Generating Training Data for Semantic Role Labeling based on Label Transfer from Linked Lexical Resources

Silvana Hartmann[◇], **Judith Eckle-Kohler**[◇], **Iryna Gurevych**[◇‡]

◇ UKP Lab, Technische Universität Darmstadt

‡ UKP Lab, German Institute for Educational Research

http://www.ukp.tu-darmstadt.de

Abstract

We present a new approach for generating role-labeled training data using Linked Lexical Resources, i.e., integrated lexical resources that combine several resources (e.g., WordNet, FrameNet, Wiktionary) by linking them on the sense or on the role level. Unlike resource-based supervision in relation extraction, we focus on complex linguistic annotations, more specifically FrameNet senses and roles. The automatically labeled training data (www.ukp.tu-darmstadt.de/knowledge-based-srl/) are evaluated on four corpora from different domains for the tasks of word sense disambiguation and semantic role classification. Results show that classifiers trained on our generated data equal those resulting from a standard supervised setting.

1 Introduction

In this work, we present a novel approach to automatically generate training data for semantic role labeling (SRL). It follows the distant supervision paradigm and performs knowledge-based label transfer from rich external knowledge sources to large corpora.

SRL has been shown to improve many NLP applications that rely on a deeper understanding of semantics, such as question answering, machine translation or recent work on classifying stance and reason in online debates (Hasan and Ng, 2014) and reading comprehension (Berant et al., 2014).

Even though unsupervised approaches continue to gain popularity, SRL is typically still solved using supervised training on labeled data. Creating such labeled data requires manual annotations by experts,[1] resulting in corpora of highly limited size. This is especially true for the task of FrameNet SRL where the amount of annotated data available is small.

FrameNet SRL annotates fine-grained semantic roles in accordance with the theory of Frame Semantics (Fillmore, 1982) as illustrated by the following example showing an instance of the *Feeling* frame including two semantic roles:

> *He*$_{Experiencer}$ *felt*$_{Feeling}$ *no sense of guilt*$_{Emotion}$ *in the betrayal of personal confidence.*

Our novel approach to training data generation for FrameNet SRL uses the paradigm of distant supervision (Mintz et al., 2009) which has become popular in relation extraction. In distant supervision, the overall goal is to align text and a knowledge base, using some notion of similarity. Such an alignment allows us to transfer information from the knowledge base to the text, and this information can serve as labeling for supervised learning. Hence, unlike semi-supervised methods which typically employ a supervised classifier and a small number of seed instances to do bootstrap learning (Yarowsky, 1995), distant supervision creates training data in a single run. A particular type of knowledge base relevant for distant supervision are linked lexical resources (LLRs): integrated lexical resources that combine several resources (e.g., WordNet, FrameNet, Wiktionary) by linking them on the sense or on the role level.

Previous approaches to generating training data for SRL (Fürstenau and Lapata, 2012) do not use lexical resources apart from FrameNet. For the task of

[1] Even though crowdsourcing has been used, it is still problematic for SRL labeling: the task is very complex, which results in manually adapted definitions (Fossati et al., 2013), or constrained role sets (Feizabadi and Padó, 2014).

Transactions of the Association for Computational Linguistics, vol. 4, pp. 197–213, 2016. Action Editor: Yuji Matsumoto.
Submission batch: 9/2015; Revision batch: 1/2016; Published 5/2016.

word sense disambiguation (WSD), recent work on automatic training data generation based on LLR has only used WordNet (Cholakov et al., 2014), not considering other sense inventories such as FrameNet.

Our distant supervision approach for automatic training data generation employs two types of knowledge sources: LLRs and linguistic knowledge formalized as rules to create data labeled with FrameNet senses and roles. It relies on large corpora, because we attach labels to corpus instances only sparsely.

We generate training data for two commonly distinguished subtasks of SRL: first, for disambiguation of the frame-evoking lexical element relative to the FrameNet sense inventory, a WSD task; and second, for argument identification and labeling of the semantic roles, which depends on the disambiguation result. Regarding the subtask of FrameNet WSD, we derive abstract lexico-syntactic patterns from lexical information linked to FrameNet senses in an LLR and recover them in large-scale corpora to create a sense (frame) labeled corpus. We address the subsequent steps of argument identification and role labeling by making use of linguistic rules and role-level links in an LLR, creating a large role-labeled corpus with more than 500,000 roles.

We extrinsically evaluate the quality of the automatically labeled corpora for frame disambiguation and role classification for verbs, using four FrameNet-labeled test-sets from different domains, and show that the generated training data is complementary to the FrameNet fulltext corpus: augmenting it with the automatically labeled data improves on using the FrameNet training corpus alone. We also evaluate our approach on German data to show that it generalizes across languages. We discuss in detail how our method relates to and complements recent developments in FrameNet SRL. The need for additional training data has also been reported for state-of-the-art systems (FitzGerald et al., 2015).

Our work has three main contributions: (i) for automatic sense labeling, we significantly extend Cholakov et al. (2014)'s distant supervision approach by using discriminating patterns and a different sense inventory, i.e., FrameNet. We show that discriminating patterns can improve the quality of the automatic sense labels. (ii) We use a distant supervision approach – building on LLRs – to address the complex problem of training data generation for FrameNet role labeling, which builds upon the sense labeling in (i). (iii) Our detailed evaluation and analysis show that our approach for data generation is able to generalize across domains and languages.

The rest of this paper is structured as follows: after introducing our approach to training data generation in section 2, we describe the automatic sense labeling (section 3) and role classification (section 4) in detail. In section 5 we apply our approach to German data. We present related work in section 6 and discuss our approach in relation to state-of-the-art FrameNet SRL in section 7, followed by discussion and outlook in section 8. Section 9 concludes.

2 Knowledge-based Label Transfer

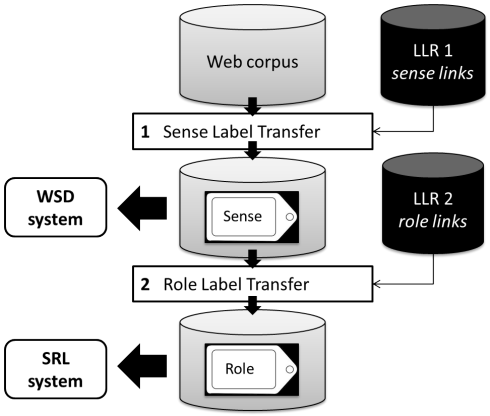

Figure 1: Automatic training data generation – overview.

Our distant supervision method for generating training data for SRL consists of two stages, first generating sense-labeled data, then extending these to role-labeled data, as shown in Fig.1. Both stages use large-scale corpora and LLRs as knowledge sources.

