [Lopez de Lacalle et al. \(2016\)](#) neither perform an extrinsic evaluation, for instance evaluating the contributions of the inferred links to semantic role labeling performance, nor a post-hoc evaluation of the new established sense and role alignments that is free of the bias to frequent senses and roles that may be incorporated in the SemLink test set. Thus their results only give an approximate evaluation of quality of newly established linkings. Nevertheless, their approach fills

an important research gap for automatic alignments on the predicate argument structure level.

Summary on linking lexical knowledge bases. In this section, we introduced the tasks of linking lexical knowledge bases on the levels of word sense and predicate argument structure. We discussed the difficulties that arise when aligning lexical knowledge bases with different granularity at the sense- or predicate-level. For automatic alignments, the different degrees to which relevant information, e.g., definition glosses and graph structure based on semantic relations, is available for different lexical knowledge bases also influences which lexical knowledge bases can be aligned automatically, and which methods should be selected for this task.

For sense-level alignments, several automatic alignment methods have been explored, e.g. gloss-based approaches, graph-based approaches and hybrid combinations of both, leading to generic methods for sense alignments of various lexical knowledge bases and resulting in a large number of aligned pairs of LKBs.

There is only few work on creating alignments on the level of predicate argument structure: SemLink, linking FrameNet, VerbNet, and PropBank, has been created manually. There are first research efforts to fill the gaps in SemLink and extend it automatically. They exploit a) existing alignments and specific properties of the linked LKBs, for instance the examples in VerbNet, and b) use automatic SRL systems to infer linkings from multiply annotated texts. Intrinsic evaluation on held-out sets indicates that the induced linkings are of high quality, but an extrinsic evaluation is still lacking.

In the next section, we present the first automatic sense-alignment between FrameNet and Wiktionary with the goals to extend the FrameNet lexicon for English and to create FrameNet resources for new languages.

2.4 Extending and Translating FrameNet using Wiktionary as Interlingua

This section gives an overview of our research on automatically extending FrameNet and inferring FrameNet resources for other languages. We automatically create a sense-level alignment between FrameNet and the English Wiktionary that is used to infer a FrameNet lexicon for German. The subsequent sections, Section 2.5 and Section 2.6, present details on the creation of the alignment and the German FrameNet lexicon. The results of this work were previously published as [Hartmann and Gurevych \(2013b\)](#).

Problem description. The automatic alignment between FrameNet and Wiktionary is motivated by the coverage bottleneck of lexical knowledge bases, in particular FrameNet, as discussed in the previous sections. Expert-built lexical knowledge bases such as FrameNet

are expensive to create. Previous cross-lingual transfer of FrameNet used corpus-based approaches, or resource alignments to multilingual *expert-built* resources, such as EuroWordNet. The latter approach indirectly also suffers from the high cost and constrained coverage of expert-built resources.

Solution: Wiktionary as interlingual connection. Our suggested solution to the problem of multilingual extension of lexical-semantic knowledge bases is to use Wiktionary, a collaboratively created dictionary, as a connection between languages. Wiktionary provides high-quality lexical information on all parts-of-speech, for instance glosses, sense relations, and syntactic subcategorization. Like Wikipedia, it is continuously extended and contains translations to hundreds of languages, including low-resource ones. To our knowledge, Wiktionary has not been utilized as an interlingual index for the cross-lingual extension of lexical knowledge bases.

The monolingual linking of FrameNet to the English Wiktionary is the first step of a novel method for the creation of bilingual FrameNet lexicons based on an alignment to Wiktionary. We demonstrate our method on the language pair English–German and present the resulting resources, a lemma-based multilingual and a sense-disambiguated German–English FrameNet lexicon.

The understanding of lexical-semantic resources and their combinations, e.g., how alignment algorithms can be adapted to individual resource pairs and different POS, is essential for their effective use in NLP and is a prerequisite for the later in-task evaluation and application. To enhance this understanding for the presented resource pair FrameNet–Wiktionary, we perform a detailed analysis of the created resource and compare it to existing FrameNet-like resources for German.

2.4.1 Method Overview

Our method consists of two steps visualized in Figure 2.20. The first step is presented in detail in Section 2.5. It creates a novel sense alignment between FrameNet and the English Wiktionary following Niemann and Gurevych (2011). Thus, the FrameNet sense of *to complete* with the frame *Activity_finish* is aligned to the sense of *to complete* in Wiktionary meaning *to finish*.

This step establishes Wiktionary as an interlingual index between FrameNet senses and lemmas in many languages, and builds the foundation for the bilingual FrameNet extension. It results in a basic multilingual FrameNet lexicon FNWK_{xx} with translations to lemmas in 283 languages. An example: by aligning the FrameNet sense of the verb *complete* with gloss *to finish* with the corresponding English Wiktionary sense, we collect 39 translations to 22 languages, e.g., the German *fertigmachen* and the Spanish *terminar*. This step additionally extends FrameNet by the linguistic information in the English Wiktionary and expands the FrameNet lexicon.

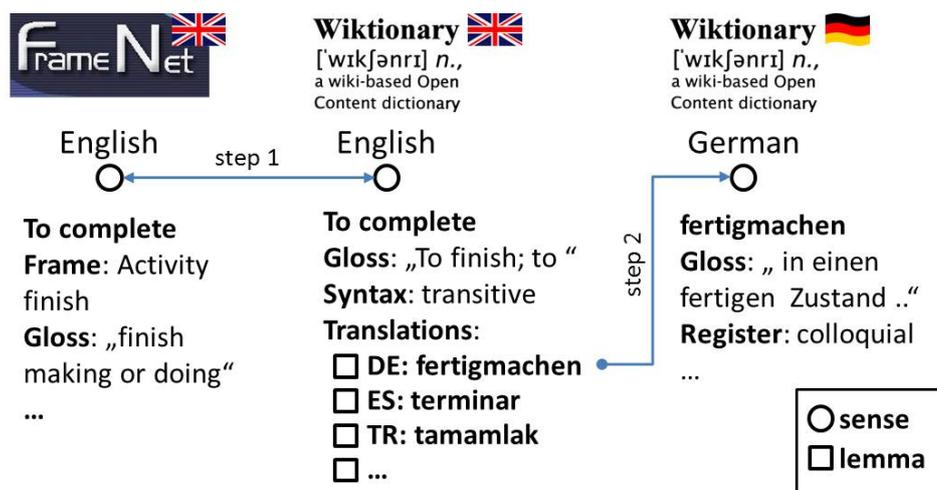


Figure 2.20: Method overview: using Wiktionary as interlingual connection to extend FrameNet to other languages.

The second step, presented in detail in Section 2.6, concerns the disambiguation of the translated lemmas with respect to the target language Wiktionary in order to retrieve the linguistic information of the corresponding word sense in the target language Wiktionary (Meyer and Gurevych, 2012a). We evaluate this step for English and German and create the bilingual FrameNet lexicon FNWKde. For the example sense of *complete*, we extract lexical information for the word sense of its German translation *fertigmachen*, for instance a German gloss, an example sentence, register information (*colloquial*), and synonyms, e.g., *beenden*. The same method is used to disambiguate targets of sense relations in the English Wiktionary, thus extending the English FrameNet with new senses.

2.4.2 Related Work

Related work concerns automatic sense alignments and the automatic creation of FrameNet resources for languages other than English. We already introduced related work on automatic sense alignments in Section 2.3.1. The main difference to our work is that previous alignments involving Wiktionary did not focus on aligning verb senses. In the following paragraph, we introduce related work on creating FrameNet resources for new languages.

Creating FrameNets in new languages. There are two main lines of research in bootstrapping a FrameNet for languages other than English.

The first, corpus-based approach is to automatically extract word senses in the target language based on parallel corpora and frame annotations in the source language. In this vein, Padó and Lapata (2005a) propose a cross-lingual FrameNet extension to German and French. Johansson and Nugues (2005) and Johansson and Nugues (2006) do the same for

Spanish and Swedish, and [Basili et al. \(2009\)](#) for Italian. These methods are introduced in greater detail in Section 3.3 that describes various methods for the generation of role-labeled training data. [Padó and Lapata \(2005a\)](#) observe that their approach suffers from polysemy errors, because lemmas in the source language need to be disambiguated with respect to all the frames they evoke. To alleviate this problem, they use a disambiguation approach based on the most frequent frame, while [Basili et al. \(2009\)](#) use distributional methods for frame disambiguation. Our approach to FrameNet translation is based on resource alignments on the sense level and therefore explicitly aims to avoid such errors.

The second line of work is resource-based: FrameNet is aligned to multilingual resources in order to extract senses in the target language. These approaches have already been introduced in detail in Section 2.3.1. Using monolingual resources, this approach has also been employed to extend FrameNet coverage for English ([Shi and Mihalcea, 2005](#); [Johansson and Nugues, 2007b](#); [Ferrandez et al., 2010](#)).

Finally, there have been suggestions to combine the corpus-based and the resource-based approaches: [Borin et al. \(2012\)](#) do this for Finnish and Swedish. They bootstrap a Finnish FrameNet based on Swedish FrameNet, Finnish and Swedish wordnets, and aligned bilingual corpora. They create a preliminary Finnish FrameNet with 2,694 senses, thus proving the feasibility of their approach.

[Mouton et al. \(2010\)](#) directly exploit the translations in the English and French editions of Wiktionary to extend the French FrameNet. They match the FrameNet senses to Wiktionary lexical entries, thus encountering the problem of polysemy in the target language. To solve this, they define a set of filters that control how target lemmas are distributed over frames, increasing precision at the expense of recall. They reach a precision P of 0.74, but the recall R is only 0.3, which leads to an F_1 score of 0.42. While their approach is in theory applicable to other languages, our approach goes beyond this by laying the ground for simultaneous FrameNet extension in multiple languages using Wiktionary as an interlingual connection to FrameNet.

Summary. This section introduced and illustrated the general concept of our approach to automatically generate FrameNet lexica for new languages using Wiktionary as an interlingual connection. In the next two sections, we present the two steps of our approach in detail and show an exemplary application to German, creating a German FrameNet lexicon. In Section 2.5, we present Step 1 of Figure 2.20, the automatic alignment of FrameNet and the English Wiktionary. Section 2.6 then presents Step 2 of Figure 2.20, which uses a sense-alignment of the English Wiktionary to the German Wiktionary to create a German FrameNet lexicon.

2.5 Creating the FrameNet – Wiktionary Alignment

This section presents the creation of an automatic alignment between FrameNet and the English Wiktionary. It thus describes Step 1 of our method to translate FrameNet to other languages as shown in Figure 2.20.

2.5.1 Automatic FrameNet – Wiktionary Alignment

The sense alignment method we use to align FrameNet and Wiktionary follows the gloss-based method introduced by Niemann and Gurevych (2011) for their alignment between WordNet and Wikipedia. They align senses in WordNet to Wikipedia entries in a supervised setting based on semantic similarity of sense glosses.

One reason to use their method is that it allows zero alignments (1-to-0) and one-to-many alignments (1-to-n). This is crucial for obtaining a high-quality alignment of heterogeneous resource pairs, such as the presented one, because their sense granularity and coverage can diverge a lot.

Alignment method. The alignment algorithm consists of two steps. The *candidate extraction* step iterates over all FrameNet senses and matches them with all senses from Wiktionary which have the same lemma and thus are likely to describe the same sense.

This step yields a set of candidate sense pairs C_{all} . In the *classification* step, a similarity score between the textual information associated with the senses in a candidate pair, e.g., their gloss, is computed and a threshold-based classifier decides for each pair whether it constitutes a valid alignment.

Niemann and Gurevych (2011) combine two different types of similarity: (i) cosine similarity on bag-of-words vectors (COS) and (ii) a personalized PageRank-based similarity measure (PPR). The PPR measure (Agirre and Soroa, 2009) maps the glosses of the two senses to a semantic vector space spanned by WordNet synsets and then compares them using the chi-square measure.

The semantic vectors \mathbf{ppr} are computed using the personalized PageRank algorithm on the WordNet graph. They determine the important nodes in the graph as the nodes that a random walker following the edges visits most frequently:

$$\mathbf{ppr} = cM\mathbf{ppr} + (1 - c)\mathbf{v}_{\mathbf{ppr}} \quad (2.1)$$

where M is a transition probability matrix between the n WordNet synsets, c is a damping factor, and $\mathbf{v}_{\mathbf{ppr}}$ is a vector of size n representing the probability of jumping to the node i associated with each \mathbf{v}_i . For personalized PageRank, $\mathbf{v}_{\mathbf{ppr}}$ is initialized in a particular way: the initial weight is distributed equally over the m vector components (i.e., synsets) associated with a word in the sense gloss, other components receive a 0 value.

	verb	noun	adjective	all POS
Cohen’s κ	0.65	0.77	0.8	0.72

Table 2.3: Inter-rater agreement for sense alignment gold standard.

For each similarity measure, [Niemann and Gurevych \(2011\)](#) determine a threshold (t_{ppr} and t_{cos}) independently on a manually annotated gold standard. The final alignment decision is the conjunction of two decision functions:

$$a(s_s, s_t) = \text{PPR}(s_s, s_t) > t_{ppr} \& \text{COS}(s_s, s_t) > t_{cos} \quad (2.2)$$

The presented method differs from [Niemann and Gurevych \(2011\)](#) in that it uses a joint training setup which determines t_{ppr} and t_{cos} to optimize classification performance directly, as proposed in [Gurevych et al. \(2012a\)](#):

$$(t_{ppr}, t_{cos}) = \text{argmax}_{(t_{ppr}, t_{cos})} F_1(a), \quad (2.3)$$

where F_1 is the maximized evaluation score and a is the decision function in Equation 2.2.

Candidate extraction. To compile the candidate set, we paired all senses from both resources that have identical lemma-POS combinations. FrameNet senses are defined by a lemma, a definition gloss, and a frame. Wiktionary senses are defined by a lemma and a gloss. We find two candidate senses in Wiktionary for the FrameNet sense *Activity_finish* of the verb *complete*, the senses *to finish* and *to make whole*. There are on average 3.7 candidates per FrameNet sense. The full candidate set C_{all} contains more than 44,000 sense pairs and covers 97% of the 11,942 senses in FrameNet.

Gold standard creation. To create a gold standard, we sampled 2,900 candidate pairs from C_{all} . The properties of the gold standard mirror the properties of C_{all} : the sampling preserved the distribution of POS in C_{all} , i.e., around 40% verbs and nouns, and 12% adjectives, and the average numbers of candidates per FrameNet sense. This ensures that highly polysemous words as well as words with few senses are selected.

Two human raters annotated the sense pairs based on their glosses. The annotation task consisted in a two-class annotation: *Do the presented senses have same meaning - (YES|NO)*. The raters received detailed guidelines and were trained on around 100 sense pairs drawn from the sample. We computed Cohen’s κ to measure the inter-rater agreement between the two annotators. It is $\kappa=0.72$ on the full set, an acceptable score according to [Artstein and Poesio \(2008\)](#). An additional expert annotator disambiguated ties.

For comparison: [Meyer and Gurevych \(2011\)](#) report $\kappa=0.74$ for their WordNet – Wiktionary gold standard, and [Niemann and Gurevych \(2011\)](#) $\kappa=0.87$ for their WordNet –

Wikipedia gold standard. These gold standards only consist of nouns, which appear to be an easier annotation task than verb senses. This is supported by our analysis of the agreement by POS – see Table 2.3: the agreement on nouns and adjectives lies between the two agreement scores previously reported on nouns. Thus our annotation is of similar quality. Only the agreement on verbs is slightly below the acceptability threshold of 0.67 (Artstein and Poesio, 2008). The verb senses are very fine-grained and thus present a difficult alignment task. To ensure high quality, an expert annotator corrected the verbal part of the gold standard set. After removing the training set for the raters, the final gold standard contains 2,789 sense pairs. 28% of these are aligned. The final gold standard is publicly available, see a list of resource links in Appendix A.

Alignment experiments. In the following paragraphs we describe the set of experiments that led to the creation of a sense alignment between FrameNet and Wiktionary. We present experiment results and provide a detailed error analysis.

Parameter setting. We determined the best setting for the alignment of FrameNet and Wiktionary in a ten-fold cross-validation on the gold standard.

Besides the parameters for the computation of the PPR vectors using the publicly available UKB tool by Agirre and Soroa (2009), the main parameter in the experiments is the textual information that is used to represent the senses. For the FrameNet senses, we used the *lemma-pos*, *sense gloss*, *example sentences*, *frame label* and *frame definition* as textual features; for the Wiktionary senses, we considered *lemma-pos*, *sense gloss*, *example sentences*, *hyponyms* and *synonyms*. The similarity scores were computed on tokenized, lemmatized and stopword-filtered texts using pre-processing tools from DKPro Core (Eckart de Castilho and Gurevych, 2014), e.g., TreeTagger for lemmatization and a list-based stop word filter.

First, we evaluated models for COS and PPR independently based on various combinations of the textual features listed above. We then used the textual features of the best-performing single models to train the model that jointly optimizes the thresholds for PPR and COS, see Equation 2.3.

Evaluation setup. For the evaluation, we compute precision P, recall R and F_1 on the positive class, i.e., *aligned=true*. Precision P is the number of pairs correctly aligned divided by all aligned pairs, recall R is the number of correctly aligned pairs divided by the number of aligned pairs in the gold standard. F_1 is the harmonic mean of precision and recall.

Experiment results. Table 2.4 shows the evaluation scores of the best single models and the best joint model on the gold standard. We achieved the highest precision and F_1 -score for COS using all available features, but excluding FrameNet *example sentences* because they

introduce too much noise. Many senses in FrameNet are equipped with a large number of example sentences covering various topics. These sentences are based on corpus instances and were not specifically selected to explain the meaning of the sense, like example sentences in Wiktionary. Adding them to the textual representation of a FrameNet sense leads to spurious recognition of similarities, i.e., false positives, when comparing them to the smaller and more concise textual representations of Wiktionary senses. Adding the *frame label* and *frame definition* to the often short glosses provides a richer sense representation for the COS measure.

The best-performing PPR configuration uses *sense gloss* and *lemma-pos*. For the joint model, we used the best single PPR configuration, and a COS configuration that uses *sense gloss* extended by Wiktionary *hypernyms*, *synonyms*, and FrameNet *frame label* and *frame definition*, to achieve the highest score, an F_1 -score of 0.739.

We compare the performance of our alignment on the gold standard to two baselines. First, there is Random-1, a baseline which randomly selects one target sense from the candidate set of each source sense. We also consider the more competitive Wiktionary first sense baseline WKT-1. This baseline is guided by the heuristic that more frequent senses are listed first in Wiktionary (Meyer and Gurevych, 2010). It is a stronger baseline with an F_1 -score of 0.65, as shown in Table 2.4.

We consider human alignment performance as the upper bound for the sense alignment task. To derive this upper bound, *UBound* in Table 2.4, we computed the F_1 score between the two annotators according to Hripcsak and Rothschild (2005): instead of the gold standard alignment and the system decisions, the alignment decisions of annotator 1 and annotator 2 are used to compute the F_1 score. This metric is symmetric and results in the same score independent of which annotator is declared to be the gold standard alignment. For more than two annotators, Hripcsak and Rothschild (2005) suggest to compute the average of the F_1 scores for all pairs of annotators.

As the evaluation set mirrors the part-of-speech distribution in FrameNet and is sufficiently large, unlike earlier alignments, which typically focus on a particular part-of-speech or ignore the part-of-speech distinction in the evaluation, an analysis of the performance by part-of-speech is possible. The BEST JOINT model performs well on nouns, slightly better on adjectives, and worse on verbs, see Table 2.4. For the baselines and *UBound* the same applies, with the difference that adjectives receive even better results in comparison. This fits in with the perceived degree of difficulty according to the observed polysemy for the parts-of-speech: for verbs we have many candidate sets with two or more candidates, i.e., we observe higher polysemy, while for nouns and even stronger for adjectives, many small candidate sets occur, which stand for an easier alignment decision. This is in line with the reported higher complexity of lexical resources with respect to verbs and greater difficulty in alignments and word sense disambiguation (Laparra and Rigau, 2010).

alignment method	verb	noun	adjective	all POS
Precision P				
BEST COS	0.639	0.778	0.706	0.703
BEST PPR	0.66	0.754	0.729	0.713
BEST JOINT	0.677	0.766	0.742	0.728
Random-1 BL	0.503	0.559	0.661	0.557
WKT-1 BL	0.620	0.664	0.725	0.66
Recall R				
BEST COS	0.658	0.758	0.754	0.715
BEST PPR	0.666	0.724	0.754	0.699
BEST JOINT	0.683	0.783	0.83	0.75
WKT-1 BL	0.581	0.65	0.75	0.64
BEST COS	0.658	0.758	0.754	0.715
F₁				
BEST COS	0.648	0.768	0.729	0.709
BEST PPR	0.663	0.739	0.741	0.706
BEST JOINT	0.68	0.775	0.784	0.739
Random-1 BL	0.487	0.552	0.672	0.549
WKT-1 BL	0.60	0.657	0.737	0.65
UBound	0.735	0.834	0.864	0.797

Table 2.4: Word sense alignment performance by POS.

The performance of BEST JOINT on all POS is $F_1=0.739$, which is significantly higher than the WKT-1 baseline: $p<0.05$ according to McNemar’s test. The performance on nouns, i.e., $F_1=0.775$, is on par with the results reported by [Niemann and Gurevych \(2011\)](#) for their alignment of nouns in Wikipedia and WordNet, e.g., $F_1=0.78$.

Error analysis. The confusion matrix from the evaluation of BEST JOINT on the gold standard shows 214 false positives and 191 false negatives. The false negatives suffer from low overlap between the glosses, which are often quite short, e.g., *contend: assert*, sometimes circular, e.g., *sinful: relating to sin*. Aligning senses with such glosses is difficult for a system based on semantic similarity. An explanation is that they do not provide much evidence to the algorithm. For definitions that only contain paraphrasing synonyms, e.g., *assert* for *contend*, there is no lexical overlap if the second definition uses a different synonym, e.g., *argue*. Similarity is zero for the gloss overlap measure *COS* in this case, and it might also be low for the *PPR* measure if either of the synonyms is not represented well in WordNet. The same applies to circular definitions that are compared to a non-circular

definition. Additionally, the circular definition does not contain any information that disambiguates one sense of the word in the definition (here: *sinful*) from potential other senses of this word.

In about 50% of the analyzed pairs, highly similar words are used in the glosses. It should be possible to further exploit this similarity to improve the similarity computation by using second-order representations of these words, for instance by expanding short glosses with the glosses of the contained words, for instance *assert* and *argue*, or via derivational similarity. Taking derivational similarity into account would resolve the lack of overlap between pairs of glosses such as *electrical energy* and *electricity* for the noun *juice*.

An alternative to the similarity measures we used in our experiments are word embeddings that have been shown to perform well for textual similarity tasks in recent years, e.g., [Pennington et al. \(2014\)](#); [Mikolov et al. \(2013\)](#). They map word vectors to a lower-dimensional space and thus avoid the sparsity problems of the gloss overlap measure. Because they can be computed on large corpora, their coverage of the words in the glosses is expected to surpass that of the WordNet-based PPR measure.

A number of false positives occur because the gold standard was developed in a very fine-grained manner: distinctions such as causative vs. inchoative, e.g., *enlarge: make large* vs. *enlarge: become large*, were explicitly stressed in the annotation guidelines, and thus annotated as different senses by the annotators. This annotation strategy is motivated by the fact that this distinction is systematically applied to distinguish frames in FrameNet, and thus occurs for many frames ([Ruppenhofer et al., 2010a](#)). The first sense of *enlarge* belongs to the frame *Expansion*, the second to *Cause_expansion*. A similarity-based approach cannot capture such differences well, because the selectional preferences of both senses are very similar: both will include a description of the item that is enlarged, leading to false positives in the classification. Information that supports to distinguish the senses is sub-categorization information, because the inchoative reading of *enlarge* is intransitive, and the causative reading is transitive.

We moreover find that wrong sense alignments can still be correct frame alignments: for some Wiktionary to FrameNet sense alignments that are considered incorrect according to the strict gloss similarity, the FrameNet frame of the FrameNet senses is still appropriate. We further analyze this observation for the resource FNWKxx that results from applying the alignment method to the full candidate set. FNWKxx is described in the next subsection.

2.5.2 Resulting Resource FNWKxx

Applying the best system configuration to the full candidate set of more than 44,000 candidates results in the intermediate resource FNWKxx. In the acronym, *FN* stands for FrameNet, *WK* for Wiktionary, and *xx* serves as a placeholder for the various languages that are available as translations in Wiktionary. By applying Step 2 of our approach, the disambiguation of translations in Wiktionary to a specific language, the *xx* is replaced by the

respective ISO 639-1 language code, for instance *de* for German, resulting in FNWKde. The application of Step 2 will be described in the next section. This section describes FNWKxx in detail and provides a post-hoc analysis of the alignment quality.

Statistics. The alignment in FNWKxx consists of 12,094 sense pairs. It covers 82% of the senses in FrameNet and 86% of the frames. It connects more than 9,800 unique FrameNet senses with more than 10,000 unique Wiktionary senses. From these numbers, we can infer that both non-alignments and 1-to-many alignments occur for some source senses from FrameNet: not all of the 11,942 senses in FrameNet are covered, which means that there are non-alignments; the total number of alignments is larger than the number of covered FrameNet senses, which means that some FrameNet senses have several alignments to Wiktionary senses, i.e., 1-to-many alignments.

Post-hoc evaluation. The threshold-based cross-validation approach used in our experiments entails the danger of over-fitting. In order to verify the quality of the alignment, we performed a detailed post-hoc analysis on a sample of 270 aligned sense pairs randomly drawn from the set of aligned senses.

Because the issue of sense granularity appeared in the error analysis – we found that some sense alignments were rated as incorrect, but the incorrectly aligned Wiktionary sense still fits the FrameNet frame of the FrameNet sense, we consider three alignment tasks in the post-hoc evaluation:

- (a) fine-grained alignment: the two glosses describe the same sense according to the annotation guidelines.
- (b) coarse-grained alignment: the causative/inchoative distinction is for instance ignored; this evaluation task is more adapted to the capabilities of the automatic alignment method than task (a).
- (c) sense-to-frame alignment: the Wiktionary sense represents the FrameNet frame, even if the glosses do not describe the same sense. This setting is even more coarse-grained than (b).

An expert annotator rated the alignment pairs in the sample as correct or incorrect according to tasks (a), (b), and (c). Based on these ratings, we computed accuracy scores on the selected sample for each task. The resulting scores are listed in Table 2.5.

The post-hoc accuracy can be compared to the *precision* of the automatic alignment method on the gold standard, as shown in Table 2.4, because both scores consider the number of correct alignments according to human annotation in relation to a set of automatically determined alignments.

evaluation setting – accuracy	verb	noun	adjective	all POS
(a) fine-grained sense alignment	0.53	0.73	0.80	0.67
(b) coarse-grained sense alignment	0.73	0.82	0.85	0.78
(c) sense to frame alignment	0.79	0.87	0.92	0.83

Table 2.5: Manual post-hoc evaluation of word sense alignment: accuracy score on a sample of automatic alignments.

According to the scores in Table 2.5, the accuracy for the fine-grained task (a) is lower than the precision over all parts-of-speech on the gold standard. The evaluation by POS shows that the accuracy for nouns and adjectives is equal or superior to the precision on the gold standard, while it is worse for verbs. This shows that over-fitting, if at all, is only a risk for the verb senses. The overall accuracy for (b) exceeds the precision on the gold standard. Verbs in particular receive much better results. This is expected, because this evaluation setting is closer aligned to the capabilities of the automatic alignment that does not make certain fine-grained distinctions.

Not all frames in FrameNet require the fine-grained distinctions made in the gold standard, some frames for instance include senses that are antonyms, which is why we also evaluate setting (c). The accuracy scores are higher for (c) than for (a) and (b), which shows that a Wiktionary sense to FrameNet frame alignment based on our sense alignment is of high quality. Therefore, a dedicated sense-to-frame alignment between Wiktionary and FrameNet might be an alternative to the proposed sense alignment. For other uses of the alignment, such as expanding the number of example sentences for a specific FrameNet sense, or transferring other sense-specific information, for instance on subcategorization, the sense level alignment produced by our approach is however required. We present an application of the alignment that uses the example sentences from Wiktionary to expand the set of FrameNet example sentences in Chapter 3 below.

This evaluation confirms the quality of the sense alignment, both with respect to creating a fine-grained sense alignment, and even more for the goal of FrameNet extension that does not require a fine-grained alignment. Our results suggest that a coarse-grained alignment from senses to frames might suffice when the main goal of the alignment is to increase the FrameNet lexicon coverage.

Monolingual FrameNet expansion. For each of the FrameNet senses in the 12,094 aligned sense pairs, we can extract additional *glosses* from Wiktionary. The FrameNet sense *Activity_finish* of the verb *complete* is, for instance, aligned to the Wiktionary sense 1 of the same verb. This means that we can add the lexical information in Wiktionary for sense 1 of *complete*, as shown in Figure 2.17, to the information for the aligned FrameNet that is shown in Figure 2.3.

Because FrameNet glosses are often very brief, the additional glosses will benefit algorithms such as frame identification for semantic role labeling. The alignment also adds 4,352 new *example sentences* from Wiktionary to FrameNet.

We can associate 2,151 new lemma-POS combinations with FrameNet frames via the synonyms of the aligned senses in Wiktionary. We also extract other related lemma-POS, for instance 487 antonyms, 126 hyponyms, and 19 hypernyms. These lemma-POS do not link to a specific target sense in the English Wiktionary, they only link to the lexical entry that may provide several senses. These can be disambiguated automatically, as proposed by Meyer and Gurevych (2012a). Using their sense-disambiguated English Wiktionary, more than 13,000 additional sense links to Wiktionary can be derived from synonyms of the aligned senses. The method by Meyer and Gurevych (2012a) is introduced in the next section, Section 2.6.

Multilingual FrameNet expansion. The alignment establishes Wiktionary as an interlingual connection between FrameNet and a large number of languages, including low-resource ones: the alignment to Wiktionary connects FrameNet senses to **translations** in 283 languages. To show some examples: the alignment allows us to translate the sense of the verb *complete* associated with the frame *Activity_finish* to the German colloquial *fertigmachen*, the Spanish *terminar*, the Turkish *tamamlamak*, and 39 other languages.

For 35 languages, we can extract more than 1,000 translations each, among them low-resource languages such as Telugu, Swahili, or Kurdish. The languages with most translations are: Finnish at 9,333, Russian at 7,790, and German at 6,871 translations. The number of Finnish translations is more than three times larger than the preliminary Finnish FrameNet by Borin et al. (2012). Likewise, we get three times the number of German lemma-POS than provided by the SALSA corpus. Table 2.6 lists the 35 languages in FNWKxx with more than thousand senses. It also shows the number of lexical entries and definitions in the respective Wiktionary language edition, which allows the reader to estimate the upper bound of the size of an automatically created FrameNet for this language.

The lemma-POS in the linked translations can be used to create a lexical entry in the new language, but it does not refer to a target sense of the linked translation, which would allow us to assign the FrameNet frame to the appropriate senses for an ambiguous word and use the information associated with this sense in the target language Wiktionary. The assignment of a target senses is done in Step 2 of our approach as shown in Figure 2.20. The next section describes Step 2 and shows how the German lexical entries can be sense-disambiguated so that their corresponding senses can be linked directly to the FrameNet senses, creating a proper sense-level linking between the German Wiktionary and FrameNet. We rely on the cross-lingual variant of Meyer and Gurevych (2012a), the approach that has also been used to disambiguate English sense relation targets.

language	# translations	# lexical entries	# definitions
Finnish	9,993	108,326	135,339
Russian	7,318	16,201	25,002
German	7,248	69,991	114,650
Dutch	6,538	60,843	77,555
French	5,701	273,943	356,350
Spanish	5,630	243,704	358,975
Japanese	5,324	48,714	73,236
Italian	4,800	487,068	613,172
Portuguese	4,206	44,085	62,478
Czech	4,146	20,156	23,418
Swedish	3,966	89,795	101,016
Polish	3,624	32,446	45,358
Hungarian	3,554	30,437	33,853
Bulgarian	3,159	36,839	44,980
Greek	2,725	24,095	41,816
Arabic	2,541	3,320	7,824
Danish	2,493	23,963	29,913
Armenian	2,350	9,007	12,490
Norwegian	2,140	6,720	7,922
Hebrew	1,972	5,263	8,005
Romanian	1,830	11,641	18,056
Korean	1,724	15,163	18,597
Turkish	1,703	13,391	16,133
Icelandic	1,638	9,648	13,135
Mandarin	1,637	56,060	113,236
Esperanto	1,631	104,230	105,315
Slovene	1,563	3,335	4,030
Macedonian	1,556	956	1,238
Serbo-Croatian	1,520	39,732	54,195
Latin	1,392	613,474	999,849
Kurdish	1,378	4,265	9,113
Catalan	1,338	56,878	73,250
Swahili	1,288	2,015	2,170
Telugu	1,214	4,863	6,065
Estonian	1,027	3,731	4,224

Table 2.6: Multilingual word-level FrameNet expansions from translations in FNWKxx.

2.6 Translating FrameNet to German via Wiktionary

In this section, we exemplarily perform Step 2 of our method to translate FrameNet to other languages, see Figure 2.20, for German. Our goal is to create a sense-disambiguated FNWKde from FNWKxx.

FNWKxx initially does not provide lexical-semantic information for the German translations: a translation links an English sense to a lexical entry in the German Wiktionary, not a target sense. In order to integrate the information attached to a German Wiktionary sense, e.g., the gloss, into our resource, the lemmas need to be disambiguated. The same applies to the new lemmas associated with FrameNet frames for English that were derived by following the sense relations in the English Wiktionary. We use the sense-disambiguated Wiktionary resulting from [Meyer and Gurevych \(2012a\)](#) for the disambiguation of relations and translations in Wiktionary to create our new bilingual German–English FrameNet lexicon FNWKde. Then, we provide a detailed analysis of our two-step approach for creating FrameNets in new languages and the resulting resource FNWKde.

2.6.1 Disambiguating German Lexical Entries

The approach by [Meyer and Gurevych \(2012a\)](#) combines information on the source sense and all potential target senses in order to determine the best target sense in a rule-based disambiguation strategy. The information is encoded as binary features, which are ordered in a back-off hierarchy: if the first feature applies, the target sense is selected, otherwise the second feature is considered, and so forth.

The most important features are:

- (a) definition overlap between source and automatically translated target definitions,
- (b) occurrence of the source lemma in the target definition,
- (c) shared linguistic information, e.g., the same register,
- (d) inverse translation relations that apply when the source lemma occurs on the translation list of the target sense,
- (e) relation overlap,
- (f) the Lesk measure between original and translated glosses in the source and target languages, and finally
- (g) backing off to the first target sense as default option. For monosemous target senses, the only available sense is used for the disambiguation.

For the gold standard evaluation of the approach we refer to [Meyer and Gurevych \(2012a\)](#): their system obtained an F_1 -score of 0.67 for the task of disambiguating translations from English to German. [Meyer and Gurevych \(2012a\)](#) report that the Wiktionary first sense baseline already provides a strong performance with $F_1=0.65$. Such a baseline could be easily implemented for other languages in Wiktionary, in order to create more

relation	# English senses per FrameNet sense	# English senses per frame
synonym	17,713	13,288
hyponym	4,818	3,347
hypernym	6,369	3,961
antonym	9,626	6,737

Table 2.7: English FrameNet expansion after relation disambiguation.

sense disambiguated resources. The quality of such a resource will, however, depend on the size and quality of the target language Wiktionary.

For disambiguation of related senses in the English Wiktionary, the same algorithm is used, but the translation of glosses in (f) is not required. Meyer and Gurevych (2012a) report an F_1 -score of 0.79 for the disambiguation of English sense relations. We use the resource resulting from the disambiguation of English sense relations provided by Meyer and Gurevych (2012a) to identify the target senses of synonyms in FNWK_{xx}.

2.6.2 Resulting Lexical Knowledge Base FNWKde

We now present FNWKde, the result of the translation disambiguation introduced in the previous subsection, in detail. FNWKde is a linked lexical knowledge base that connects FrameNet with the English and German Wiktionary lexicons. We describe resource statistics, perform a post-hoc error analysis, and then compare the created lexicon to other German FrameNet resources.

Statistics. Table 2.8 gives an overview of FNWKde in the third column. The sense alignment contains 5,897 pairs of German Wiktionary senses and FrameNet senses, i.e., 86% of the translations could be disambiguated. These sense alignments contain 4,066 unique instances of German lemma, POS, and FrameNet frame. Thus, they define 4,066 unique senses in FrameNet. The remaining 1,831 alignment pairs are results of 1-to-n alignments of FrameNet to the finer-grained German Wiktionary, where $n \in (2, \dots, 8)$. Therefore, 32% of the 4,066 German FrameNet senses have several sense glosses. Each Wiktionary sense in the alignment has a gloss, leading to 5,897 sense glosses for the German FrameNet lexicon. There also are 6,933 example sentences associated with the 5,897 alignment pairs and the 4,066 German FrameNet senses.

Based on the relation disambiguation and inference of new relations by Meyer and Gurevych (2012a), we can also disambiguate synonyms in the English Wiktionary. This leads to a further extension of the English FrameNet summarized in Table 2.7. The number of Wiktionary senses aligned to FrameNet senses is increased by 50% compared to the direct alignment in FNWK_{xx}.

	SALSA	P&L05	FNWKde
type	corpus	corpus	lexicon
creation	manual	automatic	automatic
frames (+proto-frames)	266 (+907)	468	755
senses	1,813	9,851	4,066
sense examples	24,184	1,672,551	6,933
sense definitions	-	-	5,897

Table 2.8: Frame-semantic resources for German. SALSA frames are compatible with FrameNet frames except for separately listed proto-frames in SALSA.

We also provide results for other sense relations present in Wiktionary, i.e. antonyms, hypernyms, and hyponyms. We will discuss whether and how they can be integrated as FrameNet senses in our resource below.

Post-hoc error analysis. Because the errors of the two subsequently applied automatic alignment methods, e.g., Step 1 and Step 2 in Figure 2.20, can multiply, we provide a post-hoc evaluation of the resulting alignments. Therefore, we collected the FrameNet senses for a list of 15 frames that were sampled by [Padó and Lapata \(2005a\)](#) according to three frequency bands in a large corpus.¹⁰ There are 115 senses associated with these frames in our resource. A manual evaluation of these 115 senses shows that 67% were assigned correctly to their frames. This is higher than can be expected based on the precision of the two subsequently applied alignments steps, because the errors can multiply, potentially leading to a larger number of false positives, i.e., spurious alignments. The precision of the alignment of FrameNet to the English Wiktionary was 72.8% ([Meyer and Gurevych, 2012a](#)), the precision score of the English to German Wiktionary alignment was 67%.

Further analysis revealed that both resource creation steps contribute equally to the 39 errors observed on the sample of 115 senses. Out of 39 errors, 20 (51%) result from the translation disambiguation, 19 (49%) from the FN-WKT alignment.

An interesting observation is that some alignments were established in a redundant fashion: for 17 of the evaluated sense pairs, redundancy confirms their quality: they were obtained independently by two or three alignment-and-translation paths and do not contain any alignment errors. Such information on redundancy could be used as a measure of confidence for automatic alignments established in a similar fashion in the future.

¹⁰The frames are: *Preventing*, *Communication_response*, *Giving*, *Deciding*, *Cause_change_of_scalar_position*, *Evaluative_comparison*, *Travel*, *Employing*, *Sensation*, *Judgment_communication*, *Adding_up*, *Congregating*, *Escaping*, *Suspiciousness*, and *Recovery*.

	resource r	% of r covered by FNWKde	% of FNWKde covered by r
frame level	SALSA	89	31
	P&L05	90	55
sense level	SALSA	15	5
	P&L05	10	19

Table 2.9: Overlap of FNWKde with SALSA and P&L05.

Comparison to German frame-semantic resources. To further evaluate the quality of the induced German FrameNet lexicon, we compare FNWKde to two German frame-semantic resources, the manually annotated SALSA corpus (Burchardt et al., 2006) and a resource from Padó and Lapata (2005a), henceforth P&L05. Note that both resources are frame-labeled corpora, while FNWKde is a FrameNet-like *lexicon* and contains information complementary to the corpora.

The different properties of the resources are contrasted in Table 2.8. It shows that the automatically developed resources, including FNWKde, provide a larger number of senses than SALSA. The annotated corpora contain a large number of examples, but they do not provide any glosses, which are useful for the frame identification step of SRL, nor do they contain any other lexical-semantic information.

FNWKde covers a larger number of FrameNet frames than the other two resources. 266 of the 907 frames in SALSA are connected to original FrameNet frames, the others are proto-frames p, which have been specifically developed for SALSA, see Section 2.2.3. They are shown in parentheses in Table 2.8.

Table 2.9 describes the proportion of the overlapping frames and senses to the respective resources. The numbers on frame overlap show that FNWKde covers the frames in the other resources well – 89% and 90% coverage respectively, and that it adds frames not covered in the other resources: P&L05 only covers 55% of the frames in FNWKde. The sense overlap shows that the resources have senses in common, which confirms the quality of the automatically developed resources, but also shows that they also complement each other. FNWKde, for instance, adds 3,041 senses to P&L05.

This subsection presented the creation of FNWKde from the multilingual, but not sense-disambiguated FNWKxx. In the next section, we discuss implications of applying our method to other languages to create a truly multilingual FrameNet and compare our method to previous approaches.

2.6.3 Discussion: a Multilingual FrameNet based on FNWKxx

FNWKxx builds an excellent starting point to create FrameNet lexicons in various languages: the translation counts, for instance 6,871 for German, compare favorably to the

FrameNet release 1.5, which contains 9,700 English lemma-POS. The same applies to other languages that are well-represented in Wiktionary, for instance Finnish, Russian, or Dutch, see also Table 2.6.

In this subsection we discuss the difficulties and implications of applying our method to other languages besides German, and compare it to other methods for translating FrameNet to new languages.

Translation to other languages. To create FrameNet lexicons in other languages, the translation disambiguation approach used for FNWKde – Step 2 in Figure 2.20 – needs to be adapted to other languages. The approach is in theory applicable to any language, but there are some obstacles: first, it relies on the availability of the target sense in the target language Wiktionary. For many of the top 30 languages in FNWKxx, the Wiktionary editions seem sufficiently large to provide targets for translation disambiguation, cf. Table 2.6, and they are continuously extended. Second, our approach requires access to the target language Wiktionary, but the data format across Wiktionary language editions is not standardized. Third, the approach requires machine translation into the target language. For languages, where such a tool is not available, we could default to the first-sense-heuristic, or encourage the Wiktionary community to link the translations to their target Wiktionary senses inspired by Sajous et al. (2010).

Language specificity of FrameNet. Another issue that applies to all automatic (and also manual) approaches of cross-lingual FrameNet extension is the restricted cross-language applicability of frames. Boas (2005) reports that, while many frames are largely language-independent, other frames receive culture-specific or language-specific interpretations, for example calendars or holidays. Additionally, fine-grained sense and frame distinctions may be more relevant in one language than in another language. Such granularity differences also led to the addition of proto-frames in SALSA 2 (Rehbein et al., 2012). Therefore, manual correction or extension of a multilingual FrameNet based on FNWKxx may be desired for specific applications. In this case, the automatically created FrameNets in other languages are good starting points that can be quickly and efficiently compiled.

Resource quality. The quality of the multilingual FNWKxx depends on i) the translations in the interlingual connection Wiktionary, which are manually created, controlled by the community, and therefore reliable, and ii) on the FrameNet–Wiktionary alignment. To estimate the quality of FNWKxx, we evaluated our sense alignment method in detail. The alignment reached state-of-the-art results at the time, and the analysis shows that the method is particularly suitable for a coarse-grained alignment. However, we find lower performance for verbs in a fine-grained setting. We argue that an improved alignment algorithm, for instance taking subcategorization information into account, can identify the fine-grained

distinctions. [Matuschek \(2014\)](#) later created an improved FrameNet–Wiktionary alignment using the gold standard introduced here: F_1 is 0.78 for their best configuration, an approach that combines Dijkstra-WSA with a similarity-based approach for sense alignment.

The error analysis raised questions regarding the FrameNet frame granularity. This touches questions such as: do separate frames exist for causative/inchoative alternations, e.g., *Being_dry* and *Cause_to_be_dry* for *to dry*, or do they belong to the same frame, e.g., *Make_noise* for *to creak* and *to creak something*? For the coarse-grained frames, fine-grained decisions can be merged in a second classification step. Alternatively, we could map Wiktionary senses directly to frames, and include features that cover the granularity distinctions, e.g., whether the existing senses of a frame show the semantic alternation. This feature would be true for *Make_noise*, because the causative/inchoative alternation does not require a different frame, while it would be *false* for the causative and inchoative versions of *to dry* that elicit the different frames mentioned above. We could use the same approach to assign senses to a frame which are derived via sense relations other than synonymy, i.e., for linking antonyms or hyponyms to a frame. Some frames do cover antonymous predicates, others do not. The frame *Experiencer_focus* is for instance associated with the antonyms *to love* and *to hate*, the frame *Attaching* is associated with the antonyms *to tie* and *to untie*. For the antonyms *to buy* and *to sell*, there are separate frames *Commerce_buy* and *Commerce_sell* that are not associated with the antonyms of either buy or sell. Antonyms like *to buy* and *to sell* present opposite points of view for a single event. These kinds of antonyms represent different sets of participants, e.g., the *Seller* is optional for *Commerce_buy*, but not for the *Commerce_sell*, and are therefore systematically equipped with separate frames in FrameNet ([Ruppenhofer et al., 2010a](#)).

Comparison to other approaches. Based on Wiktionary, our approach suffers less from the disadvantages of previous resource-based work, i.e., the constraints of expert-built resources ([Tonelli and Pighin, 2009](#); [Laparra and Rigau, 2010](#)) and the lack of lexical information in Wikipedia ([Tonelli and Giuliano, 2009](#); [Tonelli et al., 2013](#)). Unlike corpus-based approaches for cross-lingual FrameNet extension ([Padó and Lapata, 2005a](#); [Johansson and Nugues, 2005](#); [Basili et al., 2009](#); [Padó and Lapata, 2009](#)), our approach does not provide frame-semantic annotations for the example sentences. Our advantage is that we create a FrameNet *lexicon* with lexical-semantic information in the target language. Example annotations can be additionally obtained separately, for instance via cross-lingual annotation projection ([Padó and Lapata, 2009](#)) or the distant supervision method introduced in Chapter 3.3. The lexical information in FNWKde can be used to guide this process.

Summary of the section. In this section, we presented the automatic creation of a German FrameNet lexicon using Wiktionary as an interlingual resource. This is Step 2 of our method to automatically create FrameNet lexicons in other languages as outlined in Sec-

tion 2.4. The German FrameNet lexicon is based on the FrameNet– Wiktionary alignment FNWKxx. By disambiguating cross-language links between the English and German Wiktionary to German Wiktionary senses, German Wiktionary senses are linked to FrameNet senses in the resulting linked lexical knowledge base FNWKde. We compare FNWKde to other German FrameNet resources, SALSA and the frame-labeled corpus created by [Padó and Lapata \(2005a\)](#), and find improved coverage on the level of frames, as well as complementary sense coverage. FNWKde also introduces lexical information that is not available for SALSA or the corpus from [Padó and Lapata \(2005a\)](#), namely sense-specific examples and sense definitions extracted from Wiktionary.

We applied our method to translate FrameNet to other languages via Wiktionary, see Section 2.4, exemplary to German, because it provides two other lexical resources for comparison and quality estimation. The presented method is, however, not constrained to German, it is applicable to other languages that are represented well in Wiktionary, e.g., Russian or Finnish, and thus provides an excellent basis for bootstrapping FrameNet lexicons for new languages.

Establishing sense-level links between lexical knowledge bases is an important step towards the creation of large LLKBs and the joint use of lexical information from various lexical knowledge bases. Another important aspect is being able to access the linked information efficiently. In the next section, we present our work on standardizing linked lexical knowledge bases that pertains to this goal.

2.7 Standardizing Semantic Knowledge Bases

Another important aspect of linking lexical resources is how to represent the links, and how to access the information from the linked resources. This is crucial for *using* the LLKBs in NLP applications: different LKBs model various types of lexical information in many ways, using different, often conflicting terminology. An example for different terminology are semantic roles that are called *frame elements* in FrameNet and *roles* in VerbNet. An example for conflicting terminology is the term *lexeme*: it has a different meaning in FrameNet, referring to the components of a multiword lemma, compared to other lexicons, in which *lexeme* refers to the lemma under which a lexical entry is listed.

A difference that is even more important, because it makes it more difficult to access the same information types across LKBs, is the different structure in various LKBs. Examples for such different structures were shown earlier in Section 2.2: some lexical knowledge bases, like WordNet or GermaNet, are built around the central concept of a synset that groups synonymous senses. Others, like FrameNet and VerbNet also group word senses, but according to different criteria. For FrameNet, these criteria include different kinds of semantically related words and other parts-of-speech, grouping senses that are associated to a certain frame, whereas they include a strong focus on syntactic similarity for VerbNet.

In order to use the information in the LLKBs efficiently, the terminology needs to be unified and corresponding types of linguistic information, for instance semantic roles in VerbNet and frame elements in FrameNet, need to be mapped. The unification of terminology and lexicon structure presents a theoretical challenge. Once the structure and terminology of the LLKBs have been unified, the information contained in the LLKBs can be accessed efficiently using a unified query language or application programming interface (API). This problem touches what has been called *structural interoperability* of lexical resources – as opposed to *semantic interoperability* provided by the resource links (Ide and Pustejovsky, 2010): ideally, the same information types in the different lexical resources are represented – and thus can be accessed – in the same way, including the sense-level or predicate argument-level links between them.

The lack of *structural interoperability* between lexical knowledge bases can be addressed by standardization. In this section, we present our contributions to the standard-conformant UBY-LMF (Eckle-Kohler et al., 2012, 2013). The UBY-LMF model implements the ISO standard Lexical Markup Framework, a meta-model of linked lexical resources (Francopoulo et al., 2006), and presents a comprehensive example of a standardized model for linked lexical knowledge bases. It provides a data model for all the linguistic information in the major lexical knowledge bases, i.e., all the resources introduced above, including resource links on the level of word sense and predicate argument structure. One of the main goals of UBY-LMF is to provide interoperability while preserving the original structure of the resources. This way, combinations of resources can be used easily in NLP applications, but the contributions of single resources to a particular NLP application can also be compared directly. This is relevant for practical applications, but also important for research that explores the contribution of different LKBs to different types of NLP applications like word sense disambiguation or semantic role labeling.

The implementation of the lexicon model UBY-LMF led to the linked lexical knowledge base UBY (Gurevych et al., 2012a). It represents the linked resources in an SQL database whose tables correspond to the classes of the UBY-LMF model and provides an open-source Java API based on object-relational modeling using the Hibernate API for programmatic access. UBY also provides a user interface for visualization and exploration of the data in UBY (Gurevych et al., 2012b).

The creation of UBY-LMF and UBY was a group effort. The present author contributed significantly to the modeling of semantic lexicons like FrameNet, multiword expressions, and predicate argument structure links in UBY-LMF and UBY.

In this section, we describe the parts of UBY-LMF that are used to model semantic knowledge bases like FrameNet and predicate argument structure links like those from SemLink in detail. The UBY-LMF model of FrameNet and the conversion of FrameNet 1.5 to UBY are contributions of the present author; they were published first as part of Gurevych

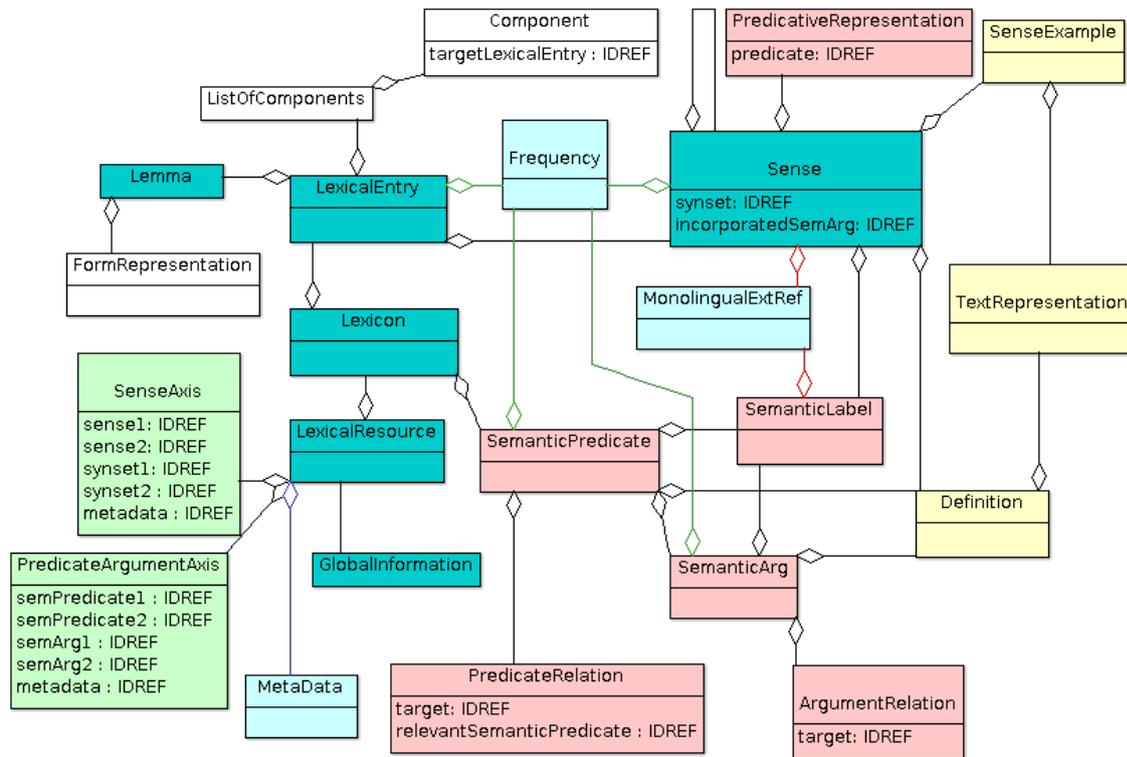


Figure 2.21: UBY-LMF classes for modeling FrameNet and SALSA.

et al. (2012a). The UBY-LMF model of predicate argument structure links is another contribution of this thesis.

2.7.1 Modeling Semantic Knowledge Bases in UBY-LMF

FrameNet is the resource with the greatest detail in modeling semantic predicate argument structure, therefore we created the UBY-LMF model with thorough care to the affordances of FrameNet. At the same time, UBY-LMF also covers the lexical information in other models of semantic predicate argument structure, like VerbNet and PropBank.

Figure 2.21 shows a subset of those UBY-LMF classes that are needed to represent the information from FrameNet. `LexicalResource` is the central class. The arrows describe an aggregation relation: a `LexicalResource` contains one or more instances of the `Lexicon` class, which contains many instances of `LexicalEntry`. Each of these classes has a unique identifier and most come with a set of attributes that contain the information associated with these classes. `LexicalEntry` has, among others, the attribute `partOfSpeech`. To ensure readability, Figure 2.21 only shows those attributes that represent associate relations to the unique identifier *IDREF* of other UBY-LMF classes. In the description of the classes below, the

attributes are introduced whenever they are required for understanding the model. For a complete overview, refer to the UBY-LMF documentation and DTD file.¹¹

Core lexicon model. The central UBY-LMF classes are shown in turquoise in Figure 2.21. As UBY-LMF aims to preserve the original resource structure, each lexical resource is added as a separate `Lexicon` to the `LexicalResource` in UBY. Thus, we add a specific `Lexicon` for `FrameNet` and another one for `VerbNet`. The `FrameNet` `Lexicon` has `LexicalEntries`, which are defined by a lemma, as modeled by the `Lemma` class, and the attribute `partOfSpeech`. The actual lemma string and potential spelling variants are contained in the class `FormRepresentation` associated with `Lemma`.

Some `LexicalEntries` are multiword expressions and use the `ListOfComponents` and `Component` classes to represent the constituent words of the multiword construction. A `LexicalEntry` together with a frame label defines a sense, or lexical unit, in `FrameNet`, modeled in the class `Sense`. The UBY-LMF class `Sense` provides the boolean attribute `transparentMeaning` and the string attribute `incorporatedSemArg` that are filled based on the information associated with a sense in `FrameNet`. The attribute `incorporatedSemArg` contains the unique id of the `SemanticArgument` representing the `FrameNet` role that is considered an incorporated role for this particular lexical unit. As an example: the verb *rise* with the frame *Motion_directional* incorporates the role *Direction* marking the upward direction of the motion. This information is retained in `incorporatedSemArg`. The attribute `transparentMeaning` is set to true if the meaning of the sense is bleached in a multiword construction.

Senses are associated with their example sentences in the class `SenseExample` which, for `FrameNet`, contains a single `TextRepresentation` with the text of the `FrameNet` example sentence. It is a design decision of UBY-LMF that UBY models lexical resource information, but does not model corpus annotations. It does, for instance, not contain the role annotations on the example sentences in the `FrameNet` lexicon. For such purposes, different standards have been developed, for instance `GrAF` (Ide and Suderman, 2007). If needed, the predicate instances in the `FrameNet` corpus can be linked to the `Senses` of UBY via their frame label or the `FrameNet` lexical unit ID, which are also represented in UBY.

Semantic information. UBY-LMF classes associated with semantic information and semantic predicate argument structure are shown in pink in Figure 2.21.

