Out-of-domain FrameNet Semantic Role Labeling

Silvana Hartmann[§†], **Ilia Kuznetsov**[†], **Teresa Martin**[§†], **Iryna Gurevych**[§†]
§Research Training Group AIPHES
†Ubiquitous Knowledge Processing (UKP) Lab
Department of Computer Science, Technische Universität Darmstadt
http://www.ukp.tu-darmstadt.de

Abstract

Domain dependence of NLP systems is one of the major obstacles to their application in large-scale text analysis, also restricting the applicability of FrameNet semantic role labeling (SRL) systems. Yet, current FrameNet SRL systems are still only evaluated on a single in-domain test set. For the first time, we study the domain dependence of FrameNet SRL on a wide range of benchmark sets. We create a novel test set for FrameNet SRL based on user-generated web text and find that the major bottleneck for out-of-domain FrameNet SRL is the frame identification step. To address this problem, we develop a simple, yet efficient system based on distributed word representations. Our system closely approaches the state-of-the-art in-domain while outperforming the best available frame identification system out-of-domain. We publish our system and test data for research purposes.[1]

1 Introduction

Domain dependence is a major problem for supervised NLP tasks such as FrameNet semantic role labeling (SRL): systems generally exhibit a strong performance drop when applied to test data from a different distribution than the training data. This prohibits their large-scale use in language technology applications.

The same problems are expected for FrameNet SRL, but due to a lack of datasets, state-of-the-art FrameNet SRL is only evaluated on a single in-domain test set, see e.g. Das et al. (2014) and FitzGerald et al. (2015).

In this work, we present the first comprehensive study of the domain dependence of FrameNet SRL on a range of benchmark datasets. This is crucial as the demand for semantic textual analysis of large-scale web data keeps growing.

Based on FrameNet (Fillmore et al., 2003), FrameNet SRL extracts frame-semantic structures on the sentence level that describe a specific situation centered around a semantic predicate, often a verb, and its participants, typically syntactic arguments or adjuncts of the predicate. The predicate is assigned a *frame* label, essentially a word sense label, that defines the situation and determines the *semantic roles* of the participants. The following sentence from FrameNet provides an example of the *Grinding* frame and its roles:

[The mill]$_{Grinding\_cause}$ **grinds**$_{Grinding}$ [the malt]$_{Patient}$ [to grist]$_{Result}$.

FrameNet SRL consists of two steps, frame identification (frameId), assigning a frame to the current predicate, and role labeling (roleId), identifying the participants and assigning them role labels licensed by the frame. The frameId step reduces the hundreds of role labels in FrameNet to a manageable set of up to 30 roles. Thus, FrameNet SRL differs from PropBank SRL (Carreras and Màrquez, 2005), that only uses a small set of 26 syntactically motivated role labels and puts less weight on the predicate sense. The advantage of FrameNet SRL is that it results in a more fine-grained and rich interpretation of the input sentences which is crucial for many applications, e.g. reasoning in online debates (Berant et al., 2014).

Domain dependence is a well-studied topic for PropBank SRL. However, to the best of our knowledge, there exists no analysis of the performance of modern FrameNet SRL systems when applied to data from new domains.

In this work, we address this problem as follows: we introduce a new benchmark dataset YAGS

[1] www.ukp.tu-darmstadt.de/ood-fn-srl

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 471–482,
Valencia, Spain, April 3-7, 2017.

(**Y**ahoo! **A**nswers **G**old **S**tandard), which is based on user-generated questions and answers and exemplifies an out-of-domain application use case. We use YAGS, along with other out-of-domain test sets, to perform a detailed analysis of the domain dependence of FrameNet SRL using `Semafor` (Das et al., 2014; Kshirsagar et al., 2015) to identify which of the stages of FrameNet SRL, frameId or roleId, is particularly sensitive to domain shifts. Our results confirm that the major bottleneck in FrameNet SRL is the frame identification step. Motivated by that, we develop a simple, yet efficient frame identification method based on distributed word representations that promise better domain generalization. Our system's performance matches the state-of-the-art in-domain (Hermann et al., 2014), despite using a simpler model, and improves on the out-of-domain performance of `Semafor`.

The contributions of the present work are twofold: 1) we perform the first comprehensive study of the domain generalization capabilities of open-source FrameNet SRL, and 2) we propose a new frame identification method based on distributed word representations that enhances out-of-domain performance of frame identification. To enable our study, we created YAGS, a new, substantially-sized benchmark dataset for the out-of-domain testing of FrameNet SRL; we publish the annotations for the YAGS benchmark set and our frame identification system for research purposes.

2 Related work

The domain dependence of FrameNet SRL systems has been only studied sparsely, however, there exists a large body of work on out-of-domain PropBank SRL, as well as on general domain adaptation methods for NLP. This section briefly introduces some of the relevant approaches in these areas, and then summarizes the state-of-the-art in FrameNet frame identification.

Domain adaptation in NLP Low out-of-domain performance is a problem common to many supervised machine learning tasks. The goal of domain adaptation is to improve model performance on the test data originating from a different distribution than the training data (Søgaard, 2013). For NLP, domain adaptation has been studied for various tasks such as POS-tagging and syntactic parsing (Daumé III, 2007; Blitzer et al., 2006). For the complex task of SRL, it is strongly associated with PropBank, because the corresponding CoNLL shared tasks promote out-of-domain evaluation (Surdeanu et al., 2008; Hajič et al., 2009). In the shared tasks, in-domain newspaper text from the WSJ Corpus is contrasted to out-of-domain data from fiction texts in the Brown Corpus. Most of the participants in the shared tasks do not consider domain adaptation and report systematically lower scores for the out-of-domain data (Hajič et al., 2009).