Knowledge sources. For the first stage, presented in detail in section 3, we use sense-level information from the LLR Uby (LLR 1 in Fig.1) and exploit the sense links between the Uby versions of FrameNet, WordNet, Wiktionary and VerbNet. More specifically, we employ (i) sense examples from FrameNet, WordNet, and Wiktionary, and (ii) VerbNet information, i.e., syntactic subcategorization frames, as well as semantic roles and selectional preference information of the arguments. It is important to note that the sense examples in FrameNet (called *lexical unit examples*) are a different resource than the FrameNet fulltext corpus.

For the second stage, presented in section 4, we use the LLR SemLink (LLR 2 in Fig.1) and exploit the role-level links between VerbNet semantic roles and FrameNet roles. SemLink (Bonial et al., 2013) contains manually curated mappings of the fine-grained FrameNet roles to 28 roles in the VerbNet role inventory, including more than 1600 role-level links.

Formalization. More formally, we can cast our distant supervision method for automatic training data generation as a knowledge-based label transfer approach. 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 $y_i \in Y$; $i = 1, \ldots, n$. For an unlabeled instance $u_j \in U$, $j = 1, \ldots, m$, 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 LLRs including sense-level links, and c considers the similarity of the patterns based on a similarity metric.

We create role-labeled data with rule-based labeling where Y is the set of role labels, r_{x_i} and r_{u_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 LLRs; here, c is fulfilled if the attribute representations match.

Experimental setup. In our distant supervision approach to training data generation, we (i) create our training data in a single run (and not iteratively) (ii) perform sparse labeling in order to create training data, i.e., we need a *very large corpus* (e.g., unlabeled web data) in order to obtain a sufficient number of training instances. We analyze the resulting labeled corpora and evaluate them extrinsically using a classifier trained on the automatically labeled data on separate test datasets from different domains. This way, we can also show that a particular strength of our approach is to generalize across domains.

In the next section we present the automatic creation of sense-labeled corpora (stage 1).

3 Automatic Labeling for Word Sense

In this work, we are the first to apply distant supervision-based verb sense labeling to the FrameNet verb sense inventory. We extend the methodology by Cholakov et al. (2014) who also exploit sense-level information from the LLR Uby for the automatic sense-labeling of corpora with verb senses, but use WordNet as a sense inventory. We use the same types of information and similarity metric (which we call sim in this paper), but our label space Y is given by the FrameNet verb sense inventory ($|Y|$=4,670), and therefore we exploit 42,000 sense links from FrameNet to WordNet, VerbNet and Wiktionary.

The similarity metric $sim \in [0..1]$ is based on Dice's coefficient and considers the common n-grams, $n = 2, \ldots, 4$:

$$(1) \qquad sim(r_{x_i}, r_{u_j}) = \frac{\sum_{n=2}^{4} |G_n(p_1) \cap G_n(p_2)| \cdot n}{norm_w}$$

where $w >= 1$ 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_{u_j}, and $norm_w$ is the normalization factor defined by the sum of the maximum number of common n-grams in the window w.[2]

Step 1A: Seed Pattern extraction and filtering. We call the sense patterns r_{x_i} generated from seed instances x_i in the LLR Uby *seed patterns* and follow Cholakov et al. (2014) for the generation of those seed patterns, i.e., we create lemma sense patterns (LSPs), consisting of the target verb lemma and lemmatized context only, and second, abstract sense patterns (ASPs), consisting of the target verb lemma and a number of rule-based generalizations of the context words. An example of each of the sense patterns 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* is:

1. **LSP**: he **feel** no sense of guilt in
2. **ASP**: PP **feel** *cognition* of *feeling* in *act*

ASPs generalize to a large number of contexts which is particularly important for identifying productively used verb senses, while LSPs serve to identify fixed constructions such as multiword expressions.

[2]Using n-grams instead of unigrams takes word order into account, which is particularly important for verb senses, as syntactic and semantic properties often correlate.

A drawback of Cholakov et al. (2014)'s method for seed pattern extraction is that it extracts a certain number of very similar (or even identical) seed patterns for *different* senses. Those seed patterns may lead to noise in the sense-labeled data. To prevent this problem, we developed an optional *discriminating filter* that removes problematic seed patterns.

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, $i, j \in (1, \ldots, n)$, 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; if we find two patterns w_i, w_j whose similarity score exceeds a filtering threshold f, we greedily discard them both.

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 discriminating filtering improves the quality of the automatically labeled corpus. Essentially, our discriminating filter integrates the goal of capturing sense distinctions into our approach. 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 was a lexicographic effort. The following example compares two ASP patterns to a CPA pattern from Popescu et al. (2014):

1. CPA: [[Human]] | [[Institution]] **abandon** [[Activity]] | [[Plan]]
2. ASP: JJ person **abandon** JJ cognition of JJ quantity
3. ASP: person **abandon** communication which VVD PP JJ in

Our abstract ASP patterns look similar, as they also abstract argument fillers to semantic classes and preserve certain function words.

Step 1B: Sense label transfer. Using the approach of Cholakov et al. (2014), we create sense patterns r_{u_j} from all sentences u_j of an unlabeled corpus that contain a target verb, for instance the sentence u_1: *I feel strangely sad and low-spirited today* for the verb *feel*. For every u_j, its sense pattern r_{u_j} is then compared to the labeled seed patterns using the similarity metric sim. From the most similar seed patterns $\{r_{x_i}, y_i\}$ that have a similarity score above a threshold t, the set of candidate labels $\{y_i\}$ is extracted.[3] The approach picks a random sense label from $\{y_i\}$ and attaches it to u_j.

If an ASP and an LSP receive the same similarity score, LSPs get precedence over ASPs, i.e., the labeled sense is selected from the senses associated with the LSP. Using this method, our example sentence u_1 receives the label *Feeling*.

This approach leads to a sparse labeling of the unlabeled corpus, i.e., many unlabeled sentences are discarded because their similarity to the seed pattern is too low. It however scales well to large corpora because it requires only shallow pre-processing, as we will show in the next section: we apply our approach to a large web corpus and analyze the resulting sense-labeled corpora.

3.1 Creating the sense-labeled corpus

Unlabeled corpora. We used parts 1 to 4 of the ukWAC corpus (Baroni et al., 2009) as input for the automatic sense label transfer for the 695 verb lemmas in our test sets (see section 3.2, Test data).

Seed Patterns. The seed patterns for Step 1A of the sense labeling were extracted from a) the FrameNet example sentences, and b) sense examples from resources linked to FrameNet in Uby, namely WordNet, Wiktionary, and VerbNet. Without a discriminating filter, this results in more than 41,700 LSP and 322,000 ASP, 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, leading to 15% LSP and 85% ASP. The number of senses with patterns decreases from 4,900 to 3,900.