The `SemanticPredicate` class models `FrameNet` frames. Each frame receives a unique identifier, the `semanticPredicateId`, a predicate label string that stores the frame label, and a `lexiconId` that refers to the source `Lexicon` of the `SemanticPredicate`. Additionally, information on whether the frame is lexicalized or perspectivalized is contained in the corresponding attributes. These attributes model information specific to `FrameNet`: the lexicalized attribute is used to mark abstract frames such as *Activity* in Figure 2.2. For

¹¹<https://www.ukp.tu-darmstadt.de/data/lexical-resources/uby/uby-lmf/>

abstract frames, `lexicalized` is set to *false*. The `perspectivalized` attribute marks whether a frame is in a *perspective_on* relation to another frame. The frames *Commerce_buy* and *Commerce_sell* are for instance perspectivalized versions of the frame *Commerce_goods-transfer*, taking the perspective of either the *Buyer* or the *Seller*.

The class `SemanticArgument` models semantic roles. Besides a unique identifier, this class reserves a string attribute `semanticRole` for the role label, and a string attribute for the information on the coreness of a role, i.e., whether it is of type *core*, *peripheral*, *extra-thematic*, or *core-unexpressed*, see also Section 2.1.1.

`SemanticArgument` also has a boolean attribute stating whether the argument is realized as an incorporated argument for that frame, and an identifier linking the role to its `SemanticPredicate`. This way, each `SemanticPredicate` in `FrameNet` is linked to one or more `SemanticArguments`.

The class `PredicativeRepresentation` links `Sense` objects to `SemanticPredicate` objects via their unique identifier. Thus it associates instances of `Sense` with their corresponding `SemanticPredicate` instances.

UBY-LMF also models the frame hierarchy: the class `PredicateRelation` represents relations between frames. For `FrameNet`, the attribute `relType` of `PredicateRelation` has the value *frameRelation* and the attribute `relName` contains the name of the relation, for instance *causative*, *inherits_from*, *subframe_of*, or *precedes*.

Relations between `SemanticArguments` are modeled in an analogous fashion in the class `ArgumentRelation`. In `FrameNet`, there are frame-specific argument relations, such as two arguments *excluding* or *requiring* each other, and relations marking so-called coreness sets. Coreness sets group `FrameNet` core roles that can be used interchangeably, i.e., the occurrence of one of the roles in a coreness set suffices to create a felicitous usage of this frame. An example is the coreness set $C = \{Goal, Source, Path\}$ for movement frames, e.g., *Motion*. Either of the three sentences “*She moved [towards the door]_{Goal}*”, “*She moved [away from the door]_{Source}*”, “*She moved [along the road]_{Path}*” uses a role from C ; using further roles from C is possible, but not required.

The `SemanticLabel` class models additional semantic information that can be attached to `Senses`, `SemanticPredicates` or `SemanticArguments`. The different types of semantic information are stored in the attribute `type`, the values in the attribute `label`. For `FrameNet`, information on the semantic type that is associated with certain frames, mostly nominal frames associated with objects, is stored in an instance of `SemanticLabel` with type *semanticCategory* that is attached to `SemanticPredicate`. The `label` attribute holds the semantic type information, including values like *State*, *Artifact*, or *Physical_object*.

`FrameNet` also provides semantic type information for some of its roles. These are basically selectional preferences of the roles, marking the role filler as, e.g., *Physical_object*, *Sentient*, or *Duration*. To represent this kind of information, another kind of `SemanticLabel` with the type label *selectionalPreference* is attached to the `SemanticArgument` class.

There are more types of `SemanticLabel` that are associated with the `Sense` class. The `SemanticLabel` of type *sentiment* associates *positive_judgment* or *negative_judgment* with an instance of `Sense`. These are for instance used to show that certain FrameNet senses associated with the *Frugality* frame convey a positive sentiment, e.g., *thrifty* and *penny-wise*, while others convey a negative sentiment, e.g., *austere* and *parsimonious*. The `SemanticLabel` of type *collocate* contains information on collocations of words that are semi-productive and thus not modeled as a multiword lemma. Examples are verbs in support verb constructions: *give* in the context of *give someone a heart attack* receives a causal reading as in *cause a heart attack* (Ruppenhofer et al., 2010c). *To give* only receives this reading as part of a support verb construction. Therefore, this reading of *to give* receives the semantic label value *Support*. Other related labels are *Bound_dependent_LU* and *Bound_LU*, that are used for other types of collocates and in part reserved for constructions not yet covered in FrameNet 1.5. For more detail see section 6.2 in Ruppenhofer et al. (2010c).

General information types. The classes colored in yellow and light blue in Figure 2.21 represent information that attaches to a large number of different classes.

The classes in yellow color are associated with example sentences modeled in the class `SenseExample` and textual descriptions of senses, e.g., lexicon glosses modeled in the class `Definition`. For FrameNet, frames and roles are attached to their definition texts using the class `Definition`. The actual definition and example texts are contained in the class `TextRepresentation`.

The classes in light-blue attach to a large number of different classes. Links to original FrameNet identifiers, e.g., *lexical unit IDs*, are present in the class `MonolingualExternalReference`. They are attached to the `Sense` and the `SemanticLabel` classes and associate the `Sense` instances with the FrameNet lexical unit ID and the `SemanticLabel` instances with their FrameNet numeric ids. The class `Frequency` is also associated with `Sense` and contains the number of instances and annotated corpus instances for a FrameNet sense as represented in the lexical unit files. The class `MetaData` attaches to lexical resource; it contains information on the creation date of a `LexicalResource`. We defined the classes `SenseAxis` and `PredicateArgumentAxis` to model linkings of FrameNet to other LKBs in UBY; these classes also receive metadata information. Details on the respective classes and the associated metadata is introduced in detail in Section 2.7.2 below.

SALSA in UBY-LMF. The introduction of SALSA in Section 2.2 showed that SALSA is very similar to FrameNet. For modeling SALSA in UBY-LMF, we use the same structure as for modeling FrameNet. There are a few differences caused by the simpler modeling of the lexicon in the SALSA corpus: In SALSA, there is no separate corpus evidence associated with the senses, i.e., there are no explicit annotated example sentences associated with the

word senses in the frame files. We instead read the sentences from the SALSA corpus for the appropriate sense and add them to the instances of `SenseExample` for this sense.

The definitions for the frames include German and English example sentences. The origin of the example sentences is not documented, and for the German sentences, it is not explicitly represented which sense of the frame they belong to. Therefore we did not include them in the `SenseExample` class. Instead, we represent the definitions with their example sentences in the `Definition` class associated with the `SemanticPredicate`.

There is an additional set of `SemanticLabel` values that note whether the frame in question stems from the FrameNet release 1.2, 1.3, or was defined as a new frame for SALSA. Because SALSA does not contain multiword lemmas, SALSA does not use the `ListOfComponents` class. Apart from this exception, we use the same UBY-LMF classes for modeling SALSA as for modeling FrameNet.¹²

VerbNet in UBY-LMF. VerbNet provides a detailed model of syntactic and semantic properties of verbs, as introduced in Section 2.2. UBY-LMF provides all the classes needed to model these properties. The conversion of VerbNet to UBY-LMF was published first as part of Gurevych et al. (2012a). In this conversion, a VerbNet sense is defined as a combination of the VerbNet lemma, class, and subcategorization frame. As a result, there are several `Sense` instances per VerbNet class. The VerbNet class label is attached to the `Sense` using the class `SemanticLabel`.

A `SemanticPredicate` is defined by the Montague-style logic string attached to each VerbNet class, as shown under *Semantics* in Figure 2.5. This string is attached as the `Definition` of the `SemanticPredicate`, the predicate label is left empty. For each `SemanticPredicate`, the VerbNet thematic roles are added as instances of `SemanticArgument`. Role labels include the VerbNet selectional preferences, e.g., `Agent[+animate| + machine]`.

VerbNet additionally requires a model of the syntax-semantics interface linking `SemanticPredicates` to syntactic frames and `SemanticArguments` to syntactic arguments. The syntax-semantics interface in UBY-LMF is described in detail in Eckle-Kohler et al. (2012).

PropBank in UBY-LMF. The PropBank frame files constitute a lexical resource that can also be represented in UBY-LMF. The lexicon model in UBY-LMF corresponds to the PropBank lexicon structure introduced in Figure 2.8 in Section 2.2. The PropBank lexicon model is simpler than the one for VerbNet, as syntactic frames are not modeled explicitly. The frame files are lemma-specific and thus build the input for a `LexicalEntry`. The different rolesets in the frame file correspond to a `Sense` in UBY-LMF. They are associated with a short `Definition` and a number of `SenseExamples`. A `SemanticPredicate` in PropBank is synonymous with a `Sense`, i.e., there is a `SemanticPredicate` for each sense. Similar to VerbNet, there is no meaningful predicate label. Each of the roles in a roleset is represented in the

¹²The conversion of the SALSA 2 corpus to UBY was also created as part of this work.

SemanticArgument class and equipped with a brief, predicate-specific Definition. The conversion of PropBank to UBY-LMF was created as part of this work.

2.7.2 Modeling Resource Links

UBY-LMF models two types of links across instances of the *Lexicon* class, for instance between FrameNet and VerbNet. The corresponding UBY-LMF classes are shown in light green in Figure 2.21. The first type, *SenseAxis*, is used to represent sense-level alignments across *Lexicon* instances, as introduced in Section 2.3.1. The second type, *PredicateArgumentAxis*, represents alignments on the level of predicate argument structures, also across *Lexicon* instances. These were introduced in Section 2.3.2. The links are represented by two UBY-LMF classes that are attached to the *LexicalResource* class and use the unique ids of the linked instances to establish the linkings.

Sense-level links. Sense-level links are represented by the *SenseAxis* class that directly attaches to *LexicalResource* and contains pairs of unique identifiers of either *Sense* or *Synset* depending on whether pairs of senses or synsets are linked.¹³ The *senseAxisType* attribute distinguishes between monolingual or cross-lingual resource links. To distinguish between several sets of sense links between the same lexicon, the *MetaData* class contains information on the creation of the set of sense links, like the version number, the date of creation, the tool used for the creation of the linking, and whether the linking was created automatically or manually. For automatic alignments, confidence scores for each aligned sense pair can also be provided in the *MetaData* class. The corresponding *MetaData* object for an instance of *SenseAxis* is accessed via an *IDREF* attribute in the *SenseAxis* class. This allows users to distinguish between different linkings of the same *Lexicons*. This detailed representation of different sense alignments is an important feature of UBY_LMF, because it provides flexibility in using the linked lexical knowledge base when several different sense alignments are available. We can, for instance, use an automatic sense alignment *A* between the same *Lexicon* instances as a back-off for the manual alignment *M*, or prefer an automatic alignment *H* with high confidence scores to an alignment *L* with lower confidence scores or without confidence scores. This leads to more efficient usage of the available sense alignments and potentially better results in the NLP applications that use these alignments.

Predicate argument structure links. Predicate argument structure links are modeled by the class *PredicateArgumentAxis*. Analogous to *SenseAxis*, *PredicateArgumentAxis* attaches to *LexicalResource*. An instance of *PredicateArgumentAxis* links pairs of *SemanticPredicate* or *SemanticArgument* based on their unique identifier *IDREF*. The *axisType* attribute can be used to mark differences between links on the semantic predicate or semantic argument

¹³UBY-LMF also contains a *Synset* class that attaches to *Lexicon* and aggregates several senses. It is used for modeling WordNet and GermaNet and is not shown in Figure 2.21.

level. Again, a reference to a `MetaData` object allows us to store additional information on the instances of `PredicateArgumentAxis` and thus enables dynamic selection of predicate argument structure links based on their metadata properties, in the same way to the selection of `SenseAxis` instances discussed above.

SemLink in UBY-LMF. If `FrameNet` and `VerbNet` are available as a UBY Lexicon, the classes `SenseAxis` and `PredicateArgumentAxis` provide the prerequisite for importing `SemLink` to UBY: `SemLink` provides a mapping of `FrameNet` lexical unit IDs to `VerbNet` class labels and lemmas. We use this information to populate the `SenseAxis` class with `Sense` instances from the `FrameNet` Lexicon and from the `VerbNet` Lexicon. `SemLink` also includes the frame labels in this mapping. This information can be used to populate a `SemanticPredicate` linking between `SemanticPredicate` instances in `VerbNet` and `FrameNet`. Because there can be several `SemanticPredicate` instances for a combination of lemma and `VerbNet` class in `VerbNet`, there are potentially multiple instances of `PredicateArgumentAxis` for one pair of frame label and `VerbNet` class in UBY. As an example: there are six senses of the verb *finish* in UBY that are associated with the `VerbNet` class *stop-55.4-1*. These six senses are associated with four different instances of `SemanticPredicate`. One of the senses is associated with the `SemanticPredicate` with ID *VN_SemanticPredicate_538* and definition *end(E, Theme) AND use(during(E), ?Agent, Instrument)*. This `SemanticPredicate` is linked to `SemanticPredicate` instances with five different frame labels in `FrameNet`, *Activity_stop*, *Activity_finish*, *Killing*, *Firing*, and *Halt*.

The role-level mapping in `SemLink` links frame-specific `FrameNet` roles to `VerbNet` thematic roles per `VerbNet` class. We convert this information to instances of `PredicateArgumentAxis` which link the corresponding `SemanticArgument` instances.

Predicate- and argument-level linkings can be distinguished at first glance by a different value for the attribute `axisType`, namely *uby_predicate_axis* and *uby_argument_axis*. Thus, the `SemanticPredicate` with the frame label *Removing* is linked to the `VerbNet` `SemanticPredicate` associated with the verb *remove* and `VerbNet` class 10.2. The corresponding roles are also linked by instances of `PredicateArgumentAxis`. Thus, the `FrameNet` roles *Theme* and *Source* are linked to the equally named roles in `VerbNet`, the `FrameNet` roles *Agent* and *Cause* are mapped to the `VerbNet` role *Agent*, and the `FrameNet` role *Goal* is mapped to the `VerbNet` role *Destination*.

Summary. This paragraph concludes the introduction of the UBY-LMF model. We presented the first model that integrates the major semantic lexicons for English, `FrameNet`, `VerbNet`, and `PropBank`, to a single unified format, and also models semantic lexicons for other languages, e.g., `SALSA`. The model integrates the lexical-semantic information from the different lexicons comprehensively and provides structural interoperability by mapping the information to a single, unified format. The extension of UBY-LMF by a new model

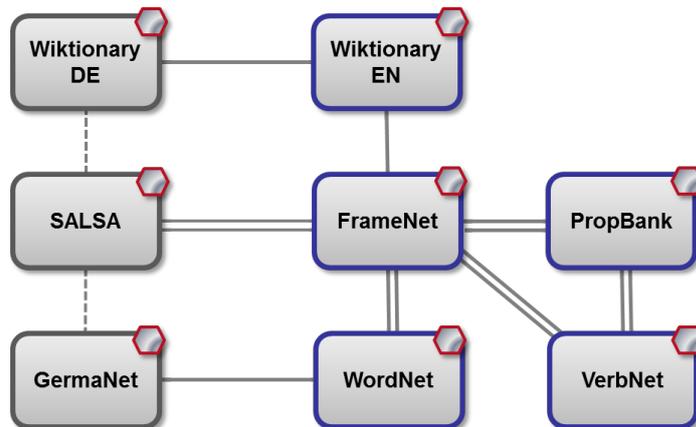


Figure 2.22: Overview of the LLKB UBY_{FN} ; Single lines represent sense-level links, double lines sense- and role-level links; dotted lines show relevant indirectly inferred links.

of links on the predicate argument structure-level allows us to model predicate argument structure alignments, such as the one provided by SemLink. Together with the existing model for sense-level links, UBY -LMF thus supports semantic interoperability between the major lexical knowledge bases. This enables us to access the semantic information types in the linked lexical resources in a uniform way. In Chapter 3, we use the uniform access to the linked lexical knowledge bases in our knowledge-based approach for the generation of frame- and role-labeled training data.

In this section, we showed how UBY -LMF models various lexical knowledge bases, and the sense alignments and predicate argument structure links between them, leading to a single, unified linked lexical knowledge base. In the next section, we introduce the linked lexical knowledge base that results from converting several presented lexical knowledge bases, the sense alignments we created in Sections 2.4 and 2.6, and other alignments on the level of word sense and predicate argument structure to the UBY -LMF format.

2.8 A Linked Lexical Knowledge Base for FrameNet

Converting the lexical resources and existing resource linkings introduced in Section 2.3 together with the newly created FrameNet–Wiktionary linking to the UBY -LMF format results in UBY_{FN} , a large linked lexical knowledge base centered around FrameNet and SALSAs. The lexical knowledge bases and links in UBY_{FN} are illustrated in Figure 2.22. This section describes UBY_{FN} and summarizes the most important statistics.

	LexicalEntry	Sense	SemanticPredicate	SemanticArgument
FrameNet	9,702	11,942	1,019	9,633
WordNet	156,584	206,978	33	60
VerbNet	4,402	29,368	608	1,337
PropBank	12,912	15,446	10,669	27,004
Wiktionary-EN	379,697	474,131	-	-
SALSA 2	648	1,826	1,023	5,024
GermaNet	107,892	121,810	104	107
Wiktionary-DE	85,575	72,752	-	-
total	757,412	934,253	13,456	43,165

Table 2.10: UBY_{FN} statistics.

Included lexicons. The linked lexical knowledge base UBY_{FN} includes FrameNet release 1.5, WordNet 3.0, VerbNet 3.2, PropBank frame files,¹⁴ the English and German Wiktionary,¹⁵ GermaNet 9.0, and a lexicon representation for SALSA release 2. The different lexical knowledge bases have already been introduced in detail in Section 2.2. UBY-LMF models of the semantic lexicons were introduced in the previous Section 2.7. Note that we did not include Wikipedia in UBY_{FN}, because we focus on verb senses when using UBY_{FN}; these are scarce in Wikipedia.

Table 2.10 shows statistics of the lexicons in UBY_{FN}. It displays the number of instances in the LexicalEntry, Sense, SemanticPredicate, and SemanticArgument classes in UBY. It showcases the advantages of the collaboratively created Wiktionary: it contributes the largest number of senses for English, and the second largest number for German. The expert-built semantic lexicons on the other hand, i.e., FrameNet, SALSA, VerbNet, and PropBank contribute their information on semantic predicates and semantic roles. Note that WordNet and GermaNet also show small numbers of very basic semantic predicates and roles that were discussed in Section 2.2.

Sense alignments. UBY_{FN} links FrameNet to several English and German lexicons on the word sense level. Figure 2.22 shows the sense- and role-level alignments between the lexicons: single lines mark sense alignments, double lines mark sense and role alignments, and dotted lines stand for *transitive alignments* that are derived from other alignments by transitive combination. The sense alignment between FrameNet and the German Wiktionary is for instance derived from two sense alignments, the alignment between FrameNet and

¹⁴the version from May 25, 2016, see <https://github.com/propbank/propbank-frames>.

¹⁵We used an English Wiktionary dump from April 3, 2010 and a German Wiktionary dump from April 6 2011.

WordNet	VerbNet	PropBank	Wiktionary-EN	SALSA	Wiktionary-DE	total
28,667	10,618	2,457	12,094	744	5,897	60,477

Table 2.11: Sense links to FrameNet in UBY_{FN} .

the English Wiktionary, and the alignment between the English and German editions of Wiktionary, exploiting a transitive chain of alignments.

The sense alignment of FrameNet to WordNet stems from the open source alignments MapNet (Tonelli and Pianta, 2009b) and WordFrameNet (Laparra and Rigau, 2010). These are automatically created alignments. We preferred them to other alignments between WordNet and FrameNet (Ferrandez et al., 2010), because they are published under open-source licenses and can therefore be republished as part of UBY_{FN} .

The sense-level links to VerbNet are part of VerbNet 3.2. The sense-level links from PropBank to VerbNet and FrameNet are provided by SemLink. The sense alignments of FrameNet to the English and German Wiktionary are provided by FNWKde, which was introduced in Section 2.6.

Additional cross-lingual links connect FrameNet to SALSA. Not displayed in the figure is a linking from WordNet to GermaNet that is based on the EuroWordNet ILLI. Transitive linkings via FrameNet and the English Wiktionary connect SALSA to the German Wiktionary. Another transitive linking step from the German Wiktionary to GermaNet connects SALSA to GermaNet. The resulting linking only covers few senses, because the number of alignments between SALSA and the German Wiktionary is small, and their overlap with the alignment of the German Wiktionary to GermaNet is even smaller. UBY_{FN} does not include Wikipedia and a linking from FrameNet to Wikipedia, because we focus on verb senses when using UBY_{FN} , which are scarce in Wikipedia. If needed, a Wikipedia lexicon and the sense-level links between Wikipedia and FrameNet from Tonelli and Giuliano (2009) or Tonelli et al. (2013) could be added easily.

The statistics of the sense-level links to FrameNet are shown in Table 2.11. The linkings connect FrameNet to more than 60,477 word senses and lead to an extension of FrameNet with more than 38,000 additional example sentences for English. The sense links result in a large increase in lexicon coverage for all corpora. Token coverage for instance increases by more than 10 percent points for all considered parts-of-speech, as shown in Table 2.12. A comparison to Table 2.1 shows that the coverage of the extended FrameNet is still below the coverage of WordNet. Further extensions can be obtained by including senses that are related to the senses linked to FrameNet, e.g., synonyms or hyponyms.

Note that the new example sentences are not labeled with semantic roles. We will introduce methods to supplement this information, for instance methods for annotation projection from labeled to unlabeled sentences, in the next chapter.

part-of-speech	BNC		ANC-written		ukWAC 1-4	
	type	token	type	token	type	token
FrameNet						
adjective, noun, verb	55.44	80.58	3.41	36.41	0.31	29.56
adjective	47.69	68.81	2.21	51.86	0.42	42.83
noun	49.48	70.51	2.45	6.30	0.18	23.32
verb	77.44	95.48	34.03	60.15	3.24	75.67
FrameNet extended via sense alignments						
adjective, noun, verb	84.52	95.26	7.34	48.36	0.73	40.89
adjective	77.31	89.05	5.18	68.36	1.19	62.48
noun	82.77	93.44	6.21	11.07	0.49	32.54
verb	95.32	99.31	49.69	70.10	5.20	82.16

Table 2.12: FrameNet lexicon coverage of several large corpora in percent when including sense alignments to WordNet, VerbNet, PropBank, and Wiktionary.

Predicate argument structure alignments. On the level of predicate argument structure, FrameNet is linked to semantic predicates and roles in VerbNet and PropBank; the source of the links is SemLink. SemLink also provides a linking between semantic predicates and roles in PropBank and VerbNet.

A linking between FrameNet and SALSA was inferred by matching the FrameNet frames and roles to those used by SALSA. For this linking, we used only those SALSA frames that could be matched to FrameNet based on identical frame and role labels. Therefore, the linking does not cover proto-frames that bear labels not contained in FrameNet. Additionally, we compiled a mapping of around 20 frames whose labels changed slightly between FrameNet release 1.3, the source of the SALSA labels, and FrameNet 1.5 in order to increase the coverage of the linking. The linking from SALSA to FrameNet can also be used to induce a role-level linking from SALSA to VerbNet, as shown by the double line in Figure 2.22.

Statistics on the predicate- and role-level links are shown in Table 2.13. There are 3,182 predicate-level alignments to SALSA, 1,828 to PropBank, and 1,530 to VerbNet. The number of predicate-level alignments between PropBank and VerbNet are particularly large, because of the fine-grained model of semantic predicates in PropBank: instances of `SemanticPredicate` in PropBank are equivalent to a word sense in PropBank.

The alignments to FrameNet cover only a quarter of the 1,019 semantic frames in FrameNet, which demonstrates the complementary coverage of the resources and the need for additional predicate argument structure linkings. One option for extending the predicate argument structure links in UBY_{FN} would be to add the automatically generated linkings from the Predicate Matrix which provides additional linkings between FrameNet, VerbNet, and PropBank (Lopez de Lacalle et al., 2016).

	VerbNet		PropBank		SALSA	
	Predicate	Argument	Predicate	Argument	Predicate	Argument
FrameNet	1,530	1,054	1,828	-	3,182	2,892
VerbNet	-	-	19,610	26,634	790	1,145

Table 2.13: Predicate argument structure links in UBY_{FN} .

Summary. In this section, we introduced UBY_{FN} , a large linked lexical knowledge base centered around FrameNet. It shows a large potential for expanding the coverage of lexical knowledge bases like FrameNet for English due to a large number of sense-level links to other LKBs, as shown in Table 2.11 and Table 2.12. The sense-level links in UBY_{FN} also allow us to connect German word senses with FrameNet frames, effectively creating a German FrameNet lexicon that covers 5,897 senses in the German Wiktionary. This lexicon covers 755 FrameNet frames, more than other German FrameNet resources, see also Table 2.8.

UBY_{FN} makes use of the extended UBY-LMF model by integrating predicate argument structure links from SemLink, and providing cross-language predicate argument structure links between FrameNet and SALSA. As a result, it is the first resource that connects a German lexical knowledge base with VerbNet roles, offering new possibilities of semantic analysis for the German language.

UBY_{FN} can easily be extended by additional LKBs and alignments on the sense and predicate argument structure level. The resource integration on the level of predicate argument structure could be improved in further work, effectively increasing the coverage of SemLink. An option would be to incorporate additional automatically created predicate argument structure links from the Predicate Matrix (Lopez de Lacalle et al., 2016). The fine-grained model of predicate argument structure alignments in UBY-LMF allows us to distinguish between reliable expert-created alignments and automatic ones, for instance using the latter as a back-off solution when the former are not available.

We publish UBY_{FN} in accordance with the licenses of the contained lexical knowledge bases. Details on the resource download are provided in Appendix A.

2.9 Summary of Chapter 2

In this chapter, we discussed the coverage bottleneck for frame-semantic resources and suggest large-scale resource integration on the sense and semantic role level as a solution. FrameNet has been in development for many years, but the coverage problems still persist for the English FrameNet, the largest frame-semantic resource available; they are even more severe for other, lower-resourced languages. We address this problem by integrating FrameNet in a large network of linked lexical knowledge bases, creating the linked lexical

knowledge base UBY_{FN} centered around FrameNet. Therefore, we first describe existing resource linkings that we use as the input for the large-scale integration.

We then propose an automatic method for extending FrameNet and translating it to other languages. It is a simple, but effective approach that uses the English Wiktionary as an interlingual representation, and as the basis for the creation of FrameNet-like resources for other languages. We validate our approach on the language pair English-German and discuss the options and requirements for creating FrameNets in further languages.

As part of this work, we created the first sense alignment between FrameNet and the English Wiktionary. The resulting resource FNWKxx connects FrameNet senses to over 280 languages. The English-German FrameNet lexicon FNWKde is a sense-disambiguated version of FNWKxx for German. FNWKde competes with manually created resources, as shown by a comparison to the SALSA corpus.

Standardization enables an efficient access to the various types of lexical-semantic information encoded in the linked lexical knowledge bases. To enable efficient access to linked lexical knowledge bases linked on the level of word sense and semantic predicate argument structure, we presented our contributions to the modeling and standardization of linked lexical knowledge bases. Our specific contribution is twofold: we present a UBY-LMF model for lexical knowledge bases that contain semantic predicate argument structure, like FrameNet, and develop an extension of UBY-LMF to model predicate argument structure links. We also introduce the large linked lexical knowledge base UBY_{FN} that results from the integration and standardization of various lexical knowledge bases.

The main contributions of this chapter are:

- Contributions to UBY-LMF: the model of FrameNet in UBY-LMF, and extending UBY-LMF to model predicate argument structure links. This includes the development of methods for the conversion of the sense and predicate argument structure alignments to UBY and API methods for accessing them.
- A novel two-step approach to bootstrap FrameNet resources in various languages using Wiktionary as an interlingual connection that results in the first sense alignment between FrameNet and Wiktionary.
- The created resources: the sense alignment gold standard between FrameNet and the English Wiktionary, and the linked lexical knowledge bases FNWKxx, FNWKde, and UBY_{FN} , see also Appendix A.

The next chapter will introduce a method that uses the linked lexical knowledge base presented in this chapter to benefit FrameNet semantic role labeling for the automatic generation of frame- and role-labeled training data.

CHAPTER 3

Knowledge-based Supervision for Semantic Role Labeling

The linking of lexical knowledge bases is not an end in itself. The goal is to use them to the benefit of NLP, which raises the main research question of this chapter: in which way can linked lexical knowledge bases be used to enhance NLP tasks such as semantic role labeling? While previous work used linked lexical knowledge bases for word sense disambiguation (Ponzetto and Navigli, 2010; Cholakov et al., 2014), or entity linking for information extraction (Moro et al., 2014), we study the novel question of using them for the complex task of training data generation for SRL.

In this chapter, we present DistantSRL, a new knowledge-based method for the automatic generation of frame- and role-labeled data: linked lexical knowledge bases are used to automatically label training data for semantic role labeling following the paradigm of distant supervision (Mintz et al., 2009). The method consists of two stages: first generating frame labels, which corresponds to a word sense disambiguation task, and second the generation of role labels on the frame-labeled data.

This new method has the potential to generate training data for any language for which a suitable LLKB exists. We evaluate it for two languages, English and German, in order to prove that it generalizes to other languages besides English. The approach is applied to verbal predicates, which are particularly important for semantic role labeling, but can also be extended to other parts-of-speech. Furthermore, it is easily extensible by linking additional lexical knowledge bases and can also be applied to other role schemata such as VerbNet or PropBank, or related semantic tasks like event extraction (Kim et al., 2009), or template-based information extraction (Sundheim, 1991). Experiments for the task of frame identification and role classification show that the generated training data are of high quality and complement the FrameNet fulltext corpus.

An alternative to using knowledge bases for the automatic generation of training data is to use annotated corpora as the information source. It is important to note the distinction between knowledge-based and corpus-based approaches: purely knowledge-based approaches do not use annotated corpora, and corpus-based approaches typically ignore the wealth of lexical-semantic information encoded in knowledge bases.

Our distant supervision-based approach creates data with different properties than other approaches to training data generation for semantic role labeling, e.g., annotation projection (Fürstenu and Lapata, 2012), because it labels data sparsely and creates noisily labeled data. The extrinsic evaluation in the tasks of frame identification and role classification shows that DistantSRL nevertheless generates high quality training data.

To further prove the usefulness of our approach to semantic role labeling, we use the automatically labeled corpus for English to train the open-source semantic role labeling system SEMAFOR. We evaluate the system on a diverse set of test sets. The results prove that our training data are of high quality: training SEMAFOR frame identification on our data results in slightly lower frame identification scores compared to SEMAFOR for three out of four test sets, but to improvements for the MASC test set. Combining the FrameNet fulltext training set with our data results in statistically significant improvements for the MASC test set. The results for role labeling are similar: training the SEMAFOR role labeling component on our data also results in lower, but reasonable scores for most test sets, but to improvements for the MASC test set.

We furthermore provide a detailed discussion of training data generation in the context of state-of-the-art semantic role labeling, and discuss previous work on using (linked) lexical knowledge bases in the context of semantic role labeling. In contrast to our approach, other recent approaches that integrate information from LLKBs directly into a SRL system did not report improved performance (Kshirsagar et al., 2015).

In the next section, we provide an overview of DistantSRL, our knowledge-based method for training data generation. It is followed by a discussion of related work and the application and experimental evaluation of DistantSRL. The chapter concludes with experiments on training SEMAFOR with our automatically labeled corpora, and a discussion section.

3.1 Knowledge-based Training Data Generation with DistantSRL

This section introduces a novel method for training data generation for SRL using LLKBs called DistantSRL. It uses knowledge-based distant supervision for the sparse labeling of large amounts of unlabeled web texts. Because of the sparse labeling, it requires a very large corpus to yield a sufficient number of training instances.

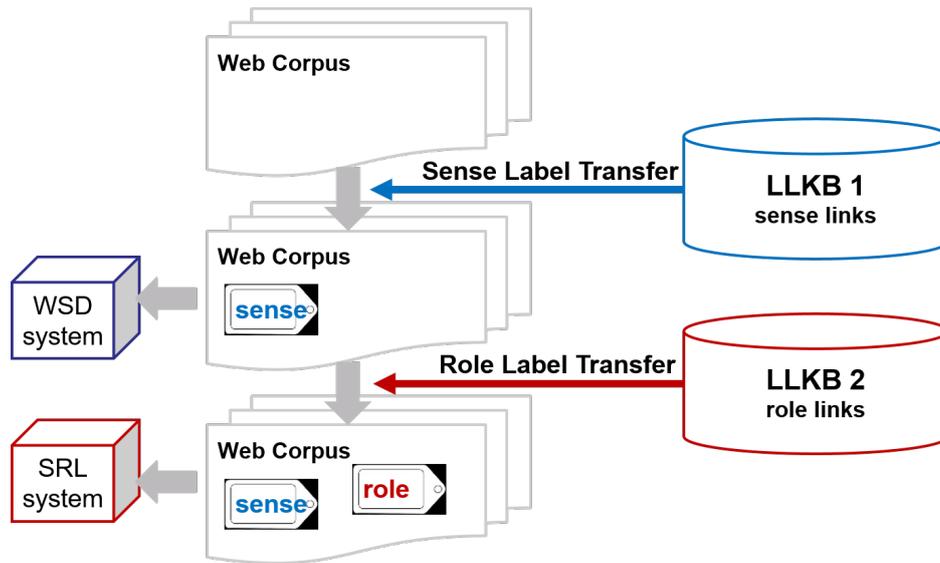


Figure 3.1: Overview of DistantSRL: automatic training data generation.

The sparse labeling is similar to annotation projection, a corpus-based method that projects frame and role labels from labeled to unlabeled corpora (Fürstenau and Lapata, 2012), but instead of using labeled sentences as information source for transferring frame and role labels, DistantSRL uses information from a set of LLKBs centered around FrameNet. It makes minimal use of frame- and role-labeled corpora. Therefore, it is appropriate for languages for which LLKBs around FrameNet have been developed or can be inferred automatically, e.g., using the approach introduced in Chapter 2, but which do not possess sufficiently large role-labeled corpora to train SRL systems. This applies, for instance, to German, Spanish, or Japanese. We apply our approach to English and German to show that it generalizes to several languages.

Unlike bootstrapping approaches, which create labels iteratively, DistantSRL labels large amounts of data in a single labeling run. This procedure is more efficient, in particular when applied to large-scale corpora, and avoids *semantic drift* that has been reported for bootstrapping approaches (Curran et al., 2007). Semantic drift refers to the problem that the results of bootstrapping diverge slightly from the input seed with each iteration, leading to the selection of unrelated instances after several iterations.

In contrast to corpus-based approaches (Fürstenau and Lapata, 2012; Exner et al., 2015), DistantSRL does not require frame- and role-labeled corpora. Instead it uses information from the linked lexical knowledge base UBY_{FN} centered around FrameNet for English and around SALSA for German.

Figure 3.1 shows the two labeling stages of DistantSRL. They follow the succession of frame and role labeling that has become standard for FrameNet SRL: in the first stage, *Stage 1*, we transfer sense labels to the unlabeled corpus. In the second stage, *Stage 2*, we expand

the sense-labeled data to role-labeled data. Both labeling stages use large-scale corpora and LLKBs as knowledge sources. The first stage, uses large web corpora from WaCky (Baroni et al., 2009) and the linked lexical knowledge base UBY_{FN} presented in Section 2.8. More specifically, it uses the linked FrameNet, WordNet, Wiktionary and WordNet for English; for German, it uses an LLKB consisting of SALSA and the German Wiktionary. The first stage is presented in detail in Section 3.1.2. The second stage, see Section 3.1.3, uses the sense labeled corpus from the first stage and the sense- and role-level links in SemLink (Bonial et al., 2013) as introduced in Section 2.3.1.

We evaluate DistantSRL extrinsically in two different setups. The first setup separately evaluates Stage 1 and Stage 2: it evaluates Stage 1 in a frame disambiguation task (FrameId, cf. Equation 1.1) and Stage 2 in a role classification task (RoleC, cf. Equation 1.3) on four FrameNet-labeled test sets from different domains. This evaluation proves that the proposed approach creates high-quality training data for the different stages of FrameNet SRL.

The second evaluation setup aims to show that the resulting corpora are also useful in a standard SRL setup: we use the automatically labeled corpora to train an open-source SRL system, SEMAFOR 3.0 (Das et al., 2014), and evaluate the resulting system on the test sets. We use SEMAFOR, because it is the most advanced open-source FrameNet SRL system available to date. For a long time, it was the state-of-the-art SRL system, and it has only recently been superseded by systems based on deep-learning (Hermann et al., 2014; FitzGerald et al., 2015). Most of the results described in this section were previously published in a journal article by Hartmann et al. (2016).

The next sections first introduce the two stages of DistantSRL in detail, and then present the application of DistantSRL to English and German, including the extrinsic evaluation.

3.1.1 Formalization

As a semi-supervised approach for the automatic labeling of FrameNet senses and roles, DistantSRL can be formalized as a knowledge-based label transfer approach (Pan and Yang, 2010). The following formalization is a citation from Hartmann et al. (2016), page 199:

“Given a set X of seed instances derived from knowledge sources and a label space Y , a set of labeled seed instances consists of pairs $\{x_i, y_i\}$, where $x_i \in X$, and $y_i \in Y$; $i = 1, \dots, n$. For an unlabeled instance $u_j \in U$, $j = 1, \dots, m$, where U is a large corpus and $U \cap X = \emptyset$, we employ label transfer from $\{x_i, y_i\}$ to u_j based on a common representation r_{x_i} and r_{u_j} using a matching criterion c . The label y_i is transferred to u_j if c is met.

For the creation of sense labeled data, we perform pattern-based labeling, where Y is the set of sense labels, r_{x_i} and r_{u_j} are sense patterns generated from corpus instances and from LKBs including sense-level links, and c considers the similarity of the patterns based on a similarity metric.

Sentence	He felt no sense of guilt in the betrayal of personal confidence.
LSP	he feel no sense of guilt in
ASP	PP feel <i>cognition of feeling</i> in <i>act</i>

Figure 3.2: Example: results of Step 1A – seed patterns.

We create role-labeled data with rule-based labeling where Y is the set of role labels, r_{x_i} and r_{y_j} are attribute representations of roles using syntactic and semantic attributes. Attribute representations are derived from parsed corpus instances and from linguistic knowledge, also including role-level links from LKBs; here, c is fulfilled if the attribute representations match.”

In the next two subsections, we introduce the methods for knowledge-based sense labeling and knowledge-based semantic role labeling in detail.

3.1.2 Knowledge-based Sense Labeling

This section introduces our method for knowledge-based label transfer of FrameNet word sense information, i.e., FrameNet frame labels, referred to as Stage 1 *Sense Label Transfer* in Figure 3.1. It extends the methodology by [Cholakov et al. \(2014\)](#), who exploit sense-level information from UBY for the automatic sense-labeling of corpora with verb senses from WordNet, in two ways: first, it provides an adaptation to FrameNet frames, and second, and more importantly, it adds a precision-enhancing discriminative filter that improves the quality of the resulting sense-labeled corpus.

The enhanced method consists of three steps: Step 1A extracts lexico-semantic patterns called seed patterns for the target word sense inventory from the LLKBs in UBY_{FN} ; the new Step 1B applies a discriminating filter to the seed patterns; Step 1C compares the filtered seed patterns to the same type of patterns in a large corpus and transfers the sense labels to similar corpus instances. Step 1A and Step 1C follow the model of [Cholakov et al. \(2014\)](#), but Step 1B extends their approach significantly, aiming to increase the precision of the sense label transfer.

Step 1A: seed pattern extraction. Step 1A extracts a set of lexico-semantic patterns called *seed patterns* from the sense examples in the LLKBs. [Cholakov et al. \(2014\)](#) propose two types of patterns: lemma sense patterns (LSPs) consist of the target verb lemma and the lemmas surrounding it in a context window of size w . Abstract sense patterns (ASPs) present a stronger generalization from the surface form: they consist of the target verb lemma and rule-based generalization of context words in a context window of size w . More specifically, the following generalizations are applied: (i) words with selected parts-of-speech, e.g.,

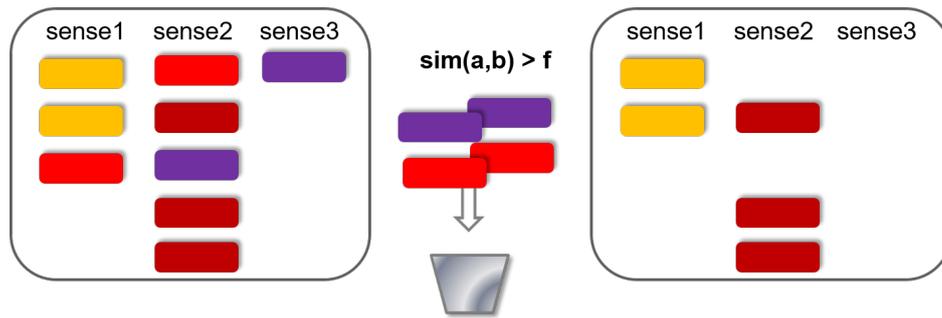


Figure 3.3: Illustration of Step 1B: seed pattern filtering. Colored boxes represent seed patterns, each column represents a seed sense for the same lemma. Patterns across senses whose similarity exceeds threshold f are discarded.

verbs, adjectives, and prepositions, are replaced by their part-of-speech tag, (ii) nouns and named-entity tags are replaced by noun classes as given by their WordNet semantic fields, and (iii) a predefined set of function words is kept as part of the pattern, others are deleted. The full lists of part-of-speech tags and function words provided by Cholakov et al. (2014) can be obtained here: <https://www.ukp.tu-darmstadt.de/data/sense-labelling-resources/verb-sense-labelling/>.

Following Cholakov et al. (2014), we create the patterns for VerbNet in a slightly different way: VerbNet does not provide example sentences for all word senses, instead it contains prototypical examples for each VerbNet class. So it does not easily lend itself to the generation of LSPs, which are omitted for VerbNet. For the generation of ASP patterns, we use information on the semantic predicate argument structure and syntactic frames in VerbNet. They provide information on the selectional preferences of arguments, which is used to derive the semantic field labels, and information on word order and prepositions.

The example in Figure 3.2 shows an LSP and an ASP pattern for the FrameNet sense *Feeling* of the verb *feel* in the sense example “*He felt no sense of guilt in the betrayal of personal confidence*”.

The example shows that ASPs provide a larger level of abstraction. As a consequence, they generalize to a larger number of contexts and serve to identify productively used verb senses, LSPs on the other hand serve to identify lexicalized, fixed constructions, for instance multiword expressions such as *carry out* or *wrap up*.

Step 1A introduces a single variable parameter: the size of the context window w . In our experiments, we follow the suggestion by Cholakov et al. (2014) and set the window size to 7 for LSPs and to 5 for ASPs.

Step 1B: seed pattern filtering. The high generalization abilities of the patterns in Step 1A ensure high recall in matching the seed patterns (see Step 1C below). But these properties also constitute a serious drawback of Cholakov et al. (2014)’s method: Step 1A extracts a

certain number of very similar (or even identical) seed patterns for *different* senses. Such seed patterns are not able to discriminate well between different senses and may lead to additional noise in the sense-labeled data. The precision of the sense labeling approach suffers. To increase the precision of the sense labeling, we developed an optional *discriminating filter* on the seed patterns that is applied after Step 1A.

The intuition behind the discriminating filter is the following: some of the ASP and LSP patterns which we extract from the seed instances *discriminate* better between senses than others; i.e., if the same or a very similar pattern is extracted for sense w_i and sense w_j of a word w , $i, j \in (1, \dots, n)$, n =number of senses of w , $i \neq j$, this pattern does not discriminate well, and should not be used when labeling new senses.

We filter the ASP and LSP patterns by comparing each pattern for sense w_i to the patterns of all the other senses w_j , $i \neq j$ using the similarity metric *sim* introduced in Equation 3.1 below; if we find two patterns w_i , w_j whose similarity score exceeds a filtering threshold f , we greedily discard them both. The process is illustrated in Figure 3.3 and formalized as follows:

For a given lemma L with k senses l_1, \dots, l_k , we extract k sets of seed patterns S_1, \dots, S_k for each sense. These build the set of seed patterns S_L for lemma L , i.e., $S_L = \{S_1, \dots, S_k\}$. Each S_j is a set of seed patterns for a sense of L and consists of a different number v of seed patterns, i.e., $|S_j| = v$. For each pair of pattern sets $(S_i, S_j) \in S_L$ such that $i \neq j$, with $|S_i| = u$ and $|S_j| = v$, we compare all pairs of patterns $(p_{i,x}, p_{j,y}) \in S_i \times S_j$, where $S_i \times S_j = \{p_{i,x}, p_{j,y} \mid p_{i,x} \in S_i \ \& \ x \in (1 \dots u) \ \& \ p_{j,y} \in S_j \ \& \ y \in (1 \dots v)\}$. We discard $p_{i,x}$ and $p_{j,y}$ if $\text{sim}(p_{i,x}, p_{j,y})$ is larger than a threshold f .

The filtering may increase precision at the cost of recall, because it reduces the number of seed patterns. Since we use the approach on large corpora, we still expect sufficient recall. Our results show that the application of the discriminating filter improves the quality of the automatically labeled corpus.

Essentially, the discriminating filter in Step 1B integrates the goal of capturing sense distinctions into the sense label transfer. The same goal is pursued by Corpus Analysis Patterns (CPA patterns, Hanks (2013)), which have been created to capture sense distinctions in word usage by combining argument structures, collocations and an ontology of semantic types for arguments. In contrast to our fully automatic approach, developing CPA patterns based on corpus evidence originally was a lexicographic effort. Now, there are first efforts at creating CPA patterns automatically in a SemEval shared task (Baisa et al., 2015). The example in Figure 3.4 contrasts two ASP patterns to a CPA pattern from Popescu et al. (2014) for the verb *abandon*.

The abstract ASP patterns look similar to CPA-patterns, as they also abstract argument fillers, the words that fill an argument position, to semantic classes and preserve certain function words. $[[Human]] \mid [[Institution]]$ in the CPA pattern matches *person* in the ASP patterns, $[[Activity]] \mid [[Plan]]$ matches *communication* and *cognition*.

CPA	[[Human]] [[Institution]] abandon [[Activity]] [[Plan]]
ASP 1	<i>person abandon communication</i> which VVD PP JJ in
ASP 2	JJ <i>person abandon</i> JJ <i>cognition</i> of JJ <i>quantity</i>

Figure 3.4: Contrasting ASP patterns to CPA patterns.

Step 1B introduces a single variable parameter f , the filtering threshold. We determine this parameter using extrinsic evaluation on a held-out set from the FrameNet corpus.

Step 1C: sense label transfer. To transfer sense labels from the filtered seed patterns to a large, unlabeled corpus C , Step 1A is applied to all sentences c_j from C which contain a target verb from the seed patterns, leading to corpus patterns p_{c_j} . For every sentence c_j , the extracted patterns p_{c_j} are compared to the labeled seed patterns using a similarity metric.

The similarity metric $sim \in [0..1]$ proposed by Cholakov et al. (2014) is based on Dice’s coefficient and considers the common n -grams between two sense patterns, $n = 2, \dots, 4$:

$$sim(r_{x_i}, r_{y_j}) = \frac{\sum_{n=2}^4 |G_n(p_1) \cap G_n(p_2)| \cdot n}{norm_w} \quad (3.1)$$

where $w \geq 1$ is the size of the window around the target verb, $G_n(p_i)$, $i \in \{1, 2\}$ is the set of n -grams occurring in r_{x_i} and r_{y_j} , and $norm_w$ is the normalization factor defined by the sum of the maximum number of common n -grams in the window w . Using n -grams instead of unigrams takes word order into account, which is particularly important for identifying verb senses, as their syntactic and semantic properties often correlate.

The most similar seed pattern whose similarity exceeds a threshold t is used to label c_j . In practice, there often are several (equally similar) most similar seed patterns. This may lead to a set of acceptable sense labels for c_j . In this case, a single sense label is selected randomly from the set. To ensure high precision, LSPs get precedence over ASPs whenever an LSP and an ASP receive the same similarity score, i.e., in this case the sense label from the LSP is used to label c_j .

The following example illustrates Step 1C: we extract an LSP and an ASP from the unlabeled sentence “*I feel strangely sad and low-spirited today*” and the verb *feel*. The LSP lp is simply *I feel strangely sad and low-spirited today*, the ASP p is *PP feel JJ JJ time*. These patterns are compared to the seed patterns, and based on the most similar ASP, the sentence is labeled with the sense label *Feeling*. To illustrate this process, we show a sample of seed ASP patterns that are compared to p and the corresponding similarity scores in Figure 3.5. In the figure, the pattern with the highest score is selected for label transfer, and its sense label (*Feeling*) is assigned to the example sentence.

Corpus pattern p	Seed pattern ASP s	Seed frame	sim(p,s)	Selected
PP feel JJ JJ <i>time</i>	PP feel JJ JJ to me PP VVP PP	Feeling	0.21	X
PP feel JJ JJ <i>time</i>	<i>state</i> PP feel JJ JJ <i>act into animal</i>	Perception_experience	0.13	-
PP feel JJ JJ <i>time</i>	JJ JJ that PP feel in <i>artifact for artifact</i>	Seeking	0.03	-
PP feel JJ JJ <i>time</i>	PP feel to VV VVD	Appearance	0.01	-
PP feel JJ JJ <i>time</i>	PP feel that PP VVZ VV PP	Opinion	0.01	-

Figure 3.5: Example: Step 1C – sense label transfer.

Step 1C introduces three variable parameters, namely n -gram size n for the similarity metric, window size w and threshold t . We follow [Cholakov et al. \(2014\)](#) in setting n to 4 and w to 7 and determine the appropriate threshold t for our problem using extrinsic evaluation on a held-out set from the FrameNet fulltext corpus.

This concludes the introduction of Stage 1, the sense label transfer stage of our approach. The next section introduces Stage 2, the role label transfer, which uses the results of the sense label transfer. Note that this approach leads to a sparse labeling of the unlabeled corpus C , since many unlabeled sentences will be discarded because their similarity to the seed patterns is too low. This effect is enhanced by the seed pattern filtering in Step 1B. The sparse labeling does not pose a problem for training data generation, because our approach scales to very large corpora. It only requires shallow preprocessing, e.g., tokenization, lemmatization, POS-tagging, named-entity recognition, and semantic tagging with WordNet semantic fields. Using very large unlabeled corpora, we can obtain sufficient recall to train a frame identification system, as shown in Section 3.4.2, which describes the application of Stage 1 to large corpora and experimental evaluation of the resulting corpora. Now, we introduce Stage 2 of DistantSRL, the knowledge-based role labeling.

3.1.3 Knowledge-based Role Labeling

This section presents the linguistically informed approach to the automated labeling of large-scale corpora with FrameNet roles introduced in [Hartmann et al. \(2016\)](#), or Stage 2 in Figure 3.1.¹⁶

This method requires rich linguistic preprocessing, i.e., dependency parsing, and uses sense- and role-level links in the LLKB SemLink. It uses the result of Stage 1, a large sense-labeled corpus, as input for automatic role label transfer. Otherwise, it only uses information from LLKBs, specifically VerbNet and SemLink, and does not rely on existing role-labeled corpora. The role label transfer consists of two steps, first the automatic, redundant labeling with VerbNet roles, and second translating the VerbNet roles to FrameNet.

¹⁶This section presents joint work. The approach for knowledge-based role labeling presented in this subsection was mainly developed by Dr. Judith Eckle-Kohler, one of the co-authors in [Hartmann et al. \(2016\)](#).

Sentence	[I] feel _{Feeling} strangely [sad and low-spirited today].	
Argument	I	strangely sad and low-spirited today
VerbNet role	<i>Experiencer, Pivot</i>	<i>Theme, Co-Theme, Topic</i>
FrameNet role	<i>Experiencer</i>	<i>Emotion</i>

Figure 3.6: Example: DistantSRL Stage 2 – creating VerbNet and FrameNet roles.

Step 2A: VerbNet role label transfer. In Step 2A, a set of deterministic rules is applied to label syntactic arguments of the sense-labeled predicate targets from Step 1C with VerbNet semantic roles. The precision-oriented rules use only information from the LKB VerbNet and build on the results of linguistic preprocessing. The approach is based on a compact list of 28 role labels from VerbNet release 3.2. Thus, a set of rules is an efficient way to generate VerbNet role labels on text; it took an expert three days to develop the rules, using a sample of the VerbNet annotations on PropBank from SemLink as a development set.

The preprocessing, performed using components from DKPro Core (Eckart de Castilho and Gurevych, 2014), includes lemmatization, POS-tagging, named-entity recognition, dependency parsing with the Stanford Parser (De Marneffe et al., 2006), and semantic tagging with WordNet semantic fields. For the semantic tagging, the most-frequent-sense heuristic is used, i.e., the semantic field associated with the most frequent sense for a target word is selected. This strategy works well for the coarse-grained semantic types represented in the WordNet semantic fields. Named-entity tags are also mapped to the semantic fields.

The approach uses the output of the dependency parser to identify the syntactic arguments of the predicate targets as the dependents of the predicate in the dependency graph. The argument phrases are represented by their syntactic heads. Argument spans are derived by computing the yield of the head of the argument phrase, i.e., collecting all the dependent words.

A chain of hierarchically organized rules is applied to the argument heads. For English, there are in total 57 rules. For German, there are 26 rules, which were adapted from the English rules. The rules consume the dependency label in combination with named-entity tags or semantic fields of the argument head, the governing predicate, or both. An example rule is: annotate the role *Experiencer* for the dependency *nsubj* if the governor’s semantic field is *perception* or *emotion*, otherwise annotate the role *Agent*. This leads to the annotation of *I* with the *Experiencer* role in the example *I feel*_{Feeling} *strangely sad and low-spirited today*. An example of a rule that uses semantic field information is the following: the dependency *prep_with* triggers the annotation of the role *Instrument* if the dependent is neither labeled with the semantic field *person*, nor with *group*.

For each argument candidate, the rules are applied sequentially. The rule-chain first assigns the roles *Location* and *Time* to arguments which are tagged as a *location* or labeled

with the semantic field value *time*. Then the other roles are annotated. Often, it is not possible to determine a single VerbNet role based on the available linguistic information. Of the 57 rules, 32 assign one role, 5 assign two roles, and 20 assign three roles. The distinction between *Theme* and *Co-Theme*, for instance, cannot be made, because it relies on information that is not available to the system, e.g., it first needs to know that there are two arguments that qualify for the *Theme* role and second needs to decide which of the two is more salient and thus receives the *Theme* label. In this case, multiple roles are annotated and transferred to the subsequent Step 2B, where some of them can be disambiguated using the VerbNet–FrameNet mapping in SemLink. The example in Figure 3.6 shows semantic arguments annotated with multiple VerbNet roles for the previously introduced example sentence “*I feel strangely sad and low-spirited today*”. The full set of rules for Step 2A is documented here: www.ukp.informatik.tu-darmstadt.de/knowledge-based-srl/.

Step 2B: mapping to FrameNet roles via SemLink. In Step 2B the VerbNet roles from Step 1A are mapped to FrameNet roles using a) information on the FrameNet sense label and b) sense- and role-level links between VerbNet and FrameNet in UBY_{FN} stemming from SemLink: the frame label filters the available role-level mappings. Admissible role linkings are those for the VerbNet senses with the given lemma that are linked to the given FrameNet sense in SemLink. The VerbNet role label resulting from Step 1A is translated to a FrameNet role label based on those linkings.

The information on the FrameNet frame label thereby constrains the one-to-many mapping of the VerbNet roles to the fine-grained FrameNet roles. The VerbNet role *Agent*, for instance, is linked to a large number of different FrameNet roles across frames, for instance *Speaker* for the frame *Request* and *Cause* for the frame *Inhibit_movement*.

Figure 3.7 shows an example mapping from SemLink for the verb *feel*. Applying this mapping leads to the annotation of two FrameNet roles in the example sentence “*I feel strangely sad and low-spirited today*”. The *predicate* *feel* has already received the frame label *Feeling* in Stage 1, so the VerbNet labels $\{Experiencer, Pivot\}$ assigned to *I* in Step 2A are mapped to the FrameNet role *Experiencer*, as shown in Figure 3.6.

In many cases, the mapping step leads to the assignment of unique FrameNet role labels. There is, however, a proportion of instances for which the mapping step leads to the annotation of multiple FrameNet roles. Examples for set labels are *Interlocutor_1* and *Interlocutor_2* for the *Discussion* frame, or *Agent* and *Cause* for the *Damaging* frame. The distinction between the two *Interlocutor* roles is somewhat arbitrary, while further disambiguation may be desired for the distinction between *Agent* and *Cause*.

The example in Figure 3.8 shows corpus instances created with DistantSRL that contain multiple FrameNet roles. We create set-valued role labels for these instances, for instance $\{Agent, Cause\}$ or $\{Theme, Co-Theme\}$.

Verb	Frame	VerbNet role	FrameNet role
feel	Appearance	<i>Experiencer</i>	<i>Perceiver_passive</i>
		<i>Stimulus</i>	<i>Phenomenon</i>
	Feeling	<i>Experiencer</i>	<i>Experiencer</i>
		<i>Stimulus</i>	<i>Emotion, Emotional_state</i>
		<i>Theme</i>	<i>Emotion</i>
	Opinion	-	-
	Perception_active	<i>Experiencer</i>	<i>Perceiver_agentive</i>
		<i>Stimulus</i>	<i>Phenomenon</i>
	Perception_experience	<i>Experiencer</i>	<i>Perceiver_passive</i>
		<i>Stimulus</i>	<i>Phenomenon</i>
	Seeking	<i>Agent</i>	<i>Cognizer_agent</i>
		<i>Location</i>	<i>Ground</i>
		<i>Theme</i>	<i>Sought_entity</i>

Figure 3.7: Example: SemLink predicate and role mappings for the verb *feel*.

As mentioned above, it should be possible to reduce some of the set-valued role labels to single labels in an additional disambiguation step. This could be done by taking additional information into account, for instance information on the semantic type of the role filler, e.g., whether the word that fills the role slot has agentive properties like *being sentient*. Such information has also been added to the VerbNet lexicon in the shape of properties associated with role labels, e.g., *Agent[+animate/+organization]*, and could be derived from the VerbNet lexicon in UBY_{FN} , but it is not included in SemLink.

The SemLink mapping is not complete, i.e., it does not cover all predicates and roles in VerbNet and FrameNet. It covers 58% of the verb lemmas in FrameNet, 49% of the FrameNet frame labels, and 27% of the fine-grained FrameNet role labels.

Because of this, the mapping in Step 2B results in partially labeled data. This means that a sentence may contain only a single predicate-role pair, even though other arguments of the predicate are present and have been labeled with a VerbNet role. The experiments in Section 3.4 below show that we can train semantic role classifiers successfully on the partially labeled data we create with DistantSRL.

Intrinsic evaluation of Step 2A and Step 2B. In [Hartmann et al. \(2016\)](#) we evaluated Step 2A and Step 2B intrinsically on held-out sets to confirm their quality. Step 2A was evaluated on a test sample of VerbNet annotations on PropBank. For Step 2A, the percentage of correctly annotated roles among all annotated roles is 96.8% under the condition that instances labeled with multiple roles are considered correct if the set of roles contains the

Sentence	Don't bend over the engine to lift it, [you] can damage _{Damaging} your back doing it this way.	
	Argument	FrameNet role
	you	<i>Agent, Cause</i>
Sentence	The fire was contained to the upper floor, although [the bar and downstairs rooms] are damaged _{Damaging} by [smoke and water].	
	Argument	FrameNet role
	the bar and downstairs rooms	<i>Patient</i>
	<i>smoke and water</i>	<i>Agent, Cause</i>
Sentence	[Landlord] discussed _{Discussion} non-payments of rent with tenant.	
	Argument	FrameNet role
	Landlord	<i>Interlocutor_1, Interlocutor_2</i>

Figure 3.8: Example: DistantSRL Step 2B – corpus instances with multiple roles.

gold label. The percentage of instances where a rule assigns at least one role was 77.4%. This makes the approach a high-precision approach, with sufficient recall.

Step 2B was evaluated on the FrameNet fulltext test set; the precision, i.e., the percentage of correctly annotated roles among all annotated roles, is 76.47%, again instances with multiple labels are considered correct if they contain the gold label.

This concludes our introduction of DistantSRL. In the next subsection, we elaborate how DistantSRL implements the paradigm of distant supervision, and discuss the advantages of distant supervision in comparison to related methods such as bootstrapping or self-training.

3.1.4 Distant Supervision

The approach for training data generation using lexical knowledge bases introduced in this chapter implements the paradigm of *distant supervision*. Distant supervision is a type of semi-supervised learning that has been introduced by Mintz et al. (2009) for relation extraction. The general principle in distant supervision is to align unlabeled text to a knowledge base using a simple and efficient matching strategy. Based on this alignment, information from the knowledge base, for instance on relation labels, can be transferred to the text and serve as labels for supervised learning.

In Mintz et al. (2009), the matching strategy is to identify two named-entities that take part in a relation r from the knowledge base Freebase in unlabeled text. This is also called *entity matching*. Then the relation label r is transferred from the knowledge base to the

instances in the text. Thus, the unlabeled text is then labeled as an instance of relation r , and can be used to train a classifier for relation identification.

The two stages of DistantSRL implement the distant supervision paradigm as follows: the formalization of DistantSRL in Section 3.1.1 referred to the matching strategy as matching criterion c . For Stage 1 of DistantSRL, the matching criterion c is the similarity-based matching of seed patterns from the LLKB to similar patterns extracted from the unlabeled corpus. For Stage 2, the matching criterion c is the equality of attribute representations derived from linguistic knowledge in the LLKB (encoded as rules) to attribute representations extracted from the unlabeled corpus.

Distant supervision implements a *transductive* learning setup (Zhu and Goldberg, 2009). In transductive learning, unlabeled data are labeled with the aim to generate additional training data. In contrast, a system based on *inductive learning* is used to label test data directly. This terminology is important to distinguish DistantSRL from alternative approaches to knowledge-based SRL introduced in Section 3.2.

Relation to bootstrapping and self-training. Distant supervision tries to remedy some of the disadvantages of other semi-supervised approaches such as bootstrapping. Bootstrapping approaches use a small number of instances and a supervised classifier or a set of heuristics to label data in several iterations. Yarowsky (1995) introduced bootstrapping for word sense disambiguation. If a sufficient number of labeled instances to train a supervised system is initially available, the method is also called self-training. Self-training can be applied in an iterative, bootstrapping setup or in a single run on a large dataset, i.e., similar to the distant supervision setup. Bejan (2009), for instance, evaluate self-training for the frame identification step of SRL and find better results in a single training run as opposed to multiple iterations.

The iterations in bootstrapping may lead to semantic drift (Curran et al., 2007) and low precision of the resulting labels. Semantic drift describes the problem that the results of bootstrapping diverge slightly from the input seed with each iteration, leading to the selection of unrelated instances after several iterations.

In contrast to iterative approaches, such as bootstrapping and bootstrapping via self-training, distant supervision avoids semantic drift by creating potentially noisy labels for large-amounts of data in a single run.

Distant supervision for linguistic categories. Distant supervision has been popular for the task of relation extraction, but it has hardly been evaluated for tasks like word sense disambiguation or semantic role labeling.

Exner et al. (2015) focus on the aspect of distant supervision that uses entity matching with subsequent label transfer, with the goal to transfer PropBank semantic role labels cross-lingually. They do not explicitly align a knowledge base to unlabeled text, but use

an automatically role-labeled corpus as the source of role labels that get transferred to an unlabeled corpus in a different language. Specifically, they parse English Wikipedia articles and label them automatically with PropBank roles. Then they identify named-entities as Wikipedia entities in the English articles and in the Swedish version of these articles. The named-entities are mapped to unique Wikidata identifiers (Vrandečić and Krötzsch, 2014). Where occurrences of these named-entities function as fillers of semantic roles in the English texts, they are used as anchors for the cross-lingual alignment of the sentences: Swedish sentences with the same named-entities as the English sentences are selected as targets for the transfer of role labels. Swedish semantic predicates are automatically identified as the governing verb of the aligned arguments. This procedure is bound to overgenerate, therefore the predicate mappings are filtered based on the frequency of the predicate mapping in the full corpus: Swedish predicates receive their sense according to their most frequent English counterpart. Then, the automatically labeled Swedish sentences are used to reconstruct PropBank frame files for Swedish and train a Swedish PropBank-style SRL system. Since this approach does not rely on mapping text to a knowledge base, which characterizes distant supervision, it could also be called an instance of cross-lingual annotation projection using weakly parallel corpora and cross-lingual alignments based on named-entity matching.

In our work, we focus on a different aspect of distant supervision which is transferring information from *knowledge bases*, specifically linked lexical knowledge bases, to text. Here, the sources of the mapping are not instances of semantic predicates and role labels, but representations of these information types in an LLKB. Henceforth, the term distant supervision is used specifically with this focus in mind.

We already introduced the approach by Cholakov et al. (2014). They propose a distant supervision strategy for automatic verb sense labeling that uses a lexical knowledge base. The process of mapping text to a knowledge base is performed based on syntactic-semantic patterns that are generated from the sense instances in the knowledge base and then mapped to the same kind of patterns extracted from the unlabeled text. They use this strategy to label large amounts of data with verb senses from WordNet and show that the resulting training data improve verb sense disambiguation. We use their work as a starting point for DistantSRL, a method that implements the paradigm of distant supervision for the complex task of SRL, creating both frame- and role labeled training data.

In the next sections, we discuss related work in knowledge-based SRL and training data generation for SRL, starting with research on SRL that uses LKBs in the next section.

3.2 Semantic Role Labeling using Knowledge Bases

There are different ways of using (linked) lexical knowledge bases for SRL. Some approaches, often explicitly called knowledge-based, aim to create a SRL system directly from the Frame-

Net lexicon (Shi and Mihalcea, 2004a; Litkowski, 2010). Others attempt to integrate information from LLKBs into a supervised SRL system that mainly relies on features extracted from the role annotations in a labeled corpus, e.g., Burchardt et al. (2005); Johansson and Nugues (2007b); Pennacchiotti et al. (2008). The goal is to create more robust semantic role labeling systems, i.e., systems with a larger lexicon coverage than the limited coverage provided by the FrameNet corpus. This section introduces related work of both kinds.

3.2.1 Knowledge-based Semantic Role Labeling

There are two types of knowledge-based SRL studied in previous work: *rule-based semantic role labeling* and *bootstrapping approaches*.

Shi and Mihalcea (2005, 2004a,b) describe a rule-based system for FrameNet SRL that builds on the results of syntactic parsing for the rule-based assignment of semantic roles to syntactic constituents. The role assignment uses rules induced from the FrameNet fulltext corpus. These rules encode sentence-level features of syntactic realizations of frames; they are combined with word-level semantic features from WordNet including the countability of nouns or attribute relations of an adjective indicating which nouns it can modify. Since the coverage of the induced rules is low, they are complemented by default rules. The accuracy of this approach for semantic role assignment is reported to be 74.5% on a held-out set of 350 sentences in Shi and Mihalcea (2004a).

Shi and Mihalcea (2005) also present an early instance of using LLKBs for SRL: the above-mentioned sense-level linking from WordNet to FrameNet is created manually and used to derive a semi-automatic linking of VerbNet verbs to FrameNet frames through the linking of VerbNet senses to WordNet senses provided by VerbNet. They also create a role-level linking of FrameNet and VerbNet in order to exploit descriptions of selectional preferences for roles in VerbNet in their system. The linking is based on the syntactic descriptions of verbs in VerbNet and syntactic information for verbs in FrameNet derived from the role annotations in the FrameNet corpus. Additionally, selectional preferences in VerbNet are linked semi-automatically to the WordNet hierarchy. This linking is used to expand role fillers with their WordNet hypernyms. Shi and Mihalcea (2004a) report that their rule-based system is more robust after integrating the various linkings.

The approach to knowledge-based semantic role labeling introduced by Litkowski (2010) does not integrate additional lexical knowledge bases. Litkowski (2010) uses a dictionary extracted from the annotated sentences in FrameNet¹⁷ to recognize and assign semantic roles. Their semantic role labeling system first performs FrameNet sense disambiguation and then tries to match syntactic constituents produced by a parser with syntactic patterns included in a dictionary built from FrameNet fulltext annotations. The system is evaluated

¹⁷Note that early versions of FrameNet— before release 1.5 — did not explicitly distinguish between annotated examples and fulltext corpus.

on the SemEval-2 task on *Linking Events and their Participants in Discourse* (Ruppenhofer et al., 2010b). The evaluation shows very low recall, which is mainly due to the low coverage of their FrameNet dictionary with regard to syntactic patterns.

Swier and Stevenson (2004, 2005) present a knowledge-based approach to VerbNet semantic role labeling using a bootstrapping strategy, first annotating unambiguous roles for a verb, then using this to train a probabilistic classifier and annotating more roles in each iteration. Swier and Stevenson (2005) use a role-level linking of FrameNet to VerbNet to create development and evaluation datasets for their systems from FrameNet-annotated data. They consider 16 of the 22 VerbNet roles available at the time. They report that their approach outperforms more complex bootstrapping models, but does not achieve the performance of the contemporary models trained in supervised fashion on PropBank data.

As already alluded by Swier and Stevenson (2005), knowledge-based semantic role labeling systems nowadays do not play a big role in contrast to increasingly complex supervised semantic role labeling systems.

In section 3.1, we presented our knowledge-based method for FrameNet semantic role labeling that is, unlike most of the related work presented above, used in a *transductive* manner. This means that the knowledge-based method is not used for role labeling directly. Instead, we use it to generate sparse and noisily labeled training data which can then be used to train a supervised semantic role labeling system.

Most of the previous work focused on a single lexical knowledge base. If linkings of knowledge bases were used, e.g., by Shi and Mihalcea (2005), they were specifically tailored to the involved lexical knowledge bases and the developed system. In contrast, the approach introduced above is able to use sense- and role-level linkings in a generic fashion. Additional sense-level links increase the available seeds for sense labeling, additional role-level links either increase the role coverage, or – if they introduce a knowledge base with a role schema – allow us to apply the approach to a new role schema.

3.2.2 Semantic Role Labeling using Linked Lexical Knowledge Bases

In the previous section, we introduced knowledge-based approaches to semantic role labeling. Besides this, there are also supervised approaches that make use of additional (linked) lexical knowledge bases. They typically either target the frame identification or the role labeling step of semantic role labeling.

Frame identification using LLKBs. Early approaches attempt to increase the FrameNet lexicon coverage and to improve frame identification using linkings to WordNet: Burchardt et al. (2005) use an ad-hoc linking to WordNet that is established by WordNet sense disambiguation of the target word to expand the FrameNet lexicon. Ide (2006) suggests a WSD-based method to link FrameNet to WordNet that uses a variant of the Lesk algorithm

(Banerjee and Pedersen, 2003) to assign WordNet senses to FrameNet senses in order to extend the lexicon coverage for various tasks (including semantic role labeling).

Johansson and Nugues (2007b) train a classifier to label WordNet lemmas (represented as distributions over their synsets) with frames, thus identifying new lemmas to extend the FrameNet lexicon. This results in increased recall at slightly reduced precision for FrameNet semantic role labeling (Johansson and Nugues, 2007a).

Pennacchiotti et al. (2008) also attempt to improve the FrameNet lexicon coverage on the frame level. They evaluate distributional and WordNet-based approaches of assigning frames to unseen words of different part-of-speech, e.g., nouns, verbs, and adjectives. The distributional approach assigns a frame label to an unknown word based on the similarity of a distributional representation of the word to the distributional representation of a frame which is defined as the set of its lexical units. The distributional representations are either word-based or syntax-based, i.e., based on dependency triples, or a combination of both. The WordNet-based approach compares the part-of-speech-specific WordNet subgraph of the lexical units of a FrameNet frame to the subgraphs of the possible WordNet senses of the unknown word, assigning the frame with the overall largest similarity to one of the WordNet senses. Pennacchiotti et al. (2008) evaluate the accuracy of their frame assignment method in a leave-one-out cross-validation setup. They report the highest accuracy in frame assignment, 0.52, for the WordNet-based method, but the highest coverage of 0.95 for the distributional approach. They propose to use the word-based distributional approach as a back-off for the WordNet-based method, and suggest to use this approach for tasks like the semi-automatic creation of new FrameNets. The WordNet-based approach by Pennacchiotti et al. (2008) works best on nouns. Later, Fürstenau and Lapata (2012) improve on this approach for verbs, see also Section 3.3.2.

There is also research using other LKBs than WordNet: Tonelli et al. (2013) infer and use a mapping of FrameNet to Wikipedia for English frame identification. Their main goal is to bootstrap FrameNet resources in other languages that are linked to Wikipedia articles via the inter-language links in Wikipedia, as discussed earlier in Section 2.3.1.

For Wikipedia-based frame identification, they apply a word sense disambiguation system for Wikipedia senses (defined as Wikipedia entries) to lexicon-based textual representations of FrameNet predicates to infer a mapping between noun predicates in FrameNet and Wikipedia senses. The textual representations of the FrameNet predicates do not make use of the example sentences, but only rely on the definition of the frame associated with the predicate and on the other lexical units available for this frame. This results in a larger lexicon coverage, as many predicates do not possess example sentences. Combining the Wikipedia WSD system and the inferred mapping to FrameNet, they create a FrameNet frame identification system: they assign a Wikipedia sense to a predicate instance and use the mapping to transfer the Wikipedia sense label to FrameNet. The resulting system shows high precision for seen predicates, i.e., the noun predicates that are part of the mapping.

The number of seen predicates is, however, small, since the mapping only covers 37% of the predicates in FrameNet. By extending the mapping, i.e., by linking additional Wikipedia senses to the senses in the mapping, they can increase the recall for unseen noun predicates compared to SEMAFOR. The main disadvantage of the approach by [Tonelli et al. \(2013\)](#) is that it acquires only noun predicates, due to the noun-centric nature of Wikipedia. As such, their approach is complementary to our method that is focused on verbs.

[Giuglea and Moschitti \(2006\)](#) compile a semi-automatic mapping between FrameNet frames and Intersective Levin Classes (ILC, [Dang et al. \(1998\)](#)), which are also used in VerbNet, using a manually created role mapping between FrameNet and VerbNet. After manual correction of the mapping, they replace FrameNet frames by the ILC classes for sense disambiguation and use the PropBank corpus labeled with VerbNet classes and the FrameNet corpus with the ILC labels as extended training data, finding that using ILC labels instead of FrameNet frames results in similar SRL performance at potentially larger lexicon coverage.

[Das and Smith \(2011\)](#) also target the frame identification step. They use an automatically created lexical resource, Lin’s distributional thesaurus ([Lin, 1998](#)), as well as the FrameNet 1.5 example sentences and FrameNet fulltext training set to enhance the FrameNet frame coverage of the SEMAFOR semantic role labeling system. They create a graph that integrates FrameNet lexical units and words from the thesaurus and use graph propagation to transfer distributions over frame labels to new lemmas in the graph. Integrated into the SEMAFOR semantic role labeling system, their strategy improves frame identification accuracy on the FrameNet fulltext test set by 15.7% for unseen words. They also report that the self-training baseline, adding over 700,000 annotated instances gathered from running SEMAFOR on 70,000 sentences of the Gigaword corpus, shows worse results than the standard frame identification of SEMAFOR.

Beyond the graph propagation method described above, SEMAFOR includes knowledge base information in the following ways: FrameNet example sentences are also used in the construction of features for the frame identification step. Argument relations in FrameNet, for instance *excludes* or *requires*, are encoded as constraints in the collective argument identification method by [Das et al. \(2014\)](#).

Role labeling using LLKBs. The related work introduced above focused on the frame disambiguation task, but there is also work addressing the role labeling part of semantic role labeling: [Matsubayashi et al. \(2009\)](#) present a method of feature expansion that adds features based on four different categories, (a) the frame and role hierarchy, i.e., exploiting frame relations, (b) common role labels between frames, (c) semantic types of role fillers, and (d) common mappings of roles to VerbNet roles in SemLink. Their goal is to exploit generalization between different roles in FrameNet to overcome training data sparsity. Their approach increases role classification F_1 by 7.42% with the largest improvements for roles with only few training instances. They report best results from combining all available

categories of feature expansion, with strongest improvements from the common role labels. The benefit of the SemLink integration is reported to be negligible because of the low coverage of the VerbNet–FrameNet linking at the time.

Kshirsagar et al. (2015) attempt to employ the sense examples and the frame hierarchy from the FrameNet lexicon and SemLink to improve the role labeling of FrameNet semantic role labeling. They did not find improvements when using SemLink, therefore, they only briefly report on their experiments: they used SemLink to translate the PropBank role labels in the SemLink corpus to FrameNet and added them as additional training data to SEMAFOR, but found that this strategy hurt role labeling performance. They credit this to the low SemLink coverage and errors in SemLink, which might be amplified by the use of a transitive linking, namely from PropBank to FrameNet via VerbNet.

They are more successful when including additional information from the FrameNet knowledge base to role labeling: they employ domain adaptation techniques to augment the feature space extracted from the FrameNet training set with features from the sense examples, increasing role labeling F_1 by 3% compared to the baseline system SEMAFOR. Second, they exploit the FrameNet hierarchy, i.e., frame relations like *inheritance* and *sub-frame*, to augment the feature space with information from related frames – similar to Matsubayashi et al. (2009).

Kshirsagar et al. (2015) additionally use features from PropBank semantic role labeling as guide features for FrameNet, making indirect use of the PropBank corpus. The results of their best system speak in favor of exploiting the FrameNet knowledge base: in this setup, they combine the use of example sentences and the FrameNet hierarchy for feature augmentation. They only evaluate on the FrameNet fulltext test set, which has become the default for the evaluation of FrameNet SRL.

Summary. Methods that have been used to improve FrameNet semantic role labeling with additional information from lexical knowledge bases are a) linkings to WordNet or other resources to increase frame coverage and thus frame identification performance, and b) feature expansion based on the FrameNet hierarchy for role labeling improvements. LLKBs like SemLink have been utilized, but they do not contribute much to the performance (Matsubayashi et al., 2009) or are even detrimental (Kshirsagar et al., 2015). It seems that the potential of lexical knowledge bases could not be exploited due to a lack in coverage or quality of those resources.

In this work, we successfully use LLKBs for the generation of frame- and role-labeled training data. To provide background on training data generation for SRL, we introduce previous work in this area in the next section.

3.3 Generating Training Data for Semantic Role Labeling

The lack of training data is frequently mentioned as one of the main obstacles to increased performance and the widespread use of FrameNet semantic role labeling (Das et al., 2014; FitzGerald et al., 2015). This lack has motivated work in automatically generating labeled training data for frame identification and role labeling. One of the main contributions of this work is a method for generating FrameNet-labeled training data using supervision from linked lexical knowledge bases. Since the frame identification step of SRL is basically a word sense disambiguation task, this section introduces previous work in generating sense- and role-labeled training data.

3.3.1 Generating Training Data for Word Sense Disambiguation

Most previous work on automatically sense-labeling corpora for word sense disambiguation focused on nouns and used WordNet as a sense inventory, e.g., Leacock et al. (1998), Mihalcea and Moldovan (1999), Martinez (2008), Duan and Yates (2010). Verbs are particularly important for SRL, and harder to disambiguate than nouns, because they are typically more polysemous. Therefore, our training data generation method targets verbs, and we primarily discuss methods targeting verbs in this subsection.

The methods for using knowledge bases to enhance FrameNet frame identification introduced in Section 3.2.2 could be used as methods for training data generation, but most of them are explored in an inductive setup, i.e., they are applied to SRL directly, not as a method for generating training data for a supervised system.

Related work on generating sense-labeled data for verbs focused on WordNet (Kübler and Zhekova, 2009; Cholakov et al., 2014). Kübler and Zhekova (2009) extract example sentences from several English dictionaries and various types of corpora, including web corpora. They use a Lesk-like algorithm to annotate target words in the extracted example sentences with WordNet senses and use them as training data for word sense disambiguation. They evaluate their method on the task of all-words WSD, but do not find performance improvements over the baseline when training on the automatically labeled data or a combination of automatically labeled and gold data. They also experiment with sampling the automatically labeled data according to the sense prior given in the manually annotated data. Only after filtering the automatically labeled data based on vector space proximity to gold standard data, i.e., only keeping the most similar automatically labeled instances, their performance level reaches the supervised setting.

The work of Cholakov et al. (2014) is a model for the method of automatically generating frame-labeled data presented in this thesis. Cholakov et al. (2014) use the sense examples provided by the LLKB UBY as gold examples for the automatic sense-labeling of unlabeled verbs. They use WordNet as their sense inventory and exploit its links to VerbNet, FrameNet, and Wiktionary to expand the set of gold examples. From this set, they extract

lexico-syntactic patterns of the verb senses in their contexts. They match these patterns to a large unlabeled web corpus to add sense labels to the verbs in the corpus. They show that a supervised word sense disambiguation system trained on the automatically labeled corpus approaches state-of-the-art results on the MASC corpus and on the Senseval-3 all-words data. We already introduced their approach and our extensions to the approach in Section 3.1.2 above.

3.3.2 Generating Training Data for Semantic Role Labeling

This section introduces different strategies for the generation of training data for semantic role labeling, from self-training approaches, over cross-lingual and monolingual annotation projection to paraphrasing-based approaches. These strategies are motivated by the sparsity of existing training data or the attempt to create an initial set of training data for a resource-poor language.

Table 3.1 contains an overview over the related work introduced in this subsection. Our work is also shown in the table for comparison purposes. The two blocks of the table distinguish monolingual and cross-lingual approaches, and the columns contain information on which role inventory was used (*SRL type*), whether training data for verb sense disambiguation or semantic role labeling were created (*labels*), which generation method was used (*method*), the source and target languages, and whether the method was evaluated extrinsically for semantic role labeling (*SRL eval*). In the next paragraphs, we introduce the related work in detail.

Self-training. In self-training, a semantic role labeling system is trained on existing training data. The resulting system is used to label new sentences which are then added to the training set, a process that can be repeated several times. This approach runs the risk of propagating errors from the original system to the newly labeled data. Accordingly, evaluations of the self-training approach for semantic role labeling led to mixed results. Different self-training approaches vary in the number of iterations that are performed, the number of sentences that are added to the training set in each iteration, and the strategy for selecting these sentences. If confidence scores are provided by a system, the labeled instances with the highest scores can be selected. Another option is to select instances based on an expected distribution of target predicates, either recreating the distribution observed in the original training data or, for cross-domain applications, emulating the distribution of predicates found in an annotated sample of the target domain.

He and Gildea (2006) evaluate self-training and co-training on a set of 15 thematic roles inferred from FrameNet-labeled data. They did not find any improvements compared to training a supervised classifier. Bejan (2009) evaluated different self-training approaches for frame identification and found improvements over the standard supervised setup with a simple one-step labeling procedure that adds labeled data in a single iteration.

reference	SRL type	labels	method	source language	target language	SRL eval
monolingual						
He and Gildea (2006)	TH	SRL	self-training	EN	EN	X
Bejan (2009)	FN	VSD	self-training	EN	EN	X
Das et al. (2014)	FN	VSD	self-training	EN	EN	X
Fürstenau and Lapata (2012)	FN	SRL	projection	EN	EN	X
Fürstenau (2011)	FN	SRL	projection	EN/DE	EN/DE	X
Gordon and Swanson (2007)	PB	SRL	paraphrasing	EN	EN	-
Woodsend and Lapata (2014)	PB	SRL	paraphrasing	EN	EN	X
Pavlick et al. (2015a)	FN	VSD	paraphrasing	EN	EN	-
Hartmann et al. (2016)	FN	SRL	knowledge-based	EN/DE	EN/DE	X
cross-lingual						
Padó and Lapata (2005b)	FN	SRL	projection	EN	DE	-
Padó and Lapata (2009)	FN	SRL	projection	EN	DE	X
Padó and Pitel (2007)	FN	SRL	projection	EN	FR	X
Johansson and Nugues (2005)	FN	SRL	projection	EN	SP	-
Johansson and Nugues (2006)	FN	SRL	projection	EN	SP	X
Basili et al. (2009)	FN	SRL	projection	EN	IT	-
Tonelli and Pianta (2009b)	FN	SRL	projection	EN	IT	-
Van der Plas et al. (2011)	PB	SRL	projection	EN	FR	X
Exner et al. (2015)	PB	SRL	projection	EN	SWE	X

Table 3.1: Overview of related work in training data generation for SRL. FN stands for FrameNet, PB for PropBank, and TH for a small set of thematic roles used by [He and Gildea \(2006\)](#).

Self-training has been used as a baseline approach for various semi-supervised approaches in semantic role labeling: [Das et al. \(2014\)](#) compare self-training of the frame identification module of SEMAFOR to the standard supervised frame identification in SEMAFOR. They label 70,000 sentences of the GigaWords corpus, leading to 711,000 frames, more than 36 times the original frame annotations, a setup they liken to the one from [Bejan \(2009\)](#). Contrary to [Bejan \(2009\)](#), they find that self-training does not improve frame identification performance of SEMAFOR compared to the standard supervised setup; for unknown target predicates the performance with self-training is even worse than the supervised setup.

[Fürstenau and Lapata \(2012\)](#) compare their annotation projection approach to three different self-training strategies, one selects and labels a number of sentences containing the target predicate, the second selects sentences based on a similarity measure for paraphrasing, the third one uses a combined measure of syntactic similarity and word similarity they also use in their annotation projection approach. They add up to 5 new labeled sen-

tences per seed sentence and evaluate role labeling performance. In their experiments they also find that self-training does not improve on the baseline classifier.

In summary, previous work on self-training for SRL does not improve, or is even detrimental to SRL performance. Therefore, we do not compare our knowledge-based method for training data generation to self-training approaches.

Cross-lingual annotation projection. Cross-lingual annotation projection aims at creating semantic role labeling training data for new languages. It uses word-level or syntax-level alignments of annotated sentences in a source language – typically English – to unlabeled sentences in a target language. Annotations from the source sentence are transferred to the corresponding words in the aligned target sentence.

The different variants of cross-lingual annotation projection are characterized by two main properties: the first one is the way the target sentences are matched to the source sentences, the second one is the method used for aligning the two sentences. A straightforward way of mapping source to target sentences is using parallel corpora, but there is also work using comparable corpora. For sentence alignment, word-based and syntax-based approaches have been evaluated.

[Padó and Lapata \(2005b\)](#) evaluate cross-lingual annotation projection based on parallel corpora. They use parallel text from the Europarl corpus to project FrameNet annotations to German. They evaluate different projection strategies based on word-level alignments using methods from machine translation, and constituent-level alignments that make use of syntactic parse information. They evaluate the projection methods intrinsically by comparing the projected roles to a manually annotated gold standard of 1,140 sentence pairs with predicates in FrameNet and SALSA. Projection based on constituent-level alignments performs significantly better than word-level alignments. Word-level alignments are considered a starting point for semi-automatic approaches for languages without adequate parsers.

[Padó and Lapata \(2009\)](#) introduce a more elaborate strategy of constituent alignment for the same task: they model the constituent alignment problem as identification of the optimal subgraph in a bipartite graph, the partitions consisting of constituents in the source and target language. Besides using gold annotations as source labels and gold syntactic parses, [Padó and Lapata \(2009\)](#) also evaluate a more realistic setup, which uses automatic parses and a semantic role labeling system for English to create the role annotations on the source sentences. Errors from automatic analysis and lack of coverage of the semantic role labeling system that only handles verbs are propagated through the system. In consequence, the evaluation performance in the realistic setup is reduced by 13% precision and more than 24% F_1 compared to the setup using oracle annotations. The best results in the evaluation are obtained for the constituent-based alignment using perfect matchings. This method certainly benefits from the similarity of English and German syntactic-semantic structure.

Padó and Pitel (2007) apply the approach by Padó and Lapata (2006) to French and find similar results. Johansson and Nugues (2005) present a projection approach for Spanish that is based on word-based alignment of parallel texts similar to the word-based approach by Padó and Lapata (2005b). Johansson and Nugues (2006) use word-based annotation projection enhanced by heuristic filters specific to the target language to create training data for a Swedish semantic role labeling system. They evaluate the projection approach extrinsically by applying the system to a small test set and find promising results with 67% precision and 47% recall.

Basili et al. (2009) and Tonelli and Pianta (2009b) evaluate approaches for projecting FrameNet labels to Italian. The approach by Basili et al. (2009) avoids syntactic analysis required for the target language by exploiting a phrase-based machine translation system for the alignment of parallel corpora and using rule-based postprocessing to improve the precision of the alignment.

Van der Plas et al. (2011) also use word-based annotation projection for French in the PropBank paradigm. Their approach exploits the similarity of semantic structures to syntactic structures in PropBank: they use the noisy PropBank-labeled data resulting from annotation projection together with syntactically labeled data to learn a joint syntactic-semantic parser and find that the joint learning smoothes over errors from the annotation projection step.

We already mentioned the approach by Exner et al. (2015). They introduce a variant of annotation projection based on comparable or loosely parallel corpora, namely different language variants of Wikipedia articles, to generate Swedish data labeled with PropBank roles. The source language is English and the source labels are created automatically using the PropBank semantic role labeling system by Björkelund et al. (2010). Exner et al. (2015) utilize entity recognition borrowed from information extraction to identify role targets in the loosely parallel texts, which they use as anchors for the cross-lingual alignment. They identify predicates in the target language as the syntactic governors of the role targets and project role labels to the role targets. For evaluation, they split the automatically labeled corpus into training, development, and test set. A semantic role labeling system trained on the training split achieves a labeled F_1 score of 52.25% on the test split. This is a good starting point, but still far from the results obtained in a supervised fashion: PropBank SRL systems trained in a supervised fashion received labeled F_1 scores between 79.71% for German and 85.63% for English in the CoNLL 2009 shared task (Hajič et al., 2009).

Monolingual annotation projection. In monolingual annotation projection, annotations are transferred from labeled data to new, unlabeled sentences in the same language (Fürstenau and Lapata, 2012; Fürstenau, 2011). In this respect, it is similar to our approach to training data generation introduced in Section 3.1 of this work, which also uses a monolingual setup – linked lexical knowledge bases in a single language – to create role-labeled data.

The motivation is similar to the one for cross-lingual projection: extending the training set for low-resource languages. Monolingual annotation projection, however, assumes the existence of a small seed set of role-labeled sentences in the target language as source data for projection.

Fürstenau and Lapata (2012) perform annotation projection of FrameNet roles for English verbs. Instead of using parallel corpora, their method integrates the selection of suitable source and target sentences from a large number of candidate pairs: they pair frame-annotated sentences from FrameNet 1.3, called seed sentences, with sentences in the British National Corpus that contain the lemma and part-of-speech of the source predicate. Then, they align the syntactic structures of the sentence pairs. If an alignment can be established that covers all roles from the source sentence, they transfer frame and role labels to the new sentence. For the sentence alignment, they introduce an alignment algorithm that takes the syntactic similarity of the sentences and distributional similarity of the role fillers into account. Both properties are integrated into a single similarity measure. They solve the alignment as a graph alignment problem, optimizing the joint syntactic-semantic similarity measure, and propose a tailored integer linear programming algorithm to exactly solve the graph alignment. If the alignment is successful, the frame and role labels can be transferred to the unlabeled target sentence. For each unlabeled sentence, the frame and role labels of the aligned seed sentence with the highest similarity score are selected for projection. For each seed sentence, newly labeled sentences with the highest similarity scores are added to the newly labeled corpus. The alignment is subject to some quality constraints to ensure a larger precision: Fürstenau and Lapata (2012) discard sentences that do not align well to their seeds and discard candidate pairs for which not all roles could be mapped. This leads to a sparse labeling, but ensures that all roles for a candidate are mapped. Such an approach does have disadvantages, e.g., a potentially lower domain variability of the corpus, since they only label sentences very similar to the seed sentences. Repeating their experiments for German, Fürstenau (2011) finds that the variety of the automatically annotated sentences decreases when a larger expansion corpus is used. This is a result of the optimization objective and the selection of few best expansion sentences. The algorithm optimizes syntactic and semantic similarity of the candidate sentence and seed sentence, and thus prefers expansion sentences that are most similar to the seed sentence. In a smaller corpus, these very similar sentences are simply not available, and the algorithm selects the best sentence that fulfills the constraints. In a larger corpus, the method has a large candidate set to select from, and consistently identifies the most similar sentences. We confirmed this bias: in our annotation projection experiments, we find that we can create a corpus based on the BNC that is very similar to the original FrameNet fulltext corpus by using the FrameNet fulltext corpus as seed set and selecting the top-1 expansion, see Section 4.4.3. This preference can be mitigated by selecting a larger k , or relaxing the constraint that all roles in the seed sentence need to be covered by the alignment.

In their experiments, [Fürstenau and Lapata \(2012\)](#) simulate a low-resource scenario by varying the number n of labeled seed sentences, i.e., random selection of $n \in (1, \dots, 10)$ seeds per FrameNet sense, and the number of expansions per seed. For each seed, they extend the labeled corpus with the best k results of projection for k ranging from 1 to 6. For evaluation, they train a supervised semantic role labeling system on the union of the seed training data and the automatically labeled data and find improvements over the unexpanded baseline and over self-training. They do not evaluate their method in a standard scenario for FrameNet semantic role labeling, which would be to expand the full FrameNet training set with a suitably large number of projected sentences and to compare to training on the unexpanded training corpus.

[Fürstenau and Lapata \(2012\)](#) evaluate their projection method in a second setup: projection to verbs not in the lexicon, i.e., identifying new lexical units and generating training data for them simultaneously. In contrast, the setup described above generates training data for verbs already in the FrameNet lexicon which are already equipped with a small number of labeled training sentences. In theory, this new variant spans up a large search space: instead of creating sentence pairs based on matching lemmas, an unlabeled sentence needs to be compared to the seed sentences for all the lexical units in FrameNet. To pare down this space, [Fürstenau and Lapata \(2012\)](#) filter potential candidate sentences using a method for the acquisition of FrameNet predicates: they extend the method for lexical unit acquisition proposed by [Pennacchiotti et al. \(2008\)](#) which uses distributional similarities or WordNet co-hyponyms to assign frames to unknown words. The WordNet-based method resulted in the highest accuracy in the experiments by [Pennacchiotti et al. \(2008\)](#), but this is mainly the result of a more elaborate approach for nouns; their method performs worse for verbs which are the target part-of-speech in the work of [Fürstenau and Lapata \(2012\)](#). Therefore, they develop a more elaborate frame assignment strategy for verbs that includes a larger number of WordNet relations and improves on the results achieved with the method from [Pennacchiotti et al. \(2008\)](#). Matching new predicates to frames with this method and projecting roles to the candidate sentences improves upon an unexpanded baseline.

[Fürstenau \(2011\)](#) also evaluates the annotation projection approach for German, using the SALSA corpus as their source of seed sentences and the large *Süddeutsche Zeitung corpus* as expansion corpus. Due to the smaller size of the seed corpus and the larger size of the expansion corpus, the improvements of training data generation to role labeling he observes for German are much more pronounced than those for English. These results show that the method is also applicable to other languages providing seed corpora.

The annotation projection approach by [Fürstenau and Lapata \(2012\)](#) is based on FrameNet-labeled training data. It does not use information from lexical knowledge bases, foregoing the valuable information these resources have to offer. In our work, we take the opposite approach: we do not assume existing role-labeled corpora, but make use of the rich semantic information in the LLKB UBY_{FN}. In Chapter 4, we compare the two approaches,

presenting an evaluation of annotation projection in a real semantic role labeling setup, and an analysis of its contributions to domain adaptation of semantic role labeling systems.

Paraphrasing-based approaches. Another variant of training data generation for semantic role labeling aims at creating variations of the existing role-labeled data. [Gordon and Swanson \(2007\)](#) present an early variant of this technique. They aim to extend the instance coverage of PropBank-labeled corpora by supplementing the instances of predicates with few labeled sentences with sentences from syntactically similar predicates. Therefore, they extract a syntactic signature of the target predicate from parses of its instances in a large unlabeled corpus. The signature is a vector representation on the parse-tree paths starting from the predicate. They match this signature to the signature of other predicates in PropBank using cosine similarity to determine candidate predicates. Then they filter the candidates: they use a semantic role labeling system trained on the existing instances of each candidate predicate to label the arguments of an instance of the target predicate. If the role labels match the gold labels, the candidate predicate is an acceptable supplement. The instances of the most similar candidate predicates are added to the training set for the target predicate.

[Woodsend and Lapata \(2014\)](#) present a more elaborate paraphrasing-based approach to training data generation that also targets PropBank semantic role labeling. They automatically extract syntactic rewrite rules using synchronous grammars derived from comparable corpora based on Wikipedia and bitext from the Paraphrase Database – a large database of paraphrases generated from bilingual parallel texts ([Ganitkevitch et al., 2013](#)). [Woodsend and Lapata \(2014\)](#) use these rules to generate paraphrases of the training sentences in the PropBank corpus. They compare the automatic methods to using manually created rewriting rules. Using the automatic rules, they multiply the number of sentences in the PropBank training set by up to 24, from 39,000 to 940,000. Using manual rules creates a set that is only 40% larger than the original training set. They observe improvements in semantic role labeling performance for both types of rewrites, but find the largest improvements for the automatic rewrites. They report that a model of the Mate-tools semantic parser ([Björkelund et al., 2009](#)) trained on automatic extension outperforms the state-of-the-art system on the CoNLL-2009 test set for in-domain and out-of-domain data.

There is also work on extending the FrameNet fulltext corpus via paraphrasing: [Pavlick et al. \(2015a\)](#) use paraphrasing rules extracted automatically from the Paraphrase Database ([Ganitkevitch et al., 2013](#)) to replace lexical units in the FrameNet fulltext corpus by new lexical units suitable to the sentence context. They use manual postprocessing via crowdsourcing to filter out the noise. The number of lexical units in the resulting resource called FrameNet+ is three times as high as the original FrameNet fulltext corpus. It supplies more than 80,000 paraphrased sentences labeled with the new lexical units. The role coverage is not affected, because [Pavlick et al. \(2015a\)](#) do not expand the role-labeled corpus with the

new lexical units. In an evaluation on a stratified sample of 300 words from the New York Times, they show that their resource increases the frame coverage by 40%. They did not evaluate the benefits of the extended corpus to frame identification or FrameNet semantic role labeling. We perform such an evaluation for frame identification in Chapter 4.

Most of the work on generating training data for semantic role labeling in this section was corpus-based. In the next section, we present the application of DistantSRL, our knowledge-based approach to training data generation, to English data.

3.4 Application of DistantSRL to English

This section describes the application of DistantSRL, our approach to knowledge-based distant supervision for SRL training data generation, to English data.

The application to English benefits from the numerous resource links of FrameNet to other lexical knowledge bases in UBY_{FN} as introduced in Section 3.4. In addition, it benefits from the availability of an annotated corpus, the FrameNet fulltext corpus, and open-source systems for training and evaluating FrameNet SRL systems. We use these resources for the evaluation of our method: we use classifiers trained on the FrameNet fulltext corpus as baseline systems and compare them to systems trained on our generated data, or combinations of both. Since DistantSRL is focused on verbs, we test on verbal predicates from a range of test sets.

In this section, we first introduce relevant datasets, including the linked lexical knowledge base used for English. Then, we present the creation of the frame-labeled corpus and its extrinsic evaluation in the frame identification task. This is followed by the creation of the role-labeled corpus and its extrinsic evaluation in a role classification task.

3.4.1 Unlabeled Corpora and Gold Standard Data

The experiments in this section rely on several gold standard training and test datasets, including the FrameNet fulltext corpus, and on the linked lexical knowledge base centered around FrameNet. The extrinsic evaluation is based on four different FrameNet-labeled datasets. Table 3.2 shows the statistics of these datasets, focusing on the verb instances we use for testing. A description of each of the datasets follows.

FNFT-test. This dataset is the test-split of the FrameNet fulltext corpus used in [Das and Smith \(2011\)](#). This dataset is a random sample of documents from the FrameNet fulltext corpus and as a result mirrors the distribution of the FrameNet fulltext training set FNFT-train, which consists of newspaper texts and fictional texts; this means it is an in-domain test set. FNFT-test currently is the only commonly used test set for the evaluation of FrameNet semantic role labeling.

dataset	verbs lemmas	sense types	average senses per verb	sense tokens	role tokens
Fate	526	725	1.4	1,326	3,490
MASC	44	143	3.3	2,012	4,142
SemEval	278	335	1.2	644	1,582
FNFT-test	424	527	1.2	1,235	3,078
FNFT-dev	490	598	1.2	1,450	3,857

Table 3.2: Test dataset statistics for verbs: lemmas, sense types, average senses per verb, number of sense and role instances (tokens).

Das and Smith (2011) additionally provide a development dataset FNFT-dev based on the FrameNet fulltext corpus that we use for an initial estimation of the variable parameters of our approach, e.g., the thresholds f and t introduced above. Because one expected benefit of DistantSRL is a larger domain variability of the generated training data, our evaluation also considers other available datasets with different domain distributions.

SemEval. The SemEval test set is based on the frame and role annotation in the trial and evaluation dataset of the SemEval 2010 shared task on Linking Events and Their Participants in Discourse (Ruppenhofer et al., 2010b). It consists of literature texts, more specifically crime stories written by Arthur Conan Doyle.

Fate. The Fate corpus contains frame annotations on the RTE-2 textual entailment challenge test set (Burchardt and Pennacchiotti, 2008). It contains 800 pairs of sentences with or without entailment relation. It is based on newspaper texts, texts from information extraction datasets such as ACE and MUC-4, texts from question answering datasets such as CLEF and TREC, and texts used for multi-document summarization of news documents. The SemEval and Fate datasets were created prior to the release of FrameNet 1.5. For those sets, only verb senses, i.e., verb-frame combinations, that still occur in FrameNet 1.5 and their roles were included in the evaluation.

MASC. The MASC WordSense sentence corpus (Passonneau et al., 2012) is a balanced corpus that contains sense annotations for 1,000 instances of 100 words from the MASC corpus. MASC mostly provides WordNet sense labels, we use a slightly smaller subset annotated with FrameNet 1.5 labels.¹⁸

Linked lexical knowledge base. The lexical knowledge base used for training data generation for the English FrameNet consists of FrameNet itself and the lexical knowledge

¹⁸At the time of writing, this subset is not part of the MASC download, but according to personal communication with the developers, it has been integrated into the FrameNet release 1.6.

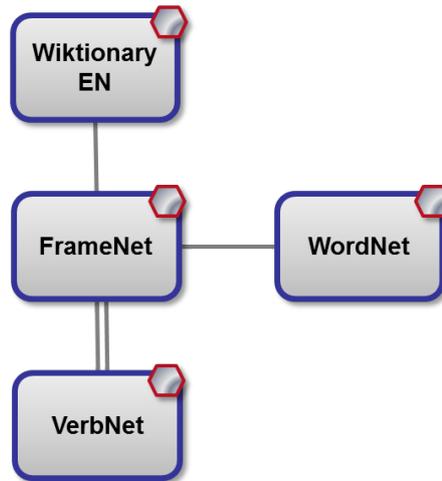


Figure 3.9: Linked lexical knowledge base centered around FrameNet.

bases linked to FrameNet as introduced in Section 2.8, namely WordNet, VerbNet, and Wiktionary, see also Figure 3.9.

PropBank is not included, because our experiments predate the integration of PropBank. Using PropBank, we could add 2,457 additional sense links, see Table 2.11, which only adds few links to the 50,000 sense links for the LLKB in Figure 3.9. Wikipedia is not included, because our method targets verbs, which are only sparsely represented in Wikipedia.

Unlabeled corpus. For the sparse labeling approach, a very large corpus is required. The experiments in this chapter are based on the first four sections of the ukWAC corpus (Baroni et al., 2009). They contain more than 14 million sentences and more than 370 million tokens. These are used as unlabeled input for the automatic sense label transfer. They are filtered for sentences containing the 695 verb lemmas in our four test sets, resulting in a subset of more than 11 million sentences and more than 320 million tokens.

Gold standard training corpus. To compare the automatically labeled data to a standard training setup, we use the training split FNFT-train of the FrameNet fulltext corpus from Das and Smith (2011). It contains 79,000 tokens and 3,526 sentences with 19,482 frame instances and 33,690 role instances for 2,913 lemmas. Filtered by the 695 verbs in our test sets, it contains 5,974 frame instances for 856 different verb senses and 575 verb lemmas. We call the filtered training corpus FNFT★. Table 3.4 below also lists the statistics of FNFT★.

3.4.2 Frame Labeling: Corpus Creation and Experiments

This subsection presents the creation of frame-labeled corpora according to Stage 1 of DistantSRL and their application to the task of frame identification, which provides an extrinsic evaluation of Stage 1.

Frame identification task. The extrinsic evaluation setup uses a standard supervised system for frame identification, which is basically verb sense disambiguation. The system is based on the verb sense disambiguation system presented by [Cholakov et al. \(2014\)](#) and uses a logistic regression classifier in the WEKA implementation ([Hall et al., 2009](#)).

For feature extraction, we use pre-processing tools from DKPro Core ([Eckart de Castilho and Gurevych, 2014](#)), more specifically the Stanford tokenizer, the TreeTagger for POS tagging and lemmatization, the StanfordNamedEntityRecognizer for named-entity recognition, and the Stanford Parser for annotating dependency structures. The frame identification system uses lexical, syntactic, and semantic features which are extracted from the various training sets for training and the four test sets for testing. The features are the same as used by [Cholakov et al. \(2014\)](#). Lexical features include the lemmas and POS tags of the two words preceding and following the target verb. Syntactic features are based on the output of the Stanford parser: for all dependency relations in which the target verb is connected to a noun, pronoun or a named-entity, the lemma or named-entity tag are combined with the type of the dependency relation to build a separate feature. Another set of features is created by replacing the lemma by its part-of-speech tag. The semantic features include all synsets found in WordNet for the nominal arguments of the verb. Personal pronouns are mapped to the noun “person” beforehand.

Evaluation metrics used in this section are precision P , the number of correct instances divided by the number of labeled instances, recall R , the number of labeled instances divided by all instances, and F_1 , the harmonic mean of P and R . All significance scores reported in this section are based on Fisher’s exact test at significance level $p < 0.05$.

Parameter estimation and corpus creation. Recall that the creation of the automatically labeled corpus consists of three steps, Step 1A, seed pattern creation, Step 1B discriminative filtering of seed patterns, and Step 1C, the labeling of the unlabeled corpus based on similarity to the seed patterns, as introduced in Section 3.1. The variable parameters in the corpus creation are the discriminative filter f and the similarity threshold t . These are determined in a tuning procedure: we perform a line search, testing several parameter configurations, on the frame identification task using FNFT-dev as validation set.

Step 1A: seed patterns. The seed patterns for Step 1A of the sense labeling were extracted from the example sentences in the FrameNet lexicon, and all the sense examples linked

parameters	t	0.07					0.1				
	f	-	0.2	0.14	0.1	0.07	-	0.2	0.14	0.1	0.07
scores	P	67.2	68.8	68.9	66.9	67.3	67.2	67.2	68.0	68.7	67.0
	R	72.3	69.2	69.2	69.2	63.9	71.2	68.8	68.8	65.4	63.4
	F ₁	69.6	69.0	69.1	69.1	65.5	69.2	68.0	68.4	67.0	65.2
parameters	t	0.14					0.2				
	f	-	0.2	0.14	0.1	0.07	-	0.2	0.14	0.1	0.07
scores	P	65.6	65.0	65.8	68.5	67.9	68.0	68.3	68.9	70.2	71.3
	R	64.2	58.1	58.1	56.2	54.3	63.3	56.6	56.6	54.4	52.6
	F ₁	65.3	61.3	61.7	61.7	60.3	65.6	61.9	62.1	61.3	60.5

Table 3.3: Parameter tuning for DistantSRL: combinations of f and t evaluated on FNFT-dev; configurations for best P, R, and F₁ in percent in **boldface**.

to FrameNet as described above. There are more than 38,000 linked sentences between FrameNet, Wiktionary, WordNet, and VerbNet.

Without a discriminating filter, this results in more than 41,700 LSP patterns and more than 322,000 ASP patterns, 11% and 89% of the total number, respectively. Adding a strict discriminating filter $f=0.07$ reduces the patterns to 39,000 LSP and 217,000 ASP. Proportionally more ASP are filtered, which is expected, since they generalize stronger from the surface text than the LSPs, leading to a proportion of 15% LSP and 85% ASP. When applying this filter, the number of senses with patterns decreases from 4,900 to 3,900.

Threshold setting for Step 1B and Step 1C. In order to determine the parameter values for the label transfer – i.e., which values for threshold t , and filter f result in a high-quality training corpus, we perform a line search using extrinsic evaluation on the FrameNet full-text development set, i.e., we evaluate a range of parameter values on the validation set.

For this purpose, we generate a set of automatically labeled corpora based on ukWAC section 1 using a range of different threshold values. Each of the values is used to train our verb sense disambiguation system and is evaluated on the development set FNFT-dev.

The evaluated range of threshold values for the discriminating filter f from Step 1A and the threshold t from Step 1B is (0.07, 0.1, 0.14, 0.2), which was suggested by [Cholakov et al. \(2014\)](#) for t . Additionally, we compared corpora with and without the discriminating filter f . Table 3.3 shows the results of the experiments on the development set.

As expected, increasing the pattern similarity threshold t at which a corpus sentence is labeled with a sense increases the precision at the cost of recall. This is shown in the first part of Table 3.3. Similarly, employing a discriminating filter f at $t=0.2$ increases precision compared to using no filter, and leads to the best precision on the validation set, as shown in the second part of the table. Note that the discriminating filter gets stricter, i.e.

corpus	t	f	sense tokens	sense types	verb types	average senses per verb	average instances per sense
WaS-XL	0.07	-	1.6·10 ⁶	1,460	637	1.8	1,139
WaS-X	0.2	-	193,000	1,249	602	1.7	155
WaS-L	0.2	0.07	109,000	1,108	593	1.5	98
FNFT★	-	-	5,974	856	575	1.5	10

Table 3.4: Sense statistics of automatically labeled corpora for English.

discriminates more, with a lower f value. Accordingly, low f values lead to the highest precision of 0.713 for the strict thresholds $t=0.2$ and $f=0.07$, indicating that precision-oriented applications can benefit from higher discrimination.

We use the following settings to create large frame-labeled corpora from ukWAC sections 1 to 4: The setting with the highest F_1 in Table 3.3 ($t=0.07$) leads to the very large sense-labeled corpus **WaS-XL**. The f and t values leading to the highest precision are used to evaluate the benefits of the discriminating filter in **WaS-L**. They are $t=0.02$ and $f=0.07$. For direct evaluation of the effect of the filter f , **WaS-X** was created using the same t as **WaS-L**, but no discriminative filter f .

Statistics of the generated corpora. Table 3.4 shows the statistics of the automatically generated corpora **WaS-L**, **WaS-X**, and **WaS-XL**. Their size ranges from 100,000 instances to 1.6 million sense instances with an average of 1.5 to 1.8 senses per verb. The number of verb senses in the FNFT★, the FrameNet fulltext corpus filtered by the 695 verbs in our test sets, is in contrast much smaller: for the 695 verbs in the four test sets, it contains less than 6,000 verb instances with an average of 1.5 senses per verb.

Comparing **WaS-L** to **WaS-X**, based on the same t , but without discriminating filter f , one can see that **WaS-L** contains 44% fewer sense instances, but only 12% fewer distinct senses. It still contains 75% of the senses that are covered by the huge **WaS-XL**. This shows that the discriminating filter causes a loss in recall, but this loss is small when related to the overall reduction of instances caused by the filter. This observation agrees with the expectation that the filter f increases the precision of the automatic sense labeling. The experiment results in the next paragraph will show whether this expectation is indeed fulfilled.

The number of instances per sense is Zipf-distributed for all automatically labeled corpora. The number of instances ranges from 1 to over 40,000 for **WaS-XL**, leading to the average of 1,139 as reported in Table 3.4.

Experimental evaluation. Following the results from the evaluation on the development set, **WaS-XL** and **WaS-L** are used to train the supervised verb sense disambiguation system. In order to compare them to manually labeled corpora, the system is also trained on the

training set	FNFT-test			Fate			MASC			SemEval		
	P	R	F ₁	P	R	F ₁	P	R	F ₁	P	R	F ₁
WaS-XL	64.7*	81.6*	72.2	62.8*	65.0*	63.9	66.0*	79.3*	72.0	66.5	76.1*	71.0
WaS-L	68.0*	61.8	64.8	66.0	50.5*	57.2	63.9	70.7*	67.1	69.4	62.0*	65.5
FNFT-train	72.9	64.3	68.3	70.0	38.0	49.3	59.8	33.9	43.3	70.6	55.0	61.8
B-WaS-XL	73.6	76.7*	75.1	68.6	61.9*	65.1	67.0*	69.9*	68.4	72.4	71.0*	71.7
U-WaS-XL	66.8*	93.5*	78.0	63.0*	68.3*	65.6	64.2*	83.3*	72.5	66.7	84.9*	74.7

Table 3.5: Verb sense disambiguation (frame identification) P, R, F₁ in percent; * marks significant differences to the system trained on FNFT-train; highest scores per dataset in **boldface**.

FrameNet fulltext training set. An additional evaluation considers combinations of the WaS corpora and the FrameNet fulltext training set, proving that they are complementary to each other. The resulting verb sense disambiguation systems are evaluated on the four test sets introduced above in Section 3.4.1. The following paragraphs discuss different aspects of the evaluation.

Impact of pattern filters. To evaluate the impact of the pattern filters on verb sense disambiguation performance, we compare the results when training on WaS-L and training on WaS-XL. The first block of Table 3.5 shows that the stricter filters in WaS-L improve the precision for three out of four test sets, which shows that stronger filtering can benefit precision-oriented applications.

Precision on the MASC corpus is lower when using a discriminating filter. This may be due to the larger polysemy in MASC. It contains on average 3.3 senses per verb (see column *average senses per verb* in Table 3.2), and it also contains rare senses. The reduction of sense instances caused by the discriminating filter leads to some loss of instances for those senses and a lower precision on MASC.

Analyzing the results in detail for the example verb *tell* shows that WaS-XL contains all 10 senses of *tell* in MASC, and WaS-L contains 9 of them. However, the number of training instances per sense for WaS-L can be lower by factor 10 or more compared to WaS-XL – e.g., tens to hundreds, hundreds to thousands, leading to only few instances per sense. The sparsity problem could either be solved by using a less strict filter, or by labeling additional instances from ukWAC, in order to preserve more training instances of the rare senses for stricter thresholds t and f .

The results also show that the noise that is added to the corpora in a low-discrimination, high-recall setting will be to a certain extent drowned out by the large number of sense instances, leading to fairly high precision for WaS-XL. The recall for WaS-XL is significantly higher than for WaS-L for all test sets, which also leads to a higher F₁ score. For recall-oriented settings, WaS-XL is the appropriate choice.