Threshold setting. In order to determine the parameter values for the label transfer, i.e., which values

[3] Unless t is very high, there is usually more than one candidate sense label.

f	t	P	R	F1
-	**0.07**	0.672	**0.723**	**0.696**
-	0.1	0.672	0.712	0.692
-	0.14	0.665	0.642	0.653
-	0.2	0.68	0.633	0.656
0.2	0.2	0.683	0.566	0.619
0.14	0.2	0.689	0.566	0.621
0.1	0.2	0.702	0.544	0.613
0.07	**0.2**	**0.713**	0.526	0.605

Table 1: Combinations of f and t evaluated on FNFT-dev; configurations for best F1, R and P in bold.

for threshold t, and filter f result in a high-quality training corpus, we perform an extrinsic evaluation on a development set: we use a set of automatically labeled corpora based on ukWAC section 1 generated with different threshold values to train a verb sense disambiguation (VSD) system. We evaluate precision P (the number of correct instances/number of labeled instances), recall R (the number of labeled instances/all instances), and F1 (harmonic mean of P and R) of the systems on the development-split FNFT-dev of the FrameNet 1.5 fulltext corpus (FNFT), used by Das and Smith (2011). A detailed description of the VSD system follows in the next section.

We varied the thresholds of the discriminating filter f (Step 1A) and the threshold t (Step 1B) on the values $(0.07, 0.1, 0.14, 0.2)$, as was suggested by Cholakov et al. (2014) for t. We also compare corpora with and without the discriminating filter f. To save space, we only report results with f for $t = 0.2$ in Table 1.

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. 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. Note that the discriminating filter gets stricter, i.e. 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.

Automatically labeled corpora. The setting with the highest F1 in Table 1 leads to the very large sense-labeled corpus **WaS_XL**. We also use f and t values with the highest precision in order to evaluate the benefits of the discriminating filter, leading to **WaS_L**.

	instances	**s**enses	**v**erbs	**s/v**	**i/s**
WaS_XL (t=0.07)	1.6*10^6	1,460	637	1.8	1,139
WaS_X (t=0.2)	193,000	1,249	602	1.7	155
WaS_L (t=0.2, f=0.07)	109,000	1,108	593	1.5	98
FNFT$\star$	5,974	856	575	1.5	10

Table 2: Sense statistics of automatically labeled corpora.

The size of these corpora ranges from 100,000 to 1.6 million sense instances with an average of 1.5 to 1.8 senses per verb, compared to 6,000 verb sense instances in FNFT$\star$, FNFT filtered by the 695 verbs in our four test sets, see Table 2.

We compare WaS_L to WaS_X, the corpus labeled with $t = 0.2$, but without filter f in order to evaluate the impact of adding the discriminating filter. Compared to the latter corpus, WaS_L contains 44% fewer sense instances, but only 12% fewer distinct senses, and 75% of the senses which are also covered by WaS_XL. The number of instances per sense is Zipf-distributed with values ranging from 1 to over 40,000, leading to the average of 1,139 reported in Table 2 for WaS_XL.

3.2 VSD experiments

To compare the quality of our automatically sense-labeled corpora to manually labeled corpora, we perform extrinsic evaluation in a VSD task. **VSD system.** We use a standard supervised setup for sense disambiguation: we extract lexical, syntactic, and semantic features from the various training sets and the test sets. For pre-processing, we use DKPro Core (Eckart de Castilho and Gurevych, 2014), e.g., Stanford tokenizer, TreeTagger for POS tagging and lemmatization, StanfordNamedEntityRecognizer for NER and Stanford Parser for dependency parsing. We train a logistic regression classifier in the WEKA implementation (Hall et al., 2009) using the same features as Cholakov et al. (2014).

Training Corpora. We trained our VSD system on **WaS_XL** and **WaS_L**, and, for comparison, on the training split **FNFT-train** of FNFT 1.5 used by Das and Smith (2011).

Test data. For evaluation, we used four different FrameNet-labeled datasets. The statistics of the test datasets are compiled in Table 3, a brief description of

	verbs	senses	s/v	inst(s)	inst(r)
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: Test dataset statistics on verbs; inst(**s/r**): number of ambiguous sense and role instances in the datasets.

each dataset follows. We use the frame and role annotations in the **Semeval 2010 task 10** evaluation and trial dataset (Ruppenhofer et al., 2010). It consists of literature texts. The **Fate corpus** contains frame annotations on the RTE-2 textual entailment challenge test set (Burchardt and Pennacchiotti, 2008). It is based on newspaper texts, texts from information extraction datasets such as ACE, MUC-4, and texts from question answering datasets such as CLEF and TREC. These two datasets were created prior to the release of FrameNet 1.5. For those sets, only senses (verb-frame combinations) that still occur in FrameNet 1.5 and their roles were included in the evaluation. The **MASC WordSense sentence corpus** (Passonneau et al., 2012) is a balanced corpus that contains sense annotations for 1000 instances of 100 words from the MASC corpus. It contains WordNet sense labels, we use a slightly smaller subset of verbs annotated with FrameNet 1.5 labels.[4] We also evaluate on the test-split **FNFT-test** of the FrameNet fulltext corpus used in Das and Smith (2011).

3.3 VSD results and analysis.

Impact of pattern filters. A comparison of results between the WaS corpora (first block of Table 4) shows that the filters in WaS_L improve 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. Due to the larger polysemy in MASC – on average 3.3 senses per verb (see s/v in Table 3), it 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.

Analysing the results in detail for the example verb *tell* shows that WaS_XL contains all 10 senses of tell in MASC; 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 instances of the rare senses for stricter thresholds t and f.

These 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.

For WaS_XL, recall is significantly higher for all test sets, leading to a higher F1. All significance scores reported in this paper are based on Fisher's exact test at significance level $p<0.05$.

Comparison to FNFT-train. We also compare the results of our WaS-corpora to a VSD system trained on the reference corpus FNFT-train (see Table 4). On Semeval, precision does not deviate significantly from the FNFT-train system. On FNFT-test, it is significantly lower. For WaS_XL, precision is significantly lower on Fate, but significantly higher on MASC. For WaS_L the precision is similar on MASC and Fate. For WaS_XL, the recall is significantly higher than for FNFT-train on all test sets, leading to a higher F1. This is the result of the larger sense coverage of the FrameNet lexicon, which provided the seeds for the automatic labeling.

Training our system directly on the FrameNet lexical unit examples is, however, not a viable alternative: it leads to a system with similar precision our WaS-corpora, but very low recall (between 0.22 and 0.37). By using the sense examples in our seed patterns, we retain their benefits on sense coverage, and improve the system recall and F1 at the same time.

Comparative analysis. In a detailed analysis of our results, we compare the performance of the WaS_XL and FNFT-train based systems on those verbs of each test set that are evaluated for both systems, i.e., the intersection I_t. On I_t, precision and F1 are higher for FNFT-train for all test sets except MASC. For MASC, precision is similar, but recall is 0.21 points higher. For the verbs in I_t, the average number of training senses in WaS_XL is two senses higher than for FNFT-train. This larger sense coverage of the WaS_XL is beneficial to recall on the

[4]This subset is currently not part of the MASC download, but according to personal communication with the developers will be published soon.