Comparison to the FrameNet fulltext training set. A comparison of the results when training on WaS-L and WaS-XL to a verb sense disambiguation system trained on the reference corpus, the FrameNet fulltext training set, in Table 3.5 shows mixed results. On SemEval, the precision for the WaS corpora does not deviate significantly from the system trained on the FrameNet fulltext training set. On FNFT-test, it is significantly lower for both WaS-L and WaS-XL. For WaS-XL, the precision is significantly lower on Fate, but significantly higher on MASC. The latter is a result of the larger sense coverage of the WaS-XL that was already discussed above. Evaluation on the less polysemous Fate test set does not benefit from the larger sense coverage in WaS-XL. For WaS-L, the precision is similar to the system trained on the FrameNet fulltext training set on MASC and Fate.

For WaS-XL, the recall is significantly higher than for the FrameNet fulltext training set on all test sets, leading to a higher F_1 . This is the result of the larger sense coverage of the FrameNet lexicon, which provides the seeds for the automatic labeling, compared to the FrameNet fulltext training set.

Training the verb sense disambiguation system directly on the FrameNet lexical unit examples is, however, not a viable alternative: it leads to a system with similar precision to the WaS-corpora, but with very low recall, ranging between 0.22 and 0.37. These results are in line with the results reported by Das and Smith (2011), who report that they evaluated their system trained on the FrameNet example sentences and did not find any improvements compared to training on the FrameNet fulltext corpus. By using the sense examples for the seed patterns in the DistantSRL approach, their positive effects on the sense coverage are retained, while recall and F_1 are improved at the same time.

Comparison to the FrameNet fulltext training set on common and disjoint sets of test verbs. This paragraph presents a detailed analysis and comparison of the performance of the systems based on WaS-XL and FrameNet fulltext training set on those verbs of the test sets that are evaluated for both systems, i.e., their *intersection*, and on those verbs that are uniquely evaluated for each system, i.e., their *set difference*.

The *intersection* is defined for each test set T as the set I of test instances x such that: $I = \{ x \in T \mid x \in \text{WaS-XL} \cap \text{FNFT-train} \}$. Those are the verb senses that occur in both training sets and are thus also evaluated for both training sets. While the intersection obviously contains the same instances for WaS-XL and FNFT-train, this is not the case for the *set difference* that is defined from the perspective of the current training set: from the perspective of training set A it is defined for a test set T and the second training set B as the set DA of test instances x such that $DA = \{ x \in T \mid x \in A \setminus B \}$. This means that the set difference DA contains all the test instances that occur in A , but not in B . Precision scores for these settings are summarized in Table 3.6.

On the intersection, precision and F_1 for the FrameNet fulltext training set are higher for all test sets with the exception of MASC. Here, precision is similar, but recall is 0.21 points

training set	evaluated subset	P
test = FNFT-test		
WaS-XL	intersection I : $\{x \in \text{test} \mid x \in \text{WaS-XL} \cap \text{FNFT-train}\}$	0.5864
FNFT-train	intersection I : $\{x \in \text{test} \mid x \in \text{WaS-XL} \cap \text{FNFT-train}\}$	0.7095
WaS-XL	difference DA : $\{x \in \text{test} \mid x \in \text{WaS-XL} \setminus \text{FNFT-train}\}$	0.7881
FNFT-train	difference DB : $\{x \in \text{test} \mid x \in \text{FNFT-train} \setminus \text{WaS-XL}\}$	0.8214
test = Fate		
WaS-XL	intersection I : $\{x \in \text{test} \mid x \in \text{WaS-XL} \cap \text{FNFT-train}\}$	0.5992
FNFT-train	intersection I : $\{x \in \text{test} \mid x \in \text{WaS-XL} \cap \text{FNFT-train}\}$	0.7004
WaS-XL	difference DA : $\{x \in \text{test} \mid x \in \text{WaS-XL} \setminus \text{FNFT-train}\}$	0.6676
FNFT-train	difference DB : $\{x \in \text{test} \mid x \in \text{FNFT-train} \setminus \text{WaS-XL}\}$	0.7
test = MASC		
WaS-XL	intersection I : $\{x \in \text{test} \mid x \in \text{WaS-XL} \cap \text{FNFT-train}\}$	0.6015
FNFT-train	intersection I : $\{x \in \text{test} \mid x \in \text{WaS-XL} \cap \text{FNFT-train}\}$	0.6060
WaS-XL	difference DA : $\{x \in \text{test} \mid x \in \text{WaS-XL} \setminus \text{FNFT-train}\}$	0.7173
FNFT-train	difference DB : $\{x \in \text{test} \mid x \in \text{FNFT-train} \setminus \text{WaS-XL}\}$	0.5385
test = SemEval		
WaS-XL	intersection I : $\{x \in \text{test} \mid x \in \text{WaS-XL} \cap \text{FNFT-train}\}$	0.6335
FNFT-train	intersection I : $\{x \in \text{test} \mid x \in \text{WaS-XL} \cap \text{FNFT-train}\}$	0.7104
WaS-XL	difference DA : $\{x \in \text{test} \mid x \in \text{WaS-XL} \setminus \text{FNFT-train}\}$	0.7262
FNFT-train	difference DB : $\{x \in \text{test} \mid x \in \text{FNFT-train} \setminus \text{WaS-XL}\}$	0.6842

Table 3.6: Detailed verb sense disambiguation (frame classification) results: precision on verbs evaluated for test instances in both training sets (intersection I) and test instances occurring only in one training set (set difference DA and DB).

higher. This is a result of the higher polysemy in MASC and the larger sense coverage of the WaS-XL corpus: for the verbs in the intersection, the number of training senses in WaS-XL is on average two senses higher than for FNFT-train. This larger sense coverage of the WaS-XL is beneficial to recall on the MASC test set which shows high polysemy.

Evaluating on the *set difference* between the systems, i.e., test verbs that remain for both training corpora after the intersection is removed, shows that the lemma coverage of WaS-XL is complementary to FNFT-train. The difference is not empty for both systems, but the number of verbs that can be evaluated additionally for WaS-XL is much larger than the one for FNFT-train. The proportion of instances only evaluated for a specific training set to all evaluated instances ranges between 11% and 48% for the WaS-XL, and between 5% and 30% for the FrameNet fulltext training set. Notably, the lowest number of additional instances for the WaS-XL corpus and the highest for the FrameNet fulltext training set are

reported for the FrameNet fulltext test set. This confirms once more that the benefits of the automatically labeled WaS corpora are more pronounced on the out-of-domain test sets. Table 3.6 shows that the precision on the set difference is high for WaS-XL, in particular for the MASC and SemEval test sets. The observations in this paragraph indicate that the automatically labeled corpora are complementary to the FrameNet fulltext training set. This assumption is further evaluated in the next paragraph.

Combining training data. The complementary nature of the automatically and manually created training sets inspired the evaluation of two combinations of these training sets: U-WaS-XL consists of the union of WaS-XL and the FrameNet fulltext training set, i.e., $\text{WaS-XL} \cup \text{FNFT-train}$; B-WaS-XL implements a back-off strategy taking into account that FNFT-train performs generally better on the labels contained in both training sets, i.e., the intersection from Table 3.6. B-WaS-XL thus consists of the FrameNet fulltext training set and those instances of WaS-XL whose lemmas are not contained in the intersection with FNFT-train, i.e., $\text{FNFT-train} \cup (\text{WaS-XL} \setminus \text{FNFT-train})$. In other words: if the FrameNet fulltext training set does not contain a particular sense for a lemma, supplement with training data from WaS-XL.

The last two lines in Table 3.5 show the results for B-WaS-XL and U-WaS-XL. Precision is higher or not significantly lower for B-WaS-XL compared to FNFT-train, while recall and F_1 are higher. U-WaS-XL leads to higher recall compared to B-WaS-XL, and overall highest F_1 scores. This proves that the automatically labeled WaS-XL is complementary to the manually labeled FrameNet fulltext training set and contributes to a better coverage on the diverse range of test sets considered in the evaluation.

Multiword verbs. Our approach of automatically generating training data also includes multiword verbs such as *carry out* or *cut short*. It treats those verb senses as additional senses of the head verb and creates sense patterns for them, i.e., the sense for *carry out* is a specific sense of *carry*.

Thus, the presented approach differs from previous approaches to multiword expressions (MWEs) in word sense disambiguation: Finlayson and Kulkarni (2011) detect MWEs before performing WSD. They credit performance improvements to the “one sense per collocation” hypothesis, the assumption that MWEs often only have a single sense. FrameNet 1.5 challenges the “one sense per collocation” hypothesis: it lists ten phrasal verbs with two or three senses, e.g., the senses *Inhibit_movement*, *Silencing* and *Become_Silent* for *shut up*. Therefore, sense-labeled training data are also valuable for multiword verbs.

WaS-XL contains over 100,000 sense instances of 194 multiword verbs, of which 35 have multiple FrameNet senses. In order to estimate the quality and potential of the DistantSRL approach for multiwords, the performance of the verb sense disambiguation system was specifically evaluated on multiwords and their head verbs from MASC. The MASC test set

corpus	role tokens	role types	sense types	average roles per sense	average instances per role
WaSR-XL-uni	549,777	1,485	809	1.8	370
WaSR-L-uni	34,678	968	597	1.6	36
WaSR-XL-set	823,768	2,054	849	2.4	401
WaSR-L-set	53,935	1,349	648	2.1	40
FNFT★	12,988	2,867	800	3.6	4.5

Table 3.7: Role statistics of automatically labeled corpora.

contains 81 relevant sense instances. The precision is 0.66 compared to 0.59 when training on the FrameNet fulltext training set, at slightly higher coverage. While the test set is too small to provide significant results, the results indicate that the automatically labeled data also contribute to the disambiguation of multiword verbs.

This observation closes our frame identification experiments. The next section describes the generation of role-labeled data and their application in the task of role classification.

3.4.3 Role Labeling: Corpus Creation and Experiments

This subsection presents the creation of a role labeled corpus, i.e., Stage 2 of DistantSRL, and its extrinsic evaluation in a role classification task.

Evaluation setup. The role classification evaluation uses a specifically developed supervised system for semantic role classification. It trains a log-linear model per verb-frame using WEKA (Hall et al., 2009) and the features described in Fürstenau and Lapata (2012).

Note that this setup does not evaluate the task of argument identification. Argument identification is performed by the rule-based VerbNet role transfer in Step 2 and follows common syntactic heuristics for argument identification based on dependency parsing. Following Zafirain et al. (2013), this section specifically considers the subtask of role classification, as the main focus is on the quality of the automatically labeled data on the semantic level. In this context, it is important that the features of our role classification system do not use span information. They include lemma and POS of the argument head, its governing word, and the words right and left of the argument head, the position of the argument relative to the predicate, and the grammatical relation between the argument head and the predicate. Pre-processing is the same as for the verb sense disambiguation experiments. The test datasets and evaluation metrics introduced for the verb sense disambiguation experiments are also used for the role experiments, see Table 3.2.

Corpus creation and statistics. The two sense-labeled corpora WaS-XL and WaS-L, the corpus with the highest sense coverage and the one leading to the highest sense labeling

precision in Section 3.4.2, serve as input for the automatic role label transfer, leading to role-labeled corpora WaSR-XL and WaSR-L. There are two variants of each of these corpora, one that only contains those role instances with a unique role label, e.g., $\text{role}=\textit{Agent}$, marked with the suffix *-uni* in Table 3.7, and one that additionally includes set-valued labels, marked with the suffix *-set*, e.g., $\text{role}=\{\textit{Agent}, \textit{Cause}\}$, see also Section 3.1.3.

For WaSR-XL, Step 2A results in 1.9 million arguments labeled with VerbNet roles. This number is reduced by 66% in Step 2B when VerbNet roles are mapped to FrameNet roles via SemLink, leading to 549,777 role instances in the WaSR-XL-uni dataset. The reduction is a result of the incomplete mapping between VerbNet and FrameNet senses and roles in SemLink. It leads to a lower role coverage in the WaSR corpora compared to the FrameNet fulltext training set, as shown in column *role types* of Table 3.7. The loss of instances during the mapping and the accompanying reduction in role coverage could be avoided by improved FrameNet to VerbNet mappings, for instance using the automatically created extension to SemLink in the Predicate Matrix (Lopez de Lacalle et al., 2016).

Table 3.7 shows that the resulting corpora WaSR-L-uni and WaSR-XL-uni contain 34,678 and 549,777 uniquely assigned role instances for the verbs in the four test sets. This is a large number compared to the 12,988 instances in FNFT★. The counts are even higher for the corpora including sets of labels: there are 53,935 and 823,768 role instances for WaSR-L and WaSR-XL respectively.

Due to the sparse labeling approach, the WaSR corpora contain on average only up to 1.8 roles per predicate, compared to an average of 3.6 roles per predicate in FNFT★. This number rises to 2.4 when instances with sets of labels are included – an instance with a set-valued label counts as a single role instance.

Role classification experiments. In the experimental evaluation, we compare the role classification system trained on WaSR-XL-(set/uni) and WaSR-L-(set/uni) to the system trained on the FrameNet fulltext training set. Again, the evaluation includes combinations of the FrameNet fulltext corpus and WaSR corpora. Additionally, it includes an evaluation of learning curves for an increasing number of training instances per verb sense sampled from the large WaSR-XL. Test datasets are the same as for the verb sense disambiguation experiments, see Table 3.2. The evaluation metrics used are also the same, e.g., we report P, R, and F_1 on all frame-verb combinations for which there is more than one role in the training data. Training the system on WaSR-XL-set and WaSR-L-set includes sets of role labels as training instances. Therefore, sets of role labels are among the predicted labels. In the evaluation, the set-valued labels are counted as correct if they contain the gold label. In this setup, we consider a single role label as a set-valued label with set size 1.

training set	FNFT-test			Fate			MASC			SemEval		
	P	R	F ₁	P	R	F ₁	P	R	F ₁	P	R	F ₁
WaSR-L-uni	62.9	27.5	38.3	55.0	23.0	32.5	56.3	23.5	33.2	66.0	33.0	44.0
WaSR-L-set	67.9	32.6	44.1	72.0	28.3	43	62.9	28.3	39.1	68.6	37.2	48.3
WaSR-XL-uni	65.8*	33.3*	44.2	61.9	28.1*	38.7	65.2*	25.3*	36.5	68.9	39.4*	51.0
WaSR-XL-set	75.0*	39.8*	59.0	73.3*	33.7*	46.2	64.8*	29.7*	48.0	72.2	44.1*	54.7
FNFT-train	74.1	83.1	78.3	65.2	64.2	64.7	72.4	52.7	61.0	75.0	62.5	66.3
B-WaSR-XL-uni	72.8*	87.8*	79.6	79.6	64.5	69.8*	71.8	57.4*	63.8	69.6	71.0*	70.3
U-WaSR-XL-uni	69.1*	88.3*	77.6	62.9	71.0*	66.3	57.9*	62.4	67.1	67.1	72.1*	69.5

Table 3.8: Role classification P, R, F₁ in percent; * marks significant differences to the system trained on FNFT-train; highest scores per dataset in **boldface**.

More formally, this means:

$$\text{correct}(S, i, g) = \begin{cases} 1, & \text{if } |S| = 1 \ \& \ g = p_1 \\ 1, & \text{if } |S| > 1 \ \& \ g \in S \\ 0, & \text{otherwise} \end{cases} \quad (3.2)$$

for a given test instance i with gold role label g and the predicted set-valued role label S that contains n role labels p_j with $j \in (1, \dots, n)$.

The next paragraphs describe the different evaluation settings in detail.

Results on WaSR corpora. Table 3.8 shows that WaSR-XL-set – as expected – leads to higher precision and recall than WaSR-XL-uni, resulting from the larger role coverage in the training set, and the lenient evaluation setting that accepts sets of labels. The comparison between WaSR-XL-(set/uni) and WaSR-L-(set/uni) shows that the benefits of the strict filtering for the sense corpora do not carry over to the role-labeled corpora: scores are lower for WaSR-L-(set/uni) on all test sets because of fewer role labeled instances in WaSR-L-(set/uni), see Table 3.7. This could be improved by creating larger corpora with the strict, high-precision filter f applied to create WaS-L. The size of WaS-L could, for instance, be doubled by labeling four additional sections of the ukWAC corpus.

Comparison to the FrameNet fulltext training set. Table 3.8 compares the results when training on WaSR-XL-(set/uni) to the role classification system trained on the FrameNet fulltext training set. Note that the lenient evaluation setting can be emulated for the FrameNet fulltext training set by retrieving the label set S_l in WaSR-XL-set for a label l predicted by the FrameNet fulltext training set system and counting l as correct if any of the labels in S_l matches the gold label. This, however, did not result in any difference to the regular

training set	evaluated subset	P
test = FNFT-test		
WaSR-XL-uni	intersection I : $\{x \in \text{test} \mid x \in \text{WaSR-XL-uni} \cap \text{FNFT-train}\}$	0.7334
FNFT-train	intersection I : $\{x \in \text{test} \mid x \in \text{WaSR-XL-uni} \cap \text{FNFT-train}\}$	0.7334
WaSR-XL-uni	difference DA : $\{x \in \text{test} \mid x \in \text{WaSR-XL-uni} \setminus \text{FNFT-train}\}$	0.5481
FNFT-train	difference DB : $\{x \in \text{test} \mid x \in \text{FNFT-train} \setminus \text{WaSR-XL-uni}\}$	0.7469
test = Fate		
WaSR-XL-uni	intersection I : $\{x \in \text{test} \mid x \in \text{WaSR-XL-uni} \cap \text{FNFT-train}\}$	0.6272
FNFT-train	intersection I : $\{x \in \text{test} \mid x \in \text{WaSR-XL-uni} \cap \text{FNFT-train}\}$	0.6747
WaSR-XL-uni	difference DA : $\{x \in \text{test} \mid x \in \text{WaSR-XL-uni} \setminus \text{FNFT-train}\}$	0.6126
FNFT-train	difference DB : $\{x \in \text{test} \mid x \in \text{FNFT-train} \setminus \text{WaSR-XL-uni}\}$	0.6342
test = MASC		
WaSR-XL-uni	intersection I : $\{x \in \text{test} \mid x \in \text{WaSR-XL-uni} \cap \text{FNFT-train}\}$	0.6494
FNFT-train	intersection I : $\{x \in \text{test} \mid x \in \text{WaSR-XL-uni} \cap \text{FNFT-train}\}$	0.7303
WaSR-XL-uni	difference DA : $\{x \in \text{test} \mid x \in \text{WaSR-XL-uni} \setminus \text{FNFT-train}\}$	0.6851
FNFT-train	difference DB : $\{x \in \text{test} \mid x \in \text{FNFT-train} \setminus \text{WaSR-XL-uni}\}$	0.7172
test = SemEval		
WaSR-XL-uni	intersection I : $\{x \in \text{test} \mid x \in \text{WaSR-XL-uni} \cap \text{FNFT-train}\}$	0.6827
FNFT-train	intersection I : $\{x \in \text{test} \mid x \in \text{WaSR-XL-uni} \cap \text{FNFT-train}\}$	0.6835
WaSR-XL-uni	difference DA : $\{x \in \text{test} \mid x \in \text{WaSR-XL-uni} \setminus \text{FNFT-train}\}$	0.7357
FNFT-train	difference DB : $\{x \in \text{test} \mid x \in \text{FNFT-train} \setminus \text{WaSR-XL-uni}\}$	0.7353

Table 3.9: Detailed role classification results: precision on verbs evaluated for test instances in both training sets (intersection I) and test instances in the set difference DA and DB .

evaluation; it appears that the labeling errors of the FrameNet fulltext training set-based system are different from the label sets resulting from DistantSRL.

The precision for WaSR-XL-uni equals the precision for the FrameNet fulltext training set when evaluating on the SemEval and Fate test sets, i.e., the differences are not significant. This is remarkable considering that only partially labeled data are available for training.

For WaSR-XL-set, the precision scores for SemEval and Fate improve over the FrameNet fulltext training set system, the differences are significant for Fate. The recall of the WaSR corpora is significantly lower overall. This is a result of the sparse, partial labeling and the lower role coverage of the automatically labeled corpora.

Comparison to the FrameNet fulltext training set on common and disjoint sets of test verbs. Similar to the verb sense disambiguation evaluation, this paragraph compares the performance of the system based on WaSR-XL-uni and the FrameNet fulltext training set-

based system on the intersection of the evaluated senses between both systems. The *intersection* I is defined for each test set T as the set of test instances x such that: $I = \{x \in T \mid x \in \text{WaS-XL} \cap \text{FNFT-train}\}$.

The *set difference* DA is defined for a test set T and two training sets A and B as the set of test instances x such that $DA = \{x \in T \mid x \in A \setminus B\}$. This means that the set difference contains all the test instances that occur in A , but not in B .

The corresponding precision scores are shown in Table 3.9. The precision of the classifier trained on the FrameNet fulltext training set is higher on the intersection I for all test sets except for SemEval, where it is similar. Moreover, the larger role coverage of FNFT-train effects the classification: the system based on FNFT-train labels on average two additional roles per sense compared to the system trained on the WaSR corpora. This is expected based on the role statistics shown in Table 3.7.

Evaluating only on the set difference, the instances not contained in the intersection, we see that WaSR-XL-uni contributes some role instances that are not covered by the FrameNet fulltext training set. This applies to a smaller extent than observed for the verb sense disambiguation experiments. The precision is lower for the additional instances in WaSR-XL-uni for FNFT-test (difference DA), as expected for an out-of-domain training set compared to the in-domain training set. For SemEval the precision is similar for difference DA and difference DB , and only slightly lower for difference DB for Fate and MASC. This indicates a good quality of the additional training instances.

The additional instances contributed by the WaSR-XL-uni constitute between 7% and 18% of the total evaluated instances, compared to 26% to 50% instances added by the FrameNet fulltext training set. The precision of WaSR-XL-uni on the intersection for MASC is high at 0.68, compared to 0.55 for FNFT-test. These results indicate that WaSR-XL-uni is complementary to the FrameNet fulltext training set, and may contribute to a better system performance when combined with FN-train, in particular on the out-of-domain test sets.

Combining training data. To give further evidence that the automatically labeled corpus is complementary to the FrameNet fulltext training set, experiments that combine WaSR-XL-uni with the FrameNet fulltext training set were performed. As for the verb sense disambiguation experiments, the combinations include U-WaSR-XL-uni, the union of the data sets, and B-WaSR-XL-uni, backing off to WaSR-XL-uni when the FrameNet fulltext training set does not provide roles for a sense.

Table 3.8 shows better results for the back-off corpus B-WaSR-XL-uni than for the union U-WaSR-XL-uni. The recall of the B-WaSR-XL-uni is significantly higher compared to the FrameNet fulltext training set, and precision values are not significantly lower except for the in-domain FrameNet fulltext test set. This demonstrates that the automatically role-labeled corpora created by DistantSRL can supplement a manually labeled corpus and benefit the resulting system.

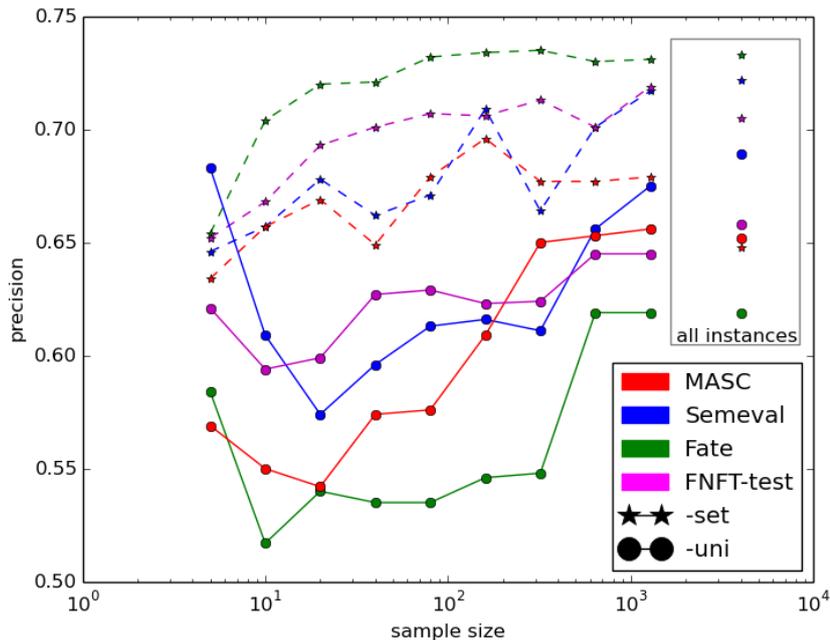


Figure 3.10: Role classification learning curves for WaSR-XL-(set/uni).

WaSR sampling. Because the WaSR corpora show a strongly Zipfian distribution of roles (i.e., there are a few roles with a very large number of instances) using all instances may lead to a non-representative distribution of roles by predicate, harming the role classification performance. To this end, sub-samples of the training sets with an increasing number of training instances per role were evaluated.

The sub-samples include nine training sets randomly sampled from WaSR-XL with a different maximal number of training instances per role s such that $s = 5 \cdot 2^i$ for $i \in \{0, 1, \dots, 8\}$. The resulting sets contain up to s instances with s ranging from 5 to 1,280. This kind of sample selection covers a wide range of set sizes, because the number s of sampled instances is doubled when i is incremented by one, e.g., $s \in \{5, 10, 20, 40, 80, 160, 320, 640, 1,280\}$.

Figure 3.10 shows the learning curves for precision on WaSR-XL-(set/uni). It shows that distributional effects occur, i.e., that certain sample sizes s lead to higher precision for a test set than using the full corpus. The MASC test set particularly benefits from the sampling: combining FNFT-train with the best sample from the WaSR-XL-set corpus (sampling 160 instances per role) results in the overall highest precision $P=0.738$ and $F_1=0.65$ on the MASC set. MASC benefits less from a highly skewed distribution of roles, because MASC, representing a lexical sample, also contains rare roles. These are better represented in an even sample. The sampling also leads to highest F_1 on SemEval and Fate.

The upper bound for this task can be measured as the human role classification performance, i.e., the human agreement scores according to the F_1 measure. These scores have

not been reported for our test sets. Recent annotation studies of FrameNet roles report role labeling agreement of 78.1% F_1 for English Twitter text (Johannsen et al., 2015). These values cannot be directly transferred to our test sets, because they were obtained on a different type of text, user-generated texts from Twitter.

3.5 Application of DistantSRL to German

This section presents the application of DistantSRL to the German language. Compared to English, German is a low-resource language. There is a FrameNet-like resource for German in the form of SALSA 2, but there are, for instance, no independently created test sets. The general setup of the experiments for German follows the one presented for English.

The German counterpart of FrameNet is SALSA 2, which is mainly known as a corpus annotated with FrameNet frames, and newly developed proto-frames, see also Section 2.2.3. As introduced in Section 2.7.1, a SALSA lexical knowledge base can be derived from the corpus and modeled in UBY-LMF. There are fewer linkings to other lexical resources than for the English linked lexical knowledge base centered around FrameNet: via the FrameNet frames that were also used in SALSA 2, SALSA 2 is implicitly linked to FrameNet. This allows us to link SALSA 2 to the German Wiktionary using the FrameNet – WiktionaryDE alignment introduced in Section 2.6.

An obvious choice for an additional lexical knowledge base to link to SALSA 2 would be GermaNet, the German WordNet. However, there is no sense-level linking between GermaNet and SALSA 2. It would also be difficult to create such a linking automatically, for instance using the sense alignment method used in Section 2.4 for FrameNet and Wiktionary, due to a lack of definition glosses in the SALSA 2 resource. The reason is that our sense alignment method relies on gloss similarity between two senses in different lexical knowledge bases to establish a sense alignment. If there is no gloss, no alignment can be established. The experiments in this section, however, show that the DistantSRL approach also works for other languages with a smaller number of available resources.

3.5.1 Automatically Generated Training Data and Test Data

Gold standard training and test data. As SALSA 2 does not provide additional lexical unit examples, the corpus was split into a training set S-train that is used for training the reference system and for the extraction of seed patterns, a development set S-dev, and a test set S-test, all consisting of verbal predicates. The proportion of train, development and test instances is 0.6, 0.2, 0.2; data statistics are shown in Table 3.10.

Linked lexical knowledge base. As already mentioned, the set of linked resources consists of the S-train part of SALSA 2 and Wiktionary, as shown in Figure 3.11. The linking between the resources is based on the implicit linking of SALSA 2 to FrameNet frames and

dataset	verbs lemmas	sense types	role types	sense tokens	role tokens
S-test	390	684	1,045	3,414	8,010
S-dev	390	678	1,071	3,516	8,139
S-train	458	1,167	1,511	9460	22,669
WaS-de ($t=0.07$)	333	920	-	602,207	-
WaSR-de-set	193	277	210	80,370	115,332
WaSR-de-uni	172	241	155	51,241	57,822

Table 3.10: German dataset statistics on verbs.

the linking of FrameNet to the German Wiktionary as introduced in Section 2.6. This leads to more than 22,900 seed patterns.

Unlabeled corpus. The unlabeled corpus is based on deWAC sections 1 to 5 (Baroni et al., 2009). These contain more than 25 million sentences and more than 440 million tokens. Like the English corpus, the deWAC corpus was filtered by the target verb lemmas prior to applying DistantSRL. The filtered corpus contains more than 11 million sentences and more than 270 million tokens.

3.5.2 Frame Labeling Experiments

This section describes the creation of frame-labeled corpora for German according to Stage 1 of DistantSRL and their evaluation in the task of verb sense disambiguation.

Evaluation setup. The evaluation setup follows the setup for English. The preprocessing of the verb sense disambiguation system is adapted to German, using the language tool segmenter, Stanford named-entity recognizer, Treetagger for POS-tagging and lemmatization (Schmid, 1995), and a module from DKPro core that connects separated particles with their main verbs (Eckart de Castilho and Gurevych, 2014).

Parameter estimation and corpus creation. The lexical knowledge base is used to generate more than 22,900 seed patterns. This is a magnitude smaller than the over 350,000 seed patterns extracted for English. A large number of the seed patterns for English stems from the ASPs generated for VerbNet. The distribution of seed patterns between ASPs and LSPs for German is roughly equal. This agrees with the statistics for English seed patterns when leaving aside the VerbNet-based patterns. For parameter estimation, the patterns are used to label deWAC sections 1-3. The thresholds t and f are determined in a verb sense disambiguation evaluation on S-dev, evaluating the same range of values as for English. The thresholds $t=0.07$ and a discriminating filter of $f=0.07$ result in best precision. Similar to English, $t=0.07$ results in the best F_1 on the development set.

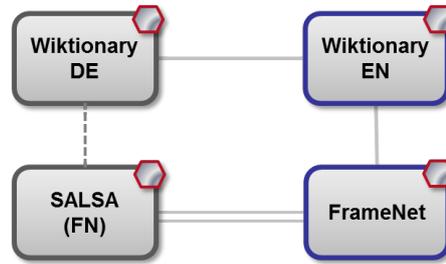


Figure 3.11: Linked lexical knowledge base centered around FrameNet for German.

training set	P	R	F ₁
WaS-de	0.672*	0.912*	0.773
B-WaS-de	0.711	0.958*	0.816
U-WaS-de	0.676*	0.961*	0.794
S-train	0.707	0.946	0.809

Table 3.11: German verb sense disambiguation (frame identification) evaluation, P, R, F₁; * marks significant differences to S-train; highest scores in **boldface**.

Corpus statistics. The corpus with $t = 0.07$ is called WaS-de. Corpus statistics are listed in Table 3.10. Compared to S-train, WaS-de contains a smaller number of unique senses, but 50 times more sense instances. The WaS-de corpus is much smaller than the English WaS corpora shown in Table 3.4.

Experimental evaluation. The evaluation compares a verb sense disambiguation system trained on S-train to systems trained on WaS-de ($t=0.07$), on U-WaS-de, which constitutes the union with S-train, and on B-WaS-de, which is the backoff-variant supplementing the training split S-train with instances from WaS-de. The test set is S-test. The results in Table 3.11 show that the performance of the system based on WaS-de is worse than the one based on S-train, but the backoff version reaches the best scores overall, indicating that our WaS-de corpora might be complementary to S-train.

3.5.3 Role Labeling Experiments

To apply DistantSRL to German, the rule-based VerbNet role labeling was adapted to the German dependencies from the Mate-tools parser (Seeker and Kuhn, 2012). Performing Steps 2A and 2B on WaS-de results in the two corpora WaSR-de-set and WaSR-de-uni. Their statistics are shown in Table 3.10.

training set	P	R	F ₁
WaSR-de-set	0.593*	0.208*	0.308
WaSR-de-uni	0.69*	0.171*	0.274
B-WaSR-de-uni	0.828	0.958	0.888
U-WaSR-de-uni	0.779*	0.958	0.859
S-train	0.828	0.956	0.887

Table 3.12: German role classification P, R, F₁; * marks significant differences to S-train; highest scores in **boldface**.

Evaluation setup. The role classification system introduced in Section 3.4.3 was again used for evaluation. The features of the role classifier are also applicable to German when German preprocessing tools are used.

Corpus statistics. The statistics for the corpora WaSR-de-set and WaSR-de-uni are shown in Table 3.10. Again, the numbers are comparatively smaller than those for English. This has two reasons. First, the number of seed patterns is much smaller, as a result of the smaller linked lexical knowledge base available for German. This already leads to a smaller sense-labeled corpus. Second, the SemLink mapping does not cover frames and roles unique to SALSA 2. Thus, the proportion of VerbNet labels from Step 2A that cannot be translated to SALSA 2 roles is larger than for English.

Experimental evaluation. Table 3.12 shows the evaluation results. Training on WaSR-de-uni results in a good precision of 0.69, but it is significantly lower than for the S-train system with 0.828. Recall is very low at 0.17. The system trained for the backoff setting (B-WaSR-de-uni) receives almost the same scores as the S-train, with minimally higher recall. It is a good sign that the additional training instances do not lead to a decrease in performance, on the other hand they only minimally impact the results in a positive way due to the low coverage of the WaSR-de corpora. We conclude that the role labeling evaluation suffers from data sparsity, as the automatically labeled corpora WaS-de and WaSR-de are much smaller than those for English, and, as shown in Table 3.10, cover fewer frame and role types than the training and test splits of the SALSA corpus.

Summary. This evaluation shows that DistantSRL can also be applied to German. For verb sense disambiguation, the automatically labeled data can be used to improve on using S-train alone. The improvements are not significant, which has several potential causes, e.g., the smaller set of lexical knowledge bases used for seed pattern extraction compared to English, and the smaller size of the automatically labeled corpora. The data sparsity also results in very low recall for the role classification.

Additional work is required to improve the quality of the German automatically labeled corpora. This includes increasing the number of linked resources and seed sentences, e.g., linking SALSA to GermaNet or OmegaWiki. Note that the expected contributions from GermaNet to German DistantSRL corpora are smaller than those from WordNet for English DistantSRL corpora, because a large number of example sentences in GermaNet have been semi-automatically extracted from the German Wiktionary (Henrich et al., 2014). On the other hand, corpora labeled with GermaNet senses like WebCAGe (Henrich et al., 2012) could be added to expand the seed set. Another point of improvement considers the SALSA-specific frames: the sense linking from FrameNet to SALSA and the role-level linking from SALSA to VerbNet via SemLink do not cover SALSA-specific frames and their role labels, because SemLink does not cover proto-frames from SALSA. This results in a decreased coverage compared to the English VerbNet to FrameNet mapping, and thus a lower role coverage of the corpora created via DistantSRL for German.

The impact of the automatically labeled training data on full semantic role labeling for German is expected to be larger than for English, because the potential for improvement is larger with only SALSA 2 as the training set. In order to be successful, it should, however, rely on larger, improved automatically labeled corpora.

3.6 Full Semantic Role Labeling with DistantSRL

The sense and role classification experiments introduced in the previous section show that DistantSRL is able to generate high-quality frame- and role-labeled training data. In order to prove the usefulness of the automatically generated data in a full SRL setup, we use them to train an open-source SRL system. This section describes the evaluation of DistantSRL using the open-source SRL system SEMAFOR 3.0 (Das et al., 2014).

For a long time, SEMAFOR has been the state-of-the-art system for FrameNet semantic role labeling. It was recently superseded by systems relying on deep learning methods (Hermann et al., 2014; FitzGerald et al., 2015). In this section, we present results from training SEMAFOR on our automatically labeled English corpora WaS and WaSR and discuss their potential benefit to state-of-the-art semantic role labeling systems.

3.6.1 Experiments with SEMAFOR

SEMAFOR training consists of the two steps of FrameNet semantic role labeling already introduced in Chapter 1, training the frame identification model and training the role labeling model that integrates argument span identification and role classification. The SEMAFOR system is described in detail in Das et al. (2014), we provide a short introduction here.

The frame identification system is a conditional log-linear model that relies on an elaborate feature set based on syntactic and lexical features, using the WordNet hierarchy as a

source of lexical information. It uses latent variables and semi-supervised lexical expansion via graph-propagation to improve the generalization across predicates.

The role labeling system of SEMAFOR is also based on a log-linear model and jointly performs argument identification and role classification. This system assigns the most suitable span for the roles of the previously assigned frame. Das et al. (2014) use optimization via dual decomposition to enforce SRL-specific constraints, for instance avoiding that several argument spans are labeled with the same role label.

In our experiments, we retrain the SEMAFOR frame identification and role labeling models using our automatically labeled corpora, and, for comparison, the FrameNet fulltext training set FNFT-train. We evaluate on the four test sets consisting of verbal predicates introduced in Section 3.4. Before presenting the experiments, we introduce the evaluation metrics typically used for the evaluation of the SEMAFOR system.

SEMAFOR evaluation metrics. We use the official SEMAFOR evaluation script (Das et al., 2014) and some additional analyses to evaluate the retrained SEMAFOR systems.

The open-source code associated with SEMAFOR provides an evaluation script that evaluates three aspects of FrameNet semantic role labeling: accuracy for frame identification (FrameId), full semantic role labeling (SRL) given frames predicted by the frame identification system, and full SRL given gold standard frames.

For frame identification it has become standard to input the gold target spans, so that the targets do not need to be identified by the system. Therefore, recall is not an issue. The SEMAFOR evaluation script reports accuracy based on the correct frame labels and the total number of frame instances in the test dataset. The script provides an exact evaluation setting that only counts exactly matching frames (exact match), and a lenient setting, that gives partial credit to frames that can be reached via frame relations in FrameNet (partial match). In all our experiments presented below, we use the exact match setting.

The second variant of evaluation concerns full semantic role labeling. It reports precision P, recall R, and F_1 scores. For this variant, correct frame and role labels are added up and compared to the corresponding counts of the gold standard data. Besides counting the correct frame labels, this includes counting the correct role labels for matching argument spans. For this evaluation, the distinction in exact match and lenient match settings is also relevant. Note that the full SRL evaluation counts predicted role labels only if the argument spans exactly match the gold standard spans. This is a strict evaluation setting: roles with overlapping spans and correct labels cannot match the gold standard. Moreover, the script gives higher weight to core roles than to non-core roles, discounting non-core labels by 50%. As a result, core roles contribute more to the overall score.

The third variant of evaluation considers full semantic role labeling *given gold standard frames*. It uses the same algorithm as the second variant, and also reports precision, recall, and F_1 scores. Because the contribution of the gold frame labels, which are correct by

training set	Fate	MASC	SemEval	FNFT-test
WaS-L	0.51124	0.56612	0.68473	0.69236
WaS-L-union	0.53372	<u>0.57071</u>	0.71798	0.73834
WaS-XL-10	0.45459	0.44013	0.52463	0.57066
WaS-XL-20	0.46088	0.44639	0.58005	0.58418
WaS-XL-40	0.47257	0.46350	0.58005	0.57471
WaS-XL-80	0.49236	0.46600	0.58621	0.62069
WaS-XL-160	0.46313	0.48602	0.54433	0.62001
WaS-XL-320	0.48291	0.50688	0.59852	0.61325
WaS-XL-640	0.48246	0.52274	0.60345	0.62813
WaS-XL-10-fntrain	0.45459	0.44013	0.52463	0.5706
WaS-XL-20-fntrain	0.50270	0.48811	0.64163	0.69033
WaS-XL-40-fntrain	0.49910	0.49353	0.63547	0.65585
WaS-XL-80-fntrain	0.50809	0.48769	0.62931	0.67343
WaS-XL-160-fntrain	0.50360	0.50396	0.62192	0.67748
FNFT-train	<u>0.55755</u>	0.52774	<u>0.73892</u>	<u>0.76876</u>

Table 3.13: Frame identification experiments with retrained SEMAFOR frame identification models: accuracy with exact match for test verbs; **boldface** marks best results in section, underline marks overall best results.

default, to the final score is not discounted from the score, this is not an evaluation of role labeling proper. This variant is useful when comparing different SRL systems on the same test set, but not sufficient when comparing the role labeling performance on different test sets with a different ratio of frame labels to role labels resulting from different annotation strategies, as done in the current section and in Chapter 4 below. It furthermore hampers the analysis of the contribution of frame identification and role labeling to the full SRL performance across test sets. We therefore add an analysis of the role labeling proper.

To compute the new role labeling scores (RoleId), we re-evaluate the output of the evaluation script when running full SRL evaluation with or without gold frames, discounting the frame labels: we analyze the output of the script to retain the original counts for role labels and compute P, R, and F_1 scores for the role labels.

Frame identification experiments with SEMAFOR. The frame identification model was trained on several corpora based on WaS-L, WaS-XL, and the FrameNet fulltext corpus for comparison. They include the union of WaS-L and FNFT-train, and several backoff variants that supplement the FrameNet fulltext training set with instances from WaS-L if the FrameNet fulltext training set contains fewer than k instances of this sense. We evaluate the following range of values for k : (0, 3, 5, 10). For $k=0$, we only supplement data from

training set	FNFT-test			Fate			MASC			SemEval		
	P	R	F ₁									
WaS-L	54.8	38.0	44.9	36.4	22.0	27.4	47.4	32.9	38.8	51.3	35.5	42.0
WaS-L-union	57.6	39.8	47.2	38.1	22.9	28.6	47.4	33.1	39.0	52.9	36.7	43.3
FNFT-train	62.0	41.0	49.4	40.9	24.1	30.32	44.7	30.9	36.5	56.2	37.7	45.1

Table 3.14: Full SRL results for test verbs with retrained SEMAFOR frame identification model and SEMAFOR/K role labeling model from [Kshirsagar et al. \(2015\)](#); **boldface** marks best results.

WaS-L if the corresponding sense is not covered by the FrameNet fulltext training set. We do not find improvements for the backoff setting and therefore do not report the results.

We also experiment with the high-recall, low precision WaS-XL corpus. SEMAFOR was not created to handle the large amounts of training instances in WaS-XL, frame identification training on WaS-XL is not feasible due to large memory requirements. Therefore, we create random samples with up to 640 instances per role, as was done in Section 3.4.3.

For comparison, we retrain the SEMAFOR frame identification model on FNFT-train. We evaluate the frame identification models on the test sets introduced in Section 3.4.1. Since the automatically labeled corpora only contain verbs, we constrain the evaluation to verbal predicates. Table 3.13 shows the results of the exact frame identification evaluation on our test datasets. Note that the results are not comparable to the frame classification results reported in Section 3.4.3. In these experiments, we only evaluated on seen predicates, filtering the test instances according to the training set. To support reproducibility of our experiments, the results in Table 3.13 represent all verb instances in the test sets. Thus, the evaluation also includes verbs not contained in the FrameNet lexicon.

The results show that the highly polysemous MASC dataset benefits from the automatically labeled corpora: the accuracy is improved for WaS-L, overall best results are achieved for the union of WaS-L and the FrameNet fulltext training set, which adds 0.043 points accuracy to the system trained on the FrameNet fulltext training set alone. This improvement is statistically significant according to Fisher’s exact test at $p < 0.05$. The other test datasets do not benefit from the additional training data. The WaS-XL samples receive lower scores than the high-precision WaS-L corpus on all test sets.

Training on WaS-L alone also improves the results for MASC, and reduces the scores for the other test sets. The results still show that SEMAFOR can train a reasonable frame identification model on our automatically generated data. The reductions are smaller for the out-of-domain test sets Fate, MASC, and SemEval, compared to the in-domain FNFT-test. FNFT-test benefits most from the in-domain training set FNFT-train.

Role labeling experiments with SEMAFOR. We evaluate two different setups, first evaluating the impact of the retrained frame identification models on role labeling using the best

available SEMAFOR role labeling model from [Kshirsagar et al. \(2015\)](#), second retraining SEMAFOR 3.0 using corpora based on WaSR-XL-uni.

Table 3.14 shows the results for full SRL using the retrained frame identification models and the best available role labeling model for SEMAFOR from [Kshirsagar et al. \(2015\)](#), called SEMAFOR/K. The results show that the improvements in frame identification for MASC contribute to enhanced role labeling, leading to an improvement of 2% F_1 for FrameNet semantic role labeling.

In Table 3.15, we show the SRL results when retraining SEMAFOR on the role-labeled data in WaSR-L. For the role classification in Section 3.4.3, we observed better results for WaSR-XL compared to WaSR-L, due to its better role coverage. Since we cannot retrain SEMAFOR with the large number of training instances in WaSR-XL, we use WaSR-L for retraining SEMAFOR in the experiments reported here.

Table 3.15 comprises results for full SRL and role labeling given a) gold frames and b) system frames predicted by the SEMAFOR retrained on FNFT-train. The results for full SRL given system frames in the first section of the table show that precision is higher than for SEMAFOR/K, but recall is low, leading to overall lower F_1 scores. The high precision is a result of the higher frame coverage, but lower role coverage of WaS-L. WaS-L contains mostly core roles. Since WaS-L does not assign labels to the non-core roles, only the few core roles per frame are assigned by the retrained SEMAFOR system. The test sets also contain a large proportion of core roles, so there are fewer non-core roles in the training set that could confound the results when training on WaS-L. The same effect leads to the surprisingly high results for role labeling and SRL given gold frames. Additionally, as we observed earlier, the SEMAFOR evaluation script discounts non-core roles, giving a higher weight to the core roles, which increases the positive effect for WaSR-L. This is an indication that it might be beneficial to train core roles independently of the non-core roles, which was also suggested by [Matsubayashi et al. \(2009\)](#).

Even though the evaluation setting benefits WaSR-L due to its bias to core roles, these results indicate that DistantSRL creates large amounts of high quality role annotations.

Summary. The experiments with SEMAFOR show that our automatically generated data can be used to train SEMAFOR successfully. Evaluation scores are lower compared to gold standard training data for most of the test sets. Improved scores on MASC show that DistantSRL can be particularly beneficial to out-of-domain test sets. This once more confirms the quality and usefulness of training data generated with DistantSRL.

We cannot exploit certain advantages of our training data, namely the large amounts of training data in the WaSR-XL corpus: SEMAFOR is not equipped to deal with such large amounts of training data. SEMAFOR was developed for the manually labeled FrameNet fulltext training set.

system	FNFT-test			Fate			MASC			SemEval		
	P	R	F ₁	P	R	F ₁	P	R	F ₁	P	R	F ₁
SRL system frame												
SEMAFOR/K	62.01	40.99	49.35	40.86	24.09	30.31	44.72	30.89	36.54	56.15	37.66	45.08
WaSR-L	71.32	30.00	42.25	50.67	19.69	28.36	48.22	20.72	28.99	69.34	28.55	40.45
SRL gold frame												
SEMAFOR/K	69.18	73.06	71.07	72.88	74.95	73.9	65.25	67.72	66.46	63.14	65.66	64.38
WaSR-L	95.02	95.91	95.46	95.74	96.35	96.04	93.86	94.86	94.36	92.27	92.65	92.46
RoleId system frame												
SEMAFOR/K	43.47	20.20	27.58	21.10	8.05	11.66	36.10	18.73	24.66	37.81	18.92	25.22
WaSR-L	33.56	2.86	5.27	17.95	1.41	2.62	25.85	2.92	5.25	47.62	5.1	9.21
RoleId gold frame												
SEMAFOR/K	52.81	57.49	55.05	59.70	62.26	60.95	47.01	49.78	48.36	45.41	48.15	46.74
WaSR-L	92.18	93.54	92.85	93.63	94.52	94.07	90.53	2.03	91.27	88.61	89.17	88.89

Table 3.15: SEMAFOR SRL results for test verbs: contrasting the role labeling model retrained on DistantSRL data to SEMAFOR/K role labeling model from [Kshirsagar et al. \(2015\)](#); P, R, and F₁ scores in percent; **boldface** marks best results per section.

To use our large and noisy training data effectively, we may need to integrate additional methods to the supervised system that support learning from noisily labeled data, e.g., [Natarajan et al. \(2013\)](#), or specifically target domain adaptation, e.g., [Chen et al. \(2011\)](#).

In the next section we discuss the potential contribution DistantSRL could add to state-of-the-art FrameNet semantic role labeling.

3.6.2 Discussion in Relation to State-of-the-art FrameNet SRL

This section discusses the potential benefits of DistantSRL to state-of-the-art FrameNet SRL. We first consider the step of frame identification, then the step of role labeling.

Frame identification. [Hermann et al. \(2014\)](#) report state-of-the-art results for FrameNet frame disambiguation. Their approach is based on distributed representations of frame instances and their arguments, i.e., embeddings, and performs frame disambiguation by mapping a new instance to the embedding space and assigning the closest frame label. This process is conditioned on the lemma for seen predicates. They report that they improve frame identification accuracy over SEMAFOR by 4% for ambiguous instances in the FNFT-test set, up to 73.39% accuracy. They also improve over the SEMAFOR system for full semantic role labeling, reporting an F₁ of 68.69% compared to 64.54% from [Das et al. \(2014\)](#).

Our results are not directly comparable to their results. We also evaluate on ambiguous instances, but we only evaluate on verb targets, which are typically more polysemous than nouns and adjectives and thus are more difficult candidates for sense disambiguation.

FrameNet semantic role labeling. For a long time, the SEMAFOR system has been the state-of-the-art FrameNet semantic role labeling system (Das et al., 2010, 2014), it is still one of the best open-source semantic role labeling systems.

Recently, several approaches were introduced that use new ways of generating training features, and neural-network based representation learning strategies. We already introduced Kshirsagar et al. (2015) and Hermann et al. (2014).

Roth and Lapata (2015) also present new strategies to enhance FrameNet role labeling, together with a new open-source FrameNet semantic role labeling system. They do not use additional lexical resources, but integrate new context-based features into their open-source semantic role labeling system Framat++, a FrameNet-adaptation of the PropBank semantic role labeling system Mateplus (Roth and Woodsend, 2014). Using SEMAFOR for frame identification, they improve results for full FrameNet semantic role labeling compared to SEMAFOR (Das et al., 2014), reporting an F_1 of 67.8%. Their new features belong to three classes: a) discourse-level features making use of coreference information for co-occurring roles and discourse newness, b) document-specific word representations, and c) global frame-based reranking. In general, this approach could benefit from using additional training data: discourse-level features and document-specific word representations could also be computed for additional training data. The reranker could also be trained on large amounts of data, if training scales appropriately.

Täckström et al. (2015) present a new dynamic program formalization for FrameNet role labeling that efficiently incorporates global constraints associated with the role labeling task, presenting a tailored solution for incorporating global constraints associated with the role labeling task that replaces the standard integer linear program solutions used in earlier work. They report state-of-the-art results in FrameNet semantic role labeling, i.e., $F_1=70.3%$, when combined with the frame identification model from Hermann et al. (2014).

The currently best-performing FrameNet semantic role labeling system is the one presented by FitzGerald et al. (2015) which builds upon Täckström et al. (2015). They develop a multitask learning setup for semantic role labeling which they evaluate for PropBank and FrameNet semantic role labeling. The setup is based on a specifically designed neural network model that embeds input and output data in a shared, dense vector space. Combined with the frame identification model from Hermann et al. (2014), their model significantly improves on the previous state-of-the-art for full FrameNet semantic role labeling, reaching F_1 of 70.9% on FNFT-test - but only when training the model jointly on FrameNet training data and PropBank-labeled data, e.g., CoNLL 2005 training data, in a multitask setup.

FitzGerald et al. (2015) report that the performance of their system on FrameNet test data suffers from the small training set available – only training on FrameNet training data yields similar results to Täckström et al. (2015). The joint training setup does not benefit PropBank semantic role labeling due to the small size of the FrameNet training set in comparison to the PropBank data. This shows that additional training data for FrameNet, for instance our automatically labeled corpora, could also benefit a state-of-the-art system. An explicit evaluation of this assumption or comparison to this system is not possible to date, because the system by FitzGerald et al. (2015) is not publicly available.

The state-of-the-art systems are optimized towards the manually labeled FrameNet full-text training set. We already mentioned above that domain adaptation methods, or methods to learn from noisy data (Chen et al., 2011; Natarajan et al., 2013) may be required to adapt conventional systems to the large and noisily labeled corpora provided by DistantSRL.

In summary, based on the discussion above, and the frame and role classification experiments that we evaluate on four test sets, the data we generate with our method are complementary to the standard FrameNet training data, and we conclude that they can be used to enhance state-of-the-art semantic role labeling systems.

3.7 Summary of Chapter 3

This chapter presented DistantSRL, a method of utilizing the LLKB UBY_{FN} created in the previous chapter for SRL via knowledge-based transfer of sense and role labels. It is motivated by the need of training data for FrameNet SRL for languages that lack the frame- and role-annotated corpora required for supervised training. The motivation also persists for the English language: state-of-the-art SRL systems for English still rely on supervised training, even when advanced methods such as deep learning are used.

We introduced DistantSRL in the context of other approaches for training data generation and other knowledge-based approaches to SRL. We then applied the method to large corpora, and evaluated it extrinsically for the tasks of frame identification and role classification, and also for full frame-semantic parsing using the SEMAFOR system. Since DistantSRL is focused on verbs, all experiments are evaluated on verbal predicates. The results show that the automatically labeled data created by DistantSRL are of high quality and complement the FrameNet fulltext corpus. For frame identification, our data lead to similar precision as a standard supervised setup, but at higher coverage. Learning curves indicate that with an even larger corpus we may be able to further improve precision. For role classification, the sparse labeling leads to a low role recall, but high precision is achieved for the covered roles. One cause for the sparse labeling is the incomplete mapping between VerbNet and FrameNet roles in SemLink. This could be remedied by extending the SemLink mapping, which would increase the coverage, and to disambiguate ambiguous role labels using selectional preferences to further increase precision.

Retraining SEMAFOR on the automatically labeled WaS corpora led to mixed results: we find slightly lower performance for frame identification for all test sets but the MASC set. The evaluation of frame identification and role labeling with the retrained frame identification model on the polysemous MASC test set shows that the automatically labeled corpora can contribute to an improved semantic role labeling performance. Best results are obtained when training on the union of the automatically labeled corpora and the FrameNet fulltext corpus. When retraining the SEMAFOR role labeling model, we see that the DistantSRL data are of high quality, leading to high role labeling precision.

To show that it generalizes to other languages, we applied DistantSRL to German. Thus, we are the first to apply knowledge-based verb sense labeling to the FrameNet verb sense inventory and to languages other than English. The results of the German experiments show that the quality of the linked lexical knowledge bases influences the outcome of the approach. For German, we expect larger coverage and precision of our approach by embedding SALSA in a larger linked lexical knowledge base, for instance including a linking to GermaNet. In summary, the main contributions of this section are:

- The extension of knowledge-based verb sense labeling and its adaptation to the FrameNet sense inventory and to the German language using the SALSA sense inventory.
- The application and evaluation of DistantSRL, a knowledge-based method for the automatic generation of role-labeled training data, on various test sets from different domains for English and German.
- The developed corpora: large sense- and role labeled corpora for English and German. We publish the corpora for research purposes, see Appendix A.

There are several directions for an extension of the presented approach: DistantSRL was developed for verbs and the English FrameNet, but can be extended to other parts-of-speech and languages. It is particularly well-suited for languages and domains for which role-labeled corpora are lacking, but linked lexical knowledge bases are available or can be created automatically.

Furthermore, DistantSRL can be adapted to other sense and role inventories covered by linked lexical knowledge bases, e.g., VerbNet and PropBank, and to related approaches to semantic parsing, e.g., QA-SRL (He et al., 2015). This would require a mapping of the role inventory to a suitable linked lexical knowledge base, for instance a mapping of the role labels in QA-SRL to SemLink. As a variation of our approach, Stage 1 of DistantSRL could be combined with other approaches for training data generation that would benefit from receiving sense-labeled data as input, e.g., monolingual annotation projection.

In the next chapter, we will compare DistantSRL to other methods of training data generation introduced in Section 3.3 and evaluate its contribution to the domain adaptation of FrameNet semantic role labeling.

CHAPTER 4

Domain Adaptation via Training Data Generation

This chapter presents the application of the knowledge-based methods for training data generation of frame- and role-labeled texts introduced in the previous chapter to domain adaptation of FrameNet semantic role labeling, more specifically to the adaptation of semantic role labeling to user-generated texts from various domains.

We study the following questions: can the knowledge-based methods improve semantic role labeling for texts from various domains, including user-generated text? And how do these methods compare to other methods for training data generation?

Before answering these questions, we first analyze the state-of-affairs: we assess the domain generalization capabilities of the open-source FrameNet SRL system SEMAFOR.

This study is motivated by a lack of research on domain adaptation for FrameNet semantic role labeling. Domain adaptation has been identified as a problem for semantic role labeling, but most related work is focused on PropBank. There have been only few evaluations of domain adaptation for FrameNet semantic role labeling, and these are now outdated. Recent work on FrameNet semantic role labeling typically only evaluates on a single in-domain test set. There is a lack of suitable out-of-domain test sets.

To enable studies on domain generalization for FrameNet SRL, we create a new, substantially sized frame- and role-labeled test set based on user-generated text and compare it to other available test sets, including those used in Chapter 3. We evaluate the open-source semantic role labeling system SEMAFOR on the various test sets to assess the need for domain adaptation for the different stages of FrameNet semantic role labeling. We find that the major bottleneck is the frame identification step: SEMAFOR frame identification performance is lower on most of the out-of-domain test sets. Once the frame labels are assigned correctly, the role labeling performance on out-of-domain test sets is similar to the performance on the in-domain data, or exceeds it.

We then evaluate the benefits of the knowledge-based approach to training data generation introduced in Chapter 3 to domain adaptation for FrameNet semantic role labeling and compare it to alternative approaches of training data generation. We find that the combination of the FrameNet fulltext corpus data with our automatically labeled WaS-L corpora leads to improved frame identification for the test sets based on user-generated text and MASC. This is mostly due the coverage increase from adding WaS-L, that provides better coverage for domain-relevant frames. In a detailed error analysis, we identify factors that lead to the lower frame identification performance on out-of-domain test set: it is a result of lower training data coverage, domain adaptation effects, and difficulties in preprocessing user-generated text. Finally, we present a detailed discussion of how to make better use of the automatically labeled corpora for frame identification, and how to further improve frame identification performance across domains.

In the next section, we motivate our study on the domain generalization capabilities of FrameNet SRL and discuss related work.

4.1 Motivation: Domain Adaptation for SRL

This section motivates our study on domain adaptation for FrameNet semantic role labeling. It first briefly introduces the concept of domain adaptation, and then describes previous work in domain adaptation for semantic role labeling in general and specifically for FrameNet SRL. Notably, previous work mostly focused on PropBank SRL. There is a lack of research on domain adaptation for FrameNet semantic role labeling.

4.1.1 Domain Adaptation

Supervised machine learning systems perform worse on data with different underlying feature distributions, i.e., different domains. The term *domain adaptation* describes the adaptation of a model to test data from a different feature distribution than seen in the training data (Søgaard, 2013).

Domain adaptation has been evaluated in detail for various NLP tasks, from part-of-speech tagging to dependency parsing. Studied domains range from different types of newspaper texts, fiction texts, web texts, informal language and speech, to texts from specific disciplines, such as Law or the Biomedical domain. A sub-type of domain adaptation relevant to this work is the adaptation to user-generated discourse (UGD), texts generated by web users in the form of blogs, web-forum posts (Biber and Conrad, 2009), or Twitter posts (Han and Baldwin, 2011). These types of text follow formal restrictions like orthography to a much lesser degree than edited text from various domains (e.g., newswire); they contain new terms and cover various topics.

In order to avoid the expensive labeling of large amounts of target domain data, methods for domain adaptation of supervised machine learning systems have been developed that aim to create systems that generalize to a specific target domain, or to variable domains (Søgaard, 2013). The datasets from other domains are either referred to as *target domain* or *out-of-domain* data; The latter assumes the perspective of the *source domain* data, the data that a conventional machine learning system is trained on (*in-domain* data),

Domain adaptation methods range from *supervised domain adaptation*, that combines labeled source domain and target domain training data, via *semi-supervised domain adaptation* (labeled source-domain data, labeled and unlabeled target domain data) and *unsupervised domain adaptation* (labeled source domain, unlabeled target domain data) to *blind domain adaptation* that attempts to create domain-independent systems – also called open-domain systems. Because of its objective to use unlabeled target-domain data and rely on fewer amounts of labeled data, domain adaptation in natural language processing often uses semi-supervised machine learning methods (Chapelle et al., 2010).

A popular method of supervised domain adaptation in NLP is the feature augmentation approach by Daumé III (2007). They use additional labeled data from the target domain to augment the feature space, encoding for each feature whether it represents the source, target, or both domains. Self-training or bootstrapping approaches are examples of semi-supervised domain adaptation, see Zhu and Goldberg (2009); these approaches use unlabeled data and small amounts of labeled data from the target domain.

As labeled target-domain data are expensive to obtain, methods of unsupervised domain adaptation that leverage unlabeled target domain data have been developed. One way is to re-weight the training data based on the distribution in the target domain data, for instance using instance weighting, another is to automatically generate target-domain training data. An increasingly successful method for unsupervised domain adaptation uses learning of low-dimensional representations (Blitzer et al., 2006).

The approach to domain adaptation explored in this chapter can be considered a variant of unsupervised domain adaptation: we use unlabeled data from the target domain (or additional domains) that is labeled automatically using various methods of training data generation, including knowledge-based label transfer as described in Chapter 3. Then a supervised system is trained on the automatically labeled data, or on a combination of these data with the source-domain training data.

4.1.2 Domain Adaptation for Semantic Role Labeling

Since Pradhan et al. (2007b) described the domain adaptation problem for semantic role labeling, domain adaptation for semantic role labeling has been mostly studied for PropBank-style semantic role labeling. The CoNLL shared tasks on semantic role labeling provide the appropriate test bed: typically newspaper text from the Wall Street Journal Corpus is con-

trasted to the target domain of fiction texts from the Brown Corpus (Carreras and Màrquez, 2005; Surdeanu et al., 2008; Hajič et al., 2009).

The CoNLL 2005 shared task on semantic role labeling (Carreras and Màrquez, 2005) is the first to include an out-of-domain test set from the Brown corpus. The results showcase the domain adaptation problem for semantic role labeling: the best-performing semantic role labeling system (Koomen et al., 2005) shows an F_1 score of 79.44% on the in-domain test set, and F_1 of 67.75% on the out-of-domain test set, a difference of more than 10 points. Carreras and Màrquez (2005) attribute this difference mostly to domain-specific preprocessing (separate preprocessing steps also perform worse on the out-of-domain dataset) and error propagation in the pipelined systems, but also mention overfitting and domain-specific features of the semantic role labeling classifiers as a potential source of error.

The CoNLL 2008 shared task is focused on the joint parsing of syntactic and semantic dependencies (Surdeanu et al., 2008). Results for semantic role labeling are reported as *labeled F_1 score*, results for syntactic parsing are reported as the *labeled attachment score (LAS)*, and for the joint syntactic and semantic parsing, they are reported as the *labeled macro F_1* .

The observations regarding the out-of-domain test set are the same as for the 2005 shared task: Surdeanu et al. (2008) report that the divergence of performance between the test sets lies between 12 and 14 points F_1 for the semantic role labeling task, while the divergence is lower (7-8 LAS points) for the dependency labeling task, indicating that “domain adaptation becomes even harder as the task to be solved gets more complex” (Surdeanu et al., 2008). The contributors of the best system, Johansson and Nugues (2008a), report a labeled macro F_1 of 85.95% on the in-domain test set, and of 75.95% on the out-of-domain test set for joint syntactic parsing and semantic role labeling. For the semantic role labeling evaluation, they report a labeled F_1 of 81.75% in-domain and 69.06% out-of-domain.

This trend continues with the CoNLL 2009 shared task (Hajič et al., 2009) that adds evaluation languages other than English: even though the overall scores get much better, the difference between in-domain and out-of-domain evaluation stays large. Most of the participants in the CoNLL shared tasks do not focus specifically on the domain adaptation task. This includes the best-performing system by Zhao et al. (2009).

Research that puts a greater focus on the domain adaptation scenario often makes use of some kind of distributional representation that is supposed to help generalization from seen training instances to unseen instances. The work described in the next paragraph recently improved on the CoNLL shared task results for domain adaptation.

PropBank semantic role labeling. The work by Huang and Yates (2010) is an early example of representation learning for domain adaptation in semantic role labeling. They aim to create an open-domain semantic role labeling system. They use only labeled data from the source domain and integrate a latent variable language model based on a Hidden Markov Model into their semantic role labeling system. The latter is trained on the

unlabeled source domain data. They evaluate their chunking-based semantic role labeling system on the CoNLL 2005 shared task dataset. [FitzGerald et al. \(2015\)](#) later improve the CoNLL results for Propbank semantic role labeling on out-of-domain sets with their neural-network based system.

[Yang et al. \(2015b\)](#) use deep belief networks for domain adaptation of PropBank semantic role labeling. They report the best results to date on the out-of-domain test set for the CoNLL 2009 shared task and the smallest performance difference between in-domain and out-of-domain training data, showing that their approach works well for domain generalization. They use the deep belief network to create a latent feature representation (called LFR) that reduces the number of features from over one million for role labeling in two steps to first 10,000 and then 5,000 features. The training of the deep belief network for role labeling uses unlabeled target domain data together with labeled source domain data. Thus, their approach is an instance of unsupervised domain adaptation.

The work by [Yang et al. \(2015b\)](#) is particularly interesting, because they compare their results ($F_1=78.75\%$ for their best system on out-of-domain data) to a number of baseline approaches. First, they compare to the system of the CoNLL 2009 shared task that performed best on the out-of-domain test set ([Zhao et al., 2009](#)), which achieves an F_1 of 74.58%. This system does not perform domain adaptation, i.e., they only use labeled in-domain data for training, but it makes use of word clusters to improve generalization capabilities. Then they compare to the baseline of training on the in-domain training data (BL1), training on out-of-domain training data that were labeled with the BL1 system, which leads to a self-training setup (BL2), and training on the combined training data from BL1 and BL2 (BL3). In addition, [Yang et al. \(2015b\)](#) compare their method to the feature augmentation approach by [Daumé III \(2007\)](#) (FA). Feature augmentation, being a supervised approach to domain adaptation, requires labeled data from the target domain. For this, [Yang et al. \(2015b\)](#) use the data created by the self-training baseline BL2. As a result, the feature augmentation setup FA is very similar to BL3 and show similar performance: they achieve an F_1 of 72.90% and 72.75% respectively.

[Yang et al. \(2015b\)](#) find that all baselines except for BL2 improve on the supervised training BL1 for the out-of-domain semantic role labeling task ($F_1=71.57\%$). The self-training baseline BL2 performs slightly worse on out-of-domain data compared to BL1 ($F_1=70.34\%$), but also additionally more than 10 percentage points F_1 worse on the in-domain data. The observation that self-training decreases semantic role labeling performance was also made for FrameNet SRL, see our discussion of self-training for FrameNet SRL in Section 3.3.2.

Another interesting finding is that BL3, simply combining gold training data with data from self-training, performs similar to the more elaborate feature augmentation approach. Even though the feature augmentation might result in larger improvements when using gold standard target domain data instead of those annotated via self-training, this observa-

tion is still encouraging with respect to our own experiments on combining gold training data with automatically labeled data which we will introduce in Section 4.4.

Besides work on the CoNLL shared tasks, there have been applications of domain adaptation to specific domains like the Biomedical domain (Tsai et al., 2006; Dahlmeier and Ng, 2010), or application to speech data (Van der Plas et al., 2009). We discuss work on adapting semantic role labeling to user-generated text in a separate paragraph below.

FrameNet semantic role labeling. The problem of domain adaptation for FrameNet semantic role labeling presents itself differently from the one for PropBank SRL, because of the two steps involved in FrameNet semantic role labeling: frame identification and role labeling. Different sets of roles are licensed for FrameNet role labeling depending on the frame label assigned during frame identification. Since PropBank role labels do not depend on the predicate label, the majority of PropBank SRL systems only deal with the task of role labeling. Domain adaptation for FrameNet semantic role labeling may in contrast be required for both steps in the semantic role labeling setup.

Domain adaptation for FrameNet semantic role labeling has so far only been evaluated sparsely. Note that we do not focus on the extension of the FrameNet frame inventory to other domains as done by Schmidt (2009) for soccer, Venturi et al. (2009) for the legal domain, and Reiter (2014) for rituals in digital humanities. We consider the domain adaptation problem that is common to many supervised machine learning tasks, i.e., the adaptation of a model to test data from a different distribution than the training data while preserving the original task.

Johansson and Nugues (2008b) evaluated the impact of different parsers on FrameNet semantic role labeling, using the Nuclear Threats Initiative (NTI) data as an out-of-domain test set. They observe low domain generalization abilities of their supervised system, but find that using dependency parsers instead of constituency parsers benefits domain adaptation. Their explanation is that the dependency-based system relies less on lexicalization than the system based on constituency parsers.

Croce et al. (2010) aim to create an open-domain FrameNet semantic role labeling system by integrating a distributional model into their semantic role labeling system. The distributional model generalizes lexicalized features for argument classification to previously unseen arguments and thus contributes to a system with similar performance on source and target domain (different of 4.5 percentage points F_1), using a similar in-domain/out-of-domain split as Johansson and Nugues (2008b). Distributional representations for generalization to unseen data have also been adopted by state-of-the-art systems, including advanced word representations obtained via deep learning (FitzGerald et al., 2015; Hermann et al., 2014). These systems, however, have not been evaluated on out-of-domain test data. Currently, FrameNet semantic role labeling benchmarks use a split of the FrameNet 1.5 fulltext corpus, that is randomly sampled and contains texts from the same sources in

training, development, and test split (Das et al., 2014). Out-of-domain evaluation is lacking, as are appropriate datasets that enable this kind of evaluation.

Recently, Kshirsagar et al. (2015) proposed to use domain adaptation methods to improve FrameNet role labeling. They use the feature augmentation approach by Daumé III (2007) to extend the feature space for FrameNet role labeling with additional data from the FrameNet lexicon examples. They, too, evaluate performance gains only on the standard in-domain FrameNet test set used by Das et al. (2014).

4.1.3 Semantic Role Labeling for User-generated Discourse

User-generated discourse is a specific text type that also requires domain adaptation strategies. The properties of UGD diverge from standard text or the newswire text ubiquitous in NLP. Popular types of user-generated discourse are tweets from Twitter, community questions and answers, or forum posts. These types of text have specific properties like lack of orthographical conventions, i.e., spelling or punctuation, and the use of colloquial language that make them appear closer to spoken language, as well as media-specific strategies, i.e., hashtags in Twitter. These divergences from standard language do have linguistic or meta-linguistic functions. Because of this, user-generated media have since been recognized as a source of various domains of their own right, for instance made up by different socio-economic strata (Eisenstein, 2013).

NLP on user-generated text makes use of normalization strategies and strategies to adapt NLP analysis tools such as POS-tagging, parsing, or named-entity recognition to user-generated text, mostly for Twitter (Han and Baldwin, 2011; Baldwin et al., 2015). There is not much previous work on SRL for user-generated text.

Initial work in semantic role labeling for user-generated texts also targets PropBank roles (Liu et al., 2010). Liu et al. (2010) present a PropBank semantic role labeling system for Twitter data in the news domain. They first define in-domain data as excerpts from edited texts that can be labeled with a standard semantic role labeling system. Then, they collect out-of-domain data as Twitter tweets on news topics with typical properties such as informal language, emoticons, etc. They project annotations from the news excerpts to the tweets and train a semantic role labeling system on the expanded training data. Liu et al. (2011) create a two-step approach for semantic role labeling on news tweets that is highly tailored to their task and exploits similarities in clusters of tweets to correct the role labels from an initial system. Liu et al. (2012) later extend this approach to noun predicates.

There are two recent papers that introduce FrameNet-labeled datasets based on user-generated texts from Twitter and Wikipedia. Søgaard et al. (2015) annotate Twitter data with FrameNet frames and roles to evaluate the benefit of syntactic and shallow semantic parsing to knowledge extraction from Twitter. They publish three versions of their dataset, each annotated by a different annotator. These datasets were not adjudicated to a single gold standard based on the argumentation that any human annotation provides reasonable

information to a machine learning system. With just over 1,000 frames and around 1,700 roles, those datasets are small compared to other test sets, for instance the FrameNet full-text test set containing more than 4,000 frames and 7,000 roles. Agreement between the annotators is high at 84.5% F_1 for frame identification and 78.1% F_1 for role labeling.

Søgaard et al. (2015) report that the frame identification performance of SEMAFOR 2.1 (Das et al., 2010) on the new test set is similar to its performance on the newswire test set from SemEval-2007 (Baker et al., 2007). For full SRL, there are large differences: F_1 reaches only 25.96% on the Twitter dataset compared to the 46.5% reported by Das et al. (2010) on the newswire set. These results show that there is ample room for improvement for SRL on the Twitter dataset, even though Søgaard et al. (2015) report that FrameNet SRL benefits their knowledge extraction task. They find that using FrameNet semantic role labeling with SEMAFOR 2.1 leads to more robust knowledge extraction, i.e., a larger percentage of extracted facts that are considered intelligible and relevant by human judges, compared to other types of preprocessing such as dependency parsing or PropBank SRL.

Johannsen et al. (2015) create labeled datasets based on texts from Twitter and Wikipedia for several languages, including English. They use these datasets to evaluate their multilingual semantic role labeling system. In order to create datasets in several languages, they use a specific preprocessing setup: frame candidates were selected based on a matching of FrameNet to the multilingual BabelNet, which was used to translate FrameNet predicates to other languages. This matching to BabelNet was also used for English in order to emulate the multilingual setting. As a result, the quality of the English frame annotations is lower than would be expected for English. The lower quality is mirrored in agreement scores of 73.4% F_1 for frame identification and 70.5% F_1 for role labeling. These scores are lower than those reported by Søgaard et al. (2015). Additional details and statistics on the two FrameNet-labeled datasets are presented in the next section.

In this section we discussed previous work on domain adaptation for SRL, with a particular focus on user-generated text. There is not much research on domain adaptation for FrameNet SRL, which may be caused by a lack of appropriate test datasets. There are only two out-of-domain test sets for FrameNet SRL that target user-generated text. One of them is small (Søgaard et al., 2015), the other of lower quality (Johannsen et al., 2015). The lack of out-of-domain test sets for FrameNet semantic role labeling motivates the creation of a new test dataset that is presented in the next section.

4.2 YAGS – a Gold Standard for User-generated Text

In the previous chapter, Chapter 3, we introduced a number of available FrameNet-labeled datasets that can be used for the training and evaluation of FrameNet semantic role labeling systems. These datasets have in common that they stem from professionally edited sources, mostly newspaper text, but also literary texts.

The screenshot shows the Yahoo! Answers interface. On the left is a navigation menu with categories like 'All Categories', 'Beauty & Style', 'Business & Finance', 'Cars & Transportation', etc. The main content area features a search bar at the top with 'Search Answers' and 'Search Web' buttons. Below the search bar, a breadcrumb trail reads 'Cars & Transportation > Car Makes > Chrysler'. A red dashed box highlights the question: 'Who now own chrysler i know daimler sold them but i want to know who is the company that bought them.?' with a 'Question' label above it. Below the question, there are two answers. The first is marked as the 'Best Answer' by Cerberus, with the text 'The Oft-Suspended RP McMurphy · 8 years ago' and 0 thumbs up/down. The second answer is by Littleball, with the text 'Daimler sold 80% of Chrysler to Cerberus Capital Management. Daimler still owns 20% of Chrysler.' and a source cited as Wikipedia. It has 1 thumbs up and 0 thumbs down. A 'Relevance' dropdown menu is visible on the right side of the answers section.

Figure 4.1: Example question from Yahoo! Answers.

As discussed in the previous section, there are only few test sets that cover the domain of user-generated text. They are based on Twitter data, and either small (Søgaard et al., 2015) or of low quality (Johannsen et al., 2015). There are no datasets that cover other user-generated text types such as blogs or forum posts.

To fill this gap, we created YAGS, a new, manually labeled gold standard dataset based on data from Yahoo! Answers, a question-and-answer website on which users ask questions and contribute answers to questions on various topics, including informational questions (see Figure 4.1) and conversational questions that are very similar to informal chats. Yahoo! Answers covers various topics from *Amusement Parks* to *Zoology*, and contains the style of user-generated language typical for spontaneous user contributions. The example in Figure 4.1 illustrates the non-standard language of the user-generated question-and-answer data:¹⁹ careless spelling (e.g., *own* instead of *owns*), omitting capitalization (e.g., *i*, *chrysler* instead of *I*, *Chrysler*) and careless use of punctuation conventions (e.g., missing question mark after *Who now own chrysler*).

In this section, we first present related work on annotation of FrameNet frames and roles and motivate the strategy we use in our annotation study, e.g., manual annotation by trained annotators. We then present the annotation study in detail and compare the resulting dataset to the Twitter datasets introduced above and the edited test sets used in our experiments in Chapter 3.

¹⁹Accessed at October 16 2015 from <https://answers.yahoo.com/question/index?qid=20080212084227AAxdpl5>.

4.2.1 FrameNet Frame and Role Annotation

For the creation of our gold standard dataset, the annotation method of choice was a classical one: given a text with automatically pre-annotated predicate targets, trained annotators, graduate-level students in linguistics, first identified the frame of the predicate, and then identified the argument spans and role labels. Each predicate instance was annotated by two annotators. Subsequently, an experienced annotator adjudicated their annotations.

Given the goal of collecting an evaluation dataset of considerable, but constrained, size with large coverage, this promised to be the most effective method. Previous annotation studies explored the option of crowdsourcing FrameNet annotations, and their experiences shed light on the difficulties associated with this strategy: for frame annotation, which is basically a word sense disambiguation task, crowdsourcing works quite well: [Hong and Baker \(2011\)](#) evaluate the feasibility of crowdsourcing frame annotations for a small number of FrameNet predicates with up to five candidate frames and test different setups. Their most successful annotation setup is the one using example sentences for each predicate sense instead of frame definitions, to which unlabeled sentences have to be assigned by the annotator. Recently, [Chang et al. \(2015\)](#) expanded their work to large-scale frame annotation and added a supervision loop that provides feedback to the crowd-annotators on gold instances which speeds up annotator training in the crowd-annotation setup and results in high-quality frame annotations.

[Fossati et al. \(2013\)](#) considered crowdsourcing for the role annotation task. For this task, crowdsourcing appears to be more difficult. It covers the detection of syntactic arguments, adjuncts and modifiers of the predicate and their labeling with a semantic role label. The difficulties are in part caused by the abstract definitions of the roles that were created by and for experts in linguistics. Therefore, [Fossati et al. \(2013\)](#) replaced the role definitions by semantic types from DBpedia associated with the role fillers. To gather these types, they use an automatic system to map FrameNet role fillers to Wikipedia pages, then they use given links from Wikipedia to DBpedia to map the pages to the DBpedia type system. This way, the `VICTIM` role of the `KILLING` frame is for instance associated with the type *Animal*. They report that presenting a list of semantic types to the annotators instead of role definitions leads to higher annotation accuracy. The reason is that these are easier to understand for lay annotators compared to the abstract labels and often complicated role definitions. Because the coverage of the linking between FrameNet role fillers and DBpedia is not complete, and because coverage and granularity of the DBpedia type system may not exactly match the one in FrameNet, this strategy is bound to omit or conflate a number of roles.

[Feizabadi and Padó \(2014\)](#) evaluate the annotation of non-local semantic roles. Non-local roles are not realized in the local sentence context, and maybe not realized at all in the discourse context – so called *implicit* roles. The phenomenon is illustrated in the following example from [Feizabadi and Padó \(2014\)](#): “*Phileas Fogg, having shut the door of*

[his house]_{Source} 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 [the Reform Club]_{Goal}”. The phrase *his house* labeled with the *Source* role is not an argument of the predicate *reached*. This is an example for the non-local instantiation of the *Source* role. Feizabadi and Padó (2014) focus on a specific subset of roles and use simplified descriptions of those roles for an annotation study focused on *Motion* and *Position* frames.

The crowdsourced approaches to role annotation have in common that they simplify and reduce the number of available roles. In order to create a comprehensive annotation and evaluate which FrameNet roles are relevant in the target domain and discourse type, our annotation study was performed in the traditional setup using trained annotators. We now introduce our annotation study, starting with the selection of a dataset to annotate.

4.2.2 Data Selection and Preparation

In order to create a new FrameNet-labeled test set on user-generated data we make use of the Yahoo! Answers Manners dataset. Our dataset consists of 55 questions and their answers randomly sampled from the 28,528 questions in the test split of the Yahoo! Answers Manners dataset used by Surdeanu et al. (2011).²⁰ The 55 questions and answers were automatically segmented, lemmatized and POS-tagged using Stanford Segmenter and TreeTagger as provided by DKPro-core (Eckart de Castilho and Gurevych, 2014). Words were automatically marked as predicate candidates for annotation if their lemma and POS are present in the FrameNet lexicon. Depending on the sentence length, up to five predicate candidates per sentences were marked. This was done for practical reasons: the constraint was motivated by the annotation interface used for the Frame and Role annotation. With a certain number of predicates per sentence, the annotation window would get too crowded. For sentences consisting of up to 15 tokens, only three predicate candidates were marked. The filter that constrained the number of candidates preferred verbal predicate candidates over adjectives and nouns, selecting verbal predicates first.

To enhance readability for the annotators, the prefix *Q:* was added to mark the beginning of questions, and the prefix *QA:* was added to mark the beginning of each answer. These markers were later removed from the final gold standard.

4.2.3 Annotation Task

The annotation task consisted of three steps, which were performed in direct succession for each predicate candidate: first, the identification of multiword predicates, second the frame labeling on the given predicate candidate, and third, if a frame could be assigned to

²⁰The Yahoo! Answers Manners dataset can be obtained via <https://webscope.sandbox.yahoo.com>. Information about the test split was obtained via personal communication with the authors of the paper.

- 1) If [you]_{Grinder} have a [mortal and pestle]_{Grinding_instrument}, **grind**_{Grinding.head} **up**_{Grinding.satellite}
 [all the ingredients]_{Undergoer} [in the order above] _{Manner} [with it]_{Instrument}.
- 2) Is there [something that [you] _{Agent} have been wanting to do] _{Event} for a long time but have been
putting_{Change_event_time.head} [it] _{Event} **off**_{Change_event_time.satellite}?

Figure 4.2: Example: FrameNet frame and role annotation in YAGS.

the candidate, the identification of the arguments and role assignment. Detailed annotation guidelines are listed in Appendix B.

Step 1: Multiword predicates. The first step entailed that for each given predicate candidate in context, the annotator was to decide whether the predicate candidate is part of a multiword construction, for instance a verb particle construction or a support verb construction. The additional parts of the multiword were annotated and each part of the multiword was marked as *head* or *satellite* of the multiword. Satellites were then linked to their heads. The example in Figure 4.2 shows two multiword predicates from YAGS, *grind up* and *putting off*.

Step 2: Frame annotation. In the second step, predicate candidates are assigned a single frame label, or the label *NF* (*Not in FrameNet*) if the given sense of the word is not listed in the FrameNet lexicon. If the meaning of the candidate could not be interpreted from the context, the annotators were asked to select the label *XX*. In the above example, the predicate *grind* receives the frame label *Grinding*, and the predicate *putting* receives the frame label *Change_event_time*.

Step 3: Argument span and role annotation. This step is only performed for predicate candidates that receive a frame label. For these instances, the argument spans are identified as syntactic arguments, adjuncts or any other kind of modifier (mostly noun phrases, prepositional phrases, sometimes subordinate clauses) and assigned a role label from the set of roles of the given frame. If semantic arguments can be identified in the sentence context without being direct arguments, these are also annotated, for instance *mortal and pestle* in the example. If no appropriate role label is available, the annotators are asked to select *NF* as the label. In the above example, *you* is identified as an argument and labeled with the *Grinder* role of the *Grinding* frame, and the span *mortal and pestle* is labeled with the *Instrument* role.

Annotators were also encouraged to annotate anaphoric referents of arguments in certain circumstances. For instance, if the argument is a pronoun, but the phrase the pronoun

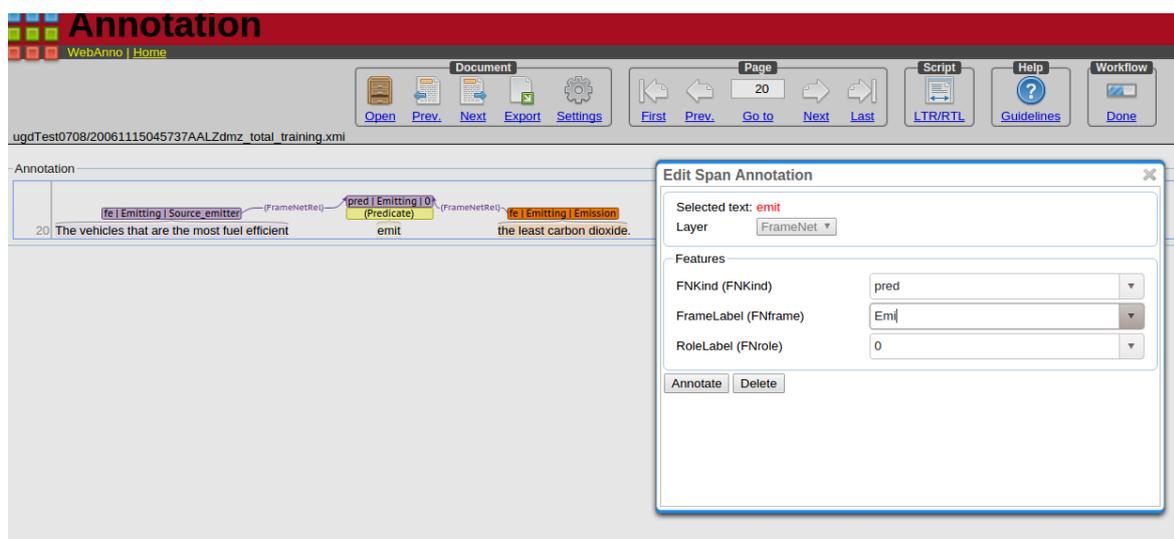


Figure 4.3: Screenshot of WebAnno annotation interface used for frame and role annotations.

refers to can be clearly identified in the context, the phrase is also annotated as the argument of the candidate predicate. The motivation for this is that we would like to collect semantic information on the fillers of arguments in addition to their mere lexical representations.

4.2.4 Annotation Study

Five annotators, all of them linguistics students experienced in semantic annotation tasks, were equipped with the annotation guidelines (see Appendix B) and trained on a training document. All documents were annotated by pairs of those five annotators and adjudicated by an expert annotator. A small subset of the available documents was annotated by all annotators to compute the overall agreement.

Annotation tool. The tool used for the annotation was WebAnno version 2.0.0 (Yimam et al., 2014). It supports the annotation of argument spans and linking arguments to their predicates with relation links that are visualized as arcs. Frame and role labels for the predicates and arguments can be selected from a drop-down list that is filtered upon entering the first letters of a label. The setup is illustrated in the screenshot in Figure 4.3.

Annotation scenario. Each of the 55 questions from Yahoo! Answers and their answers were presented to the annotators in a single document. The documents contain between 6 and 490 predicate candidates. Predicate candidates were pre-annotated and highlighted in the annotation interface.

As a reference to FrameNet 1.5 frame and role definitions, the annotators used the FrameNet Explorer tool²¹ with the data from FrameNet release 1.5. The annotators used the FrameNet Explorer tool to consult the frame and role definitions for the lemma of the predicate candidate in FrameNet.

The annotators followed the three annotation steps described above. First, they identify multiword predicates. Second, they label a predicate with a frame label that can be selected from a drop-down list. Then they identify the argument spans and label the arguments with their role label and the label of the corresponding frame. As part of this, they link arguments to their predicate with a relation link.

To make the annotation task easier for the annotators, the linking step is optional in most cases: explicit links between predicates and their arguments only had to be created if two predicates with the same frame label occurred in the same sentence. If this condition applies, the assignment of arguments to their predicate can not be performed automatically, as it is not uniquely defined. Otherwise, the matching of arguments to the predicates can be performed automatically based on the correspondence of their frame labels.

Curation scenario. The annotations were curated by an expert adjudicator. Therefore, the WebAnno annotation interface shows the annotations by the two annotators next to each other. Annotations for which the annotators agree are merged automatically and presented in a third window that shows the curated version. The adjudicator can accept the merged suggestions, or delete them, and add annotations from the annotators to the curated version by clicking on their annotation or by creating new annotations.

Postprocessing. To create a final gold standard, the curated instances were postprocessed in the following manner: first, the prefixes *Q:* and *QA:* were removed; second, implicit links between predicates and their arguments were made explicit and the WebAnno annotations converted to the DKPro type *SemanticPredicate* for the frame annotation and the type *SemanticArgument* for the role annotation. *SemanticPredicate* contains a list of *SemanticArgument* instances and thus explicitly links predicates to their arguments; third, a manual disambiguation step was added to distinguish between discontinuous argument spans and coreferent argument spans. It turned out that a number of discontinuous argument spans were annotated. These could not be distinguished automatically from the coreferent spans. In the final gold standard, discontinuous arguments should persist, while coreference targets should be marked as instances of coreference to a role target instead.

Therefore, the 400 frame instances that have several arguments labeled with the same role were manually analyzed. Out of the full set, the majority of 384 frame instances is annotated twice with the same role label, for 16 frame instances we observed triple annotations of the same role label. We categorized them into four classes based on whether they

²¹<http://www.clres.com/FNExplorer.html>

(1) display a discontinuous argument span, or (2) an instance of coreference, (3) whether the duplicate instances of the same role label are actually justified, or (4) whether they highlight an annotation error. For instances of (1) and (3), we kept the role annotation as is. Class (3) covers a small number of instances for which the assumed rule that each role only occurs once for a frame does not hold. These are typically non-core roles and describe several attributes of the predicate that are of the same type, e.g., temporal information.

For coreference instances of type (2) we mark the real argument and the coreferent. Thereby, 173 coreferents were identified and associated with their target, 169 targets in total. For these instances, the *SemanticArgument* label was removed. To preserve the coreference information, a DKPro *Coreference* annotation pointing from the coreferent to the *SemanticArgument* was added. For (4), we observed that the redundancy resulted from annotation errors and double annotation of the same label on the same argument span. These errors were subsequently fixed manually. In the end, around 170 role instances remain that have discontinuous argument spans or valid double annotations.

4.2.5 Inter-rater Agreement

To estimate the quality of the annotations, we measure the inter-rater agreement between pairs of annotators. Inter-rater agreement is evaluated on three levels,

- on the predicate level, evaluating the agreement of the frame labels assigned to the predicate heads,
- on the argument level, evaluating agreement of the argument spans independently of the labels, and
- on the role level, evaluating the agreement of the role labels given the argument span.

Inter-rater agreement for frame labels is Krippendorff’s $\alpha=0.76$; agreement for role labels given matching spans is $\alpha=0.62$, and Krippendorff’s α unitizing agreement for role spans is 0.7. This is a good result for such a difficult task on this type of text. Note that our gold standard contains a large number of verbs that are highly polysemous and includes both core and non-core role labels.

To compare to other annotation studies, we also computed F_1 agreement (Hripcsak and Rothschild, 2005). Average pairwise F_1 is 96.0% for frame labels, 65.0% for argument spans, and 54.0% for role labels given argument spans. Sogaard et al. (2015) report F_1 agreement for frame labels of 84.5%, Johannsen et al. (2015) report frame F_1 of 0.73% and role F_1 of 0.71% for their Twitter datasets, and role F_1 of 78.1% for the study by Sogaard et al. (2015). The higher role agreement for both studies may result from the different annotation setting: they ask their annotators to annotate the dependency head of the argument, not a full span. Additionally, annotators in Johannsen et al. (2015) mostly annotated core roles, which reduces potential confusion.

	all	valid labels	invalid labels	frame types	role types	frame-role types
frame labels	3,553	3,091	463	409	-	-
role labels	6,146	6,144	2	400	433	1,547

Table 4.1: Statistics on the YAGS annotation for frame and role labels. *Invalid labels* marks instances that are not represented in the FrameNet lexicon; Columns on *types* refer to unique label types as opposed to corpus instances.

x roles per instances	0	1	2	3	4	5
# predicate instances with x roles	95	764	1,500	563	80	10

Table 4.2: YAGS statistics on number of roles per predicate.

4.2.6 Gold Standard

The gold standard set YAGS contains 1,415 sentences annotated with 3,091 instances of frames and 6,144 instances of role-labeled arguments, including 1,130 instances of non-core roles. It covers 767 verb lemmas, 109 multiword lemmas, 409 unique frame labels, and 433 unique role labels. We publish the YAGS annotations in a stand-off format linking to the positions in the original dataset under an open-source license. Thus, researchers who license the original Yahoo! Answers Manner dataset will be able to easily add the role annotations from YAGS to their dataset. For details and download links see Appendix A.

Table 4.1 contains detailed statistics from the annotation study that include numbers for the invalid labels *NF* and *XX*. There are 463 predicate candidates that could not be labeled with a FrameNet frame and therefore received an invalid label. This shows that the FrameNet frame inventory does not cover all word senses for the predicates in YAGS. Only two of the 6,144 argument spans identified by the annotators could not be labeled with a role label. Table 4.1 also shows that there are 409 unique frames in YAGS. Thus, it covers 40% of the frame inventory in FrameNet. These occur in combination with 433 unique role labels, leading to 1,547 unique pairs of frame and role label, i.e., 1,547 different roles. YAGS contains between 0 and 5 role instances per frame instance with a median of 2 unique roles per frame instance, see Table 4.2.

Table 4.3 contains statistics on the final gold standard and contrasts them to the numbers of other test sets. We provide a detailed comparison below.

4.2.7 Comparison to Other Test Datasets

To better understand the properties of the new test set YAGS, we compare its statistics to those of other available test sets. Table 4.3 provides an overview on the number of sentences, frames for different parts-of-speech, roles and core roles in the different datasets.

dataset	sentences	frames	% adj frames	% noun frames	% verb frames	roles	% core roles
YAGS	1,415	3,091	5	18	75	6,081	74
TW ₁	236	1,085	10	47	40	1,704	77
TW ₂	236	1,027	11	46	39	1,614	79
TW ₃	236	1,038	11	47	39	1,399	89
ANY ₁	500	3,921	12	25	24	5,275	5
ANY ₂	400	2,360	13	30	27	3,255	5
Fate	1,686	4,473	4	38	50	11,699	73
MASC	8,444	7,226	25	42	33	11,214	78
SemEval	535	1,824	11	36	45	3,710	84
FNFT-test	2,420	4,458	12	42	33	7,172	83

Table 4.3: Evaluation dataset statistics. Subscripts for TW and ANY indicate the respective annotators, as these sets do not provide a single gold standard annotation.

The in-domain test based on the FrameNet fulltext corpus, called FNFT-test, was first introduced by [Das and Smith \(2011\)](#). It is a heldout set removed from the FrameNet fulltext corpus for evaluation purposes. While the FrameNet fulltext corpus covers data from various sources and domains, the test-split is an in-domain test set: all data sources for the test data are also represented in the training split. We already introduced the test sets based on Fate, MASC, and SemEval in Section 3.4.1. Like FNFT-test, they are mostly based on edited corpora.

The lack of FrameNet datasets from other domains and new genres, e.g., user-generated web data, motivated the creation of YAGS. By basing our data on the Yahoo! Answers question-and-answer forum, we cover a different domain than the two available Twitter-based datasets, the English dataset from [Johannsen et al. \(2015\)](#), called ANY, and TW from [Søgaard et al. \(2015\)](#). These two datasets have some very distinctive properties, they are, for instance, not adjudicated into a gold standard, leading to several separate datasets with subscripts to indicate the different annotators in Table 4.3. This leads to a high variance in role annotations: TW₃, the annotator with the lowest number of role annotations for TW, annotated only 82% of the number of roles of the annotator with the highest number of roles, TW₁, see Table 4.3. Therefore, evaluation scores are reported as averages from the annotations of two or three annotators. Both datasets have in common that they did not annotate spans for the semantic arguments, but argument heads. The ANY dataset contains mostly non-core roles, while annotators were encouraged to focus on core roles in the annotation of the TW dataset, see Table 4.3.

A remarkable difference between YAGS and all the other test sets is the higher proportion of verbal predicates and lower proportion of adjective and noun predicates in YAGS.

This is a result of the preference of verbs in the creation of YAGS. The next paragraphs compare YAGS to the other test sets in more detail.

Comparison of YAGS to TW and ANY. The TW datasets are fairly small. Note that the number of sentences represents the number of tweets for TW and ANY. A tweet contains up to 140 characters and thus may contain several sentences. YAGS contains three times more frames and roles than TW, approximating the size of FNFT-test.

The annotator of the ANY₁ dataset annotated more data than the second annotator ANY₂, which leads to the large difference in sentences, and frame and role annotations. Comparing our dataset to ANY₁, the more prolific annotator from [Johannsen et al. \(2015\)](#), it contains almost twice as many frame annotations and 2.5 times as many role annotations. As already mentioned, the ANY dataset contains mostly non-core roles, which sets it apart from all the other test sets. The low agreement suggests a lower quality of the ANY datasets. Together with the unexpected distribution of core and non-core roles, this motivates our decision to exclude the ANY dataset from the experiments in Section 4.3 and Section 4.4.

Comparison of YAGS to MASC, Fate, and SemEval. The MASC dataset is of substantial size, but it constitutes a lexical sample and therefore a slightly artificial evaluation setup. It contains the most frame annotations of all the datasets. The Fate dataset contains a very large number of role instances compared to the frame instances. The SemEval test set is much smaller than the other two test sets and also smaller than YAGS. The proportion of core roles is almost the same for YAGS and Fate, while MASC and SemEval contain a larger proportion of core roles.

Comparison of YAGS to FNFT-test. YAGS is only 30% smaller than FNFT-test considering the number of role instances, and 24% smaller when considering the number of frame instances. YAGS contains more verbal predicates and fewer nouns and adjectives than FNFT-test. The proportion of core roles in YAGS is slightly smaller compared to FNFT-test, i.e., YAGS contains relatively more non-core roles. The same applies to the TW sets.

Polysemy and coverage statistics. We also try to estimate the difficulty of the datasets with respect to the frame identification and role labeling tasks. We therefore analyze the test sets in regard to two aspects: first, their sense polysemy, because highly polysemous words present a harder disambiguation task, and second the training data coverage, because it is difficult for the system to assign unseen senses. A variant of the sense polysemy problem also applies to the role level: assigning the correct role is more difficult for a frame that lists 20 role labels compared to one that only lists 5.

dataset	max. # senses	average senses	standard deviation	total senses	lemmas	lemmas	senses	monosemous
					∕ lexicon	∕ FNFT-train	∕ FNFT-train	lemmas ∕ FNFT-train
YAGS	13	3.94	2.55	3,091	2.79%	13.33%	30.36%	27.02%
TW ₁	12	3.34	2.47	1,085	1.01%	17.51%	36.03%	26.73%
TW ₂	12	3.22	2.41	1,027	1.27%	17.91%	51.25%	27.07%
TW ₃	12	3.28	2.42	1,038	1.25%	17.24%	35.65%	27.17%
ANY ₁	13	1.88	2.12	3,921	23.18%	39.14%	60.44%	25.76%
ANY ₂	12	1.81	2.05	2,360	25.00%	40.00%	62.12%	27.96%
Fate	12	2.01	2.13	4,473	27.15%	39.17%	57.19%	27.64%
MASC	10	3.46	2.17	7,250	7.45%	21.72%	51.25%	23.51%
SemEval	13	2.78	2.13	1,824	5.36%	22.31%	38.21%	35.28%
FNFT-test	13	2.60	2.30	4,458	2.59%	9.99%	14.03%	53.99%

Table 4.4: FrameNet lexicon statistics indicate frame identification difficulty: statistics on available FrameNet frames (= senses) per lemma instance and training data coverage.

Polysemy and coverage for frames. The frame identification task is more difficult for polysemous lemmas, i.e., lemmas that are associated with more frames than others. Table 4.4 gives an overview on the lemma polysemy and thus the expected frame identification difficulty for each test set. In further experiments we particularly focus on semantic predicates that are verbs. Therefore, Table 4.5 provides verb-specific statistics, showing the same information for verbs.

The left-hand parts of the two tables show statistics on the sense polysemy, the number of available senses (or frame labels) per predicate instance. All statistics are computed on the token-level, i.e., for each predicate instance in the test set, not on the type level, i.e., for each unique predicate type. The maximum number of training senses per predicate, i.e., the highest polysemy, ranges between 10 and 13 for all datasets. YAGS, TW, and MASC show a larger average number of senses compared to the other test sets, which suggests they might be slightly more difficult. For verbal predicates, this difference to the other test sets persists for YAGS and TW, even though all test sets show a higher average number of senses per predicate instance for verbal predicates. This agrees with the notion that verbs are harder to disambiguate.

The right-hand parts of Tables 4.4 and 4.5 show to which degree the sense instances of the respective test sets are represented in the training dataset FNFT-train. The percentage of unseen senses, shown in the last but one column in the respective tables, is much lower for the in-domain test set FNFT-test; it is more than twice as high for YAGS, TW, and SemEval in Table 4.4, three times as high for MASC, and more than four times as high for Fate and the ANY test sets. For verbs, as shown in Table 4.5, the number of unseen senses is smaller for all test sets except for FNFT-test and YAGS, which contains a large

dataset	max. # senses	average senses	standard deviation	total senses	lemmas	lemmas	senses	monosemous
					∉ lexicon	∉ FNFT-train	∉ FNFT-train	lemmas ∈ FNFT-train
YAGS	13	4.45	2.61	2,313	2.82%	17.30%	30.06 %	25.77%
TW ₁	12	4.57	2.77	431	0.46%	9.74%	22.27 %	24.12%
TW ₂	12	4.35	2.71	401	0.50%	10.97%	23.94 %	26.93%
TW ₃	12	4.47	2.73	404	0.50%	9.16%	21.53 %	26.49%
ANY ₁	13	3.14	2.56	924	6.17%	27.48%	51.08 %	24.89%
ANY ₂	12	3.16	2.43	644	5.12%	27.02%	56.99 %	22.05%
Fate	12	2.68	2.20	2,226	15.86%	26.46%	46.55 %	31.05%
MASC	9	3.74	2.09	2,398	0.92%	4.59%	34.65 %	36.16%
SemEval	13	3.72	2.31	822	2.82%	17.3%	30.06 %	25.76%
FNFT-test	12	3.44	2.43	1,484	2.49%	8.49%	14.35%	40.50%

Table 4.5: FrameNet lexicon statistics indicate frame identification difficulty: statistics on available FrameNet frames (= senses) per **verb** lemma instance and training data coverage.

proportion of verbs and therefore shows equal numbers. Nevertheless, a large difference to the in-domain FNFT-test remains; for all of the out-of-domain test sets except TW, the percentage of unseen senses is at least twice as large as for FNFT-test. It is very difficult for a supervised frame identification system to assign unseen senses. Therefore, we expect lower frame identification scores for the out-of-domain test sets when training a system on FNFT-train.

Additionally, the percentage of monosemous words for verbs is more than 50% higher in FNFT-test compared to the UGD-based test sets YAGS, TW, and ANY, as shown in the last column in Table 4.5. For all parts-of-speech, the difference in the proportion of monosemous words is even higher, as shown in the last column in Table 4.4: the proportion of monosemous lemmas is twice as large for FNFT-test compared to the out-of-domain test sets. Monosemous words will by default receive the single available sense label from an automatic classifier, so they are easy to classify.²²

Statistics and coverage for roles. To gauge the difficulty for role labeling on the different test datasets, Table 4.6 contrasts the number of candidate roles, the number of roles that is available for a frame instance according to the FrameNet lexicon, for the different test sets. Our assumption is that role labeling is easier for frames with a smaller number of roles compared to a frame with many potential role labels. The statistics are collected in the same way as the statistics on sense polysemy for predicate instances in Table 4.4. Table 4.6 additionally distinguishes between statistics for all available roles and for core roles that

²²Note that errors can still occur for monosemous words: for instance, if there is a single sense A for a test verb v in the training set, but the gold label for v is a different sense $B \neq A$.

dataset	max. # roles	median roles	average roles	standard deviation	roles \in FNFT-train
all roles					
YAGS	32	11	10.97	5.11	83.00%
TW ₁	25	6	9.33	4.94	84.06%
TW ₂	25	6	9.30	4.93	85.39%
TW ₃	25	6	9.14	4.86	86.96%
ANY ₁	32	4	8.51	4.89	6.49%
ANY ₂	25	4	8.31	4.90	6.52%
Fate	32	13	10.31	5.91	75.63%
MASC	25	7	7.49	5.01	69.88%
SemEval	32	10	9.93	5.85	83.50%
FNFT-test	32	6	9.55	5.09	94.85%
core roles					
YAGS	11	2	3.24	1.54	91.51%
TW ₁	11	3	2.99	1.65	91.67%
TW ₂	11	3	2.96	1.58	92.24%
TW ₃	11	3	2.89	1.47	90.42%
ANY ₁	11	2	2.91	1.78	91.67%
ANY ₂	11	3	3.02	1.80	90.50%
Fate	11	2	2.90	2.01	93.84%
MASC	10	2	2.60	1.52	85.64%
SemEval	11	3	2.77	1.69	94.91%
FNFT-test	11	2	2.92	1.58	97.23%

Table 4.6: FrameNet lexicon statistics indicate role labeling difficulty: number of candidate roles per frame instance according to the FrameNet lexicon.

are associated with obligatory semantic arguments. The reasons for this are twofold: the first is to highlight the differences between the distribution of core and non-core roles; the second reason is that the standard evaluation script for FrameNet semantic role labeling particularly focuses on core roles. The statistics are again collected on the token level, i.e., based on matching test instances in the different test sets to the FrameNet fulltext corpus.

Table 4.6 displays the maximum number of available roles for a frame instance in the test sets, considering all roles in the first block, and only the core roles in the second block. It also shows the median and average number of roles per frame instance, and the observed standard deviation. The maximum and average number of roles per frame, and the standard deviations are much higher and more diverse when considering all roles, compared to considering core roles only. Labeling core roles thus may be an easier task. The statistics in Table 4.3 showed that FNFT-test and SemEval contain a larger proportion of core roles

than the other test sets. This indicates that they may be easier datasets for role annotation. Moreover, the training data coverage is higher for core roles than for non-core roles, as shown in the last column in Table 4.3. This also explains the conspicuous difference in training data coverage for the ANY datasets: they contain 95% non-core roles, and many of the frame and role combinations in ANY have not been annotated in the FrameNet full-text corpus. Most of the non-core roles can be used for many different frames, for instance adding temporal or spatial information to the event described by the frame, but most of these combinations have not been annotated as part of the FrameNet fulltext corpus.

In this section, we compared several tests sets for FrameNet SRL from various domains and highlighted different properties with respect to size and training data coverage. In the next section, we use the test sets to assess the domain generalization capabilities of open-source FrameNet semantic role labeling (Das et al., 2014).

4.3 Assessing the Domain Generalization of FrameNet SRL

In this section, we assess the domain generalization capabilities of FrameNet SRL using a contemporary, open-source system (Das et al., 2014). This is motivated by a lack of recent work on domain adaptation for FrameNet semantic role labeling, in particular for user-generated text, as described in Section 4.1. We expect that the domain adaptation problem presents itself differently for FrameNet compared to PropBank semantic role labeling, because PropBank semantic role labeling typically does not include a predicate labeling step.

In our analysis, we run the contemporary open-source semantic role labeling system SEMAFOR (Das et al., 2014) on the available out-of-domain test sets. We measure precision, recall, and F_1 on these test sets and present a detailed error analysis.

4.3.1 Experimental Setup

This section briefly summarizes the open-source system, the evaluation metrics, and the test data used for the evaluation.

Semantic role labeling system. The SEMAFOR semantic role labeling system follows the two typical steps of FrameNet semantic role labeling, frame identification and role labeling as defined in Chapter 1. We already introduced the system in Section 3.6. For frame identification, we use a retrained version of the SEMAFOR frame identification model (Das et al., 2014).²³ We used the default parameters presented by the system, not performing any parameter tuning. For role labeling, we use the best role labeling model from Kshirsagar et al. (2015) which was kindly provided by the authors. Recall that the systems receive

²³<https://github.com/Noahs-ARK/semafor>

text pre-annotated with frame targets as input, which has become the standard in recent evaluations.

Evaluation metrics. The SEMAFOR evaluation script (Das et al., 2014) measures accuracy for frame identification and precision, recall, and F_1 for full FrameNet semantic role labeling. For both, a parameter is available to select between evaluation with exact frame match or partial frame match. Full SRL can additionally be evaluated with and without using gold frames instead of the predicted frames. We additionally compute precision, recall, and F_1 scores for role labeling only. The evaluation script and our additional evaluations were already introduced in more detail in Section 3.6.

The results reported below are based on the evaluation results obtained with the exact frame match setting. We report separate scores for frame identification, full semantic role labeling given frames predicted by the system, full semantic role labeling given gold frames, role labeling given system frames, and role labeling given gold frames.

Test sets. We evaluate on the test sets introduced previously: FNFT-test, MASC, Fate, and SemEval were already introduced in Section 3.4.1; we also report results on two test sets based on user-generated text, our own YAGS and the Twitter-based test sets TW from Søgaard et al. (2015). We do not evaluate on the ANY datasets by Johannsen et al. (2015), because of the lower agreement for frames and the idiosyncratic annotation behavior that appears biased towards non-core roles. For the Twitter-based test sets from Søgaard et al. (2015), we report the average of the results obtained from the three distinct annotations, which we call *TW-av*. We report test results on verbal predicates and on all predicates.

4.3.2 Experiment Results

Table 4.7 shows the evaluation results for the exact match evaluation of frame identification, full SRL, and role labeling. We also report the exact match results for verbal predicates, shown in Table 4.8, because they will be relevant for the comparison of the training data generation methods in the next section. Both tables show the accuracy of the frame identification step (FrameId) in the second column. The other columns show precision P, recall R, and F_1 score as provided by the SEMAFOR evaluation scripts for full SRL or role labeling given gold frames or system frames, i.e., frames predicted by the system with an accuracy as shown in the FrameId column.

The main result of this analysis is that SEMAFOR performance is, as expected, lower on test sets from other domains and specifically user-generated discourse. Role labeling does not constitute the performance bottleneck for FrameNet SRL on out-of-domain test sets. The main problem is the lower frame identification performance which, because frames pre-select a certain set of roles, negatively affects the subsequent role labeling.

test set	FrameId	SRL system frame			SRL gold frame			RoleId system frame			RoleId gold frame		
	Acc	P	R	F ₁	P	R	F ₁	P	R	F ₁	P	R	F ₁
	YAGS	59.62	47.39	30.64	37.22	70.68	74.59	72.58	30.65	13.35	18.60	54.68	59.50
TW-av	62.17	46.52	32.74	38.44	73.53	80.24	76.74	24.05	11.89	15.91	57.28	66.25	61.45
Fate	50.18	36.54	22.52	27.87	75.65	78.68	77.14	16.55	06.53	09.36	62.43	66.38	64.35
MASC	39.52	34.86	24.90	29.05	69.54	72.69	71.08	28.59	14.75	19.46	49.89	53.72	51.74
SemEval	70.71	55.06	38.30	45.18	68.52	71.05	69.76	35.66	18.01	23.93	50.29	53.30	51.75
FNFT-test	82.09	67.57	46.94	55.40	71.15	75.28	73.16	46.35	22.27	30.08	52.71	57.93	55.20

Table 4.7: Domain generalization experiments: exact Acc, P, R, F₁ scores in percent for SEMAFOR on test sets from different domains. The highest scores per column are shown in **boldface**.

In the upcoming paragraphs, we discuss the results for the different evaluated SRL sub-tasks in detail, first testing on all parts-of-speech, and then testing on verbs only.

Full SRL scores on all parts-of-speech. The third column in Table 4.7, labeled *SRL system frame*, shows that SEMAFOR performs best on FNFT-test for full SRL, scores for the other test sets are distinctly lower. This is mostly due to the lower frame identification performance on the out-of-domain test sets, which becomes evident when comparing the full SRL scores given system frames to the to full SRL scores given gold frame labels in column four: the differences between the in-domain and out-of-domain test sets vanish, the scores for TW-av and Fate are even higher than those for the in-domain FNFT-test. A break-down by SRL subtask helps to understand the effects that lead to these results.

Frame identification scores on all parts-of-speech. Frame identification accuracy, shown in column *FrameId* in Table 4.7, is at least 19 points lower for the test sets based on user-generated text, TW-av and YAGS, compared to FNFT-test. Note that the difference is smaller, only 14 points F₁, for the partial match setting that gives partial credit to related frames (not shown in Table 4.7); a quarter of the difference in F₁ scores results from predicting a wrong frame that is related to the gold frames. Nevertheless, frame identification performance in the partial match setting is also distinctly lower for the out-of-domain test sets based on user-generated text compared to FNFT-test. The results are also far from the upper bound as given by the F₁ agreement scores across annotators for TW-av and YAGS, as reported in Section 4.2.5: F₁ agreement for frame labeling is 96.0% for YAGS and 84.5% for TW-av. For the other test sets suitable agreement information is not available.

For YAGS, accuracy is 59.62% compared to 82.09% for FNFT-test. Accuracy is slightly higher for TW-av. The difference may be a result of the slightly lower polysemy of the instances in the TW sets, see Table 4.4. The training data coverage of the TW sets is even

slightly lower than the coverage of YAGS, while the number of monosemous instances is similar.

The frame identification scores for the edited out-of-domain test sets, Fate, MASC, and SemEval are also lower than the scores for the in-domain FNFT-test. The highest scores are reported for the SemEval test set. This set shows similar polysemy to FNFT-test and a fairly high proportion of monosemous lemmas, see Table 4.4. The lowest frame identification accuracy is observed for the MASC set. This can be explained by a combination of the fairly high sense polysemy of MASC, the low sense coverage, and the low proportion of monosemous words in the MASC corpus, see Table 4.4. YAGS shows even higher polysemy than MASC. Its higher FrameId scores can be explained by the higher sense coverage.