	FNFT-test			Fate			MASC			Semeval		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
WaS_XL	0.647*	0.816*	0.722	0.628*	0.65*	0.639	0.66*	0.793*	0.72	0.665	0.761*	0.71
WaS_L	0.68*	0.618	0.648	0.66	0.505*	0.572	0.639	0.707*	0.671	0.694	0.62*	0.655
FNFT-train	0.729	0.643	0.683	0.7	0.38	0.493	0.598	0.339	0.433	0.706	0.55	0.618
B-WaS_XL	0.736	0.767*	0.751	0.686	0.619*	0.651	0.67*	0.699*	0.684	0.724	0.71*	0.717
U-WaS_XL	0.668*	0.935*	0.78	0.63*	0.683*	0.656	0.642*	0.833*	0.725	0.667	0.849*	0.747

Table 4: VSD P, R, F1; * marks significant differences to the system trained on FNFT-train.

MASC test set which shows high polysemy.

Evaluating on the set difference between the systems (test verbs that remain after the intersection is removed), we see that the lemma coverage of WaS_XL is complementary to FNFT-train. The difference D_t 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 train set to all evaluated instances ranges between 11% and 48% for WaS_XL, and between 5% and 30% for FNFT-train.

Combining training data. The complementary nature of the sets led us to evaluate two combinations of training sets: U-WaS_XL consists of the union of WaS_XL and FNFT-train, B-WaS_XL implements a backoff strategy and thus consists of FNFT-train and those instances of WaS_XL whose lemmas are not contained in the intersection with FNFT-train (i.e., if FNFT-train does not contain enough senses for a lemma, supplement with WaS_XL).

The third block in Table 4 shows that precision is higher or not significantly lower for B-WaS_XL compared to FNFT-train, recall and F1 are higher. U-WaS_XL leads to higher recall compared to B-WaS_XL, and allover highest F1. This proves that our automatically labeled corpus WaS_XL is complementary to the manually labeled FNFT-train and contributes to a better coverage on diverse test sets.

Multiword verbs. Our approach of training data generation also includes multiword verbs such as *carry out*. We treat those verb senses as additional senses of the head verb, for which we also create sense patterns, i.e., the sense for *carry out* is a specific sense of *carry*. As a result, we do not need to rely on additional multiword detection strategies for VSD.

Our WaS_XL contains more than 100,000 sense instances of 194 multiword verbs, of which 35 have multiple FrameNet senses. We specifically evaluated the performance of our VSD system on multiwords and their head verbs from MASC which contains 81 relevant sense instances. The precision is 0.66 compared to 0.59 when training on FNFT-train, at slightly higher coverage. While the test set is too small to provide significant results, there is an indication that the automatically labeled data also contribute to the disambiguation of multiword verbs.

This analysis concludes our section on automatic sense labeling. In the next section, we will describe our method for automatically adding FrameNet role labels to the WaS-corpora.

4 Automatic Labeling for Semantic Roles

In this section, we present our linguistically informed approach to the automated labeling of FrameNet roles in arbitrary text. Our method builds on the results of rich linguistic pre-processing including dependency parsing and uses role-level links in the LLR SemLink and the sense labels from section 3. First, a set of deterministic rules is applied to label syntactic arguments with VerbNet semantic roles. Then, we map the VerbNet semantic roles to FrameNet roles based on role-level links in SemLink and the automatically created sense labels.

Step 2A: VerbNet role label transfer. Our precision-oriented deterministic rules build on the results of linguistic pre-processing. Our pre-processing pipeline[5] performs lemmatization, POS tagging, named-entity-recognition and parsing with the Stanford Parser (de Marneffe et al., 2006), as well as semantic tagging with WordNet semantic fields.[6] Step

[5]We used the components from DKPro Core (Eckart de Castilho and Gurevych, 2014).

[6]We used the most-frequent-sense disambiguation heuristic which works well for the coarse-grained semantic types given by the WordNet semantic fields. Named-entity tags are also mapped

1 provides FrameNet sense labels for the target verbs.

Dependency parsing annotates dependency graphs, linking a governor to its dependents within a sentence. Governors and dependents are represented by the heads of the phrases they occur in. For verbal governors, the dependency graphs correspond to predicate argument structures with the governor being the predicate and the dependents corresponding to the argument heads.

Our rules attach VerbNet role labels to dependent heads of their verbal governors. We can then derive argument spans by expanding the dependent heads by their phrases. The semantic role inventory as given by VerbNet is our label space Y ($|Y|$=28). Rule-based role labeling can be seen as label transfer where a corpus instance u_j is given by the dependent of a verbal governor and its sentential context, including all linguistic annotations. Then r_{u_j} is compared to a prototypical attribute representation r_{x_i} of a semantic role, derived from linguistic knowledge.[7]

More specifically, we iterate over the collapsed dependencies annotated by the Stanford parser and apply a hierarchically organized chain of 57 rules to the dependents of all verbal governors. In this rule chain, *Location* and *Time* roles are assigned first, in case a dependent is a *location* or has the semantic field value *time*. Then, the other roles are annotated. This is done based on the dependency type in combination with named entity tags or semantic fields, either of the dependent or the verbal governor or both. An example rule is: for the dependency *nsubj*, the role *Experiencer* is annotated if the governor's semantic field is *perception* or *emotion*, and the role *Agent* otherwise. This way, *I* in our example *I* $\underline{feel}_{Feeling}$ *strangely sad and low-spirited today* is annotated with the *Experiencer* role.

Some rules also check the semantic field of the dependent, e.g., the dependency *prep_with* triggers the annotation of the role *Instrument*, if the dependent is neither a *person* nor a *group*. Often, it is not possible to determine a single VerbNet role based on the available linguistic information (32 rules assign one role, 5 rules assign 2 roles, and 20 rules assign 3 roles), e.g., the distinction between *Theme* and *Co-Theme* can not be made. In such cases, multiple roles are annotated, which are all considered in the subsequent Step 2B. Evaluated on a test sample of VerbNet annotations on PropBank, the percentage of correctly annotated roles among all annotated roles is 96.8% – instances labeled with multiple roles are considered correct if the set of roles contains the gold label. The percentage of instances where a rule assigns at least one role was 77.4%.

Step 2B: Mapping VerbNet roles to FrameNet roles. Finally, the annotated VerbNet roles are mapped to FrameNet roles using (i) the automatically annotated FrameNet sense and (ii) the SemLink mapping of VerbNet roles to FrameNet roles for this FrameNet sense (frame).

The information on the FrameNet frame is crucial to constrain the one-to-many mapping of VerbNet roles to fine-grained FrameNet roles. For example, the VerbNet role *Agent* is mapped to a large number of different FrameNet roles across all frames. While the SemLink mapping allows unique FrameNet roles to be assigned in many cases, there are still a number of cases left where the rule-based approach annotates a set of FrameNet roles. Examples are *Interlocutor_1* and *Interlocutor_2* for the *Discussion* frame, or *Agent* and *Cause* for the *Cause_harm* frame. For the former the distinction between the roles is arbitrary, while for the latter further disambiguation may be desired. As the SemLink mapping is not complete, our approach results in partially labeled data, i.e., a sentence may contain only a single predicate-role pair, even though other arguments of the predicate are present. Our experiments show that we can train semantic role classifiers successfully on partially labeled data.

We used the training set from Das and Smith (2011) (annotated with FrameNet roles) as a development set. Evaluated on the test set from Das and Smith (2011), the percentage of correctly annotated roles among all annotated roles is 76.74%.[8]

4.1 Creating the role-labeled corpus

We use the two sense-labeled corpora WaS_XL and WaS_L as input for the automatic role label transfer, creating role-labeled corpora WaSR_XL and WaSR_L. We distinguish two variants of these cor-

to WordNet semantic fields.

[7] It took a Computational Linguist 3 days to develop the rules, using a sample of the VerbNet annotations on PropBank from SemLink as a development set.

[8] As in Step 2A, instances labeled with a set of roles are considered correct if the set contains the gold label.

	instances	roles	senses	r/s	i/r
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 5: Role statistics of automatically labeled corpora.

pora, one that only contains those role instances with a unique role label, marked with the suffix -uni in Table 5, and one that additionally includes sets of labels, marked with the suffix -set.

For WaSR_XL, Step 2A results in 1.9 million arguments labeled with VerbNet roles. This number is reduced by 66% in Step 2B as a result of the incomplete mapping between VerbNet and FrameNet senses and roles in SemLink.

Table 5 shows that the resulting corpora contain 34,000 (WaSR_L-uni) and 549,000 (WaSR_XL-uni) uniquely assigned role instances for the verbs in our test sets, a lot compared to the 13,000 instances in FNFT⋆, FNFT filtered by the 695 verbs in our four test sets. The counts are even higher for the corpora including sets of labels.

Due to the sparse labeling approach, our WaSR corpora contain on average 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 added.

4.2 Role classification experiments

Role classification system. We trained a supervised system for semantic role classification as a log-linear model per verb-frame using the features described in Fürstenau and Lapata (2012).

Note that we do not evaluate the task of argument identification. Argument identification is performed by our rule-based VerbNet role transfer and follows common syntactic heuristics based on dependency parsing. Following Zapirain et al. (2013), we specifically consider the subtask of role classification, as we focus on the quality of our data on the semantic level. In this context it is important that the features of our role classifier 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 VSD.

Training and test data. We compare our role classifier trained on WaSR_XL-(set/uni) and WaSR_L-(set/uni) to the one based on FNFT-train. Test datasets are the same as for VSD, see Table 3.

4.3 SRL results and analysis

Results on WaSR corpora. We evaluate P, R, and F1 on all frame-verb combinations for which there is more than one role in our training data. Training the system on WaSR_XL-set and WaSR_L-set include training instances with sets of role labels. Therefore, sets of role labels are among the predicted labels. In the evaluation, we count the label sets as correct if they contain the gold label.

As expected, WaSR_XL-set 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, see Table 6.

We skip the WaSR_L-* corpora in Table 6, because 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-* on all test sets because of a smaller number of role-labeled instances in the WaSR_L-* corpora (see Table 5).

Comparison to FNFT-train. Table 6 compares the results of WaSR_XL-* to the system trained on FNFT-train. Note that we emulated the lenient evaluation setting for FNFT-train by retrieving the label set S_l in WaSR_XL-set for a label l predicted by the FNFT-train system and counting l as correct if any of the labels in S_l matches the gold label. We, however, did not find any difference to the regular evaluation; it appears that the labeling errors of the FNFT-train-based system are different from the label sets resulting from our labeling method.

The precision for WaSR_XL-uni matches the precision for FNFT-train for the Semeval and Fate test sets (the difference is 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 FNFT-train system, significantly for Fate. Recall of the WaSR corpora is significantly lower allover, as a result of the sparse, partial labeling and the lower role coverage of our automatically labeled corpora.

	FNFT-test			Fate			MASC			Semeval		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
WaSR_XL-uni	0.658*	0.333*	0.442	0.619	0.281*	0.387	0.652*	0.253*	0.365	0.689	0.394*	0.501
WaSR_XL-set	0.705*	0.398*	0.509	0.733*	0.337*	0.462	0.648*	0.297*	0.408	0.722	0.441*	0.547
FNFT-train	0.741	0.831	0.783	0.652	0.642	0.647	0.724	0.527	0.61	0.705	0.625	0.663
B-WaSR_XL-uni	0.728*	0.878*	0.796	0.645	0.698*	0.67	0.718	0.574*	0.638	0.696	0.71*	0.703
U-WaSR_XL-uni	0.691*	0.883*	0.776	0.629	0.701*	0.663	0.677*	0.579*	0.624	0.671	0.721*	0.695

Table 6: Role classification P, R, F1; * marks significant differences to the system trained on FNFT-train.

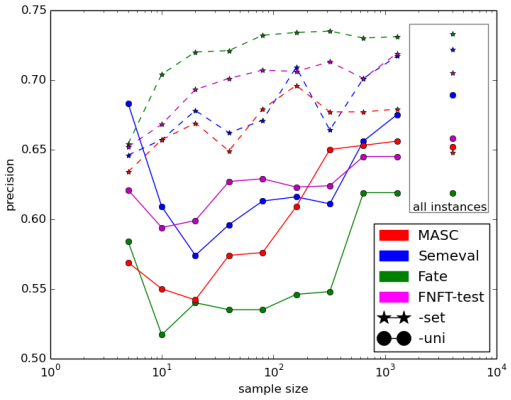

Figure 2: Role classification learning curves.

Comparative analysis. We compare the performance of our WaSR_XL-uni and the FNFT-train based system on the intersection of the evaluated senses between both systems. Precision of FNFT-train is higher on the intersection, except for Semeval, where it is similar. FNFT-train evaluates on average two more roles per sense than the WaSR. Evaluating only on the difference, the instances not contained in the intersection, we see that WaSR_XL-uni contributes some instances that are not covered by FNFT-train. These constitute between 7% and 18% of the total evaluated instances, compared to 26% to 50% instances added by FNFT-train. The precision of WaSR_XL-uni on the intersection for MASC is high at 0.68, compared to 0.55 for FNFT-test (not shown in Table 6). These results indicate that our WaSR_XL-uni is complementary to FNFT-train.

Combining training data. To give further evidence of the complementary nature of the automatically labeled corpus, we run experiments that combine WaSR_XL-uni with FNFT-train. We again use the union of the datasets (U-WaSR_XL-uni) and backing off to WaSR_XL-uni when FNFT-train does not provide enough roles for a sense (B-WaSR_XL-uni).

Table 6 shows better results for the backoff corpus than for the union. Recall is significantly higher compared to FNFT-train, and precision values are not significantly lower except for FNFT-test. This demonstrates that our automatically role-labeled corpora can supplement a manually labeled corpus and benefit the resulting system.