Comparing the frame identification results to the perceived difficulty of the test sets according to the polysemy of the contained predicates (see column *average senses* in Table 4.4) shows that the highly polysemous YAGS and MASC receive lower frame identification scores. The Fate dataset is an outlier. It shows very low average sense polysemy. For Fate, the low accuracy is the result of insufficient sense coverage (see column *senses \notin FNFT-train* in Table 4.4); it may be caused by some frame mismatches between the FrameNet versions used for labeling the Fate corpus and the training data from FrameNet 1.5.

FNFT-test receives the overall highest scores, see Table 4.4. Besides being an in-domain test set, which is expected to correlate with high sense coverage in the training data, it contains a large proportion of monosemous words, which also contributes to higher frame identification scores.

Role labeling on all parts-of-speech. The low accuracy of frame identification negatively affects role labeling for the out-of-domain test sets, which can be seen in column five of Table 4.7 (*RoleId system frame*): both precision and recall are lower for the out-of-domain test sets compared to FNFT-test, leading to overall lower role labeling F_1 . For YAGS, for instance, role labeling F_1 is 18.60% compared to 30.08% for FNFT-test. Note that considering the partial match setting, which unlike the exact match setting gives credit to related frames, improves scores for all test sets with slightly higher improvements for the out-of-domain test sets (not shown in Table 4.7).

When using gold frame labels, as shown in columns four and six of Table 4.7, the performance of role labeling and full SRL improves dramatically for all test sets, including FNFT-test. These results show that the performance of role labeling is not the bottleneck for semantic role labeling on the out-of-domain test sets. It is the frame identification that needs to be improved.

For the out-of-domain test sets, full SRL and role labeling scores exceed the results for FNFT-test when using gold frames. The same applies to the Fate test set. MASC and SemEval still receive lower scores, but the distance to FNFT-test is much smaller than when using system frames. For MASC, the difference can in part be explained with a lower role

coverage of both core and non-core roles in FNFT-train, see Table 4.6. The lower scores for SemEval show that the percentage of core roles and coverage in the training data do not suffice as predictors for difficulty: SemEval has a high percentage of core roles, see Table 4.3 and a high level of role coverage, see Table 4.6, but it nevertheless receives lower scores. This indicates that additional domain effects may influence the system performance.

Evaluation on verbs. Because the training data generation methods evaluated in this section mostly target verbs, we also report the results for verbal predicates, as summarized in Table 4.8. The overall results are very similar to the results on all test instances. For YAGS the results change only minimally, as this dataset contains a large proportion of verbal predicates. For TW-av and FNFT-test the frame identification accuracy decreases compared to evaluation on all parts-of-speech, by more than 3 percentage points for TW-av and more than 5 for FNFT-test. This can be explained by the different coverage statistics of verbs: for FNFT-test the proportion of seen monosemous lemmas is smaller for the verbal predicates, as shown in Table 4.5. Additionally, as expected, the average polysemy is higher for verbal predicates in all test sets, which agrees with the notion that verbs are harder to disambiguate.

For the remaining datasets (Fate, MASC, and SemEval) frame identification accuracy however increases, by more than 10 points for MASC. This can also be explained by coverage statistics: there are fewer unseen sense instances for verbs in Fate, MASC, and SemEval. Thus, the large performance increase for MASC may be caused by a much higher lexicon coverage for verbal predicates, and a larger proportion of monosemous verb lemmas in Table 4.5 compared to Table 4.4. The same applies to Fate, but to a lower degree. For SemEval, the sense coverage is also increased, but the proportion of monosemous instances is lower for verbs than for all parts-of-speech. Both, improved lexicon coverage and larger proportion of monosemous lemmas for verbs can explain the improved performance on verbs for the such-affected test sets.

The change in frame identification performance influences the results of the full SRL performance (column 3 in Table 4.8), i.e., for test sets, for which frame identification improves, the SRL scores given system frames also improve. Full semantic role labeling and role labeling performance given gold frames are lower for all datasets compared to evaluation on all parts-of-speech in Table 4.7, which indicates that role labeling by itself is slightly more difficult for verbal predicates. This is expected, because verbal predicates typically license more roles than, for instance, nominal predicates, due to the richer subcategorization of verbs compared to nouns.

In order to further analyze the effects of training data coverage and polysemous senses on SRL performance, and to understand what kinds of errors remain, when these effects are factored out, we performed a detailed error analysis.

test set	FrameId Acc	SRL system frame			SRL gold frame			RoleId system frame			RoleId gold frame		
		P	R	F ₁	P	R	F ₁	P	R	F ₁	P	R	F ₁
YAGS	59.40	48.33	30.39	37.32	69.83	72.90	71.33	33.66	14.18	19.96	54.10	57.80	55.89
TW-av	58.48	40.83	27.31	32.73	72.78	77.81	75.21	17.91	08.40	11.43	57.90	64.33	60.93
Fate	55.76	40.86	24.09	30.31	72.88	74.95	73.90	21.10	08.05	11.66	59.70	62.26	60.95
MASC	52.77	44.72	30.89	36.54	65.25	67.72	66.46	36.10	18.73	24.66	47.01	49.78	48.36
SemEval	73.89	56.15	37.66	45.08	63.14	65.66	64.38	37.81	18.92	25.22	45.41	48.15	46.74
FNFT-test	76.88	62.01	40.99	49.35	69.18	73.06	71.07	43.47	20.20	27.58	52.81	57.49	55.05

Table 4.8: Domain generalization experiments for verbs: exact Acc, P, R, F₁ scores in percent for SEMAFOR on diverse test sets. The highest scores per column are shown in **boldface**.

Error analysis. We perform a detailed error analysis to better understand the sources of the errors in frame identification. In the following we present statistics on different error types, e.g., whether errors occur because predicates do not occur in the training data, or despite being represented in the training data. We also study frequent confusions, i.e., pairs of wrong system frame and gold frame that are often assigned by the system across the different test sets. A different distribution of confusions across test sets from different domains can indicate domain effects, i.e., indicate that the test domain is not represented well in the training data. Note that the error analysis is performed for the evaluation on verbal predicates.

Error analysis: error types. Because of the incomplete sense coverage of the FNFT-train, we can identify two different types of errors, those that result from insufficient sense coverage, and those that result from misclassifications of instances for which the correct sense has been seen in the training data. The latter are relevant to a more detailed error analysis.

Table 4.9 shows coverage statistics, i.e., the percentage of test instances seen in the training data, and statistics on different error types, i.e., the proportion of misclassified instances that do or do not occur in FNFT-train. It also displays the same kind of statistics for the correctly labeled instances.

The first line in Table 4.9 shows the coverage statistics of all test instances. We can see that the coverage of the in-domain test set is higher than the coverage of the other test sets. The coverage is particularly low for Fate and MASC. These two test sets also received the lowest frame identification scores. Despite lower coverage, SemEval receives much better scores than TW-av and YAGS, both of which show higher lexicon coverage. This can be explained: the domain of the SemEval set is much closer to the domain of FNFT-train than the user-generated text that comprises TW-av and YAGS.

YAGS		TW-av		Fate		MASC		SemEval		FNFT-test							
c	∈ T	%	c	∈ T	%	c	∈ T	%	c	∈ T	%						
1		75.9	1		77.4	1		53.5	1		65.3	1		69.9	1		85.6
1		59.4	1		58.5	1		54.8	1		52.8	1		73.7	1		76.9
1	1	52.9	1	1	51.4	1	1	45.1	1	1	50.2	1	1	58.1	1	1	71.5
1	0	6.5	1	0	7.1	1	0	9.7	1	0	2.5	1	0	15.6	1	0	5.4
0		40.6	0		41.5	0		45.2	0		47.2	0		26.3	0		23.1
0	1	23.0	0	1	26.0	0	1	8.4	0	1	15.1	0	1	11.8	0	1	14.1
0	0	17.6	0	0	15.5	0	0	36.8	0	0	32.1	0	0	14.5	0	0	9.0

Table 4.9: Study on domain generalization of SEMAFOR: different error types based on occurrence of test instances in training set ($\in T$); c stands for *instance labeled correctly* with boolean values $\{0,1\}$, column % contains the percentage of test instances for this error type.

According to the numbers in the fifth column of Table 4.9, assigning the correct frame for a predicate that did not occur in the training data works particularly well for SemEval that has a fairly low proportion of monosemous verbal predicates. This column shows that SEMAFOR is to a certain degree able to generalize to unseen predicates.

Error analysis: frequent confusions. For each test run, the pairs of gold frames and wrongly-assigned system frames can be used to create a confusion matrix and analyze frequent confusions. Because this matrix is very sparse, we do not display a full matrix, but report on frequent confusions. We observe differences in the distribution of confusions between test sets based on user-generated text and the test sets based on edited text: many confusions are shared across all test sets, for instance confusions based on closely related frames such as *Capability* and *Possibility*, others only occur for the user-generated test sets. Some of these appear to be a result of a different domain-specific preference of the test sets, others are examples of frame pairs that are easily confused as shown above. A few examples of the former from YAGS are: a) the system assigns a generic *Cause_change* to the more specific test frame *Cooking_creation*, b) the system assigns the generic *Cause_change_position_on_a_scale* to the more specific test frame *Cutting*. The *Cooking* domain is not well-represented in FNFT-train, but relevant to the YAGS test set. This analysis further supports our assumption that both training data coverage and domain effects lead to the lower performance of SEMAFOR frame identification on out-of-domain test sets.

Summary. In this section, we performed a detailed analysis of the domain generalization capabilities of open-source FrameNet SRL based on the SEMAFOR system. The evaluation results show that there is a need for domain adaptation in the frame identification step in semantic role labeling: SEMAFOR performs worse on most out-of-domain test sets, in particular the user-generated test sets. This affects the role labeling step, but our experimental

analysis shows that the role labeling scores given gold frames are similar, or even higher, on the out-of-domain test sets. Our error analysis shows that the inadequate coverage of domain specific predicates in the training set is one reason for the lower frame identification performance of SEMAFOR on out-of-domain data.

In the next section, we analyze whether automatic training data generation on web texts and user-generated texts, including the knowledge-based DistantSRL introduced in Chapter 3, can contribute to better coverage and domain generalization of FrameNet SRL.

4.4 Experiments on Training Data Generation for Domain Adaptation

The previous section motivated the need for domain adaptation of FrameNet semantic role labeling, more specifically of the frame identification step. This section presents experiments that evaluate automatically generated training data in a domain adaptation setting. We analyze the impact of automatically generated training data to SRL performance on the diverse test sets introduced above. We evaluate and compare variants of three methods for training data generation that have been proposed for semantic role labeling: the first is our own DistantSRL that we introduced in Chapter 3, the second is monolingual annotation projection (Fürstenau and Lapata, 2012), and the third the paraphrasing-based dataset created by Pavlick et al. (2015b). We use the first two methods to create several new training datasets, and for the third method use the published data.

In the next subsection, we introduce the different methods of training data generation and resulting datasets. Then we present semantic role labeling experiments using those datasets and evaluating on our battery of six test sets from different domains.

4.4.1 Methods of Training Data Generation

We evaluate three different methods of training data generation that have unique properties, the already introduced DistantSRL, monolingual annotation projection (Fürstenau and Lapata, 2012), and FrameNet+, a resource based on paraphrasing the FrameNet fulltext corpus (Pavlick et al., 2015b). Two of them, DistantSRL and FrameNet+, extend the training data coverage in comparison to FNFT-train, while FrameNet+ additionally adds new lemmas to the lexicon. A different pair of methods, DistantSRL and annotation projection, can be applied to unlabeled corpora from various domains. In the following paragraphs we briefly introduce the different methods.

DistantSRL. We use the large frame-labeled corpora WaS-L and WaS-XL introduced in Section 3.1 as training data. Recall that these are based on an automatic labeling of ukWAC

training set	sense types	sense tokens	verb senses tokens	roles types	roles tokens	verb roles tokens
FNFT-train	3,391	19,482	5,508	448	33,690	12,364
ap-bnc-top1	981	4,826	4,826	448	10,393	10,393
ap-bnc-top5	981	23,832	23,832	448	50,071	50,071
ap-bnc-top10	981	46,367	46,367	448	97,262	97,262
ap-ukwac-top1	986	4,933	4,933	453	10,517	10,517
ap-ukwac-top5	986	24,138	24,138	453	51,194	51,194
ap-ukwac-top10	986	47,478	47,478	453	100,291	100,291
ap-ya-glove-top1	930	4,434	4,434	435	9,154	9,154
ap-ya-glove-top5	930	40,702	40,702	435	83,228	83,228
ap-ya-glove-top10	930	76,396	76,396	435	154,662	154,662
ap-ya-glove-top1- α 02	852	4,445	4,445	416	9,205	9,205
ap-ya-glove-top5- α 02	852	40,702	40,702	416	83,232	83,232
ap-ya-glove-top10- α 02	852	76,324	76,324	416	154,534	154,534
FN+all	17,944	76,436	21,404	-	-	-
FN+4.0	9,971	32,629	9,155	-	-	-
FN+5.0	1,935	3,253	897	-	-	-
WaSR-L	967	89,550	89,550	153	35,336	5,336
WaSR-XL	1,445	1,408,703	1,408,703	175	610,431	610,431
WaSR-XL-10	1,445	13,501	13,501	162	6,119	6,119
WaSR-XL-20	1,445	25,895	25,895	167	11,731	11,731
WaSR-XL-40	1,445	48,554	48,554	172	21,869	21,869
WaSR-XL-80	1,445	87,929	87,929	171	39,105	39,105
WaSR-XL-160	1,445	153,435	153,435	172	68,219	68,219
WaSR-XL-320	1,445	256,403	256,403	174	111,537	111,537
WaSR-XL-640	1,445	404,582	404,582	174	171,434	171,434

Table 4.10: Statistics of automatically generated training corpora: sense and role types and tokens.

with DistantSRL. The seed patterns used for the frame labeling in DistantSRL were generated from the FrameNet example sentences and the example sentences linked to them in UBY_{FN} . Because the coverage of the FrameNet example sentences is larger than the coverage of the FrameNet fulltext training set, we expect small improvements in sense coverage. Note that this method provides additional training data for word senses existing in FrameNet. It does not increase the lexicon coverage: it does not expand the FrameNet lexicon with new lemmas or word senses.

Table 4.10 shows an overview of the sizes of the automatically created training datasets. The automatically created corpora are much larger than FNFT-train: WaSR-XL contains more than 1,000 training instances for 20% of the senses, up to 40,000 instances per sense. The distributions of senses is very skewed in WaSR-XL, and the large numbers of train-

ing instances impose a large computational cost to the classifier. Therefore, we randomly sample smaller training sets from WaS-XL which contain up to 640 training instances per sense. The selection of sample size follows the strategy used for the role classification in Section 3.4.3. We also evaluate the combinations of FNFT-train and the corpora based on DistantSRL, similar to the evaluation of DistantSRL in Section 3.4.

Annotation projection. We use a reimplementation of the annotation projection approach by Fürstenu and Lapata (2012) to create labeled training data on three unlabeled expansion corpora, the BNC, as was done by Fürstenu and Lapata (2012), ukWAC, and, to create an in-domain training set for the YAGS test set, the training section of the Yahoo! Answers Manners dataset according to the split by Surdeanu et al. (2011).²⁴ The projection approach was already briefly introduced in Section 3.3.2. Like DistantSRL, the annotation projection approach creates labeled training data for verbal predicates.

The core step of the projection algorithm is that labeled seed sentences are aligned to unlabeled sentences on the level of syntactic dependency structures. An integer linear programming algorithm (ILP) is used to derive an optimal alignment by optimizing a similarity score that combines lexical and syntactic similarities of the two sentences subject to some constraints. The optimized objective is the following:

$$score(M, N) = 1/C \cdot \left(\sum_{x \in M, y \in N} LexSim(x, y) + \alpha \cdot \sum_{d \in M, e \in N} SynSim(d_{x_1}^{x_2}, e_{y_1}^{y_2}) \right) \quad (4.1)$$

where M and N are the dependency graphs of the seed and expansion sentence, C is a normalizing constant, $LexSim$ is the lexical similarity between the heads of a pair of dependency nodes in source and target sentences, and $SynSim$ is the syntactic similarity of pairs of edges d, e in the two dependency graphs that connect pairs of aligned nodes x_1 and x_2 in the source graph and y_1 and y_2 in the target graph. In other words: $SynSim(d_{x_1}^{x_2}, e_{y_1}^{y_2})$ is the similarity score of the dependency edge d that connects dependency nodes x_1 and x_2 in the source sentence, and the edge e that connects dependency nodes x_1 and x_2 in the target sentence. The objective score in Equation 4.1 sums over all pairs of nodes and edges in the two dependency graphs. The parameter α determines the relative weight between syntactic and lexical similarities.

The constraints enforce that each dependency node in the source graph is aligned to at most one dependency node in the target graph, and vice versa. Additional constraints enforce that only dependencies between aligned nodes can be aligned, and that the two predicate nodes are aligned by default. For additional details, refer to the journal article by Fürstenu and Lapata (2012).

Syntactic similarity $SynSim$ is set to 1 if the dependency labels between two dependency nodes match, to 0 otherwise. Lexical similarity $LexSim$ compares the dependency heads ac-

²⁴The presented results are based on joint work with Ilia Kutsnetsov, a co-author of Hartmann et al. (2017a).

ording to a lexical similarity measure. Fürstenu and Lapata (2012) use a distributional similarity measure for their experiments. In our experiments, we use word vectors based on Glove (Pennington et al., 2014) to compute lexical similarity defined as the cosine of two word vectors. This means we replace the approach to lexical similarity used by Fürstenu and Lapata (2012) by a state-of-the-art method: dense embedding vectors like those provided by Pennington et al. (2014) have been shown to improve over conventional distributional representations for the word similarity task and can capture semantic phenomena such as word analogies (Pennington et al., 2014; Mikolov et al., 2013).

If the alignment is successful, all semantic role labels from the seed sentence can be transferred – or projected – to the expansion sentence. The objective score resulting from the ILP optimization is used to rank the expansions for a seed sentence. The top k expansion sentences for each seed are added to the generated training set.

Because annotation projection is based on the alignment of complete dependency structures, it is strongly biased towards projecting to syntactically similar expansion sentences. To alleviate this effect, we vary two parameters of the method, k and α . Adding a larger number of expansions k to our corpus, means that not only the expansion from the optimal alignment, but also from less optimal expansions are added to the automatically labeled corpus. We explore different values for $k \in (1, 5, 10)$.

Using a low value for α reduces the weight of the syntactic similarity in the objective score, which should result in more lexically-driven alignments and potentially provide more versatile expansions, i.e., expansions less similar to the seed sentence. In our first set of experiments, we use the α value of 0.55 proposed by Fürstenu and Lapata (2012), but we also experiment with a lower α of 0.2.

In the projection variant we use, seeds and expansion sentences are paired if their lemma and part-of-speech match. Because we base the seed set on the FrameNet fulltext training set, we do only generate additional training data for senses in the FrameNet fulltext training set. We do not expand the lexicon for training. This could be done by using the FrameNet example sentences as seeds, like we did for DistantSRL, or by relaxing the restriction that seed and expansion sentences need to have the same lemma. Fürstenu and Lapata (2012) also evaluate this method of performing lexical acquisition and annotation projection at the same time. Because this method creates a huge set of candidate pairs (potentially matching all expansions with all seeds), they use lexical similarities between the predicates in the seed and expansion sentences to filter the huge candidate set.

We automatically generate several training datasets by applying annotation projection to three different expansion corpora, the British National Corpus BNC, the web-based ukWAC that we also used for the experiments with DistantSRL, and the training split of the Yahoo! Answers Manners corpus as reported by Surdeanu et al. (2011). The names of the corpora are abbreviated as *bnc*, *ukwac*, and *ya* in Table 4.10 that gives an overview on the created datasets. Since we project labels for the top k alignments from the FrameNet full-

text training set to the unlabeled corpora, the automatically labeled corpora have roughly the same size for each of the unlabeled corpora.

Table 4.10 shows an overview of the evaluated corpora. They are named based on the expansion corpus (*bnc*, *ukwac*, and *ya*), the top k used for the corpus, and, if a different α from the default 0.55 was chosen, the α score. We publish projected corpora based on the open-licensed ukWAC corpus for research purposes, see Appendix A.

FrameNet+. FrameNet+ was created by Pavlick et al. (2015b) by replacing predicates in the FrameNet fulltext corpus by their paraphrases, as introduced in Section 3.3.2. By creating paraphrases of the predicate, they extend the FrameNet lexicon and create frame-labeled training data for previously unseen predicates. They do not provide role-labeled data, but we could infer them by matching the paraphrases to the FrameNet fulltext corpus. Pavlick et al. (2015b) do not evaluate the benefit of their expanded corpus to the task of FrameNet semantic role labeling.

Unlike the two previous approaches, FrameNet+ (short FN+) covers all parts-of-speech in the FrameNet lexicon. Pavlick et al. (2015b) used crowdsourcing to manually verify the validity of the automatically generated paraphrases. They rank the paraphrases by quality according to the crowdworkers’ judgments, providing the quality ratings 3, 4, and 5 for the accepted paraphrases. We use these scores to create variants of the FrameNet+ corpus, using all instances (FN+all), or only those with a confidence score of at least 4 (FN+4.0) or 5 (FN+5.0). We additionally removed those instances from FN+ that are paraphrases of instances in FNFT-test to be able to evaluate on FNFT-test without bias. Statistics on the three variants are shown in Table 4.10.

4.4.2 Experimental Setup

Our experimental setup consists of the following steps: (1) generating training data (in the case of FN+ preprocessing the existing corpus), (2) training a supervised frame identification model, (3) evaluation on the in-domain and out-of-domain test data.

SRL system. We again use the open-source system SEMAFOR (Das et al., 2014) to train models for frame identification (see also Section 4.3).

Test data. We evaluate our systems on the out-of-domain test sets from YAGS, TW, ANY, and MASC, as in Section 4.3. We compare the results to the in-domain FNFT-test.

Training data. We use the frame-labeled corpus from FrameNet+, and the frame- and role-labeled corpus created with DistantSRL (see Chapter 3) as additional training data. We additionally create training data using annotation projection following Fürstenauf and Lapata (2012), as described above, including an in-domain training set for YAGS by projecting

training set	YAGS	TW-av	Fate	MASC	SemEval	FNFT-test
high-precision WaS-L						
WaS-L	0.55979	0.5715133	0.51124	0.56612	0.68473	0.69236
WaS-L-fntrain	<u>0.60702</u>	<u>0.59551</u>	0.53372	<u>0.57071*</u>	0.71798	0.73834
high-recall WaS-XL samples						
WaS-XL-10	0.40511	0.40509	0.45459	0.44013	0.52463	0.57066
WaS-XL-20	0.43154	0.45671	0.46088	0.44639	0.58005	0.58418
WaS-XL-40	0.41941	0.4274133	0.47257	0.46350	0.58005	0.57471
WaS-XL-80	0.44541	0.4475367	0.49236	0.46600	0.58621	0.62069
WaS-XL-160	0.43154	0.4567533	0.46313	0.48602	0.54433	0.62001
WaS-XL-320	0.44931	0.47996	0.48291	0.50688	0.59852	0.61325
WaS-XL-640	0.45797	0.49888	0.48246	0.52274	0.60345	0.62813
WaS-XL samples combined with FNFT-train						
WaS-XL-10-fntrain	0.40511	0.40506	0.45459	0.44013	0.52463	0.5706
WaS-XL-20-fntrain	0.52166	0.5450667	0.50270	0.48811	0.64163	0.69033
WaS-XL-40-fntrain	0.50217	0.5089933	0.49910	0.49353	0.63547	0.65585
WaS-XL-80-fntrain	0.52340	0.5330367	0.50809	0.48769	0.62931	0.67343
WaS-XL-160-fntrain	0.51083	0.5314567	0.50360	0.50396	0.62192	0.67748
FNFT-train	0.59402	0.5847867	<u>0.55755</u>	0.52774	<u>0.73892</u>	<u>0.76876</u>

Table 4.11: Frame identification accuracy on test verbs for training data expansion with DistantSRL; **boldface** marks best results in section, underline marks overall best results, * marks significant improvements to FNFT-train.

on the training split of the Yahoo! Answers Manners dataset used by Surdeanu et al. (2011), see also Section 4.2.2. Using the training split, there is no overlap to the YAGS test set.

These methods lead to corpora with different properties that we compare for our task. Besides training only on the automatically created corpora, we combine them with the FrameNet fulltext training set. For these combinations, we add the suffix *-fntrain* to the training set names that were used in Table 4.10.

4.4.3 Experiment Results

We aim to estimate the influence of the different methods of training data expansions introduced above on domain adaptation for frame identification. Therefore, we train the SEMAFOR frame identification model on the different corpora introduced above and compare the performance on the various out-of-domain test sets. This section presents the results of these experiments. Since two of the compared approaches, DistantSRL and annotation projection, are focused on verbs, we evaluate on verbal predicates only.

DistantSRL results. Table 4.11 shows the frame identification results when training on the corpora created with DistantSRL and evaluating on our diverse set of test sets. We experiment with the two different corpora presented in Section 3.4, WaS-L and WaS-XL. Both corpora are much larger than the commonly used FNFT-train. WaS-L contains more than four times the number of predicate instances of FNFT-train, and WaS-XL more than 70 times. Since SEMAFOR was not developed for training on large amounts of training data, training on the full WaS-XL is infeasible due to large memory requirements. We therefore train SEMAFOR on randomly sampled subsets of WaS-XL.

We use the same sampling strategy as in Section 3.4.3 to sample up to 640 instances per sense from WaS-XL. Training corpus statistics are shown in Table 4.10. We additionally evaluate the union of the WaS-L and WaS-XL corpora with FNFT-train, marked by suffix *-fntrain*. In contrast to Section 3.4.3, the union of the corpora produces better results than a back-off setting. We did not observe improvements in a back-off setting and therefore do not report the results obtained for this setting.

The best results for the automatically generated corpora are observed for WaS-L, as shown in the first block of the table. For TW-av, the accuracy scores approach the results for the standard setup, i.e., training only on FNFT-train. For MASC results improve upon the standard setup by 0.038. For the other test sets, the accuracy scores are between 0.034 and 0.076 lower. The largest differences to the standard setup are observed for the edited test sets SemEval and the in-domain FNFT-test.

Overall best results are obtained when combining WaS-L with FNFT-train, i.e., for WaS-L-fntrain. For the user-generated test sets YAGS and TW-av, we observe small improvements compared to the standard setup: the accuracy is increased by 0.006 for YAGS and by 0.01 for FNFT. We observe even larger improvements for MASC, an increase of 0.04, that significantly improves over the standard setup according to Fisher’s exact test with $p < 0.05$, leading to the overall best result on MASC. The frame identification performance on the MASC set seems to benefit from the additional sense coverage provided by WaS-L. There also is a small decrease in accuracy for Fate and SemEval, and a slightly larger decrease of 0.03 for the in-domain test sets FNFT-test. In line with the results observed for our frame classification experiments in Section 3.4.3, we find that the automatically labeled corpora contribute to improved frame identification performance.

The frame identification scores for the WaS-XL samples do not reach the performance of the high-precision WaS-L corpus. There, however, is a tendency for the accuracy scores to increase with larger samples. This can be observed for five out of the six test sets, all sets except Fate. This indicates that including larger samples might further improve the performance. Based on the lower results for training on the WaS-XL samples, it is expected that the results for WaS-XL samples in combination with FNFT-train do not reach the results for WaS-L-fntrain, as shown in the third block of Table 4.11.

training set	YAGS	TW-av	Fate	MASC	SemEval	FNFT-test
projection to BNC						
ap-bnc-top1	0.59012	0.5594233	0.55531	0.51815	0.73768	0.75997
ap-bnc-top5	0.59445	0.5966533	0.54856	0.54693	0.71305	0.75659
ap-bnc-top10	0.59359	0.5930233	0.55800	0.55152	0.70443	0.76606
ap-bnc-top1-fntrain	0.59272	0.5849033	0.56115	0.55444	0.72783	0.77079
ap-bnc-top5-fntrain	0.59099	0.5930633	0.56115	0.54819	0.71675	0.75997
ap-bnc-top10-fntrain	0.59792	0.59057	0.56205	0.55111	0.70813	0.76673
projection to ukWAC						
ap-ukwac-top1	0.60312	0.58674	0.54406	0.54485	0.73030	0.75321
ap-ukwac-top5	0.59489	0.5873533	0.53867	0.56112	0.71921	0.74577
ap-ukwac-top10	0.60485	0.5849533	0.55621	0.56487	0.70936	0.75862
ap-ukwac-top1-fntrain	0.58492	0.59927	0.55126	0.55736	0.71059	0.75794
ap-ukwac-top5-fntrain	0.59489	0.5873533	0.53867	0.56112	0.71921	0.74577
ap-ukwac-top10-fntrain	0.60269	0.5750797	0.56205	0.55903	0.71182	0.75456
projection to Yahoo! Answers Manners						
ap-ya-glove-top1	0.59272	0.5654167	0.54182	0.50271	0.72537	0.73969
ap-ya-glove-top5	0.59272	0.57037	0.54137	0.53150	0.69458	0.73631
ap-ya-glove-top10	0.58925	0.5874033	0.53867	0.53859	0.68473	0.72752
ap-ya-glove-top1-fntrain	0.58752	0.5912733	0.55621	0.54568	0.71921	0.76268
ap-ya-glove-top5-fntrain	0.59619	0.5792467	0.55935	0.54735	0.70813	0.76335
ap-ya-glove-top5-fntrain- α 2	0.59185	0.58183	0.55755	0.54944	0.71798	0.75727
ap-ya-glove-top10-fntrain	0.59402	0.5774033	0.55935	0.53859	0.71059	0.72752
ap-ya-glove-top10-fntrain- α 2	0.59749	0.5843833	0.56295	0.54777	0.71798	0.75254
FNFT-train	0.59402	0.5847867	0.55755	0.52774	0.73892	0.76876

Table 4.12: Frame identification accuracy on test verbs – training data generated via annotation projection; **boldface** marks best results in section, underline marks overall best results.

These results again prove the quality of the automatically labeled corpus created with DistantSRL in the high-precision setting, further corroborating the results of Chapter 3.

Annotation projection results. Table 4.12 presents the results of our experiments with corpora based on annotation projection. It is important to note that Fürstenau and Lapata (2012) evaluated the projection approach in a different way: they use a low-resource scenario, only using a small number of seeds per predicate and comparing performance for an increasing number of seeds and expansions, and they particularly focus on role labeling, not frame identification. This is the first evaluation of annotation projection for FrameNet frame identification at a large scale, i.e., using a larger number of seeds and comparing the performance to a system based on the full training set.

We evaluate a number of different corpora created via annotation projection based on three corpora, BNC, ukWAC, and Yahoo! Answers. The latter aims to create an in-domain training setup for the YAGS test set. Each of the base corpora receives its own section in Table 4.12. We vary the number of top k expansions selected to create the automatically labeled corpora with $k \in (1, 5, 10)$, and show results for combinations of the automatically labeled corpora with FNFT-train. The corpora based on the single best expansion are slightly smaller than verbal part of FNFT-train. The corpora based on top 10 expansions contain more than 70,000 predicate instances and thus are of a similar order of magnitude as the WaS-L corpus.

The experiments with the corpora based on the BNC are shown in the first section of Table 4.12. The evaluation results on the in-domain FNFT-test show that annotation projection generates high-quality training data that closely approximate the original FNFT-train: when training on ap-bnc-top1, accuracy approximates the accuracy of the standard setup that is shown for comparison in the last line of the table. Scores get even closer to the standard setup for the larger ap-bnc-top10, that is expected to generalize from the seed corpus, FNFT-train, due to the selection of the top 10 expansion sentences instead of selecting only the best expansion. Overall best results on FNFT-test are, however, observed for the combination of ap-bnc-top1 with FNFT-train. Accuracy scores for the other corpora reach the scores for FNFT-train (SemEval), or are even slightly higher (YAGS, TW-av, Fate, MASC). The largest improvements are observed for MASC.

The experiments with the corpora based on the ukWAC corpus are shown in the second section of Table 4.12. The scores are higher than those for the corpora based on BNC for the user-generated test sets and for MASC, leading to the best results in this table for YAGS, TW-av, and MASC. Best results for Fate are the same as for the corpora based on BNC. Accuracy scores for the in-domain FNFT-test and the SemEval set, that receives best results when training on FNFT-train, are slightly lower than for BNC. These results indicate that annotation projection to the web corpus ukWAC leads to training data that are better-suited for the out-of-domain training sets.

The third section in Table 4.12 shows experiments with corpora based on the Yahoo! Answers Manners dataset. For these corpora, we evaluate an additional setting: we relaxed the importance of the syntactic similarity in the optimized objective function by choosing a lower α . While α is set to 0.55 per default, we also experiment with setting α to 0.2, marked by the suffix $\alpha 2$ in the table.

With training corpora based on Yahoo! Answers, we aim to create an in-domain training dataset. There are, however only small improvements compared to training on FNFT-train. The highest accuracy score for ap-ya-glove-top10-fntrain- $\alpha 2$ is still lower than the best results obtained for YAGS when training on the ukWAC-based corpora.

The scores for the lower - $\alpha 2$ show that reducing the impact of the syntactic similarity on the alignment (by reducing α to 0.2) increases the quality of the training data for all

test sets: accuracy increases when training on `ap-ya-glove-top10-fntrain- α 2` compared to training on `ap-ya-glove-top10-fntrain`. This seems to be a promising direction for further exploration of the annotation projection approach.

When comparing results for the different unlabeled corpora, we find that the BNC-based corpora appear to reproduce the original FNFT-train most closely. This leads to good results on SemEval and the in-domain FNFT-test, that also receive the highest frame identification scores when training on FNFT-train. Best results for the out-of-domain test sets are observed when projecting to the ukWAC corpus.

For TW-av, there is no clear tendency for a best-suited expansion corpus. This might result from averaging over results for three different annotators. It seems that improvements for the TW-av set are gained from smaller, high-precision expansions: an increase of 0.01 accuracy compared to the standard setup is achieved when training on `ap-ukwac-top1-fntrain`. For the in-domain test set FNFT-test, we also observe small improvements for small, high-precision expansions, for instance for `ap-bnc-top1-fntrain`.

In contrast to the corpora based on DistantSRL, annotation projection corpora achieve very high frame identification scores without the addition of FNFT-train: overall best results for three out of six test sets do not require the addition of FNFT-train, and the results for the other three test sets approximate the best results.

We do not observe improved results for YAGS when training on an in-domain training set based on Yahoo! Answers Manners: best results for YAGS, and the other test set based on user-generated data, TW-av, are reported for corpora based on ukWAC. We conclude that simple projection to an in-domain corpus does not lead to a domain-adapted training set. We assume that this is due to the strict selection of expansions in the annotation projection that has a tendency to reproduce the original training set. We expect improved scores, when relaxing the constraints that enforce this tendency, for instance by choosing even lower alpha scores. Other ways to adapt annotation projection to create data that generalize better could be to use a syntactic similarity measure that is more flexible than the binary measure used, for instance based on dependency embeddings, or to relax the constraint that all of the seed roles need to be aligned to their counterparts in the expansion sentence.

FN+ results. Results for frame identification when training on variants of FN+ are shown in Table 4.13. When using all instances from FN+, the size of the training corpus is in a similar range as the corpora for annotation projection or WaS-L. Using only high-confidence instances from FN+ leads to smaller corpora, as shown in Table 4.10. Recall that we removed those paraphrases from FN+ that are based on instances in FNFT-test to be able to evaluate on FNFT-test.

When training on FN+, we observe very low performance for the user-generated test sets YAGS and TW-av compared to the standard setup, training on FNFT-train. The same applies to MASC. The lexicon expansions in FN+ do not benefit the out-of-domain test

training set	YAGS	TW-av	Fate	MASC	SemEval	FNFT-test
FN+all	0.50607	0.4991317	0.52518	0.43221	0.58744	0.67140
FN+4.0	0.51863	0.4693033	0.51754	0.41218	0.57882	0.68830
FN+5.0	0.42244	0.3957	0.50180	0.39758	0.55788	0.63624
FN+all-fntrain	0.57929	0.5791933	<u>0.55935</u>	<u>0.54360</u>	0.70567	0.77011
FN+4.0-fntrain	0.59315	0.5825867	0.55081	0.53484	0.70320	0.76133
FN+5.0-fntrain	0.59055	0.55818	0.55576	0.53859	0.70197	0.75118
FNFT-train	<u>0.59402</u>	<u>0.5847867</u>	0.55755	0.52774	<u>0.73892</u>	0.76876

Table 4.13: Frame identification accuracy for training data from FrameNet+ on test verbs; **bold-face** marks best results in section, underline marks overall best results.

sets. Scores are lower for all test sets for the smaller, high-confidence subsets FN+4.0 and FN+5.0, due to reduced coverage. In combination with FNFT-train, there are improvements over FNFT-train alone for FNFT-test, MASC, and Fate. The MASC test set seems to benefit from the increased sense coverage when adding FN+, but less than observed for annotation projection or DistantSRL.

In summary, we do not observe much improvement when adding FN+ to FNFT-train. Since only the predicates in the FrameNet fulltext training set are replaced during the creation of FN+, but no new sentences are added, we do not expect large effects on domain generalization from FN+. The increased coverage in FN+ improves the frame identification performance for Fate, MASC, and FNFT-test, but the coverage increase does not result in improvements for the test sets based on user-generated text, TW-av and YAGS.

Best results. To determine the overall best results, we compare the results in the Tables 4.11, 4.12, and 4.13. This comparison furthermore confirms the quality of the WaS-L corpus: using this corpus to expand FNFT-train leads to the best frame identification results for YAGS and MASC, accuracy of 0.60702 and 0.57071 respectively, and the second best result for TW-av, accuracy of 0.59551. The improvements compared to training on the FrameNet fulltext training set alone are statistically significant for MASC. Tests sets based on edited corpora from similar domains to the FNFT-train (Fate, SemEval, FNFT-test) do not benefit from extensions via DistantSRL. MASC and the user-generated corpora TW-av and YAGS also benefit from annotation projection. A particularly successful setup is using ukWAC as the expansion corpus and the top 10 expansions.

For the corpora based on annotation projection, combination with FNFT-train is not required for all out-of-domain test sets to achieve best results, unlike for the WaS-L corpus. The reason for this is that corpora resulting from annotation projection are very similar to their seed corpus FNFT-train. An annotation projection corpus based on the BNC combined

with FNFT-train also leads to the highest accuracy score for the in-domain FNFT-test. At 0.77079, the score is higher than when training on FNFT-train alone.

The improvements for WaS-L are due to the increase in lexicon coverage provided by WaS-L and the addition of web-based training data that lead to a wider representation of different domains in the training data. WaS-L is based on ukWAC, and the improved domain representation also comes into effect when using ukWAC as extension corpus for annotation projection, leading to best results for TW-av (0.59927).

FN+ increases the size of the lexicon, but does not change the domain properties of the training data, because it only paraphrases FNFT-train. As a result, training on FN+ shows only small performance gains on our test sets, and no gains on the tests sets from the user-generated domain. With accuracy of 0.77011 for FN+all-fntrain, we however observe second highest accuracy for the in-domain FNFT-test.

Summing up, WaS-L combines improved lexicon coverage and improved domain generalization that leads to the overall best results on two out of five out-of-domain test sets, MASC and YAGS. The improvements are statistically significant for MASC. In this setup, the performance for TW-av is only slightly lower than for annotation projection. The performance on the in-domain test set is reduced by 0.03 points, but remains high with an accuracy score of 0.738. Thus, WaS-L is competitive to the corpora based on annotation projection that reach best results for TW-av, Fate and the in-domain FNFT-train. This shows the potential of both methods for training data expansion.

Since the performance on most of the out-of-domain test sets is still more than 0.16 points accuracy smaller than the performance on the in-domain FNFT-test for all data expansion methods, there is ample room for improvement of the frame identification performance. We perform a detailed error analysis to identify error sources.

4.4.4 Error Analysis

We perform a detailed error analysis to understand the lower performance of SEMAFOR on the out-of-domain test sets, even when expanded training data are used. We particularly focus on the new test set based on user-generated text, YAGS.

Error analysis: error types. We again distinguish two different types of errors, those that result from insufficient sense coverage, and those that result from misclassifications of instances for which the correct sense has been seen in the training data. In Table 4.14 we contrast the different error types when training SEMAFOR on the different training sets and evaluating on our battery of test sets. The table contains the percentage of correct ($c=1$) and incorrect instances ($c=0$) that have ($\in \text{train}=1$) or have not been seen ($\in \text{train}=0$) in the respective training set. Empty table fields indicate that both values apply. This means that the first row in Table 4.14 reports the percentage of test instances that has been seen in the training data, independently of whether they have been labeled correctly.

YAGS		TW-av		Fate		MASC		SemEval		FNFT-test	
c	∈ T	%	c	∈ T	%	c	∈ T	%	c	∈ T	%
T = WaS-L-fntrain											
1		85.6	1		84.9	1		61.8	1		82.2
1		60.7	1		59.6	1		52.4	1		57.0
1	1	56.6	1	1	54.2	1	1	47.3	1	1	56.9
1	0	4.1	1	0	5.4	1	0	5.1	1	0	0.01
0		39.3	0		40.4	0		47.6	0		43.0
0	1	29.0	0	1	30.7	0	1	14.5	0	1	25.3
0	0	10.3	0	0	9.7	0	0	33.1	0	0	17.6
1		87.7	1		87.7	1		87.7	1		87.7
1		71.5	1		71.5	1		71.5	1		71.5
1	1	66.4	1	1	66.4	1	1	66.4	1	1	66.4
1	0	5.1	1	0	5.1	1	0	5.1	1	0	5.1
0		28.5	0		28.5	0		28.5	0		28.5
0	1	21.3	0	1	21.3	0	1	21.3	0	1	21.3
0	0	7.1	0	0	7.1	0	0	7.1	0	0	7.1
T = FNFT-train											
1		75.9	1		77.4	1		53.5	1		65.3
1		59.4	1		58.5	1		54.8	1		52.8
1	1	52.9	1	1	51.4	1	1	45.1	1	1	50.2
1	0	6.5	1	0	7.1	1	0	9.7	1	0	2.5
0		40.6	0		41.5	0		45.2	0		47.2
0	1	23.0	0	1	26.0	0	1	8.4	0	1	15.1
0	0	17.6	0	0	15.5	0	0	36.8	0	0	32.1
1		69.9	1		69.9	1		69.9	1		69.9
1		73.7	1		73.7	1		73.7	1		73.7
1	1	58.1	1	1	58.1	1	1	58.1	1	1	58.1
1	0	15.6	1	0	15.6	1	0	15.6	1	0	15.6
0		26.3	0		26.3	0		26.3	0		26.3
0	1	11.8	0	1	11.8	0	1	11.8	0	1	11.8
0	0	14.5	0	0	14.5	0	0	14.5	0	0	14.5

Table 4.14: Domain generalization experiments: different error types based on occurrence of test instances in training set ($\in T$); c stands for *instance labeled correctly* with boolean values $\{0,1\}$, column % contains the percentage of test instances for this error type.

Table 4.14 shows that the training data coverage increases for all test sets when using WaS-L-fntrain. The coverage of YAGS and TW-av now reaches the coverage of the FNFT-test for the unexpanded training set, which is 85.6%. The largest improvements in coverage are observed for the MASC and SemEval test sets. For MASC, the coverage is increased up to 82.2%. The frame identification scores for MASC, that also contains rare senses, benefit the most from the improved coverage. The increased coverage does not help to improve the frame identification scores for the in-domain test set and SemEval, that appears to be very similar to the in-domain set. Both of these sets already show high frame identification accuracy when training on FNFT-train alone, as can be seen in Table 4.13.

While the coverage of the out-of-domain test sets increases, the increases in frame identification performance are small. There are several potential reasons for this: a) certain domain-specific properties of the data may not be represented well in the training sets, and b) preprocessing errors due to properties of the user-generated content may cause errors for the test sets TW-av and YAGS. The lower performance on the MASC test set, that should not suffer from b), but also includes rare senses due to being a lexical sample, points to a third potential reason c), namely fine-grained sense distinctions for certain predicates that

training set	most-frequent-sense baseline					FNFT-test
	YAGS	TW-av	Fate	MASC	SemEval	
FNFT-train	53.38	53.0	46.42	55.42	59.51	73.38
WaS-L-fntrain	56.28	55.0	47.08	55.88	66.38	73.99

Table 4.15: Most-frequent-sense baseline for WaS-L-fntrain and FNFT-train; accuracy scores in %.

may be hard to distinguish automatically. We perform a detailed error analysis of misclassification below to determine the influence of these factors.

According to Table 4.14, the percentage of correctly labeled instances that do not occur in the training data decreases for WaS-L-fntrain compared to FNFT-train. At the same time, the percentage of incorrectly labeled instances that do occur in the training data increases. Both can be expected when increasing the training data coverage. The incorrectly labeled instances that do occur in the training data can provide more information on the reason for the misclassification. We study these instances in a detailed manual error analysis below.

The percentage of errors decreases for YAGS, TW-av, and MASC when training on WaS-L-fntrain compared to training on FNFT-train. We analyzed the changes in errors for the instances in both training setups, i.e., contrasting for each instance whether it is labeled correctly for FNFT-train and WaS-L-fntrain or not. We find that the remaining errors are different between both setups. This means that WaS-L-fntrain does not simply reduce the number of errors in for these test sets, but introduces new errors, while remedying others. This is the result of a different sense distribution. The improvements might be the result of a better suiting most-frequent-sense (MFS) baseline for the test sets that benefit from the automatically labeled training data. To analyze this effect, we contrasted the MFS baselines for both training sets.

Most-frequent-sense baseline. Table 4.15 contrasts the most-frequent-sense baselines (MFS) for FNFT-train and WaS-L-fntrain. The MFS baseline improves for the test sets YAGS, TW-av, and SemEval when adding the automatically generated training data in WaS-L-fntrain. There are modest improvements for MASC. The SEMAFOR frame identification results improve significantly for MASC when adding WaS-L to FNFT-train, and there are modest improvements for YAGS and TW-av. These changes are not correlated to changes in the MFS baseline, so the observed increases and decreases in frame identification performance for the different test sets cannot be accounted to a changed MFS baseline.

Error analysis: frequent confusions. Looking at the frequent confusions when training on WaS-L-fntrain in comparison to FNFT-train alone, we find that most of the confusions that occurred across the user-generated test sets, such as *Cause_change* – *Cooking_creation*, do not occur for the expanded training set. Some of the frequent confusions across all

datasets do not occur for the extended training set (*Becoming* vs. *Arriving*), while others persist, e.g., *Attempt* vs. *Trying_out* and *Capability* vs. *Possibility*. The reduced number of confusions that are unique to the user-generated test sets also demonstrate that the expanded test sets better represent the domains of the user-generated data.

Error analysis: YAGS verbal predicates. We perform a detailed manual error analysis on a sample of 100 misclassified instances from the YAGS test set that occur in the training set WaS-L-fntrain. The errors can be classified into several categories that overlap. A large proportion, 47%, of the analyzed errors result from the assignment of incorrect frame labels that correspond to the most-frequent-sense baseline. This indicates a different label distribution in the domain of YAGS compared to the training set. The fact that they are the most frequent sense for a frame may, however, not be the only reason why these labels were selected. Moreover, there are a few instances, 4% of the errors, for which the predicted label is also acceptable. A fairly large proportion, 19% of the errors, are related frames that are typically very similar to the gold frame. In 10% of the misclassifications, the predicted frame could be considered correct in a figurative reading of the frame, for instance, “*building efforts now [went] towards secular*” could be considered a figurative reading of the *Motion* frame. For these, a more appropriate gold frame has been assigned, for instance *Becoming* in the above example. In 6% of the analyzed instances, the FrameNet granularity was very fine-grained or the distinctions were not clearly defined between gold frame and predicted frame: the FrameNet lexicon examples for *Capability* and *Possibility* for instance show a strong overlap.

There are 10% of errors that could be a results of spelling errors or lack of punctuation, which indicates that properties of the user-generated text affect the frame identification with SEMAFOR. In the sentence “*I am trying to figure out a way that i can [make] a wreath out of christmas balls*”, *wreath* is misspelled as *wrath*, inducing an abstract representation of the corresponding predicate *make*, which may lead the system to assigning it the frame label *Causation* instead of the appropriate *Building*. Lack of punctuation appears to cause the system to interpret *there* as the location description for *take* in the sentence “*...if you are [taking] a pill there might me something in it causing your hair to become weak...*”. Thus, *taking* is labeled with the frame label *Bringing* instead of the appropriate *Ingest_substance*.

Other properties of the user-generated question answer data in YAGS can also explain some of the errors: some sentences are very short, lacking appropriate context to disambiguate, or represent a fixed expression, such as the phrase “[*Worked*] for me”. In this sentence, *worked* represents the *Usefulness* frame, but was labeled *Working_on* by the system. To the system, it is not clear, whether the omitted argument, i.e., *who or what worked for me*, refers to an action or a person, which would disambiguate the sentence either to the *Usefulness* frame or to the *Working_on* frame.

Unlike the Twitter-based test set TW, YAGS contains only few UGD-specific terms such as words marked with hashtags. Another difference to TW is that the sentences in YAGS get very long, either because users create long and complex sentences, or because they omit punctuation. This is not the case for the TW test set, because Twitter constrains the length of the tweets to 140 characters. This leads to shorter sentences in TW.

We also observe a frequent use of modal verbs such as *can*, *need*, and *should*, which leads to long-distance dependencies between predicates and their arguments, for instance the *Agent* in the subject position of a raising construction and its predicate. An example for the long distance between a predicate and its direct object can be found in the following sentence: “*Chicharrones de pollo is a typically South American dish, [variations of which] can be [found] all over that continent and up into the Central American peninsula*”. Here, the predicate *found* was misclassified as *Coming_to_believe* instead of *Locating*. The frequent use of modals is a domain-specific property of the YAGS set, because the Yahoo! Answers Manners dataset on which YAGS is based contains *How-to* questions, i.e., users asking for advice, that typically elicit answers containing these modal verbs.

Summary: domain generalization and coverage increase. We find that the increased coverage contributes to improved frame identification performance on out-of-domain test sets. This improves the representation of certain frames relevant to the out-of-domain test sets, such as *Cooking_creation* that only receives a single instance in FNFT-train, but 498 instances in WaS-L. Other domain-specific properties of the datasets could not be captured by our training data generation methods, such as different frame distributions for different domains, or the strong preference for modals in YAGS. Other errors for the UGD-based test sets could most likely be solved by using preprocessing specific to user-generated text, such as spelling correction and improved segmentation in the face of reduced punctuation. We provide a detailed discussion of our results in the next section.

4.5 Discussion

In Section 4.3, we presented the first study that analyzes the domain generalization capabilities of contemporary FrameNet SRL, breaking down the evaluation by subtask and diligently analyzing the results of each subtask.

The evaluation of SEMAFOR on the test sets from various domains revealed that domain adaptation is mainly a problem for the frame identification task, not for the role labeling task: full SRL and role labeling scores on the diverse test sets are similar, or even higher than the scores for the in-domain FNFT-test when gold frames are given, see Table 4.8. This can be explained by the inherent properties of the FrameNet SRL task: FrameNet contains a large number of frame-specific role labels, but once a frame is determined, the number of roles available to the role labeling system are reduced to up to 32, on average less than

11 roles, see Table 4.6, which results in an easier role labeling task, and, if the assigned frame label is correct, improved role labeling scores. The evaluated test sets contain a large proportion of core roles, see Table 4.3, which further reduces the number of roles to choose to an average of 3 and up to 11 core roles. Our evaluation of the SEMAFOR SRL system on various out-of-domain test sets showed that other potential sources of error play a smaller role: the preprocessing required for role labeling, such as parsing, appears to work well on the out-of-domain test sets.

Based on our analysis, low domain generalization is a problem of the frame identification task. The problems are amplified by the low coverage of the test sets by the FrameNet fulltext corpus. As a potential solution to this problem, we studied the benefits of different methods for training data generation to domain generalization for frame identification. The results of this evaluation show that the different methods for training data generation can benefit domain adaptation for semantic role labeling. DistantSRL and annotation projection approaches result in improvements of frame identification accuracy for the out-of-domain test sets. DistantSRL increases the lexicon coverage, providing training instances for previously unseen predicates, and it also increases the instance coverage, providing additional training instances for seen predicates on corpora from different domains. Annotation projection does not increase the lexicon coverage, but increases the instance coverage on corpora from different domains. DistantSRL showed the largest improvements overall, resulting in the highest frame identification scores for YAGS and MASC, and second highest scores for TW-av.

We also observed that relaxing the constraints on syntactic similarity for annotation projection improves the usefulness of this training data generation approach across all test sets. Creating large expansion corpora from annotation projection via relaxed syntactic similarity and by selecting larger top k expansion instances is a promising direction for future work. This might also help with creating training corpora that generalize better to the target domain instead of reproducing the seed corpus, which we observed for annotation projection to three different corpora. Thus another possible route to improvement would be to apply DistantSRL to unlabeled corpora in the target domain, and to use the resulting corpora in combination with explicit domain adaptation methods, such as [Blitzer et al. \(2006\)](#) or [Daumé III \(2007\)](#).

An open issue for future work is to what degree normalization strategies help to improve frame identification on the tests sets based on user-generated data, YAGS and TW-av.

Relation to state-of-the-art frame identification. The state-of-the-art in frame identification is the system by [Hermann et al. \(2014\)](#) which is based on specifically trained embeddings of frame instances and their arguments. The accuracy of their best system is 88.41% for FNFT-test.²⁵ This system has not been evaluated on out-of-domain data, so it is not

²⁵Based on the errata version in <http://www.aclweb.org/anthology/P/P14/P14-1136v2.pdf>.

clear whether it would generalize to the other test sets evaluated here. [Hermann et al. \(2014\)](#) however report that the frame identification model by [Das et al. \(2014\)](#) is likely to overgeneralize on FNFT-test because of its use of a latent variable to establish relationships between predicates and for smoothing over frames for ambiguous lexical units. [Hermann et al. \(2014\)](#) state that their system avoids such generalization. We on the other hand assume that a stronger generalization could be useful for applying the system to out-of-domain data such as our test sets. In order to confirm this assumption, we would need to evaluate the model by [Hermann et al. \(2014\)](#) on our test sets.

In an extension to the work presented in this chapter, [Hartmann et al. \(2017a\)](#) created SimpleFrameId, a frame identification system based on distributed word representations that improves on the SEMAFOR scores out-of-domain and approaches the scores from [Hermann et al. \(2014\)](#) on the in-domain test set, despite using a simpler model. In the next paragraph, we present their system and discuss its domain generalization capacities in comparison to the results reported in this chapter.

Out-of-domain evaluation of state-of-the-art frame identification. The original goal of [Hartmann et al. \(2017a\)](#) was to re-implement the model by [Hermann et al. \(2014\)](#) in order to perform out-of-domain tests, but initial attempts were not successful: it turned out that the input feature space of their embeddings based on syntactic paths is very sparse, which harmed the system performance. Instead, [Hartmann et al. \(2017a\)](#) developed a simpler model and found that it approaches the in-domain performance of [Hermann et al. \(2014\)](#) and improves on SEMAFOR in the out-of-domain evaluation on YAGS, TW-av, and MASC.

Their system SimpleFrameId uses pre-trained word embeddings to represent predicates and creates embeddings based on the FrameNet fulltext training set to represent the *contexts* of predicates. Predicate and context embeddings are concatenated and used as input to the classifier, a two-layer neural network, thus creating a simple sense representation. The best-performing setting uses simple word-based context embeddings and does not require syntactic parsing to create embeddings.

[Hartmann et al. \(2017a\)](#) report accuracy of 87.63% on the in-domain test set, compared to 82.09% reported for SEMAFOR for all parts-of-speech. For the out-of-domain test sets, they also report a performance increase compared to their SEMAFOR model. Accuracy increases by 2.5 percentage points for YAGS, 6.5 percentage points for TW-av, and 15.47 percentage points for MASC, resulting in accuracy of 62.51%, 68.67%, and 55.09% respectively. Note that they test on all parts-of-speech and split YAGS into a development split YAGS-dev comprising 1,000 predicates and a test split YAGS-test comprising 2,093 predicates. Thus, their results on YAGS-test cannot be compared directly to our results on the full YAGS presented in Sections 4.3.2 and 4.4.3.

To provide such a comparison, we test the version of SimpleFrameId that uses word-based context embeddings on verbal predicates. We also test the best configuration of SE-

MAFOR frame identification, i.e., extended with DistantSRL, on the smaller YAGS-test. The performance of SimpleFrameId on test verbs is 2.1 percentage points higher for YAGS-test, 2.5 percentage points higher for TW-av, and 3.7 percentage points higher for MASC compared to the retrained SEMAFOR. The smaller performance increase for MASC on test verbs (compared to the larger one reported for all parts-of-speech) can be explained by the higher frame identification performance of the base SEMAFOR system for verbal predicates in MASC, see Tables 4.7 and 4.8. The performance improvements of SimpleFrameId to the unexpanded SEMAFOR are roughly twice as high as the improvements reached via training data expansion with DistantSRL. This shows that the deep learning-based model shows higher domain generalization than our SEMAFOR model expanded with DistantSRL, but there is large room for improvement for both systems: the out-of-domain frame identification performance as reported by Hartmann et al. (2017a) is still considerably lower than the in-domain performance, and also lower than the human performance observed for YAGS and TW, see Section 4.2.5.