WaSR sampling. Because our WaSR corpora show a Zipfian distribution of roles (there are a few roles with a very large number of instances), we randomly sample nine training sets from WaSR_XL with a different maximal number of training instances per role s such that $s = 5 \cdot 2^i$ for $i \in \{0, 1, .., 8\}$, i.e., s ranges from 5 to 1280. Fig. 2 shows the learning curves for precision on WaSR_XL-*. 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 allover highest precision (0.738) and F1 (0.65).

5 German Experiments

To show that our method generalizes to other languages, we applied it to German data.

We used the SALSA2 corpus (Burchardt et al., 2006) as a source of German data with FrameNet-like labels. As SALSA2 does not provide additional lexical unit examples, we split the corpus into a training set S-train that is also used for the extraction of seed patterns, a development set S-dev, and a test set S-test. The proportion of train, development and test instances is 0.6, 0.2, 0.2; data statistics are shown in Table 7. The unlabeled corpus used is based on deWAC sections 1-5 (Baroni et al., 2009).

VSD corpus and evaluation. The LLR used to generate more than 22,000 seed patterns consists

	verbs	senses	roles	inst(s)	inst(r)
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	155	80,370	115,332
WaSR-de-uni	172	241	210	51,241	57,822

Table 7: German dataset statistics on verbs.

of S-train and the German Wiktionary based on the linking by Hartmann and Gurevych (2013).

DeWAC is labeled based on those patterns, and the thresholds t and f are determined in a VSD task on S-dev using a subset of the corpus based on sections 1-3. The threshold t=0.07 together with a discriminating filter of f=0.07 result in the best precision, and t=0.07 in the best F1 score. Therefore, we perform extrinsic evaluation in a VSD task on S-test with WaS-de (t=0.07) and on the combinations U-WaS-de (union with S-train) and B-WaS-de (backoff-variant).

The results in Table 8 show that the performance of the WaS-de-based system is worse than the S-train-based one, but the backoff version reaches best scores allover, indicating that our WaS-de corpora are complementary to S-train.

	P	R	F1
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 8: German VSD P, R, F1; * marks significant differences to S-train.

SRL corpus and evaluation. We adapt the rule-based VerbNet role-labeling to German dependencies from the mate-tools parser (Seeker and Kuhn, 2012), and perform Steps 2A and 2B on WaS-de, resulting in WaSR-de-set/uni (see Table 7).

We train our role classification system on the corpora in order to evaluate them extrinsically. Training on WaSR-de-uni results in precision of 0.69 – better than for English, but still significantly lower than for the S-train system with 0.828. Recall is very low at 0.17. This is due to the low role coverage of the WaSR corpora shown in Table 7.

The evaluation shows that our approach can be applied to German. For VSD, the automatically labeled data can be used to improve on using S-train alone; improvements in precision are not significant, which has several potential causes, e.g., the smaller set of LLRs used for seed pattern extraction compared to English, and the smaller size of the resulting corpora. The smaller corpora also result in very low recall for the role classification.

Future work could be to extend the German dataset by adding additional resources to the LLR, for instance GermaNet (Hamp and Feldweg, 1997). Extending the SemLink mapping to frames unique to SALSA should additionally contribute to an improved role coverage.

6 Related Work

Relevant related work is research on (i) the automatic acquisition of sense-labeled data for verbs, (ii) the automatic acquisition of role-labeled data for FrameNet SRL, and (iii) approaches to FrameNet SRL using lexical resources and LLRs, including rule-based and knowledge-based approaches.

Automatic acquisition of sense-labeled data. Most previous work on automatically sense-labeling corpora for WSD focussed on nouns and WordNet as a sense inventory, e.g., Leacock et al. (1998), Mihalcea and Moldovan (1999), Martinez (2008), Duan and Yates (2010). In this section, we describe work that specifically considers verbs. Besides the already introduced work by Cholakov et al. (2014), which we extended by discriminating patterns and adapted to the FrameNet verb sense inventory, this includes work by Kübler and Zhekova (2009), who 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 sentences with WordNet senses and use them as training data for WSD. They evaluate on an all-words task and do not find performance improvements when training on the automatically labeled data alone or on a combination of automatically labeled and gold data.

Automatic acquisition of role-labeled data. Previous work in the automatic acquisition of role-labeled data uses annotation projection methods, i.e., aligning a role-annotated sentence to a new sentence on the syntactic level and transferring the role annotations to the aligned words.

The goals of Fürstenau and Lapata (2012)'s work are most similar to our work. They perform annotation projection of FrameNet roles for English verbs. For this, they pair sentences in the British National Corpus with frame-annotated sentences, align their syntactic structures (including arguments), and project annotations to the new sentences. They simulate a "low-resource" scenario that only provides few training instances (called seed sentences) by varying the number of seed sentences and added labeled sentences. They use the automatically labeled data together with seed training data to train a supervised system and find improvements over self-training.

A main difference to our approach is that Fürstenau and Lapata (2012) do not use external information from LLRs or other lexical resources like WordNet. Like our approach, their approach creates a sparse labeling by a) discarding sentences that do not align well to their seeds, and b) discarding candidate pairs for which not all roles could be mapped. This leads to a high-precision approach that does not allow partially labeled data. 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. In our approach, the ASP patterns generalize from the seed sentences (cf. section 3), leading us to assume that our knowledge-based approach could be more generous with respect to such variability; we already successfully evaluated it on four datasets from various domains, but would like to further confirm our assumption in a direct comparison.

Another approach for training data generation for PropBank-style semantic role labeling is described in Woodsend and Lapata (2014). Using comparable corpora they extract rewrite rules to generate paraphrases of the original PropBank sentences. They use a model trained on PropBank as the seed corpus to filter out noise introduced by the rewrite rules. A model trained on the extended PropBank corpus outperforms the state of the art system on the CoNLL-2009 dataset. Recently, Pavlick et al. (2015) presented a similar method to expand the FNFT corpus through automatic paraphrasing. Noise was filtered out using crowdsourcing and the resulting frame-labeled corpus showed a lexical coverage three times as high as the original FNFT. However, they did not evaluate the augmented corpus as training data for semantic role classification.

FrameNet SRL using lexical resources. Similar to our approach of automatically creating role-labeled data in section 4, there are other rule-based approaches to FrameNet SRL that rely on FrameNet and other lexical resources (Shi and Mihalcea, 2004; Shi and Mihalcea, 2005). Both 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 approach to SRL introduced by Litkowski (2010) uses a dictionary built from FrameNet fulltext annotations to recognize and assign semantic roles. Their system first performs frame disambiguation and then tries to match syntactic constituents produced by a parser with syntactic patterns included the generated dictionary. Their system is evaluated on the SemEval-2 task for linking events and their participants in discourse. It shows very low recall, which is mainly due to the low coverage of their FrameNet dictionary with regard to syntactic patterns.

Our approach differs from previous rule-based approaches to SRL in that we do not use the rule-based system directly, but use it to create labeled training data for training a supervised system. This transductive semi-supervised learning setup should be able to deal better with the noise introduced by the rule based system than the inductive rule-based approaches.