During the experiments for Hartmann et al. (2017a), we also found that simply adding the automatically labeled corpora to the FrameNet fulltext training set does not lead to performance improvements. This is another indication that specialized methods may be required to efficiently use large-scale, noisily labeled training data.

Large amounts of training data. Training the SEMAFOR frame identification system on the huge WaS-XL set that contains up to 40,000 training instances per sense is prohibited because it would require too many resources. The system is not prepared to deal with such large amounts of data. We therefore randomly sample smaller subsets of WaS-XL with up to k instances per sense for k up to 640, finding that frame identification accuracy improves with a larger number of instances.²⁶

There are other ways to reduce the large amount of training instances. In order to better represent the target domain, we could subsample the training instances to create a corpus with a similar perplexity to the target domain. This approach has been successfully evaluated by Pavlick et al. (2015a) for domain adaptation applied to paraphrasing.

Another option would be to evaluate approaches to frame identification that are equipped to deal with large amounts of training data. Deep learning methods can for instance be used to create dense feature representations from large amounts of training data. Our experiments while working towards Hartmann et al. (2017a) showed that simply adding large, noisily labeled training data to the SimpleFrameId system does not improve results. Instead, more complex deep learning representations of frames and senses may be required. In Hartmann et al. (2017a), we propose using advanced context representations from *con-*

²⁶The accuracy scores obtained for the high-precision WaS-L corpus that does not require sampling due to smaller size are still higher than the scores for the WaS-XL samples, but we could create a larger version of the WaS-L corpus that would also require us to select a sample by using a larger seed corpus.

text2vec (Melamud et al., 2016) or adapting embedding representations to particular domains (Taghipour and Ng, 2015).

Following a similar notion, Chen et al. (2014) use a neural network-based language model, the skip-gram model (Mikolov, 2012), to learn sense-based representations. They evaluate these models for several word sense disambiguation tasks based on WordNet, including the evaluation of domain-specific word sense disambiguation. They find that their model outperforms state-of-the-art supervised models in domain-specific word sense disambiguation for WordNet. Similar approaches could be applied to FrameNet frame identification to improve the domain generalization.

4.6 Summary of Chapter 4

In this chapter, we presented an analysis of the domain generalization capabilities of contemporary FrameNet semantic role labeling. While the need for domain adaptation is widely acknowledged for PropBank semantic role labeling, cross-domain performance is for instance evaluated in shared tasks, there are no recent analyses of the domain adaptation capabilities of FrameNet semantic role labeling. FrameNet SRL systems are evaluated on the same, in-domain test set. Therefore, we evaluated the SEMAFOR SRL system on test sets from a various domains, including the sub-domain of user-generated text, and presented a detailed analysis of the cross-domain performance of the two main subtasks of FrameNet SRL, frame identification and role labeling, and its effects on the full SRL task.

FrameNet test datasets from other domains are scarce. As a prerequisite for our analysis, we therefore collected available FrameNet-labeled test sets from various domains and created the new test set YAGS, a large frame- and role-labeled test set based on user-generated question-and-answer data from Yahoo! Answers. We then evaluated the open-source system SEMAFOR on the different test sets and discovered that domain adaptation is mostly a problem for the frame identification step: when evaluating role labeling using gold frame labels, the performance of SEMAFOR on the out-of-domain test sets is similar to the performance on the in-domain test set.

The problem of domain adaptation goes hand in hand with the problems of training data coverage, instance coverage, and lexicon coverage, because many test instances in the out-of-domain test sets do not occur in the training sets. Therefore, we evaluated the benefits of different methods of training data generation, some of which extend the lexicon, i.e., the paraphrasing-based FN+, some of which only extend the number of training instances, i.e., annotation projection, and some of which do both, i.e., our knowledge-based approach DistantSRL. This includes the first evaluation of the FN+ dataset from Pavlick et al. (2015b) in a frame identification task, and the first large-scale evaluation of the annotation projection approach by Fürstenaу and Lapata (2012), i.e., evaluating their method as compared to a system trained on a reference training set instead of a low-resource evaluation scenario.

To perform the domain generalization experiments, we retrained the SEMAFOR frame identification model on the automatically generated training data, alone and in combination with the FrameNet fulltext training set, and evaluated the resulting frame identification systems on our battery of test sets from different domains. We find that the automatically labeled corpora can help to improve frame identification performance on the out-of-domain test sets, in particular for the user-generated test sets and the polysemous MASC test set. These experiments further prove the benefits of the DistantSRL approach introduced in Chapter 3: training on the high-precision WaS-L, which we created via DistantSRL, in combination with the FrameNet fulltext training set leads to the best frame identification results on the out-of-domain test sets YAGS and MASC. Frame identification performance is nevertheless worse for the out-of-domain test sets compared to the in-domain test set, best accuracy scores for the test sets based on user-generated text are still 0.16 points lower than scores for FNFT-test. Further work is required to close this gap. We propose to evaluate frame identification models that deal with large amounts of automatically generated training data, for instance based on deep learning, in future work, and to use our automatically generated data in combination with traditional domain adaptation methods. Initial experiments by [Hartmann et al. \(2017a\)](#) show that deep learning methods, which can create dense feature representations from large amounts of data, may additionally provide improved domain generalization.

In summary, the main contributions of this chapter are:

- YAGS, a new, substantially-sized FrameNet-labeled test set based on user-generated community questions and answers, see Appendix A; the annotation guidelines used for the creation of YAGS are reported in Appendix B.
- A detailed analysis of the domain generalization capabilities of open-source FrameNet semantic role labeling that identified the frame identification step as critical with respect to domain adaptation.
- Experiments on the contributions of various methods for training data generation to the domain adaptation of frame identification. This includes the first experimental evaluation of FrameNet+ ([Pavlick et al., 2015b](#)) and the first large-scale evaluation of monolingual annotation projection ([Fürstenu and Lapata, 2012](#)) for the task of frame identification. We also publish some of the projected corpora, see Appendix A.

Our experiments showed that additional domain adaptation is required to improve the domain generalization of frame identification. This will also enhance the performance of full SRL that depends on the frame identification step for FrameNet semantic role labeling. In the next chapter, we summarize the findings of our experiments and discuss directions of future work towards improved domain generalization of FrameNet SRL.

CHAPTER 5

Conclusion

This dissertation presents a series of experiments on knowledge-based methods to FrameNet semantic role labeling and their application to domain adaptation. The presented work is motivated by a lack of FrameNet coverage observed for English – and more severely for other languages – which hampers the large-scale application of semantic role labeling in tasks that require advanced semantic analysis and natural language understanding, like machine translation, or question answering. The main research question is: can the comprehensive integration of lexical knowledge bases benefit semantic role labeling in the contexts of domain adaptation and adaptation to other languages?

In this chapter, we first summarize the work we undertook to answer this question, starting with the above-mentioned motivation, followed by our work on a) automatically integrating and standardizing lexical-semantic knowledge bases, b) using the resulting linked lexical knowledge base for the generation of frame- and role-labeled training data for English and German, and c) studying the benefits of the generated training data to domain adaptation for FrameNet semantic role labeling for English.

We then present concrete suggestions for the extension of our work and discuss open issues, giving an outlook on how to improve FrameNet semantic role labeling in the direction of open-domain systems: since this work presents solutions to the insufficient lexicon coverage and training data coverage of FrameNet (that hampers the performance of FrameNet semantic role labeling), a major remaining question is how to deal with the insufficient *model coverage* of FrameNet, i.e., the problem that many real-world situations and domains are not yet represented in FrameNet. We discuss possible solutions to this problem and show how they could build upon the resources and methods developed in this work.

5.1 Summary and Contributions

In Chapter 1, we motivate our work on using linked lexical knowledge bases to enhance the automatic semantic analysis of texts, and more specifically FrameNet semantic role labeling: lexical knowledge bases are valuable knowledge sources and we do not want to forgo the knowledge painstakingly encoded in them by experts, in particular for semantic information types like semantic predicate argument structure and semantic tasks like semantic role labeling. At the same time, FrameNet semantic role labeling suffers from a lack of coverage with respect to the senses represented in the lexicon and with respect to the training instances available for supervised training of semantic role labeling systems. These coverage problems are particularly severe for languages other than English. For some of these, lexical knowledge bases like FrameNet exist, but they are typically smaller, and are often not equipped with large frame- and role-labeled corpora.

In this thesis, we propose a series of steps to solve these issues from a knowledge-based perspective: first linking lexical knowledge bases on the level of senses and predicate argument structure and standardizing them to enhance their coverage, which allows us to translate FrameNet to other languages, second using the resulting linked lexical knowledge base in a new method for large-scale training data generation, and third using the created training data to facilitate domain adaptation for FrameNet semantic role labeling. We dedicated a chapter of this thesis to each of the steps.

Chapter 2 builds the foundations for our work on knowledge-based supervision for semantic role labeling: in this chapter, we present our contributions to linking lexical knowledge bases and resource standardization. Sense-level links between lexical resources like FrameNet and VerbNet provide semantic interoperability between these resources: word senses in one resource can be enriched with (often complementary) information from another resource. By exploiting the information on synonyms in one resource, like WordNet, the lexicon coverage of the other linked resource, like FrameNet, can be extended.

The main contribution of this chapter results in such a coverage extension for English and for German: we present a novel method for creating a FrameNet in various languages based on the automatic alignment of FrameNet and the English Wiktionary, using Wiktionary as an interlingual connection for bootstrapping FrameNet lexica for other languages. We evaluate this method on the example of German, effectively creating a larger FrameNet knowledge base for English and German called FNWKde. To provide representational interoperability between semantic knowledge bases that – like FrameNet – provide models of semantic predicate argument structure, we develop a standardized model of lexical-semantic knowledge bases like FrameNet as part of UBY-LMF. This model includes a representation of frame- and role-level links between different lexicons. We convert the linked lexical knowledge bases we created, e.g., FNWKde, and existing lexical knowledge

bases to this representation, resulting in UBY_{FN} , a linked lexical knowledge base centered around FrameNet.

In Chapter 3, we use UBY_{FN} to address the lack of labeled training data for FrameNet semantic role labeling for English and German. We present DistantSRL, a knowledge-based method for the creation of sense- and role labeled training data that follows the paradigm of distant supervision: the lexical information in the linked lexical knowledge base is used to guide large-scale transfer of frame and role labels to unlabeled corpora. In experiments on frame identification, role classification, and by retraining SEMAFOR on the resulting training data, we show that DistantSRL creates training sets of high quality that are complementary to the manually labeled data, e.g., the FrameNet fulltext corpus and SALSA. The presented method is evaluated for verbal predicates in English and German, and for the FrameNet role inventory, but can be adapted to other parts-of-speech, languages, and role inventories.

In Chapter 4, we analyze the benefits of automatic training data generation to domain adaptation for FrameNet semantic role labeling for English. While domain adaptation has been studied for a long time for PropBank semantic role labeling, it has been neglected for FrameNet semantic role labeling: the current measuring-stick for FrameNet semantic role labeling is an in-domain test set, basically a held-out set from the FrameNet fulltext corpus, and shows a similar distribution to the training corpus. There are only few out-of-domain test sets for FrameNet semantic role labeling. Therefore, we create a large frame- and role-labeled test set based on user-generated question-and-answer data from Yahoo! Answers. We use this test set together with the other available test sets to assess the domain generalization capabilities of FrameNet semantic role labeling: our analysis shows that the performance of the open-source semantic role labeling system SEMAFOR is significantly worse on out-of-domain data for the task of frame identification, but the task of role labeling does not suffer from domain adaptation. We then experiment with different methods of training data generation with respect to their potential for mitigating the domain adaptation problems for frame identification. This study includes a comparison of DistantSRL, monolingual annotation projection based on structural alignment (Fürstenaу and Lapata, 2009), and paraphrasing-based lexicon extension in FrameNet+ (Pavlick et al., 2015b). We retrain SEMAFOR on the aforementioned training data and evaluate it on the various in-domain and out-of-domain test sets. The results show that the coverage extensions provided by DistantSRL and by annotation projection improve frame identification performance, in particular on the test sets based on user-generated texts, further proving the benefits of DistantSRL.

In the following section, we discuss potential extensions of our work, and present ideas on how to tackle the coverage problem associated with FrameNet that we did not address in this work, namely the problem of model coverage, i.e., the problem that many real-world situations and domains are not yet represented in FrameNet. We also discuss how the

proposed solutions to extend the model coverage could build upon the methods and linked lexical knowledge bases presented in this thesis.

5.2 Open Issues and Outlook

In each chapter of this thesis we already briefly mentioned opportunities and challenges for future work. There are several directions worth exploring. We first propose concrete extensions to the methods and experiments introduced in the course of this dissertation that promise to further enhance the coverage and quality of the proposed methods. We then target future work on a larger scale: research that could build upon the results of this work to address problems that are outside the scope of this thesis. We discuss how the proposed methods could be used to solve the unsolved problem of model coverage for FrameNet, which includes a discussion of using linked lexical knowledge bases as a basis for sense representations based on deep learning.

5.2.1 Directions for Further Work

Automatic alignments on the predicate argument structure-level. To further enhance the coverage of the LLKB UBY_{FN} , we propose automatic alignments on the level of predicate argument structure, i.e., frames and roles, which would also be helpful to increase the role coverage of DistantSRL: the automatic linkings could be used to expand the predicate- and role-level linking between FrameNet and VerbNet, and the linking between VerbNet and proto-frames and their roles from SALSA. Thus, it could contribute to further improve the quality and coverage of training data generation with DistantSRL, see Chapter 3. This is an extension to our work on automatic linking of lexical knowledge bases on the sense level to enhance the coverage of FrameNet in Chapter 2.

We evaluated a prototype for such an alignment that is inspired by the annotation projection approach and based on optimizing the similarity of frames and their roles between FrameNet and VerbNet. The approach operates on the similarity of the frame and role descriptions in the resources and role fillers in FrameNet-labeled texts. The difficulty of such an approach is tuning the trade-off between precision and recall. Besides the evaluated resource-based linking approaches, automatic linking approaches using instance-based alignment methods like monolingual annotation projection could be used: by projecting, for example, VerbNet roles onto FrameNet-labeled corpora, new predicate- and role-level links could be inferred. A similar approach has been employed by [Lopez de Lacalle et al. \(2016\)](#), who use automatic SRL systems instead of the proposed projection approach to increase the number of predicate and argument links in SemLink.

Extensions to DistantSRL. Our method for knowledge-based training data generation, DistantSRL, that was presented in Chapter 3 could be extended to support:

- training data generation for other parts-of-speech. Our implementation was focused on verbs, as we expect the largest gains for semantic role labeling from adding verbal training data.
- training data generation for other semantic role inventories that are linked to VerbNet via SemLink, e.g., PropBank, or could be linked to VerbNet, e.g., QA-SRL (He et al., 2015).
- training data generation for other semantic tasks that follow a template-based structure similar to predicate argument structure, for instance template-filling in information extraction in the TAC and MUC shared tasks (Ellis et al., 2015; Sundheim, 1991), and related tasks in event extraction (Kim et al., 2009).
- training data generation for other languages that are represented well in FNWKxx, for instance Finnish, Russian, Swedish, or Spanish.

The experiments with DistantSRL for German led to the conclusion that a larger linked lexical knowledge base centered around SALSA is needed to improve the coverage of the automatically labeled corpus for German. An obvious resource for expansion would be GermaNet, the German counterpart to WordNet. Another opportunity for expansion would be to link the SALSA-specific frames and roles to other sense and role inventories, to enhance the coverage of DistantSRL for German. In Hartmann et al. (2017b) we created a prototype for a SemLink-like resource for German that provides such additional links. This prototype consists of a small corpus labeled in parallel with FrameNet-style, VerbNet-style, and PropBank-style role labels.

Furthermore, there are interesting options for combining the different methods for training data generation: the automatic frame labeling with DistantSRL can be used as a filtering step for other training data generation approaches, for instance for adapting annotation projection to deal with previously unknown predicates. Instead of using lexical similarity based on WordNet, as proposed by Fürstenau and Lapata (2012) to reduce the number of candidate pairs, the automatic frame labeling stage of DistantSRL could be used. Additionally, sentences from FrameNet+ (Pavlick et al., 2015b) could be used as seeds for DistantSRL or annotation projection approaches. This is motivated by the observation that not all expansions in FrameNet+ are of high quality: many of the paraphrases that received the lower confidence rating of 3 out of 5 in the crowdsourcing evaluation (see Section 4.4.1) do not appear natural to the reader. The evaluation results in Chapter 4 also suggested that the high-confidence paraphrases result in better frame identification on the test sets. To enhance the quality of the training sentences for the lower-confidence paraphrases, we could

use DistantSRL or annotation projection to gather further, real-life example sentences for the new lexical units provided by these paraphrases in FrameNet+.

Adaptation to automatically labeled training data. The training data generation methods introduced in Chapter 3 are able to generate large amounts of training data. The properties of the generated data vary. DistantSRL, for instance, generates large amounts of noisy data that are only partially labeled with roles.

Conventional open-source supervised semantic role labeling systems like SEMAFOR can, on the one hand, benefit from the added training data – additional training data would also be beneficial for state-of-the-art semantic role labeling systems (FitzGerald et al., 2015), on the other hand, the open-source systems do not expect large amounts of noisily labeled training data. As a result they do not process them efficiently. We first consider the problem of dealing with large amounts of training data and propose two ways to solve this problem:

- Subsampling of the training data according to the sense distribution of the target domain, similar to the method suggested by Pavlick et al. (2015a) for domain adaptation in paraphrasing.
- Using models of frame identification and role labeling that efficiently deal with large amounts of training data, for instance via representation learning of dense, low-dimensional feature representations, also called embeddings, of frames and senses, similar to the models proposed by Yang et al. (2015b) for PropBank semantic role labeling. For frame identification, the method by Chen et al. (2014) that uses neural network-based representations to enhance cross-domain WordNet WSD could be adapted to FrameNet frames.

Further work that relates to the second solution are the deep learning approaches to frame identification by Hermann et al. (2014) and Hartmann et al. (2017a), and the approach to create sense embeddings for WordNet proposed by Rothe and Schütze (2015) called AutoExtend. The method by Hermann et al. (2014) incorporates the syntactic structure of the training examples into an embedding model: for each instance of a FrameNet predicate, they create a structured vector that contains a slot for each dependency available for the predicate. This slot is filled with an embedding representation of the slot filler. We anticipate that such a model may overfit to the given training data, and therefore suggest to use models that rely less on the syntactic structure of the training data and focus on a purely semantic representation of word senses. This for instance applies to AutoExtend (Rothe and Schütze, 2015) and SimpleFrameId (Hartmann et al., 2017a).

AutoExtend is a method to create sense embeddings from word embeddings exploiting the lexical relations in the sense inventory. Rothe and Schütze (2015) use these as additional features to improve word sense disambiguation for WordNet. This method could also be adapted to FrameNet. Besides the frame hierarchy in FrameNet, sense-level links to other

resources in the linked lexical knowledge base UBY_{FN} and their relational hierarchy could be exploited in this setup.

SimpleFrameId uses a simple, but efficient frame identification model that relies on word-based embedding representations and a two-layer neural network classifier. To create predicate-specific embeddings, Hartmann et al. (2017a) concatenate word embeddings for semantic predicate lemmas and word-based context embeddings learned from the predicate instances in the FrameNet fulltext training set. They report the best frame identification results to date on out-of-domain test sets and results competitive to the system by Hermann et al. (2014) on the in-domain test set. During our experiments for Hartmann et al. (2017a), we found that simply adding the automatically labeled training data to the FrameNet fulltext training set does not increase frame identification accuracy. This result indicates that specialized methods may be required to efficiently use the automatically labeled corpora for semantic role labeling.

A characteristic property of the automatically labeled data generated with DistantSRL is that they are sparsely and noisily labeled. Current FrameNet SRL systems are developed on a small set of high-quality manually labeled data. There are machine learning methods that can be employed to achieve better classification performance on noisily labeled data, e.g., Natarajan et al. (2013) or Chen et al. (2011). These problems are closely related to domain adaptation, since the sparse and noisily labeled data also represent a different label distribution than the application domain, i.e., the automatically labeled data are *biased*. For the role labeling step in DistantSRL, we for instance observe a bias towards core FrameNet roles. We expect greater contributions of DistantSRL to SRL performance, once the automatically labeled data are used efficiently in an adapted machine learning setup.

Improved domain adaptation. In Chapter 4, we used the training data created automatically with DistantSRL to improve the performance of SEMAFOR frame identification on out-of-domain test sets based on user-generated text. The performance does, however, not reach the accuracy of the in-domain evaluation. There are two main reasons, the need for additional domain adaptation, and errors caused by preprocessing user-generated test sets. The former could be addressed by employing explicit domain adaptation methods, e.g., applying DistantSRL to unlabeled corpora in the target domain and using the resulting corpora in combination with explicit domain adaptation methods, such as Blitzer et al. (2006) or Daumé III (2007). The latter could be addressed by strategies to normalization and domain specific preprocessing of user-generated text (Eisenstein, 2013; Baldwin et al., 2015). Both proposed strategies in combination should lead to improved domain adaptation of FrameNet frame identification, thus improving full FrameNet SRL.

We expect that state-of-the-art deep learning methods can be harnessed to further improve the domain generalization of frame identification. While Hartmann et al. (2017a) also report a large disparity between in-domain and out-of-domain frame identification perfor-

mance for their recently introduced frame identification system based on deep learning, they do not find improvements when applying spelling correction to the YAGS test set, indicating that frame identification based on embeddings can generalize across the characteristic properties of user-generated text.

The out-of-domain performance of embedding-based systems could further be improved by advanced methods for integrating noisily labeled data, as discussed in the previous paragraph, using more complex embedding representations (Melamud et al., 2016), and efficient use of embeddings for frame identification (Chen et al., 2014; Iacobacci et al., 2016), for instance by using domain-adapted predicate embeddings (Taghipour and Ng, 2015).

5.2.2 Outlook: FrameNet Model Coverage

In this subsection, we describe open issues and potential extensions of our work that follow from the original research questions and motivation for this thesis, but consider a wider horizon of future work than the concrete suggestions from the previous subsection. They are concerned with the unsolved problem of FrameNet extension on the ontology level, the level of modeling real-world concepts, and with further work in utilizing linked lexical knowledge bases for NLP.

Extension of model coverage. This work dealt with the coverage extension of FrameNet on the lexicon level (Chapter 2) and on the level of frame- and role-labeled training data (Chapter 3). To further improve semantic role labeling performance and progress towards domain-independent systems for FrameNet semantic role labeling, automatic extension on the level of the covered situations would be required, which means attempting to solve the model coverage problem.

This could be done by integrating FrameNet or UBY_{FN} in a larger semantic knowledge base that provides additional coverage. There are several potential candidates for such an integration: a) large, user-generated semantic knowledge bases like Wikidata or Freebase as suggested by Sergieh and Gurevych (2016), b) automatically created or induced semantic knowledge bases created via frame induction (Poon and Domingos, 2009; Titov and Klementiev, 2012; Cheung et al., 2013) or the automatic acquisition of n-ary relations (Titov and Klementiev, 2012; Krause et al., 2015).

Ideally, such an integration would not only cover the sense level, but also the level of semantic predicates and roles. For knowledge bases modeling semantic relations like Wikidata, Freebase or automatically acquired relation databases, FrameNet roles can be mapped to *relations*. Methods for frame induction do not provide meaningful labels for frames and roles and could benefit from the acquisition of meaningful labels for parts of their inventory that can be obtained by integrating FrameNet with their induced predicates and roles.

The integration of FrameNet with these knowledge bases could build upon the work presented in this thesis: methods for the integration of FrameNet with larger semantic knowledge bases could include 1) advanced variants of the automatic approaches for sense alignment used in Chapter 2, for instance [Pilehvar and Navigli \(2014\)](#), 2) the automatic alignment on the predicate and role level as proposed in the previous subsection, but also instance-based alignments of corpora labeled with FrameNet labels and semantic relations from the other knowledge bases. The latter could use monolingual annotation projection following [Fürstenaу and Lapata \(2009\)](#) to project labels from each of the knowledge bases to the same unlabeled corpus. The resulting parallelly labeled corpus could be used to infer alignments between the projected labels.

Linked lexical knowledge bases for deep learning. In this work, we used linked lexical knowledge bases for a distant supervision approach to training data generation. There are different other ways of using linked lexical knowledge bases for natural language processing, and a promising one is to use them as a basis for structured embeddings of knowledge base information. Structured embeddings of knowledge bases have been evaluated for knowledge bases representing facts, e.g., Freebase, with the goal to support knowledge base completion in the context of relation extraction ([Bordes et al., 2013](#); [Yang et al., 2015a](#)). While these approaches also used a lexical knowledge base, WordNet, to evaluate their methods, previous work in this area did not focus on semantic information like semantic predicates and roles, which we consider a promising direction for future research. UBY_{FN} as a linked lexical knowledge base with a rich set of alignments provides a good foundation for further exploration of structured embedding approaches for lexical-semantic information.

5.3 Closing Remarks

This dissertation assumes that lexical knowledge bases are invaluable resources for natural language processing and enable complex semantic analysis such as semantic role labeling. At the same time, coverage problems and lack of training data hamper the success of automatic natural language processing systems based on these resources, for instance FrameNet semantic role labeling. We take the stance that the impact of lexical knowledge bases increases if they are integrated on several semantic levels to form large, high-coverage linked lexical knowledge bases, which we call LLKBs.

In this dissertation, we showed that LLKBs can be used to enhance complex semantic tasks like semantic role labeling without relying on manually labeled corpora. We covered all the steps involved in this undertaking from standardizing and integrating semantic knowledge bases, to using them in a distant supervision setup for the generation of automatically labeled training data, and finally to evaluating their potential to enhance semantic role labeling and domain adaptation for semantic role labeling.

This journey also opened up new research questions that we discussed in the current chapter. These include how to efficiently use the large and noisily labeled automatically labeled data for semantic role labeling, how to further improve the integration of linked lexical knowledge bases on various semantic levels, and how to use linked lexical knowledge bases to complement increasingly popular machine learning methods that are based on large-scale data analysis, e.g., deep learning methods. These are interesting and challenging questions to solve. We hope that other researchers will build upon the results presented in this dissertation in the future, and make use of the open-licensed datasets and software created during the work on this dissertation.

CHAPTER A

Appendix A: List of Resources

This appendix lists resources, i.e., datasets and lexical knowledge bases, and open-source software that we created or contributed as part of this thesis. It provides a brief description and links to the corresponding websites.²⁷

Linked Lexical Knowledge Bases

During the work on this dissertation, we created or contributed to the modeling and the creation of several large linked lexical knowledge bases.

UBY. The linked lexical knowledge base UBY contains the major lexical knowledge bases for English and German and connects them via sense-level alignments, see also Section 2.8. The creation of UBY was a group effort (Gurevych et al., 2012a; Eckle-Kohler et al., 2012). The present author’s contributions to UBY and the lexicon model UBY-LMF are described in Section 2.7. The following URLs link to previously published UBY databases and documentation of the UBY-LMF model.

<https://www.ukp.tu-darmstadt.de/data/lexical-resources/uby/>

<https://www.ukp.tu-darmstadt.de/data/lexical-resources/uby/uby-lmf/>

FNWKxx and FNWKde. FNWKxx and FNWKde are contained in a single database in UBY-LMF format. FNWKxx contains the sense-level alignment of FrameNet and the English Wiktionary presented in Section 2.5, FNWKde adds a sense-level alignment to the German Wiktionary, allowing us to link German Wiktionary senses to FrameNet. For a description of FNWKde, see Section 2.6 or Hartmann and Gurevych (2013b). The following website also contains additional materials that show example entries in FNWKde.

<https://www.ukp.tu-darmstadt.de/fnwkde/>

²⁷Alternative access link to all datasets: <http://www.silvanahartmann.de/publications/dissertation/data/>.

UBY_{FN}. UBY_{FN} integrates FNWKde with other lexical knowledge bases, WordNet, VerbNet, PropBank, SALSA, and GermaNet. It is also the first UBY database that contains predicate argument structure links extracted from SemLink. They connect VerbNet, FrameNet, PropBank, and SALSA on the level of semantic predicates and roles. We publish a UBY database containing those resources in UBY_{FN} that are subject to open licenses and provide conversion scripts to create the full database.

<http://www.silvanahartmann.de/data/ubyfn/>

Gold Standard Datasets

During the work on this dissertation, we created two large gold standard datasets. We publish the annotated datasets for research purposes.

Word sense alignment between FrameNet and Wiktionary. We created a gold standard of 2,789 sense pairs between FrameNet and the English Wiktionary. For each pair, annotators decided whether the senses in the pair represent the same meaning, and should therefore be aligned, or not. The gold standard was used in the creation of a sense alignment between FrameNet and the English Wiktionary (Hartmann and Gurevych, 2013b), see also Section 2.5.

<https://www.ukp.tu-darmstadt.de/fnwkde/>

YAGS – FrameNet-labeled test dataset based on user-generated questions and answers. YAGS is a FrameNet-labeled gold standard based on user-generated community questions and answers that was published as part of Hartmann et al. (2017a). It contains 3,091 frame and 6,144 role annotations on 55 questions and answers from the Yahoo! Answers Manner Questions dataset published by the Yahoo! Webscope program.²⁸ The creation of YAGS is described in detail in Section 4.2. The annotation guidelines used to create YAGS are shown in Appendix B. To agree with the Webscope license, we do not publish the question and answer texts, but only the added annotations in a stand-off format. We provide JAVA routines to connect them to the corresponding tokens in the text.

<https://www.ukp.tu-darmstadt.de/out-of-domain-framenet-srl/>

²⁸<https://webscope.sandbox.yahoo.com>

Automatically Labeled Corpora

We also publish several of the large automatically labeled training corpora labeled with FrameNet frames and roles.

Corpora created with DistantSRL. We publish the automatically frame- and role-labeled corpora WaS-L, WaS-XL, and WaSR-XL created with DistantSRL as part of [Hartmann et al. \(2016\)](#) and presented in Sections 3.4 and 3.5. The following page also links to the code for role-based VerbNet role labeling used in the creation of the corpora.

<https://www.ukp.tu-darmstadt.de/knowledge-based-srl/>

Corpora created with annotation projection. We publish some of the corpora created with annotation projection in Section 4.4, specifically the corpora based on ukWAC.

<https://www.ukp.tu-darmstadt.de/automatically-labeled-srl-corpora/>

Open-source Software

The software for the creation and the API access to the linked lexical knowledge base UBY, as well as the UBY-LMF model DTDs are contained in a single GitHub repository. Methods for reading and writing specific datasets are contributed to DKPro Core.

<https://dkpro.github.io/dkpro-uby/>

<https://dkpro.github.io/dkpro-core/>

CHAPTER B

Appendix B: Annotation Guidelines of FrameNet Annotation Study

This Appendix contains the annotation guidelines that were used to create the YAGS test dataset, a FrameNet-labeled dataset based on data from the community question-and-answer forum Yahoo! Answers. They were originally titled *FrameNet Annotation on User-generated Content Guidelines (2014/08/08)* and contain contributions by Orin Hargraves who supervised the annotators and performed adjudication on the study. The guidelines contain directions for annotating frames and roles that are closely intertwined with instructions on how to perform annotations in WebAnno 2.0. They assume linguistic experience of the annotators and some experience with related annotation tasks such as word sense disambiguation or PropBank annotation. Note that there is an enhanced version of WebAnno, WebAnno 3.0 (Eckart de Castilho et al., 2016), the creation of which was informed by the experiences from the presented annotation study. This new version of WebAnno offers better support for frame and role annotation which is mirrored in increased annotation speed and reduced need for postprocessing. We recommend to use the enhanced version of WebAnno in future studies and to adapt the annotation guidelines presented below accordingly.

The following guidelines reflect the original study. Frame and role labeling are performed in a single step by the same annotator, because decisions on available roles inform and can potentially correct the decision for the frame labels.

1 General Information

The goal of this annotation study is to create a gold standard dataset of FrameNet annotations on a particular type of text, namely user-generated questions and answers from the web. The annotation task includes the classic setup for FrameNet-based role labeling: given a word highlighted as a predicate candidate, first decide whether the predicate candidate

is associated with a FrameNet frame. Basically, this is word sense disambiguation using FrameNet as a sense inventory. If you can identify a frame, you need to identify the representations of the semantic roles of the frame in the text (they are called “frame elements” in FrameNet terminology). This is a two-step approach: first, the phrases that represent the semantic roles need to be identified, then the appropriate role labels need to be assigned to them. The phrases carrying roles typically are arguments of the predicate, but can also be adjuncts or other modifiers. In the example in Figure B.1 below, optional roles (often attached to adjuncts) are marked yellow with dotted lines, while obligatory roles are marked green; the frame is marked light blue with a dotted background:



Figure B.1: Example frame and role annotation.

This task is similar to other role labeling tasks – with the exception that FrameNet has a very large inventory of frame-specific roles and often makes fine distinctions between the roles. This makes FrameNet annotations challenging. It also makes it necessary to refer to the definitions of FrameNet frames and frame elements when annotating the frame and role labels.

The genre of text we are working on contains some special properties. First, it contains more questions than other text types, which leads to unfamiliar situations in annotations. You may, for instance, want to annotate the question word “How” as a representation of the “Manner” role of the predicate “deliver” with frame “Delivery” in the sentence “How do they deliver newspapers in Germany?”. Second, the user-generated text contains the use of colloquial language, abbreviations and means for expressing emphasis typically used by web-users. It may also contain unusual orthography and grammar, or omissions. The goal is to annotate as many of the presented targets as possible, even those containing errors. However, if you cannot understand a sentence because of missing words or spelling errors, the annotation schema provides a way to mark this.

2 Using WebAnno

- WebAnno can be accessed via web interface.
- Please use Google-Chrome or Safari as a web-browser, as WebAnno is optimized to work with them.

- Each annotator will receive their own login name and password to login to WebAnno.
- WebAnno url: <url>
- Your web browser may complain about an invalid certificate of the website. Ignore this complaint and access the website anyway.

FrameNet Annotation in WebAnno

3.1 Annotation Workflow

- Log in to WebAnno.
- Click on “Annotation” on the welcome screen.
- Select <projectname> as “Project” (there should be only one project) and a “Document” to annotate, click on “open”.
- Adjust the settings as described in 3.2 and perform annotations as described in 3.3.
- When you need to interrupt the annotation, you can simply logout. WebAnno will store your progress in the background (in a database on the web-server) and load your previous annotations when you open the document again.
- When you’re finished with annotating a document, click on “Done” in the “Workflow” box in the upper right. (Use “Done” with care: you will not be able to edit the document after this step. If you click on “Done” by mistake, the project administrator (<Admin>) will be able to undo this.)

3.2 Settings

Before starting the annotation, you can adjust some display settings, as follows:

- Click on “Settings” in the “Document” box.
- Untick all annotation layers except for “Predicate”, “PredicateRel”, “FrameNet”, and “FrameNetRel”.
- Enter the number of sentences that you would like to see displayed on your screen at the same time. (The whole document may not fit on the screen.). Displaying more than one sentence allows you to see the wider context of the sentence in question, which may help disambiguation. You may want to start with 5 or 6 sentences and see how that works.

- It is recommended that you untick “auto-scroll document”.
- Decide whether you want to use “the same color for tags in a layer”. You may want to play with this setting to see what you prefer.

3.3 Annotation Layers

- There are two span annotation layers relevant for the annotation study: “Predicate” and “FrameNet”. And there are two arc annotation layers that can be used to link annotations of the span layers (“FrameNetRel” and “PredicatePart”). Their purpose and their properties are described in the following.
- This project uses FrameNet 1.5, the downloadable version of FrameNet. It does not use the online version of FrameNet, so do not consult FrameNet online for frames and frame elements. Instead, use the FrameNet Explorer software and downloadable FrameNet loaded into it. You have received instructions about this already.

Predicate

- Predicate is already annotated in your document.
- It annotates words that have been identified as predicate candidates (i.e., as words that evoke FrameNet frames).
- It can be used to annotate constituents of multiword predicates.
- Features of Predicate:
 - Predicate has one feature: **isHead**: values = (**head**, **satellite**).
 - For multiword predicates, isHead shows whether the annotation is on the head of the multiword (“head”) or on a non-head constituent of the multiword (“satellite”).
 - Only set the feature value for multiword predicates, otherwise just leave it empty.
 - Example: “the electric current was [cut] off during daylight hours”. There is an initial “Predicate” annotation on “cut”. The particle “off” can be identified, and in fact, there is a FrameNet entry for the verb “cut off” in FrameNet 1.5. In this case, also create a Predicate annotation for “off” and set its isHead feature to “satellite”. Set the isHead feature of “cut” to head.
 - An example of a multiword noun predicate would be “weapon of mass destruction”, where weapon is the head.

- You may encounter a phrasal verb that is not listed as a phrasal verb in the FrameNet explorer. In this case, also annotate it as a multiword predicate. If the meaning of this phrasal verb is covered by one of the senses of the base verb, annotate the appropriate frame label. If not, annotate it as “NF”.
- As an example: the verbs “wake” and “wake up” both belong to the “Waking_up” frame. Even if “wake up” were not in the FrameNet list of predicates for “Waking up”, it could be identified as belonging to the same frame, because its base verb “wake” is associated with this frame, and because the meaning of “wake up” (given the appropriate context) fits the “Waking_up” frame.

PredicatePart

- PredicatePart can be used to link a Predicate annotation that is part of a multi-word predicate (the “satellite”) to the Predicate annotation of the head of the multiword.
- FrameNet contains multiword predicates, such as phrasal verbs (“kick off”, “wait for”), multiword nouns (“weapon of mass destruction”), or proverbs (“kick the bucket”). They should appear in the FrameNet explorer when searching for the base lemma. Phrasal verbs are usually indexed with an underscore, e.g., “take_up.v”.
- In case of multi-word predicates with three or more constituents, link each of the satellites to the head with a new PredicatePart annotation (e.g., “kicked the bucket” would get three Predicate annotations (one for each word), and the annotations for “the” and “bucket” would have isHead=satellite, and they would be linked with PredicatePart to the Predicate annotation for “kicked”, which has isHead=head).

FrameNet

- The FrameNet annotation is used to
 1. annotate predicates with a frame label
 2. annotate semantic arguments of predicates with a role label
- Note that a semantic argument of a predicate can be either a syntactic argument, an adjunct or any other kind of modifier (mostly noun phrases, prepositional phrases, sometimes subordinate clauses). For the list of available roles (frame elements) of a frame, refer to the FrameNet Explorer.
- Features of FrameNet:
 - FrameNet has three features, **FNkind**, **FNframe**, **FNrole**.

- **FNkind:** values = (**pred, arg**)
 - * Annotate whether this FrameNet annotation is a predicate (i.e., does it evoke a frame?) or an argument (does it depend on the predicate?).
 - * Note that the **arg** label also applies to semantic arguments that are not syntactic arguments of the predicate, for instance temporal adjuncts, adverbial modifiers, etc.
- **FNframe:** values = (frame labels, too many to list exhaustively, e.g., *Commerce_buy, Activity_finish,...*)
 - * The values of FNframe are the frame labels available in FrameNet plus three additional values.
 - * The value to enter depends on the FNkind value.
 - * **If FNkind = arg**
 - don't enter a frame label, but select 0.
 - * **If FNkind = pred**
 - look up the lemma and part-of-speech of the predicate in the FrameNet lexicon using FrameNet Explorer (use the "lexical units" tab) to get a list of frames for this predicate.
 - If there is a frame for the observed word sense of the predicate, select the appropriate frame label (more info on how to navigate the long lists of frame labels and role labels will be presented below).
 - If there is not a frame for the observed word sense of the predicate, enter NF ("sense not in the FrameNet lexicon").
 - If the sense cannot be understood (for instance due to poor grammar and lack of context), enter XX.
- **FNrole:** values = (role labels, too many to list exhaustively, e.g., *Buyer, Seller, Location,...*)
 - * The value to enter depends on the FNkind value.
 - * *If FNkind = pred*
 - don't enter a role label, but select 0.
 - * *If FNkind = arg*
 - refer to FrameNet Explorer to review the list of role labels ("frame elements") for the frame of the corresponding predicate. (Also: the annotated example sentences in FrameNet might help to determine the appropriate role, as FrameNet role labels are fine-grained and their descriptions can be very abstract.)

- If you find an appropriate role label for this arg, select the appropriate role label.
 - If there is no appropriate role label for the observed arg, select NF (“not in the FrameNet lexicon”).
 - (If you have problems selecting any appropriate roles for the arguments of a predicate, you might want to reconsider the “frame” label.)
- Note: If the same span participates as an argument in two different predicates, you should create a new FrameNet argument annotation on the span for each of the two predicates (even if they have the same role). Example sentence “Alex ate and drank”, with a FrameNet.pred annotation on “ate”, and a FrameNet.pred annotation on “drank”. There should be two FrameNet.arg annotations on “Alex”, one as agent of “ate” and the other as agent of “drank”. This makes 4 FrameNet annotations and 2 FrameNetRel annotations for this sentence.

FrameNetRel

- FrameNetRel is used to link FrameNet annotations for semantic arguments to the annotation of the corresponding predicate.
- You can link several FrameNet arguments to the same FrameNet predicate.
- The motivation for this annotation is to avoid any mismatches between predicates and their arguments for sentences with several predicates.
- Link a FrameNet annotation with FNkind.arg to the corresponding FrameNet annotation with FNkind.pred as soon as you created the annotations. This avoids mismatching arguments and predicates when there are several predicates in the sentence.
- By using your mouse you can draw the FrameNetRel annotations that will show up as arcs on the screen. This is explained below.

3.4 Creating Annotations

- Creating new span annotations:
 - Span annotations covering a single word: Double-click on the word you would like to annotate. A popup box should appear. Now select the annotation layer of this annotation and select the appropriate feature values. Confirm by clicking “Annotate” or pressing “Enter”.
 - After the confirmation a new colored box will appear on top of the word containing the feature values of the annotation.

- Span annotations covering multiple words: highlight the text that the annotation should cover as you would in a text editor (i.e., draw the mouse over the text with left button pressed, then release). A popup box should appear. Now select the annotation layer of this annotation and select the appropriate feature values. Confirm by clicking “Annotate” or pressing “Enter”.
- Changing existing span annotations:
 - If you want to change an existing annotation, double click on the annotation. A popup box should appear. You can now change the feature values of this annotation. Save the changes by clicking “Annotate” or pressing “Enter”. To close the box without changes, click the “x” in the upper right of the box. To delete the annotation, click “Delete”.
 - After changing an annotation, the new values should be displayed in the box.
 - Entering feature values (for span annotations):
 - Lists of available features will appear in a drop-down list when you select the feature value field (by navigating with TAB or by mouse-clicking on the field) Some features have long lists of values. For these, you can enter the first letters of the value you want to enter. The list of values will be sorted automatically based on the entered letters and will allow you to select the value.
- Creating new arc annotations:
 - Arc annotations are used to create links between annotations, for instance linking the FrameNet annotation of a predicate to the FrameNet annotation of its arguments.
 - The links are displayed in the user interface as arrows. The direction of the arrows does not matter for this annotation task!
 - Arc annotations can only be created between annotations of the same layer (FrameNet to FrameNet; Predicate to Predicate).
 - To create an arc, left-click on one annotation, draw the mouse button to the target annotation and release. During this process, you can already see an arc.
 - A popup should appear. There is only one appropriate layer type for the arc annotation. The arcs don’t have features, so just confirm by clicking “ok” or pressing “enter”.
 - Note that arc annotations can cross sentence boundaries. (This, for instance, happens when there are errors in text segmentation or when you’re annotating anaphoric references.)

- Changing existing arc annotations:
 - The only way to change an arc annotation without features, is to delete it.
 - Therefore, double click on the arc annotation. Then click on “delete”.
 - The arc should be gone.
- Predicate annotations:
 - Predicate annotations are already annotated initially to mark the predicate candidates.
 - There are up to 5 Predicate annotations per sentence.
 - You should only change the predicate annotation if you encounter a multiword predicate, for instance a phrasal verb (“give up”).
 - In this case: add another Predicate annotation for each constituent of the multiword and create PredicatePart arcs to link the constituents to their heads.
- FrameNet annotations:
 - For each Predicate annotation, create a FrameNet annotation based on the guidelines in 3.3.
 - If a FrameNet annotation is FNkind.pred and has a FrameNet frame label, create new FrameNet.arg annotations for the arguments of that predicate. Argument annotations typically cover several words (e.g., noun phrases or sentential complements).
 - (If the value of FrameNet.FNframe is XX or NF, you don’t annotate the arguments of that predicate.)
 - Link all the FrameNet.arg annotations to their corresponding FrameNet.pred annotation using FrameNetRel.
- Navigating annotations using the keyboard:
 - This may be more convenient and speed up the annotation process:
 - In the span annotation popup window, you can use TAB and SHIFT-TAB to navigate between the feature value fields.
 - Use the “up” and “down” arrows on your keyboard to select feature values (in particular for features with a small number of values).
 - If you want to select “0” as a value for the FNframe or FNrole features of a “FrameNet” annotation, press the down-arrow and then up-arrow to access the 0 directly. (“0” is the first item in the value list for both features.)

- You can use “Enter” to finish a new annotation (instead of clicking “annotate” or “ok”) or to confirm changes (instead of clicking “annotate”).

3.5 Additional Information

- Questions and Answers in the text are marked with a prefix. Questions are marked with “Q:”. The beginning of each new answer is marked with “QA:”.
- This information should support the readability of the texts. An answer, for instance, may contain a reference to the question text (rather than to the directly preceding answer).
- Predicate candidates can be nouns or adjectives, but will be mostly verbs. Identifying semantic arguments of nouns may seem unusual, but the FrameNet example sentences may help. It is particularly important to use nouns or adjectives as predicates for sentences in which the verb is a copula (e.g., be, become, remain, etc.).
- Anaphora: also annotate the referents of anaphoric arguments. For example, in a sentence like “it worked for Holland,” where “it” is an anaphora for “build dikes,” annotate both, the anaphora (“it”), the antecedent (“build dikes”) with a FrameNet role annotation.

3.6 Planning Annotation Work

- This section provides some information on how to plan your annotation work in WebAnno. The documents that are made accessible to you in WebAnno vary in size and in the number of initial “Predicate” annotations.
- This means that, when planning your amount of work per week, you cannot use a number n of documents you want to annotate per week as a goal, but should rather use the number of sentences in the document – or the number of initial “Predicate” annotations in the document – as a goal.
- Therefore, the following table lists the document name as displayed on the upper left in the WebAnno annotation view together with the number of sentences and initial “Predicate” annotations in that document.
- (Note that you cannot access/see all of these documents in WebAnno. This is okay. More documents will be made accessible/visible as you go along with the annotation work. Please notify <Adjudicator> when you’re finished with a batch of accessible documents and need to get new ones assigned.)

The following list gives an overview on the documents to annotate:²⁹

Document name	# sentences	# Predicate annotations (in the document)
20061107232131AAR7UaS	67	160
20061116193351AAkDkOB	9	28
20061119050639AAqIkNK	41	108
20061115141609AAIaTJW	19	58
20061105132616AAXHEHn	14	16
20061125101443AAdx9by	29	86
20061124194515AA4r83C	56	134
20061112080733AAUTarn	19	44
20061112141254AAQdgpm	18	53
20061121212907AAzy0DQ	27	96
20061105094853AABXIAQ	16	53
20061115073404AAG0Ryw	17	57
20061124125544AA17v7a	62	182
20061109115954AAOOug9	4	10
20061105175513AAECYEz	23	77
20061125171418AAkIrG5	9	21
20061123073007AAaUdrl	9	20
20061116061702AAaLb8o	27	62
20061101195550AAZp2vI	40	122
20061107114407AAWTmZ0	31	86
20061114100944AA2CZ8X	3	6
20061114054910AAc1QLn	11	27
20061106170925AA5anvt	42	109
20061125101404AA1pcqh	52	133
20061115080658AAHzf4C	11	23
20061122072030AAXGUee	20	62
20061103104550AAaNJl5	11	30
20061117111656AAXFob3	22	51
20061106144125AAIqPks	19	57
20061127180135AA3ldjl	7	16
20061119101536AASalEi	27	89

²⁹The document names correspond to the question ids in the Yahoo! Answers Manners dataset.

4 FrameNet Explorer

- FrameNet Explorer (FNE)³⁰ is a simple tool for looking at the contents of FrameNet 1.5 (the downloadable version of FrameNet). You can search on frames, frame elements, or lexical units (i.e., the words that actually evoke frames). You will find it most useful as follows.
- Keep FNE open while you are annotating in WebAnno. When you find a predicate that is a candidate for annotation, look it up in the Lexical Units tab.
- If the word (including POS) is used in more than one Frame, determine which frame is the best fit for the example you are looking at.
- Go to the Frames tab and look up the frame you have decided on. When you bring it into the display you will see all of the frame elements of that frame.
- If the frame elements for the frame fit the sentence you are looking at, do the frame annotation for the predicate.
- Then, return to FNE and study the frame elements to determine how you should annotate the arguments of your predicate in WebAnno.