The work by Kshirsagar et al. (2015) uses lexical resources to enhance FrameNet SRL. They also use the FrameNet sense examples and SemLink, but in a completely different manner. Regarding the sense examples, they employ domain adaptation techniques to augment the feature space extracted from the FrameNet training set with features from the sense examples, thereby increasing role labeling F1 by 3% compared to the baseline system SEMAFOR.

We use the FrameNet example sentences only indirectly: as seed sentences for the frame label transfer (cf. Step 1), they provide distant supervision for the automatic frame labeling. Our approach is complementary to the one by Kshirsagar et al. (2015) who use the sense examples for role labeling.

Kshirsagar et al. (2015) only briefly report on their experiments using SemLink. They used the translation of PropBank labels to FrameNet in the SemLink corpus as additional training data, but found that this strategy hurt role labeling performance. They credit this to the low coverage and errors in SemLink, which might be amplified by the use of a transitive linking (from PropBank to FrameNet via VerbNet). In this work, we successfully employ SemLink: we use the VerbNet-FrameNet (sense- and role-level) linking from SemLink in our role label transfer approach (Step 2). The resulting automatically role-labeled training data improve role classification in combination with the FN-train set (cf. section 4.3). We assume that the large-scale generation of training data smoothes over the noise resulting from errors in the SemLink mapping.

Kshirsagar et al. (2015) additionally use features from PropBank SRL as guide features and exploit the FrameNet hierarchy to augment the feature space, a method complementary to our approach. Their best results combine the use of example sentences and the FrameNet hierarchy for feature augmentation. They only evaluate on the FNFT-test set, as has become standard for FrameNet SRL evaluation. Our distantly supervised corpus might be useful for domain adaptation to other datasets, as our role classification evaluation shows.

According to our above analysis, our strategy is complementary to the approach by Kshirsagar et al. (2015). It would be interesting to evaluate to what degree our automatically labeled corpus would benefit their system.

7 Relation to FrameNet SRL

In this section, we discuss the potential impact of our work to state-of-the-art FrameNet SRL.

Our experimental setup evaluates frame disambiguation and role classification separately, which is a somewhat artificial setup. We show that our automatically generated training data are of high quality and contribute to improved classification performance. This section motivates that the data can also be useful in a state-of-the-art SRL setting.

For a long time, the SEMAFOR system has been the state-of-the-art FrameNet SRL system (Das et al., 2010; Das et al., 2014). Recently, systems were introduced that use new ways of generating training features and neural-network based representation learning strategies. We already introduced (Kshirsagar et al., 2015). Hermann et al. (2014) use distributed representations for frame disambiguation. Others integrate features based on document-level context into a new open-source SRL system Framat++ (Roth and Lapata, 2015), or present an efficient dynamic program formalization for FrameNet role labeling (Täckström et al., 2015). They all report improvements on SEMAFOR results for full FrameNet SRL.

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 (embeddings) and performs frame disambiguation by mapping a new instance to the embedding space and assigning the closest frame label (conditioned on the 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 SRL, reporting an F1 of 68.69% compared to 64.54% from Das et al. (2014).

Our frame disambiguation results are not directly comparable to their results. We also evaluate on ambiguous instances, but only on verbal predicates, which are typically more polysemous than nouns and adjectives and more difficult to disambiguate.

The currently best-performing FrameNet SRL system is the one presented by FitzGerald et al. (2015). They present a multitask learning setup for semantic role labeling which they evaluate for PropBank and FrameNet SRL. The setup is based on a specifically designed neural network model that embeds input and output data in a shared, dense vector space. Using the frame identification model from Hermann et al. (2014), their results significantly improve on the previous state-of-the-art for full FrameNet SRL, reaching F1 of 70.9% on FNFT-test – but only when training the model jointly on FrameNet training data and PropBank-labeled 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 SRL 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 left to future work.

Based on the discussion above, and on the frame and role classification experiments evaluated on four test sets, we expect that the data we generate with our method are complementary to the standard FrameNet training data and can be used to enhance state-of-the-art SRL systems. We leave empirical evaluation of this claim to future work. By publishing our automatically labeled corpora for research purposes, we support efforts by other researchers to analyze them and integrate them into their systems.

8 Discussion and Outlook

The evaluation shows our purely knowledge-based approach for automatic label transfer results in high-quality training data for English SRL that is complementary to the FNFT corpus.

For VSD, our data lead to similar precision to a standard supervised setup, but at higher recall. 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; in future work we would like to extend the SemLink mapping automatically to enhance the coverage of our method, and to disambiguate ambiguous labels to further increase precision.

As a knowledge-based approach, our method is particularly well-suited for languages and domains for which role-labeled corpora are lacking, but LLRs are available or can be created automatically. We therefore applied our approach to German data; the resulting sense-labeled corpus is complementary to the training data from SALSA2. The role classification evaluation should improve with a larger corpus.

State-of-the-art SRL systems still rely on supervised training, even when advanced methods such as deep learning are used. In section 7, we discussed in detail how our method relates to and complements the most recent developments in FrameNet SRL. It would be interesting to evaluate the benefits that our automatically labeled data can add to an advanced SRL system. We expect particularly strong benefits in the context of domain adaptation: currently, FrameNet SRL systems are only evaluted on in-domain test data.

Our method can be adapted to other sense and role inventories covered by LLRs (e.g., VerbNet and PropBank) and to related approaches to SRL and semantic parsing (e.g., QA-SRL (He et al., 2015)); the latter requires a mapping of the role inventory to a suitable LLR, for instance mapping the role labels in QA-SRL to SemLink. We would also like to evaluate our approach in comparison to other methods for training data generation, for instance methods based on alignments (Fürstenau and Lapata, 2012), or paraphrasing (Woodsend and Lapata, 2014).

9 Conclusion

We presented a novel approach to automatically generate training data for FrameNet SRL. It follows the distant supervision paradigm and performs knowledge-based label transfer from rich external knowledge sources to large-scale corpora without relying on manually labeled corpora.

By transferring labels to a large, diverse web-corpus (ukWAC) the potential of our approach for generating data for different domains becomes apparent. By applying it to German data, we showed that our approach is applicable across languages. As a further result of our work, we publish the automatically labeled corpora and release our implementation for knowledge-based role labeling (cf. Step 2A in section 4) as open source software.

Automatic label transfer using linked resources has become popular in relation extraction (Mintz et al., 2009) and has been applied to VSD (Cholakov et al., 2014), but not to SRL. In this work, we showed that knowledge-based label transfer from LLRs to large-scale corpora offers great opportunities also for complex semantic tasks like SRL.

Acknowledgments

This work has been supported by the German Research Foundation under grant No. GU 798/9-1, grant No. GU 798/17-1, and grant No. GRK 1994/1. We thank the action editors and anonymous reviewers for their thoughtful comments. Additional thanks go to Nancy Ide and Collin Baker for providing the MASC dataset.

References

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, 43(3):209–226.

Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, and Christopher D. Manning. 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.