³⁰<http://www.clres.com/FNExplorer.html>

List of Tables

2.1	FrameNet lexicon coverage of several large corpora	25
2.2	Frame-semantic resources for languages other than English	27
2.3	Inter-rater agreement for sense alignment gold standard	56
2.4	Word sense alignment performance by POS	59
2.5	Manual post-hoc evaluation of word sense alignment	62
2.6	Multilingual word-level FrameNet expansions in FNWKxx	64
2.7	English FrameNet expansion after relation disambiguation	66
2.8	Frame-semantic resources for German	67
2.9	Overlap of FNWKde with SALSA and P&L05	68
2.10	UBY _{FN} statistics	81
2.11	Sense links to FrameNet in UBY _{FN}	82
2.12	Lexicon coverage of the extended FrameNet	83
2.13	Predicate argument structure links in UBY _{FN}	84
3.1	Overview of related work in training data generation for SRL	109
3.2	Test dataset statistics for verbs	116
3.3	Parameter tuning for DistantSRL	119
3.4	Sense statistics of automatically labeled corpora for English	120
3.5	Frame identification experiments with DistantSRL	121
3.6	Frame classification precision for verbs on union and set difference of test set	123
3.7	Role statistics of automatically labeled corpora	125
3.8	Role classification experiments with DistantSRL	127
3.9	Role classification precision for verbs on union and set difference of test set	128
3.10	German dataset statistics on verbs	132
3.11	German frame identification experiments with DistantSRL	133
3.12	German role classification experiments with DistantSRL	134
3.13	Frame identification experiments with retrained SEMAFOR	137
3.14	Retrained SEMAFOR frame identification– full SRL scores	138

3.15 Retrained SEMAFOR role labeling model – SRL scores	140
4.1 YAGS annotation statistics	160
4.2 YAGS statistics on role labels per predicate	160
4.3 Evaluation dataset statistics.	161
4.4 FrameNet lexicon statistics: number of frames per lemma	163
4.5 FrameNet lexicon statistics: number of frames per verb lemma	164
4.6 FrameNet lexicon statistics on the number of roles per frame	165
4.7 Domain generalization experiments with SEMAFOR	168
4.8 Domain generalization experiments with SEMAFOR for verbs	171
4.9 SEMAFOR study on domain generalization – error types	172
4.10 Statistics of automatically generated training corpora	174
4.11 Frame identification accuracy – DistantSRL	178
4.12 Frame identification accuracy – annotation projection	180
4.13 Frame identification accuracy – training data expansion by FrameNet+	183
4.14 Domain generalization experiments - error types	185
4.15 Most-frequent-sense baseline	186

List of Figures

1.1	Example: FrameNet-annotated sentence	3
1.2	Example: user-generated text from Yahoo! Answers	10
1.3	Thesis overview diagram	13
2.1	FrameNet lexicon structure	21
2.2	Example: FrameNet frame hierarchy	22
2.3	Example: FrameNet lexicon entry for the verb <i>complete</i>	23
2.4	VerbNet lexicon structure	29
2.5	Example: VerbNet lexicon entry for the verb <i>complete</i>	30
2.6	Example: FrameNet frame and VerbNet class linkings	31
2.7	Example: FrameNet, VerbNet, and PropBank role labels in comparison	32
2.8	PropBank lexicon structure	33
2.9	Example: PropBank lexicon entry for the verb <i>complete</i>	34
2.10	SALSA lexicon structure	35
2.11	Example: SALSA lexicon entry for the verb <i>aufhören</i>	36
2.12	WordNet lexicon structure	37
2.13	Example: WordNet lexicon entry for the verb <i>complete</i>	37
2.14	GermaNet lexicon structure	39
2.15	Example: GermaNet lexicon entry for the verb <i>aufhören</i>	40
2.16	Wiktionary lexicon structure	41
2.17	Example: Wiktionary lexicon entry for the verb <i>complete</i>	41
2.18	Example: word sense alignment	42
2.19	Example: predicate argument structure alignment	43
2.20	Method overview: using Wiktionary as interlingua for FrameNet translation	53
2.21	UBY-LMF classes for modeling FrameNet and SALSA.	73
2.22	Overview of the LLKB UBY _{FN}	80
3.1	Method overview DistantSRL: automatic training data generation	89

3.2	Example: results of Step 1A – seed patterns	91
3.3	Illustration: Step 1B – seed pattern filtering	92
3.4	Contrasting ASP patterns to CPA Patterns	94
3.5	Example: Step 1C – sense label transfer	95
3.6	Example: DistantSRL Stage 2 – creating VerbNet and FrameNet roles	96
3.7	Example: SemLink predicate and role mappings	98
3.8	Example: DistantSRL Step 2B – corpus instances with multiple roles	99
3.9	Linked lexical knowledge base centered around FrameNet	117
3.10	Role classification learning curves for WaSR-XL	130
3.11	Linked lexical knowledge base for German	133
4.1	Example: question from Yahoo! Answers	153
4.2	Example: FrameNet frame and role annotation in YAGS	156
4.3	Screenshot of WebAnno annotation interface	157
B.1	Example frame and role annotation for annotation study	210

Bibliography

- Adler, B. T., de Alfaro, L., Mola-Velasco, S. M., Rosso, P., and West, A. G. (2011). Wikipedia Vandalism Detection: Combining Natural Language, Metadata, and Reputation Features. In *Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science, pages 277–288. Springer, Berlin, Heidelberg. (Cited on page 28)
- Agirre, E. and Soroa, A. (2009). Personalizing PageRank for Word Sense Disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 33–41, Athens, Greece. (Cited on pages 55 and 57)
- Artstein, R. and Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596. (Cited on pages 56 and 57)
- Baisa, V., Bradbury, J., Cinkova, S., El Maarouf, I., Kilgarriff, A., and Popescu, O. (2015). Semeval-2015 task 15: A cpa dictionary-entry-building task. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval)*, pages 315–324, Denver, CO, USA. (Cited on page 93)
- Baker, C. F., Ellsworth, M., and Erk, K. (2007). SemEval-2007 Task 19: Frame Semantic Structure Extraction. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval)*, pages 99–104, Prague, Czech Republic. (Cited on pages 3, 26 and 152)
- Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL)*, pages 86–90, Montréal, Canada. (Cited on pages 2 and 21)
- Baldwin, T., de Marneffe, M.-C., Han, B., Kim, Y.-B., Ritter, A., and Xu, W. (2015). Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 126–135, Beijing, China. (Cited on pages 151 and 201)
- Banerjee, S. and Pedersen, T. (2003). Extended Gloss Overlaps As a Measure of Semantic Relatedness. In *Proceedings of the 18th International Joint Conference on Artificial*

- Intelligence (IJCAI)*, pages 805–810, Acapulco, Mexico. (Cited on page 104)
- Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226. (Cited on pages 24, 25, 90, 117 and 132)
- Basili, R., De Cao, D., Croce, D., Coppola, B., and Moschitti, A. (2009). Cross-Language Frame Semantics Transfer in Bilingual Corpora. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, volume 5449 of *Lecture Notes in Computer Science*, pages 332–345. Springer Berlin Heidelberg. (Cited on pages 54, 70, 109 and 111)
- Bejan, C. A. (2009). *Learning Event Structures From Text*. Dissertation, The University of Texas at Dallas, Dallas, TX, USA. (Cited on pages 100, 108 and 109)
- Berant, J., Srikumar, V., Chen, P.-C., Vander Linden, A., Harding, B., Huang, B., Clark, P., and Manning, C. D. (2014). Modeling Biological Processes for Reading Comprehension. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1499–1510, Doha, Qatar. (Cited on page 1)
- Biber, D. and Conrad, S. (2009). *Register, Genre, and Style*. Cambridge University Press, Cambridge, UK. (Cited on pages 10 and 146)
- Björkelund, A., Bohnet, B., Hafdell, L., and Nugues, P. (2010). A high-performance syntactic and semantic dependency parser. In *Coling 2010: Demonstrations*, pages 33–36, Beijing, China. (Cited on page 111)
- Björkelund, A., Hafdell, L., and Nugues, P. (2009). Multilingual Semantic Role Labeling. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL): Shared Task*, pages 43–48, Boulder, CO, USA. (Cited on page 114)
- Blitzer, J., McDonald, R., and Pereira, F. (2006). Domain Adaptation with Structural Correspondence Learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 120–128, Sydney, Australia. (Cited on pages 147, 189 and 201)
- Boas, H. C. (2005). Semantic Frames as Interlingual Representations for Multilingual Lexical Databases. *International Journal of Lexicography*, 18(4):445–478. (Cited on pages 26 and 69)
- Bonial, C., Babko-Malaya, O., Choi, J. D., Hwang, J., and Palmer, M. (2010). PropBank Annotation Guidelines. Technical report, CLEAR, University of Colorado Boulder, Boulder, CO, USA. (Cited on page 32)
- Bonial, C., Bonn, J., Conger, K., Hwang, J. D., and Palmer, M. (2014). PropBank: Semantics of New Predicate Types. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, pages 3013–3019, Reykjavik, Iceland. (Cited on page 33)

- Bonial, C., Stowe, K., and Palmer, M. (2013). Renewing and Revising SemLink. In *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL): Representing and linking lexicons, terminologies and other language data*, pages 9–17, Pisa, Italy. (Cited on pages [6](#), [30](#), [31](#), [44](#), [48](#), [49](#) and [90](#))
- Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., and Yakhnenko, O. (2013). Translating Embeddings for Modeling Multi-relational Data. In *Advances in Neural Information Processing Systems 26 (NIPS)*, pages 2787–2795, Stateline, NV, USA. (Cited on page [203](#))
- Borin, L., Forsberg, M., Johansson, R., Muhonen, K., Purtonen, T., and Voionmaa, K. (2012). Transferring Frames: Utilization of Linked Lexical Resources. In *Proceedings of the NAACL-HLT Workshop on the Induction of Linguistic Structure*, pages 8–15, Montréal, Canada. (Cited on pages [54](#) and [63](#))
- Brants, S., Dipper, S., Eisenberg, P., Hansen, S., König, E., Lezius, W., Rohrer, C., Smith, G., and Uszkoreit, H. (2004). TIGER: Linguistic Interpretation of a German Corpus. *Journal of Language and Computation*, 2:597–620. (Cited on page [34](#))
- Burchardt, A., Erk, K., and Frank, A. (2005). A WordNet detour to FrameNet. *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen*, 8:408–421. (Cited on pages [6](#), [38](#), [102](#) and [103](#))
- 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 the 5th International Conference on Language Resources and Evaluation (LREC)*, pages 969–974, Genoa, Italy. (Cited on pages [27](#), [34](#) and [68](#))
- Burchardt, A. and Pennacchiotti, M. (2008). FATE: a FrameNet-Annotated Corpus for Textual Entailment. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*, pages 539–546, Marrakech, Morocco. (Cited on page [116](#))
- 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 (CoNLL)*, pages 152–164, Ann Arbor, Michigan, USA. (Cited on pages [9](#), [33](#) and [148](#))
- Chang, N., Paritosh, P., Huynh, D., and Baker, C. (2015). Scaling Semantic Frame Annotation. In *Proceedings of The 9th Linguistic Annotation Workshop (LAW)*, pages 1–10, Denver, CO, USA. (Cited on page [154](#))
- Chapelle, O., Schölkopf, B., and Zien, A. (2010). *Semi-Supervised Learning*. The MIT Press, Cambridge, MA, USA. (Cited on page [147](#))
- Chen, M., Weinberger, K. Q., and Blitzer, J. (2011). Co-training for Domain Adaptation. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2456–2464, Granada,

- Spain. (Cited on pages 140, 142 and 201)
- Chen, X., Liu, Z., and Sun, M. (2014). A Unified Model for Word Sense Representation and Disambiguation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1025–1035, Doha, Qatar. (Cited on pages 192, 200 and 202)
- Cheung, J. C. K., Poon, H., and Vanderwende, L. (2013). Probabilistic Frame Induction. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 837–846, Atlanta, Georgia, USA. (Cited on page 202)
- Chiarcos, C., Hellmann, S., Nordhoff, S., Cimiano, P., McCrae, J., Brekle, J., Eckle-Kohler, J., Gurevych, I., Hartmann, S., Matuschek, M., Meyer, C. M., and Littauer, R. (2012a). The Working Group for Open Data in Linguistics. In *Sprache als komplexes System. Proceedings der 34. Jahrestagung der DGfS*, pages 284–285, Frankfurt, Germany. (Cited on page 16)
- Chiarcos, C., Hellmann, S., Nordhoff, S., Moran, S., Littauer, R., Eckle-Kohler, J., Gurevych, I., Hartmann, S., Matuschek, M., and Meyer, C. M. (2012b). The Open Linguistics Working Group. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, pages 3603–3610, Istanbul, Turkey. (Cited on page 16)
- Cholakov, K., Eckle-Kohler, J., and Gurevych, I. (2014). Automated Verb Sense Labelling Based on Linked Lexical Resources. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 68–77, Gothenburg, Sweden. (Cited on pages 87, 91, 92, 94, 95, 101, 107, 118 and 119)
- Croce, D., Giannone, C., Annesi, P., and Basili, R. (2010). Towards Open-Domain Semantic Role Labeling. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 237–246, Uppsala, Sweden. (Cited on pages 9 and 150)
- Curran, J. R., Murphy, T., and Scholz, B. (2007). Minimising Semantic Drift with Mutual Exclusion Bootstrapping. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLIC)*, pages 172–180, Melbourne, Australia. (Cited on pages 89 and 100)
- Dahlmeier, D. and Ng, H. T. (2010). Domain adaptation for semantic role labeling in the biomedical domain. *Bioinformatics*, 26(8):1098–1104. (Cited on page 150)
- Dang, H. T., Kipper, K., Palmer, M., and Rosenzweig, J. (1998). Investigating Regular Sense Extensions based on Intersective Levin Classes. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL)*, pages 293–299, Montréal,

- Canada. (Cited on page [105](#))
- Das, D., Chen, D., Martins, A. F. T., Schneider, N., and Smith, N. A. (2014). Frame-Semantic Parsing. *Computational Linguistics*, 40:1:9–56. (Cited on pages [3](#), [4](#), [24](#), [90](#), [105](#), [107](#), [109](#), [135](#), [136](#), [140](#), [141](#), [151](#), [166](#), [167](#), [177](#) and [190](#))
- Das, D., Schneider, N., Chen, D., and Smith, N. A. (2010). Probabilistic Frame-Semantic Parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 948–956, Los Angeles, CA, USA. (Cited on pages [3](#), [141](#) and [152](#))
- 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 (NAACL-HLT)*, pages 1435–1444, Portland, OR, USA. (Cited on pages [105](#), [115](#), [117](#), [122](#) and [161](#))
- Daumé 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. (Cited on pages [147](#), [149](#), [151](#), [189](#) and [201](#))
- De Cao, D., Croce, D., Pennacchiotti, M., and Basili, R. (2008). Combining Word Sense and Usage for Modeling Frame Semantics. In Bos, J. and Delmonte, R., editors, *Semantics in Text Processing. STEP 2008 Conference Proceedings*, volume 1 of *Research in Computational Semantics*, pages 85–101. College Publications, London, UK. (Cited on page [45](#))
- De Marneffe, M.-C., MacCartney, B., and Manning, C. D. (2006). Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of the 5th Edition of the International Conference on Language Resources and Evaluation (LREC)*, pages 449–454, Genoa, Italy. (Cited on page [96](#))
- De Melo, G. and Weikum, G. (2009). Towards a Universal Wordnet by Learning from Combined Evidence. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM)*, pages 513–522, New York, NY, USA. (Cited on page [47](#))
- Dowty, D. R. (1986). Thematic Roles and Semantics. In *Proceedings of the Twelfth Annual Meeting of the Berkeley Linguistics Society*, pages 340–354, Berkeley, CA, USA. (Cited on page [32](#))
- Duan, W. and Yates, A. (2010). Extracting Glosses to Disambiguate Word Senses. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 627–635, Los Angeles, CA, USA. (Cited on page [107](#))
- Eckart de Castilho, R. and Gurevych, I. (2014). A broad-coverage collection of portable NLP components for building shareable analysis pipelines. In *Proceedings of the*

- Workshop on Open Infrastructures and Analysis Frameworks for HLT (OIAF4HLT) at COLING 2014*, pages 1–11, Dublin, Ireland. (Cited on pages [57](#), [96](#), [118](#), [132](#) and [155](#))
- Eckart de Castilho, R., Mújdricza-Maydt, E., Yimam, S. M., Hartmann, S., Gurevych, I., Frank, A., and Biemann, C. (2016). A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 76–84, Osaka, Japan. (Cited on page [209](#))
- Eckle-Kohler, J., Gurevych, I., Hartmann, S., Matuschek, M., and Meyer, C. M. (2012). UBY-LMF - A Uniform Model for Standardizing Heterogeneous Lexical-Semantic Resources in ISO-LMF. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, pages 275–282, Istanbul, Turkey. (Cited on pages [6](#), [15](#), [72](#), [77](#) and [205](#))
- Eckle-Kohler, J., Gurevych, I., Hartmann, S., Matuschek, M., and Meyer, C. M. (2013). UBY-LMF - Exploring the Boundaries of Language-Independent Lexicon Models. In Francopoulo, G., editor, *LMF Lexical Markup Framework*, chapter 10, pages 145–156. ISTE - HERMES - Wiley, London, UK. (Cited on pages [15](#) and [72](#))
- Eisenstein, J. (2013). What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 359–369, Atlanta, GA, USA. (Cited on pages [151](#) and [201](#))
- Ellis, J., Getman, J., Fore, D., Kuster, N., Song, Z., Bies, A., and Strassel, S. (2015). Overview of Linguistic Resources for the TACK KBP 2015 Evaluations: Methodologies and Results. In *Proceedings of the 2015 Text Analysis Conference (TAC)*, Gaithersburg, MA, USA. (Cited on page [199](#))
- Erk, K. and Padó, S. (2006). SHALMANESER – A Toolchain For Shallow Semantic Parsing. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pages 527–532, Genoa, Italy. (Cited on page [35](#))
- Exner, P., Klang, M., and Nugues, P. (2015). A Distant Supervision Approach to Semantic Role Labeling. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 239–248, Denver, CO, USA. (Cited on pages [8](#), [89](#), [100](#), [109](#) and [111](#))
- Falk, I., Gardent, C., and Lamirel, J.-C. (2012). Classifying French verbs using French and English lexical resources. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 854–863, Jeju Island, Korea. (Cited on page [31](#))
- Feizabadi, P. S. and Padó, S. (2014). Crowdsourcing Annotation of Non-Local Semantic Roles. In *Proceedings of the 14th Conference of the European Chapter of the Association*

- for *Computational Linguistics, volume 2: Short Papers*, pages 226–230, Gothenburg, Sweden. (Cited on pages [154](#) and [155](#))
- Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, USA. (Cited on page [36](#))
- Ferrandez, O., Ellsworth, M., Munoz, R., and Baker, C. F. (2010). Aligning FrameNet and WordNet based on Semantic Neighborhoods. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, pages 310–314, Valletta, Malta. (Cited on pages [46](#), [54](#) and [82](#))
- Fillmore, C. J. (1976). Frame Semantics and the Nature of Language. In *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, volume 280, pages 20–32. New York Academy of Sciences, New York, NY, USA. (Cited on pages [5](#) and [21](#))
- Fillmore, C. J. and Baker, C. F. (2010). A Frame Approach to Semantic Description. In Heine, B. and Narrog, H., editors, *Oxford Handbook of Linguistic Analysis*. Oxford University Press, New York, NY, USA. (Cited on page [21](#))
- Finlayson, M. and Kulkarni, N. (2011). Detecting Multi-Word Expressions Improves Word Sense Disambiguation. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 20–24, Portland, OR, USA. (Cited on page [124](#))
- FitzGerald, N., Täckström, O., Ganchev, K., and Das, D. (2015). Semantic Role Labeling with Neural Network Factors. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 960–970, Lisbon, Portugal. (Cited on pages [3](#), [6](#), [7](#), [10](#), [90](#), [107](#), [135](#), [141](#), [142](#), [149](#), [150](#) and [200](#))
- Fossati, M., Tonelli, S., and Giuliano, C. (2013). Frame semantics annotation made easy with dbpedia. In *Proceedings of the 1st International Workshop on Crowdsourcing the Semantic Web co-located with 12th International Semantic Web Conference (ISWC)*, pages 23–32, Zurich, Switzerland. (Cited on page [154](#))
- Francopoulo, G., Bel, N., George, M., Calzolari, N., Monachini, M., Pet, M., and Soria, C. (2006). Lexical Markup Framework (LMF). In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pages 233–236, Genoa, Italy. (Cited on pages [20](#) and [72](#))
- Fürstenau, H. (2011). *Semi-supervised Semantic Role Labeling via Graph Alignment, volume 32 of Saarbrücken Dissertations in Computational Linguistics and Language Technology*. PhD thesis, German Research Center for Artificial Intelligence and Saarland University, Saarbrücken, Germany. (Cited on pages [109](#), [111](#), [112](#) and [113](#))

- Fürstenaу, H. and Lapata, M. (2009). Semi-supervised Semantic Role Labeling. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 220–228, Athens, Greece. (Cited on pages 197 and 203)
- Fürstenaу, H. and Lapata, M. (2012). Semi-Supervised Semantic Role Labeling via Structural Alignment. *Computational Linguistics*, 38(1):135–171. (Cited on pages 8, 88, 89, 104, 109, 111, 112, 113, 125, 173, 175, 176, 177, 180, 192, 193 and 199)
- Gabrilovich, E. and Markovitch, S. (2007). Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1606–1611, Hyderabad, India. (Cited on page 28)
- Ganitkevitch, J., Van Durme, B., and Callison-Burch, C. (2013). PPDB: The Paraphrase Database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 758–764, Atlanta, GA, USA. (Cited on pages 8 and 114)
- Gildea, D. and Jurafsky, D. (2002). Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288. (Cited on page 3)
- Giuglea, A.-M. and Moschitti, A. (2006). Semantic Role Labeling via FrameNet, VerbNet and PropBank. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (COLING-ACL)*, pages 929–936, Sydney, Australia. (Cited on page 105)
- Gordon, A. and Swanson, R. (2007). Generalizing Semantic Role Annotations Across Syntactically Similar Verbs. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 192–199, Prague, Czech Republic. (Cited on pages 109 and 114)
- Gurevych, I., Ecker-Kohler, J., Hartmann, S., Matuschek, M., Meyer, C. M., and Nghiem, T.-D. (2013). UBY - A Large-Scale Lexical-Semantic Resource. In *Book of Abstracts of the 23rd Meeting of Computational Linguistics in the Netherlands: CLIN 2013*, page 81, Enschede, The Netherlands. (Cited on pages 15 and 16)
- Gurevych, I., Ecker-Kohler, J., Hartmann, S., Matuschek, M., Meyer, C. M., and Wirth, C. (2012a). Uby - A Large-Scale Unified Lexical-Semantic Resource Based on LMF. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 580–590, Avignon, France. (Cited on pages 6, 45, 46, 56, 72, 77 and 205)
- Gurevych, I., Matuschek, M., Nghiem, T.-D., Ecker-Kohler, J., Hartmann, S., and Meyer, C. M. (2012b). Navigating Sense-Aligned Lexical-Semantic Resources: The Web

- Interface to UBY. In *Proceedings of the 11th "Konferenz zur Verarbeitung natürlicher Sprache" (KONVENS)*, pages 194–198, Vienna, Austria. (Cited on pages [16](#) and [72](#))
- Hajič, J., Ciaramita, M., Johansson, R., Kawahara, D., Martí, M. A., Màrquez, L., Meyers, A., Nivre, J., Padó, S., Štěpánek, J., Straňák, P., Surdeanu, M., Xue, N., and Zhang, Y. (2009). The CoNLL-2009 Shared Task: Syntactic and Semantic Dependencies in Multiple Languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL): Shared Task*, pages 1–18, Boulder, CO, USA. (Cited on pages [9](#), [33](#), [111](#) and [148](#))
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1):10–18. (Cited on pages [118](#) and [125](#))
- Hamp, B. and Feldweg, H. (1997). GermaNet - a Lexical-Semantic Net for German. In *Proceedings of the ACL workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15, Madrid, Spain. (Cited on page [38](#))
- Han, B. and Baldwin, T. (2011). Lexical Normalisation of Short Text Messages: Makn Sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pages 368–378, Portland, OR, USA. (Cited on pages [11](#), [146](#) and [151](#))
- Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations*. MIT Press, Cambridge, MA, USA. (Cited on page [93](#))
- Hartmann, S., Eckle-Kohler, J., and Gurevych, I. (2016). Generating training data for semantic role labeling based on label transfer from linked lexical resources. *Transactions of the Association for Computational Linguistics (TACL)*, 4:197–213. (Cited on pages [15](#), [90](#), [95](#), [98](#), [109](#) and [207](#))
- Hartmann, S. and Gurevych, I. (2013a). Acquisition of Multiword Lexical Units for FrameNet. Presentation at The International FrameNet Workshop 2013. (Cited on page [16](#))
- Hartmann, S. and Gurevych, I. (2013b). FrameNet on the Way to Babel: Creating a Bilingual FrameNet Using Wiktionary as Interlingual Connection. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pages 1363–1373, Sofia, Bulgaria. (Cited on pages [15](#), [51](#), [205](#) and [206](#))
- Hartmann, S., Kuznetsov, I., Martin, T., and Gurevych, I. (2017a). Out-of-domain FrameNet Semantic Role Labeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers (EACL)*, pages 471–482, Valencia, Spain. (Cited on pages [15](#), [175](#), [190](#), [191](#), [193](#), [200](#), [201](#) and [206](#))

- Hartmann, S., Mújdricza-Maydt, É., Kuznetsov, I., Gurevych, I., and Frank, A. (2017b). Assessing SRL Frameworks with Automatic Training Data Expansion. In *Proceedings of the 11th Linguistic Annotation Workshop (LAW)*, pages 115–121, Valencia, Spain. (Cited on pages [16](#) and [199](#))
- Hartmann, S., Szarvas, G., and Gurevych, I. (2012). Mining Multiword Terms from Wikipedia. In Paziienza, M. T. and Stellato, A., editors, *Semi-Automatic Ontology Development: Processes and Resources*, pages 226–258. IGI Global, Hershey, PA, USA. (Cited on page [16](#))
- Hasan, K. S. and Ng, V. (2014). Why are You Taking this Stance? Identifying and Classifying Reasons in Ideological Debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 751–762, Doha, Qatar. (Cited on page [1](#))
- Hautli-Janisz, A., King, T. H., and Ramchand, G. (2015). Encoding event structure in Urdu/Hindi VerbNet. In *Proceedings of The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 25–33, Denver, CO, USA. (Cited on page [31](#))
- He, L., Lewis, M., and Zettlemoyer, L. (2015). Question-Answer Driven Semantic Role Labeling: Using Natural Language to Annotate Natural Language. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 643–653, Lisbon, Portugal. (Cited on pages [143](#) and [199](#))
- He, S. and Gildea, D. (2006). Self-training and Co-training for Semantic Role Labeling: Primary Report. TR 891. Technical report, University of Colorado at Boulder, Boulder, CO, USA. (Cited on pages [108](#) and [109](#))
- Henrich, V. and Hinrichs, E. (2010). GernEiT – The GermaNet Editing Tool. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, pages 2228–2235, Valletta, Malta. (Cited on page [38](#))
- Henrich, V., Hinrichs, E., and Vodolazova, T. (2012). WebCAGe – A Web-Harvested Corpus Annotated with GermaNet Senses. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 387–396, Avignon, France. (Cited on page [135](#))
- Henrich, V., Hinrichs, E., and Vodolazova, T. (2014). Aligning GermaNet Senses with Wiktionary Sense Definitions. In Vetulani, Z. and Mariani, J., editors, *Human Language Technology Challenges for Computer Science and Linguistics*, volume 8387 of *Lecture Notes in Computer Science*, pages 329–342. Springer International Publishing. (Cited on pages [39](#), [43](#), [46](#) and [135](#))

- Hermann, K. M., Das, D., Weston, J., and Ganchev, K. (2014). Semantic Frame Identification with Distributed Word Representations. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1448–1458, Baltimore, MD, USA. (Cited on pages [10](#), [90](#), [135](#), [140](#), [141](#), [150](#), [189](#), [190](#), [200](#) and [201](#))
- Hoffart, J., Suchanek, F. M., Berberich, K., and Weikum, G. (2013). YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia. *Artificial Intelligence*, 194:28 – 61. Artificial Intelligence, Wikipedia and Semi-Structured Resources. (Cited on page [47](#))
- Hong, J. and Baker, C. F. (2011). How Good is the Crowd at "real" WSD? In *Proceedings of the 5th Linguistic Annotation Workshop (LAW)*, pages 30–37, Portland, OR, USA. (Cited on page [154](#))
- Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006). Ontonotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers (NAACL-HLT)*, pages 57–60, New York City, NY, USA. (Cited on pages [38](#) and [48](#))
- Hripcsak, G. and Rothschild, A. S. (2005). Agreement, the F-Measure, and Reliability in Information Retrieval. *Journal of the American Medical Informatics Association*, 12(3):296–298. (Cited on pages [58](#) and [159](#))
- Huang, F. and Yates, A. (2010). Exploring Representation-Learning Approaches to Domain Adaptation. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 23–30, Uppsala, Sweden. (Cited on page [148](#))
- Iacobacci, I., Pilehvar, M. T., and Navigli, R. (2016). Embeddings for word sense disambiguation: An evaluation study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 897–907, Berlin, Germany. (Cited on page [202](#))
- Ide, N. (2006). Making Senses: Bootstrapping Sense-Tagged Lists of Semantically-Related Words. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, volume 3878 of *Lecture Notes in Computer Science*, pages 13–27. Springer Berlin Heidelberg. (Cited on page [103](#))
- Ide, N. and Pustejovsky, J. (2010). What Does Interoperability Mean, Anyway? Toward an Operational Definition of Interoperability for Language Technology. In *Proceedings of the Second International Conference on Global Interoperability for Language Resources*, Hong Kong, China. (Cited on pages [42](#) and [72](#))
- Ide, N. and Suderman, K. (2007). GrAF: A Graph-based Format for Linguistic Annotations. In *Proceedings of the Linguistic Annotation Workshop (LAW)*, pages 1–8, Prague, Czech Republic. (Cited on page [74](#))

- Johannsen, A., Martínez Alonso, H., and Søgaard, A. (2015). Any-language frame-semantic parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2062–2066, Lisbon, Portugal. (Cited on pages [131](#), [152](#), [153](#), [159](#), [161](#), [162](#) and [167](#))
- Johansson, R. and Nugues, P. (2005). Using Parallel Corpora for Automatic Transfer of FrameNet Annotation. In *Proceedings of the 1st ROMANCE FrameNet Workshop*, Cluj-Napoca, Romania. (Cited on pages [53](#), [70](#), [109](#) and [111](#))
- Johansson, R. and Nugues, P. (2006). A FrameNet-Based Semantic Role Labeler for Swedish. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions (COLING-ACL)*, pages 436–443, Sydney, Australia. (Cited on pages [53](#), [109](#) and [111](#))
- Johansson, R. and Nugues, P. (2007a). LTH: Semantic Structure Extraction using Nonprojective Dependency Trees. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval)*, pages 227–230, Prague, Czech Republic. (Cited on page [104](#))
- Johansson, R. and Nugues, P. (2007b). Using WordNet to extend FrameNet coverage. In *Proceedings of the Workshop on Building Frame-semantic Resources for Scandinavian and Baltic Languages, at NODALIDA*, pages 27–30, Tartu, Estonia. (Cited on pages [6](#), [38](#), [54](#), [102](#) and [104](#))
- Johansson, R. and Nugues, P. (2008a). Dependency-based Syntactic–Semantic Analysis with PropBank and NomBank. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning (CoNLL)*, pages 183–187, Manchester, UK. (Cited on page [148](#))
- Johansson, R. and Nugues, P. (2008b). The Effect of Syntactic Representation on Semantic Role Labeling. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 393–400, Manchester, UK. (Cited on pages [9](#) and [150](#))
- Kim, J.-D., Ohta, T., Pyysalo, S., Kano, Y., and Tsujii, J. (2009). Overview of BioNLP’09 Shared Task on Event Extraction. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pages 1–9, Boulder, CO, USA. (Cited on pages [87](#) and [199](#))
- Kipper, K., Korhonen, A., Ryant, N., and Palmer, M. (2006). Extending VerbNet with Novel Verb Classes. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pages 1027–1032, Genoa, Italy. (Cited on page [28](#))
- Kipper-Schuler, K. (2005). *VerbNet: A Broad-coverage, Comprehensive Verb Lexicon*. Dissertation, University of Pennsylvania, Philadelphia, PA, USA. (Cited on pages [2](#), [28](#) and [29](#))

- Kipper-Schuler, K., Korhonen, A., Ryant, N., and Palmer, M. (2008). A large-scale classification of English verbs. *Language Resources and Evaluation*, 42(1):21–40. (Cited on page 28)
- 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 (CoNLL)*, pages 181–184, Ann Arbor, MI, USA. (Cited on page 148)
- Krause, S., Hennig, L., Gabryszak, A., Xu, F., and Uszkoreit, H. (2015). Sar-graphs: A Linked Linguistic Knowledge Resource Connecting Facts with Language. In *Proceedings of the 4th Workshop on Linked Data in Linguistics: Resources and Applications*, pages 30–38, Beijing, China. (Cited on page 202)
- Kshirsagar, M., Thomson, S., Schneider, N., Carbonell, J., Smith, N. A., and Dyer, C. (2015). Frame-Semantic Role Labeling with Heterogeneous Annotations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 218–224, Beijing, China. (Cited on pages 6, 49, 88, 106, 138, 139, 140, 141, 151 and 166)
- Kübler, S. and Zhekova, D. (2009). Semi-Supervised Learning for Word Sense Disambiguation: Quality vs. Quantity. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP)*, pages 197–202, Borovets, Bulgaria. (Cited on page 107)
- Laparra, E. and Rigau, G. (2009). Integrating WordNet and FrameNet using a Knowledge-based Word Sense Disambiguation Algorithm. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP)*, pages 208–213, Borovets, Bulgaria. (Cited on pages 6, 7 and 46)
- Laparra, E. and Rigau, G. (2010). eXtended WordFrameNet. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, pages 1214–1419, Valletta, Malta. (Cited on pages 46, 58, 70 and 82)
- Leacock, C., Miller, G. A., and Chodorow, M. (1998). Using Corpus Statistics and WordNet Relations for Sense Identification. *Computational Linguistics*, 24(1):147–165. (Cited on page 107)
- Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago, IL, USA. (Cited on pages 5 and 28)
- Lin, D. (1998). Automatic Retrieval and Clustering of Similar Words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL)*, pages 768–774, Montréal, Canada. (Cited on page 105)

- Litkowski, K. (2010). CLR: Linking Events and Their Participants in Discourse Using a Comprehensive FrameNet Dictionary. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval)*, pages 300–303, Los Angeles, CA, USA. (Cited on page 102)
- Liu, X., Fu, Z., Wei, F., and Zhou, M. (2012). Collective Nominal Semantic Role Labeling for Tweets. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI)*, pages 1685–1691, Toronto, Canada. (Cited on page 151)
- Liu, X., Li, K., Han, B., Zhou, M., Jiang, L., Xiong, Z., and Huang, C. (2010). Semantic Role Labeling for News Tweets. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pages 698–706, Beijing, China. (Cited on pages 11 and 151)
- Liu, X., Li, K., Zhou, M., and Xiong, Z. (2011). Collective Semantic Role Labeling for Tweets with Clustering. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1832–1837, Barcelona, Spain. (Cited on page 151)
- 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 (IWCS)*, Tilburg, the Netherlands. (Cited on page 31)
- Lopez de Lacalle, M., Laparra, E., Aldabe, I., and Rigau, G. (2016). Predicate Matrix: automatically extending the semantic interoperability between predicate resources. *Language Resources and Evaluation*, 50(2):263–289. (Cited on pages 49, 50, 83, 84, 126 and 198)
- Lopez de Lacalle, M., Laparra, E., and Rigau, G. (2014). Predicate Matrix: extending SemLink through WordNet mappings. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, pages 903–909, Reykjavik, Iceland. (Cited on pages 46, 48 and 49)
- Martinez, D. (2008). On the Use of Automatically Acquired Examples for All-Nouns Word Sense Disambiguation. *Journal of Artificial Intelligence Research*, 33:79–107. (Cited on page 107)
- Matsubayashi, Y., Okazaki, N., and Tsujii, J. (2009). A Comparative Study on Generalization of Semantic Roles in FrameNet. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP)*, pages 19–27, Suntec, Singapore. (Cited on pages 105, 106 and 139)
- Matuschek, M. (2014). *Word Sense Alignment of Lexical Resources*. Dissertation, Technische Universität Darmstadt, Darmstadt, Germany. (Cited on page 70)
- Matuschek, M. and Gurevych, I. (2013). Dijkstra-WSA: A Graph-Based Approach to Word Sense Alignment. *Transactions of the Association for Computational Linguistics (TACL)*,

1:151–164. (Cited on page 46)

Matuschek, M., Meyer, C. M., and Gurevych, I. (2013). Multilingual Knowledge in Aligned Wiktionary and OmegaWiki for Translation Applications. *Translation: Corpora, Computation, Cognition (TC3)*, 3(1):87–118. (Cited on page 45)

Melamud, O., Goldberger, J., and Dagan, I. (2016). context2vec: Learning Generic Context Embedding with Bidirectional LSTM. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning (CoNLL)*, pages 51–61, Berlin, Germany. (Cited on pages 192 and 202)

Merlo, P. and van der Plas, L. (2009). Abstraction and Generalisation in Semantic Role Labels: PropBank, VerbNet or both? In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP)*, pages 288–296, Suntec, Singapore. (Cited on page 31)

Meyer, C. M. (2013). *Wiktionary: The Metalexigraphic and the Natural Language Processing Perspective*. Dissertation, Technische Universität Darmstadt, Darmstadt, Germany. (Cited on pages 28 and 40)

Meyer, C. M. and Gurevych, I. (2010). How Web Communities Analyze Human Language: Word Senses in Wiktionary. In *Proceedings of the Second Web Science Conference*, Raleigh, NC, USA. (Cited on page 58)

Meyer, C. M. and Gurevych, I. (2011). What Psycholinguists Know About Chemistry: Aligning Wiktionary and WordNet for Increased Domain Coverage. In *Proceedings of 5th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 883–892, Chiang Mai, Thailand. (Cited on pages 45 and 56)

Meyer, C. M. and Gurevych, I. (2012a). To Exhibit is not to Loiter: A Multilingual, Sense-Disambiguated Wiktionary for Measuring Verb Similarity. In *Proceedings of COLING 2012: Technical Papers (COLING)*, pages 1763–1780, Mumbai, India. (Cited on pages 53, 63, 65, 66 and 67)

Meyer, C. M. and Gurevych, I. (2012b). Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography. In Granger, S. and Paquot, M., editors, *Electronic Lexicography*, pages 259–291. Oxford University Press, Oxford, UK. (Cited on page 40)

Mihalcea, R. and Moldovan, D. (1999). An Automatic Method for Generating Sense Tagged Corpora. In *Proceedings of the American Association for Artificial Intelligence (AAAI)*, pages 461–466, Orlando, FL, USA. (Cited on page 107)

Mikolov, T. (2012). *Statistical Language Models Based on Neural Networks*. Dissertation, Brno University of Technology, Brno, Czech Republic. (Cited on page 192)

- Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26 (NIPS)*, pages 3111–3119, Stateline, NV, USA. (Cited on pages [60](#) and [176](#))
- Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (2009). Distant Supervision for Relation Extraction Without Labeled Data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP)*, pages 1003–1011, Suntec, Singapore. (Cited on pages [8](#), [12](#), [87](#) and [99](#))
- Moro, A., Raganato, A., and Navigli, R. (2014). Entity Linking meets Word Sense Disambiguation: A Unified Approach. *Transactions of the Association for Computational Linguistics (TACL)*, 2:231–244. (Cited on page [87](#))
- Mousser, J. (2010). A large coverage verb taxonomy for arabic. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC)*, pages 2675–2681, Valletta, Malta. (Cited on page [31](#))
- Mouton, C., de Chalendar, G., and Richert, B. (2010). FrameNet Translation Using Bilingual Dictionaries with Evaluation on the English-French Pair. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, pages 20–27, Valletta, Malta. (Cited on page [54](#))
- Mújdricza-Maydt, É., Hartmann, S., Gurevych, I., and Frank, A. (2016). Combining Semantic Annotation of Word Sense & Semantic Roles: A Novel Annotation Scheme for VerbNet Roles on German Language Data. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, pages 3031–3038. (Cited on pages [16](#) and [31](#))
- Narayanan, S. and Harabagiu, S. (2004). Question Answering Based on Semantic Structures. In *Proceedings of the 20th international conference on Computational Linguistics - COLING '04*, pages 693–701, Geneva, Switzerland. (Cited on page [1](#))
- Natarajan, N., Dhillon, I. S., Ravikumar, P. K., and Tewari, A. (2013). Learning with Noisy Labels. In *Advances in Neural Information Processing Systems 26 (NIPS)*, pages 1196–1204, Stateline, NV, USA. (Cited on pages [140](#), [142](#) and [201](#))
- Navigli, R. and Ponzetto, S. P. (2010). BabelNet: Building a Very Large Multilingual Semantic Network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden. (Cited on page [6](#))
- Navigli, R. and Ponzetto, S. P. (2012). BabelNet: The Automatic Construction, Evaluation and Application of a Wide-coverage Multilingual Semantic Network. *Artificial Intelligence*, 193:217–250. (Cited on page [47](#))

- Niemann, E. g. W. and Gurevych, I. (2011). The People’s Web meets Linguistic Knowledge: Automatic Sense Alignment of Wikipedia and WordNet. In *Proceedings of the International Conference on Computational Semantics (IWCS)*, pages 205–214, Singapore. (Cited on pages [45](#), [52](#), [55](#), [56](#) and [59](#))
- Padó, S. and Erk, K. (2005). To Cause Or Not To Cause: Cross-Lingual Semantic Matching for Paraphrase Modelling. In *Proceedings of the Workshop on Cross-Linguistic Knowledge Induction at EUROLAN*, Cluj-Napoca, Romania. (Cited on page [26](#))
- Padó, S. and Lapata, M. (2005a). Cross-lingual Bootstrapping of Semantic Lexicons: The Case of FrameNet. In *Proceedings of the 20th national conference on Artificial intelligence (AAAI)*, pages 1087–1092, Pittsburgh, PA, USA. (Cited on pages [53](#), [54](#), [67](#), [68](#), [70](#) and [71](#))
- Padó, S. and Lapata, M. (2005b). Cross-linguistic Projection of Role-Semantic Information. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP)*, pages 859–866, Vancouver, Canada. (Cited on pages [109](#), [110](#) and [111](#))
- Padó, S. and Lapata, M. (2006). Optimal Constituent Alignment with Edge Covers for Semantic Projection. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*, pages 1161–1168, Sydney, Australia. (Cited on page [111](#))
- Padó, S. and Lapata, M. (2009). Cross-lingual Annotation Projection for Semantic Roles. *Journal of Artificial Intelligence Research*, 36:307–340. (Cited on pages [8](#), [70](#), [109](#) and [110](#))
- Padó, S. and Pitel, G. (2007). Annotation précise du franchaptercais en sémantique de rôles par projection cross-linguistique. In *Proceedings of TALN-07*, pages 271–280, Toulouse, France. (Cited on pages [109](#) and [110](#))
- Palmer, A. and Sporleder, C. (2010). Evaluating FrameNet-style semantic parsing: the role of coverage gaps in FrameNet. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pages 928–936, Beijing, China. (Cited on pages [5](#), [24](#) and [26](#))
- Palmer, M. (2009). Sem Link: Linking PropBank, VerbNet and FrameNet. In *Proceedings of the Generative Lexicon Conference (GenLex-09)*, Pisa, Italy. (Cited on pages [6](#) and [48](#))
- Palmer, M., Gildea, D., and Kingsbury, P. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106. (Cited on pages [2](#) and [32](#))
- Palmer, M., Gildea, D., and Xue, N. (2009). *Semantic Role Labeling*. Synthesis Lectures on Human Language Technologies. Morgan and Claypool Publishers, Williston, VT, USA. (Cited on page [1](#))

- Pan, S. J. and Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359. (Cited on page 90)
- Passonneau, R. J., Baker, C. F., Fellbaum, C., and Ide, N. (2012). The MASC Word Sense Corpus. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, pages 3025–3030, Istanbul, Turkey. (Cited on page 116)
- Pavlick, E., Ganitkevitch, J., Chan, T. P., Yao, X., Van Durme, B., and Callison-Burch, C. (2015a). Domain-Specific Paraphrase Extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 57–62, Beijing, China. (Cited on pages 8, 109, 114, 191 and 200)
- Pavlick, E., Wolfe, T., Rastogi, P., Callison-Burch, C., Dredze, M., and Van Durme, B. (2015b). FrameNet+: Fast Paraphrastic Tripling of FrameNet. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 408–413, Beijing, China. (Cited on pages 173, 177, 192, 193, 197 and 199)
- Pennacchiotti, M., De Cao, D., Basili, R., Croce, D., and Roth, M. (2008). Automatic induction of FrameNet lexical units. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 457–465, Honolulu, HI, USA. (Cited on pages 102, 104 and 113)
- Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. (Cited on pages 60 and 176)
- Pianta, E., Bentivogli, L., and Girardi, C. (2002). MultiWordNet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, pages 292–302, Mysore, India. (Cited on page 45)
- Pilehvar, M. T. and Navigli, R. (2014). A Robust Approach to Aligning Heterogeneous Lexical Resources. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 468–478, Baltimore, MD, USA. (Cited on pages 47 and 203)
- Ponzetto, S. P. and Navigli, R. (2009). Large-Scale Taxonomy Mapping for Restructuring and Integrating Wikipedia. In *Proceedings of the 21st International Joint Conference on AI*, pages 2083–2088, Pasadena, CA, USA. (Cited on page 45)
- Ponzetto, S. P. and Navigli, R. (2010). Knowledge-Rich Word Sense Disambiguation Rivaling Supervised Systems. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1522–1531, Uppsala, Sweden. (Cited on page 87)

- Poon, H. and Domingos, P. (2009). Unsupervised Semantic Parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1–10, Singapore. (Cited on page 202)
- Popescu, O., Palmer, M., and Hanks, P. (2014). Mapping CPA Patterns onto OntoNotes Senses. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, pages 882–889, Reykjavik, Iceland. (Cited on page 93)
- Pradhan, S. S., Hovy, E. H., Marcus, M. P., Palmer, M., Ramshaw, L. A., and Weischedel, R. M. (2007a). OntoNotes: A Unified Relational Semantic Representation. In *Proceedings of the First IEEE International Conference on Semantic Computing (ICSC)*, pages 517–526, Irvine, CA, USA. (Cited on page 48)
- Pradhan, S. S., Ward, W., and Martin, J. (2007b). Towards Robust Semantic Role Labeling. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 556–563, Rochester, NY, USA. (Cited on pages 9 and 147)
- Rehbein, I., Ruppenhofer, J., Sporleder, C., and Pinkal, M. (2012). Adding nominal spice to SALSA - frame-semantic annotation of German nouns and verbs. In *Proceedings of the 11th Conference on Natural Language Processing (KONVENS)*, pages 89–97, Vienna, Austria. (Cited on pages 26, 34, 35 and 69)
- Reiter, N. (2014). *Discovering Structural Similarities in Narrative Texts using Event Alignment Algorithms*. Dissertation, Ruprecht-Karls-Universität Heidelberg, Heidelberg, Germany. (Cited on page 150)
- Roth, M. and Lapata, M. (2015). Context-aware Frame-Semantic Role Labeling. *Transactions of the Association for Computational Linguistics (TACL)*, 3:449–460. (Cited on page 141)
- Roth, M. and Woodsend, K. (2014). Composition of Word Representations Improves Semantic Role Labelling. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 407–413, Doha, Qatar. (Cited on page 141)
- Rothe, S. and Schütze, H. (2015). AutoExtend: Extending Word Embeddings to Embeddings for Synsets and Lexemes. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 1793–1803, Beijing, China. (Cited on page 200)
- Ruiz-Casado, M., Alfonseca, E., and Castells, P. (2005). Automatic Assignment of Wikipedia Encyclopedic Entries to WordNet Synsets. In *Advances in Web Intelligence*, volume 3528 of *Lecture Notes in Computer Science*, pages 380–386. Springer, Berlin Heidelberg. (Cited on pages 6 and 45)

- Ruppenhofer, J., Ellsworth, M., Petruck, M. R. L., Johnson, C. R., and Scheffczyk, J. (2010a). FrameNet II: Extended Theory and Practice. Technical report, ICSI, University of California, Berkeley. (Cited on pages 24, 60 and 70)
- Ruppenhofer, J., Sporleder, C., Morante, R., Baker, C., and Palmer, M. (2010b). SemEval-2010 Task 10: Linking Events and Their Participants in Discourse. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval)*, pages 45–50, Uppsala, Sweden. (Cited on pages 3, 103 and 116)
- Ruppenhofer, J., Sunde, J., and Pinkal, M. (2010c). Generating FrameNets of Various Granularities: The FrameNet Transformer. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, pages 2736–2743, Valletta, Malta. (Cited on pages 26 and 76)
- Sajous, F., Navarro, E., Gaume, B., Prévot, L., and Chudy, Y. (2010). Semi-automatic Endogenous Enrichment of Collaboratively Constructed Lexical Resources: Piggybacking onto Wiktionary. In *Advances in Natural Language Processing: 7th International Conference on NLP (iceTAL)*, pages 332–344. Springer, Berlin, Heidelberg. (Cited on page 69)
- Salaberri, H., Arregi, O., and Zafirain, B. (2014). First approach toward Semantic Role Labeling for Basque. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, pages 1387–1393, Reykjavik, Iceland. (Cited on page 31)
- Schmid, H. (1995). Improvements in Part-of-Speech Tagging with an Application to German. In *Proceedings of the ACL SIGDAT-Workshop*, pages 47–50, Dublin, Ireland. (Cited on page 132)
- Schmidt, T. (2009). The Kicktionary – A Multilingual Lexical Resource of Football Language. In Boas, H. C., editor, *Multilingual FrameNets in computational lexicography : methods and applications*, pages 101–134. Mouton de Gruyter, Berlin, Boston. (Cited on page 150)
- Seeker, W. and Kuhn, J. (2012). Making Ellipses Explicit in Dependency Conversion for a German Treebank. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, pages 3132–3139, Istanbul, Turkey. (Cited on page 133)
- Sergieh, H. M. and Gurevych, I. (2016). Enriching Wikidata with Frame Semantics. In *Proceedings of the 5th Workshop on Automated Knowledge Base Construction (AKBC) 2016 held in conjunction with NAACL 2016*, pages 29–34, San Diego, CA, USA. (Cited on page 202)
- Shen, D. and Lapata, M. (2007). Using Semantic Roles to Improve Question Answering. In *Proceedings of the Conference on Empirical Methods in NLP and on Computational Natural Language Learning*, pages 12–21, Prague, Czech Republic. (Cited on page 1)

- Shi, L. and Mihalcea, R. (2004a). An Algorithm for Open Text Semantic Parsing. In *3rd Workshop on Robust Methods in Analysis of Natural Language Data at COLING 2004*, pages 59–67, Geneva, Switzerland. (Cited on page 102)
- Shi, L. and Mihalcea, R. (2004b). Open Text Semantic Parsing Using FrameNet and WordNet. In *Demonstration Papers at HLT-NAACL 2004, HLT-NAACL–Demonstrations '04*, pages 19–22, Boston, MA, USA. (Cited on page 102)
- 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: 6th International Conference (CICLing)*, volume 3406 of *Lecture Notes in Computer Science*, pages 100–111. Springer, Berlin Heidelberg. (Cited on pages 6, 38, 54, 102 and 103)
- Silberer, C. and Frank, A. (2012). Casting Implicit Role Linking as an Anaphora Resolution Task. In **SEM 2012: 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 (SemEval 2012)*, pages 1–10, Montréal, Canada. (Cited on page 31)
- Søgaard, A. (2013). *Semi-Supervised Learning and Domain Adaptation in Natural Language Processing*, volume 6 of *Synthesis Lectures on Human Language Technologies*. Morgan and Claypool Publishers, Williston, VT, USA. (Cited on pages 146 and 147)
- Søgaard, A., Plank, B., and Martínez Alonso, H. (2015). Using Frame Semantics for Knowledge Extraction from Twitter. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 2447–2452, Palo Alto, CA, USA. (Cited on pages 11, 151, 152, 153, 159, 161 and 167)
- Sundheim, B., editor (1991). *Third Message Understanding Evaluation and Conference (MUC-3): Phase 1 Status Report*. (Cited on pages 87 and 199)
- Surdeanu, M., Ciaramita, M., and Zaragoza, H. (2011). Learning to Rank Answers to Non-Factoid Questions from Web Collections. *Computational Linguistics*, 37(2):351–383. (Cited on pages 11, 155, 175, 176 and 178)
- 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 (CoNLL)*, pages 159–177, Manchester, UK. (Cited on pages 9, 33 and 148)
- Swier, R. S. and Stevenson, S. (2004). Unsupervised Semantic Role Labelling. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 95–102, Barcelona, Spain. (Cited on page 103)

- Swier, R. S. and Stevenson, S. (2005). Exploiting a Verb Lexicon in Automatic Semantic Role Labelling. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP)*, pages 883–890, Vancouver, Canada. (Cited on pages 31 and 103)
- Täckström, O., Ganchev, K., and Das, D. (2015). Efficient Inference and Structured Learning for Semantic Role Labeling. *Transactions of the Association for Computational Linguistics (TACL)*, 3:29–41. (Cited on pages 3, 141 and 142)
- Taghipour, K. and Ng, H. T. (2015). Semi-supervised word sense disambiguation using word embeddings in general and specific domains. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 314–323, Denver, CO, USA. (Cited on pages 192 and 202)
- Taulé, M., Martí, M. A., and Borrega, O. (2010). AnCorra-Net: Mapping the Spanish AnCorra-Verb lexicon to VerbNet. In *Workshop on Verbs. The Identification and Representation of Verb Features*, Pisa, Italy. (Cited on page 31)
- Titov, I. and Klementiev, A. (2012). Semi-Supervised Semantic Role Labeling: Approaching from an Unsupervised Perspective. In *Proceedings of COLING 2012: Technical Papers (COLING)*, pages 2635–2652, Mumbai, India. (Cited on page 202)
- Tonelli, S., Bryl, V., Giuliano, C., and Serafini, L. (2012). Investigating the Semantics of Frame Elements. In *Knowledge Engineering and Knowledge Management*, volume 7603 of *Lecture Notes in Computer Science*, pages 130–143. Springer Berlin Heidelberg. (Cited on page 46)
- Tonelli, S., Giuliani, C., and Tymoshenko, K. (2013). Wikipedia-based WSD for multilingual frame annotation. *Artificial Intelligence*, 194:203–221. (Cited on pages 46, 70, 82, 104 and 105)
- Tonelli, S. and Giuliano, C. (2009). Wikipedia as Frame Information Repository. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 276–285, Singapore. (Cited on pages 6, 46, 70 and 82)
- Tonelli, S. and Pianta, E. (2009a). A novel approach to mapping FrameNet lexical units to WordNet synsets. In *Proceedings of the 8th International Conference on Computational Semantics (IWCS), short papers*, pages 342–345, Tilburg, The Netherlands. (Cited on pages 7 and 45)
- Tonelli, S. and Pianta, E. (2009b). Three Issues in Cross-Language Frame Information Transfer. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP)*, pages 441–448, Borovets, Bulgaria. (Cited on pages 82, 109 and 111)

- Tonelli, S. and Pighin, D. (2009). New Features for FrameNet – WordNet Mapping. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL)*, pages 219–227, Boulder, CO, USA. (Cited on pages 46 and 70)
- Tsai, R. T.-H., Chou, W.-C., Lin, Y.-C., Sung, C.-L., Ku, W., Su, Y.-S., Sung, T.-Y., and Hsu, W.-L. (2006). BIOSMILE: Adapting Semantic Role Labeling for Biomedical Verbs. In *Proceedings of the HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology*, pages 57–64, New York, NY, USA. (Cited on page 150)
- Van der Plas, L., Henderson, J., and Merlo, P. (2009). Domain Adaptation with Artificial Data for Semantic Parsing of Speech. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), Companion Volume: Short Papers*, pages 125–128, Boulder, CO, USA. (Cited on page 150)
- Van der Plas, L., Merlo, P., and Henderson, J. (2011). Scaling up Automatic Cross-Lingual Semantic Role Annotation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 299–304, Portland, OR, USA. (Cited on pages 109 and 111)
- Venturi, G., Lenci, A., Montemagni, S., Vecchi, E. M., Sagri, M. T., Tiscornia, D., and Agnoloni, T. (2009). Towards a FrameNet Resource for the Legal Domain. In *Proceedings of the Third Workshop on Legal Ontologies and Artificial Intelligence Techniques (LOAIT)*, pages 67–76, Barcelona, Spain. (Cited on page 150)
- Vossen, P., editor (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Dordrecht, The Netherlands. (Cited on pages 39 and 44)
- Vrandečić, D. and Krötzsch, M. (2014). Wikidata: A Free Collaborative Knowledgebase. *Communications of the ACM*, 57(10):78–85. (Cited on page 101)
- Woodsend, K. and Lapata, M. (2014). Text Rewriting Improves Semantic Role Labeling. *Journal of Artificial Intelligence Research*, 51:133–164. (Cited on pages 8, 109 and 114)
- Xu, W., Han, B., and Ritter, A., editors (2015). *Proceedings of the Workshop on Noisy User-generated Text (WNUT)*. Beijing, China. (Cited on page 11)
- Yang, B., tau Yih, W., He, X., Gao, J., and Deng, L. (2015a). Embedding Entities and Relations for Learning and Inference in Knowledge Bases. In *Proceedings of the International Conference on Learning Representations (ICLR)*, San Diego, CA, USA. (Cited on page 203)
- Yang, H., Zhuang, T., and Zong, C. (2015b). Domain Adaptation for Syntactic and Semantic Dependency Parsing Using Deep Belief Networks. *Transactions of the Association for Computational Linguistics (TACL)*, 3:271–282. (Cited on pages 9, 149 and 200)

- Yarowsky, D. (1995). Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (COLING)*, pages 189–196, Cambridge, MA, USA. (Cited on page [100](#))
- Yi, S.-t., Loper, E., and Palmer, M. (2007). Can Semantic Roles Generalize Across Genres? In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 548–555, Rochester, NY, USA. (Cited on page [31](#))
- Yimam, S. M., Eckart de Castilho, R., Gurevych, I., and Biemann, C. (2014). Automatic Annotation Suggestions and Custom Annotation Layers in WebAnno. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL). System Demonstrations*, pages 91–96, Baltimore, MD, USA. (Cited on page [157](#))
- Zapirain, B., Agirre, E., and Màrquez, L. (2008). Robustness and Generalization of Role Sets: PropBank vs. VerbNet. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pages 550–558, Columbus, OH, USA. (Cited on page [31](#))
- Zapirain, B., Agirre, E., Màrquez, L., and Surdeanu, M. (2013). Selectional Preferences for Semantic Role Classification. *Computational Linguistics*, 39(3):631–663. (Cited on page [125](#))
- Zhao, H., Chen, W., Kazama, J., Uchimoto, K., and Torisawa, K. (2009). Multilingual Dependency Learning: Exploiting Rich Features for Tagging Syntactic and Semantic Dependencies. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL): Shared Task*, pages 61–66, Boulder, CO, USA. (Cited on pages [148](#) and [149](#))
- Zhu, X. and Goldberg, A. B. (2009). *Introduction to Semi-Supervised Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan and Claypool Publishers, Williston, VT, USA. (Cited on pages [100](#) and [147](#))

Ehrenwörtliche Erklärung[†]

Hiermit erkläre ich, die vorgelegte Arbeit zur Erlangung des akademischen Grades „Dr. rer. nat.“ mit dem Titel „*Knowledge-based Supervision for Domain-adaptive Semantic Role Labeling*“ selbstständig und ausschließlich unter Verwendung der angegebenen Hilfsmittel erstellt zu haben. Ich habe bisher noch keinen Promotionsversuch unternommen.

Darmstadt, den 3. August 2016

Silvana Hartmann, Dipl. Ling.

[†] Gemäß § 9 Abs. 1 der Promotionsordnung der TU Darmstadt

Wissenschaftlicher Werdegang der Verfasserin[‡]

2002–2010 Diplomstudiengang Linguistik (Computerlinguistik)

Nebenfach: Informatik

Diplomarbeit: „Task-targeted Discriminative Language Modeling: Query Re-ranking“

Gutachter: Prof. Hinrich Schütze, Ph.D.

am Institut für Maschinelle Sprachverarbeitung

Universität Stuttgart

2010–2016 Wissenschaftliche Mitarbeiterin

am Ubiquitous Knowledge Processing Lab

Technische Universität Darmstadt

[‡]Gemäß § 20 Abs. 3 der Promotionsordnung der TU Darmstadt

Publikationsverzeichnis der Verfasserin

- Hartmann, S., Kuznetsov, I., Martin, T., and Gurevych, I. (2017). Out-of-domain FrameNet Semantic Role Labeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers (EACL 2017)*, pages 471–482, Valencia, Spain.
- Hartmann, S., Mújdricza-Maydt, É., Kuznetsov, I., Gurevych, I., and Frank, A. (2017). Assessing SRL Frameworks with Automatic Training Data Expansion. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 115–121, Valencia, Spain.
- Eckart de Castilho, R., Mújdricza-Maydt, É., Yimam, S. M., Hartmann, S., Gurevych, I., Frank, A., and Biemann, C. (2016). A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 76–84, Osaka, Japan.
- Mújdricza-Maydt, É., Hartmann, S., Gurevych, I., and Frank, A. (2016). Combining Semantic Annotation of Word Sense & Semantic Roles: A Novel Annotation Scheme for VerbNet Roles on German Language Data. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3031–3038, Portorož, Slovenia.
- Hartmann, S., Eckle-Kohler, J., and Gurevych, I. (2016). Generating Training Data for Semantic Role Labeling based on Label Transfer from Linked Lexical Resources. In *Transactions of the Association for Computational Linguistics*, vol. 4, pages 197–213.
- Hartmann, S. and Gurevych, I. (2013). FrameNet on the Way to Babel: Creating a Bilingual FrameNet Using Wiktionary as Interlingual Connection. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pages 1363–1373, Sofia, Bulgaria.
- Gurevych, I., Eckle-Kohler, J., Hartmann, S., Matuschek, M., Meyer, C. M., and Nghiem T.-D. (2013). UBY – A Large-Scale Lexical-Semantic Resource. In Theune, M., Nijholt A., Dadvar, M., Hondorp, H., Trieschnigg, D. and Truong K., editors, *Book of Abstracts of the 23rd Meeting of Computational Linguistics in the Netherlands: CLIN 2013*, page 81, Enschede, Netherlands.
- Eckle-Kohler, J., Gurevych, I., Hartmann, S., Matuschek, M., and Meyer, C. M. (2013). UBY-LMF – Exploring the Boundaries of Language-Independent Lexicon Models. In: Francopoulo, G., editor, *LMF Lexical Markup Framework*, chapter 10, p. 145–156, ISTE - HERMES - Wiley, London, UK.

- Gurevych, I., Matuschek, M., Nghiem, T.-D., Eckle-Kohler, J., **Hartmann, S.**, and Meyer C. M. (2012). Navigating Sense-Aligned Lexical-Semantic Resources: The Web Interface to UBY. In: *Proceedings of the 11th "Konferenz zur Verarbeitung natürlicher Sprache" (KONVENS 2012)*, p. 194–198, Vienna, Austria.
- Eckle-Kohler, J., Gurevych, I., **Hartmann, S.**, Matuschek, M., and Meyer, C. M. (2012). UBY-LMF - A Uniform Model for Standardizing Heterogeneous Lexical-Semantic Resources in ISO-LMF. In: Calzolari, N. et al., editors: *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, p. 275–282, Turkey, Istanbul.
- Chiarcos, C., Hellmann, S., Nordhoff, S., Moran, S., Littauer, R., Eckle-Kohler, J., Gurevych, I., **Hartmann, S.**, Matuschek, M., and Meyer C. M. (2012). The Open Linguistics Working Group. In: Calzolari, N. et al., editors: *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, p. 275–282, Turkey, Istanbul.
- Gurevych, I., Eckle-Kohler, J., **Hartmann, S.**, Matuschek, M., Meyer, C. M., and Wirth, C. (2012). UBY – A Large-Scale Unified Lexical-Semantic Resource Based on LMF. In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, p. 580–590, Avignon, France.
- Chiarcos, C., Hellmann, S., Nordhoff, S., Cimiano, P., McCrae, J., Brekle, J., Eckle-Kohler, J., Gurevych, I., **Hartmann, S.**, Matuschek, M., Meyer, C. M., and Littauer, R. (2012). The Working Group for Open Data in Linguistics. In: *Sprache als komplexes System. Proceedings der 34. Jahrestagung der DGfS.*, p. 284–285, Frankfurt am Main, Germany.
- Hartmann, S.**, Szarvas, G., and Gurevych, I. (2012). Mining Multiword Terms from Wikipedia. In Paziienza, M. T., and Stellato, A., editors: *Semi-Automatic Ontology Development: Processes and Resources*, p. 226–258, IGI Global, Hershey, PA, USA.
- Alfonseca, E., Hall, K., and **Hartmann, S.** (2009). Large-scale Computation of Distributional Similarities for Queries. In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), Companion Volume: Short Papers*, p. 29–32, Boulder, CO, USA.
- Hartmann, S.** (2010). Task-targeted Discriminative Language Modeling: Query Reranking. Diplomarbeit, Universität Stuttgart, Germany.