Claire Bonial, Kevin Stowe, and Martha Palmer. 2013. Renewing and Revising SemLink. In *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and linking lexicons, terminologies and other language data*, pages 9 – 17, Pisa, Italy.

Aljoscha Burchardt and Marco Pennacchiotti. 2008. FATE: a FrameNet-Annotated Corpus for Textual Entailment. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 539–546, Marrakech, Morocco.

Aljoscha Burchardt, Kathrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. 2006. The SALSA Corpus: a German Corpus Resource for Lexical Semantics. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 969–974, Genoa, Italy.

Kostadin Cholakov, Judith Eckle-Kohler, and Iryna Gurevych. 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 2014)*, pages 68–77, Gothenburg, Sweden.

Dipanjan Das and Noah A. Smith. 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*, pages 1435–1444, Portland, Oregon, USA.

Dipanjan Das, Nathan Schneider, Desai Chen, and Noah A. Smith. 2010. Probabilistic Frame-Semantic Parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 948–956, Los Angeles, CA, USA.

Dipanjan Das, Desai Chen, André F. T. Martins, Nathan Schneider, and Noah A. Smith. 2014. Frame-Semantic Parsing. *Computational Linguistics*, 40(1):9–56.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of the 5th Edition of the International Conference on Language Resources and Evaluation*, pages 449–454, Genoa, Italy.

Weisi Duan and Alexander Yates. 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*, HLT '10, pages 627–635, Los Angeles, CA, USA.

Richard Eckart de Castilho and Iryna Gurevych. 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. Association for Computational Linguistics and Dublin City University.

Parvin Sadat Feizabadi and Sebastian Padó. 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.

Charles J. Fillmore, 1982. *Linguistics in the Morning Calm*, chapter Frame Semantics, pages 111–137. Hanshin Publishing Company, Seoul, South Korea.

Nicholas FitzGerald, Oscar Täckström, Kuzman Ganchev, and Dipanjan Das. 2015. Semantic Role Labeling with Neural Network Factors. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 960–970, Lisbon, Portugal.

Marco Fossati, Claudio Giuliano, and Sara Tonelli. 2013. Outsourcing FrameNet to the Crowd. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 742–747, Sofia, Bulgaria.

Hagen Fürstenau and Mirella Lapata. 2012. Semi-Supervised Semantic Role Labeling via Structural Alignment. *Computational Linguistics*, 38(1):135–171.

Hagen Fürstenau. 2011. *Semi-Supervised Semantic Role Labeling via Graph Alignment, volume 32 of*

Saarbrücken Dissertations in Computational Linguistics and Language Technology. German Research Center for Artificial Intelligence and Saarland University, Saarbrücken, Germany.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1):10–18.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet - a Lexical-Semantic Net for German. In *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15, Madrid, Spain.

Patrick Hanks. 2013. *Lexical Analysis: Norms and Exploitations*. MIT Press, Cambridge, MA, USA.

Silvana Hartmann and Iryna Gurevych. 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.

Kazi Saidul Hasan and Vincent Ng. 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.

Luheng He, Mike Lewis, and Luke Zettlemoyer. 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.

Karl Moritz Hermann, Dipanjan Das, Jason Weston, and Kuzman Ganchev. 2014. Semantic Frame Identification with Distributed Word Representations. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1448–1458, Baltimore, Maryland, USA.

Meghana Kshirsagar, Sam Thomson, Nathan Schneider, Jaime Carbonell, Noah A. Smith, and Chris Dyer. 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 (Volume 2: Short Papers)*, pages 218–224, Beijing, China.

Sandra Kübler and Desislava Zhekova. 2009. Semi-Supervised Learning for Word Sense Disambiguation: Quality vs. Quantity. In *Proceedings of the International Conference RANLP-2009*, pages 197–202, Borovets, Bulgaria.

Claudia Leacock, George A. Miller, and Martin Chodorow. 1998. Using Corpus Statistics and WordNet Relations for Sense Identification. *Computational Linguistics*, 24(1):147–165.

Ken Litkowski. 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 '10)*, pages 300–303, Los Angeles, CA, USA.

David Martinez. 2008. On the Use of Automatically Acquired Examples for All-Nouns Word Sense Disambiguation. *Journal of Artificial Intelligence Research*, 33:79–107.

Rada Mihalcea and Dan Moldovan. 1999. An Automatic Method for Generating Sense Tagged Corpora. In *Proceedings of the American Association for Artificial Intelligence (AAAI 1999)*, Orlando, Florida, USA.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 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*, pages 1003–1011, Suntec, Singapore.

Rebecca J. Passonneau, Collin F. Baker, Christiane Fellbaum, and Nancy Ide. 2012. The MASC Word Sense Corpus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3025–3030, Istanbul, Turkey.

Ellie Pavlick, Juri Ganitkevitch, Tsz Ping Chan, Xuchen Yao, Benjamin Van Durme, and Chris Callison-Burch. 2015. 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 (Volume 2: Short Papers)*, pages 57–62, Beijing, China.

Octavian Popescu, Martha Palmer, and Patrick Hanks. 2014. Mapping CPA Patterns onto OntoNotes Senses. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 882–889, Reykjavik, Iceland.

Michael Roth and Mirella Lapata. 2015. Context-Aware Frame-Semantic Role Labeling. *Transactions of the Association for Computational Linguistics*, 3:449–460.

Josef Ruppenhofer, Caroline Sporleder, Roser Morante, Collin Baker, and Martha Palmer. 2010. SemEval-2010 Task 10: Linking Events and Their Participants in Discourse. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 45–50, Uppsala, Sweden.

Wolfgang Seeker and Jonas Kuhn. 2012. Making Ellipses Explicit in Dependency Conversion for a German Treebank. In Nicoletta Calzolari et al., editor, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3132–3139, Istanbul, Turkey.

Lei Shi and Rada Mihalcea. 2004. Open Text Semantic Parsing Using FrameNet and WordNet. In *Demonstration Papers at HLT-NAACL 2004*, HLT-NAACL–Demonstrations '04, pages 19–22, Stroudsburg, PA, USA.

Lei Shi and Rada Mihalcea. 2005. Putting Pieces Together: Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing. In *Computational Linguistics and Intelligent Text Processing*, pages 100–111. Springer Berlin Heidelberg.

Oscar Täckström, Kuzman Ganchev, and Dipanjan Das. 2015. Efficient Inference and Structured Learning for Semantic Role Labeling. *Transactions of the Association for Computational Linguistics*, 3:29–41.

Kristian Woodsend and Mirella Lapata. 2014. Text Rewriting Improves Semantic Role Labeling. *Journal of Artificial Intelligence Research*, 51:133–164.

David Yarowsky. 1995. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 189–196, Cambridge, Massachusetts, USA.

Benat Zapirain, Eneko Agirre, Lluis Marquez, and Mihai Surdeanu. 2013. Selectional Preferences for Semantic Role Classification. *Computational Linguistics*, 39(3):631–663.