Representation learning has been successfully used to improve on the CoNLL shared task results (Huang and Yates, 2010; FitzGerald et al., 2015; Yang et al., 2015). Yang et al. (2015) report the smallest performance difference (5.5 points in F_1) between in-domain and out-of-domain test data, leading to the best results to date on the CoNLL 2009 out-of-domain test. Their system learns common representations for in-domain and out-of-domain data based on deep belief networks.

Domain dependence of FrameNet SRL The FrameNet 1.5 fulltext corpus, used as a standard dataset for training and evaluating FrameNet SRL systems, contains texts from several domains (Ruppenhofer et al., 2010). However, the standard data split used to evaluate modern systems (Das and Smith, 2011) ensures the presence of all domains in the training as well as test data and cannot be used to assess the systems' ability to generalize. Moreover, all the texts in the FrameNet fulltext corpus, based on newspaper and literary texts, are post-edited and linguistically well-formed. The FrameNet test setup thus cannot provide information on SRL performance on less edited out-of-domain data, e.g. user-generated web data.

There are few studies related to the out-of-domain generalization of FrameNet SRL. Johansson and Nugues (2008) evaluate the impact of different parsers on FrameNet SRL 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 is beneficial in the out-of-domain scenario. Croce et al. (2010) use a similar in-domain/out-of-domain split to evaluate their approach to open-domain FrameNet SRL. They integrate a distributional model into their SRL system to generalize lexicalized features to previously unseen arguments and thus create an SRL system with a smaller performance gap between in-domain and out-of-domain test data (only 4.5 percentage points F_1).

Note that they only evaluate the role labeling step. It is not transparent how their results would transfer to the current state-of-the-art SRL systems that already integrate methods to improve generalization, for instance using distributed representations.

Palmer and Sporleder (2010) analyze the FrameNet 1.3 training data coverage and the performance of the Shalmaneser SRL system (Erk and Padó, 2006) for frame identification on several test sets across domains, i.e. the PropBank and NTI parts of the FrameNet fulltext corpus and the fictional texts from the SemEval-2007 shared task (Baker et al., 2007). Having observed that the majority of errors results from coverage gaps in FrameNet, they suggest to focus on developing frame identification systems that generalize well to new domains. Our observations support their findings and show that the problem still persists even when modern SRL methods and the extended FrameNet 1.5 lexicon are used.

Søgaard et al. (2015) annotate 236 tweets with FrameNet labels to apply SRL to knowledge extraction from Twitter. They report that the frameId performance of `Semafor 2.1` (Das et al., 2010) on the new test set is similar to its performance on the SemEval-2007 newswire test set (Baker et al., 2007). For full SRL, there are large differences: F_1 reaches only 25.96% on the Twitter set compared to the 46.5% reported by Das et al. (2010) on the in-domain set. These results show that there is ample room for improvement for SRL on Twitter data.

Recent FrameNet SRL systems are not evaluated in the context of their domain dependence: Kshirsagar et al. (2015) use the domain adaptation approach from Daumé III (2007) to augment the feature space for FrameNet SRL with FrameNet example sentences; FitzGerald et al. (2015) and Hermann et al. (2014) adopt deep learning methods, including learning representations that may generalize better to unseen data, to present state-of-the-art results for FrameNet SRL. All of the former only use the already introduced split of the FrameNet fulltext corpus for testing, as does the long-time state-of-the-art system `Semafor` (Das et al., 2014). Out-of-domain evaluation is lacking, as are datasets that enable this kind of evaluation.

Frame identification Current state of the art in frame identification is the approach by Hermann et al. (2014), further referred to as `Hermann-14`, followed by the previous state-of-the art model `Semafor` (Das et al., 2014).

The frame identification system of `Semafor` relies on an elaborate feature set based on syntactic and lexical features, using the WordNet hierarchy as a source of lexical information, and a label propagation-based approach to take unknown predicates into account. `Semafor` is not specifically designed for out-of-domain use: the WordNet coverage is limited, and the quality of syntactic parsing might drop when the system is applied to out-of-domain data, especially in case of non-standard user-generated texts.

`Hermann-14` uses distributed word representations augmented by syntactic information. General-purpose distributed word representations (such as word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014)) are beneficial for many NLP tasks: word representations are calculated on a large unlabeled corpus, and then used as input for high-level tasks for which training data is scarce, such as syntactic parsing, word sense disambiguation, and SRL. In the syntax-augmented representations of `Hermann-14`, a region of the input vector, a *container*, is reserved for each syntactic path that can connect predicates to their arguments. This container is populated with a corresponding argument word representation, if the argument on this path is found in the training data. `Hermann-14` uses the WSABIE algorithm (Weston et al., 2011) to map input and frame representations to a common latent space. WSABIE uses WARP loss and gradient-based updates to minimize the distance between the latent representations of the predicate target and the correct frame, while maximizing the distance to all the other irrelevant frames. During testing, cosine similarity is used to find the closest frame given the input. One advantage of this approach is that similar frames are positioned close to each other in the latent space which allows information to be shared between similar predicates and similar frames. This system is the current state-of-the-art for in-domain frame identification, but has not been applied in an out-of-domain setting.

3 Out-of-domain FrameNet test data

This section describes available in-domain and out-of-domain FrameNet test sets and the creation of YAGS, a new out-of-domain FrameNet test set.

FrameNet test sets FrameNet SRL is typically evaluated on **das-test**, the test set first introduced by Das and Smith (2011). It is a held-out set randomly sampled from the FrameNet 1.5 fulltext cor-

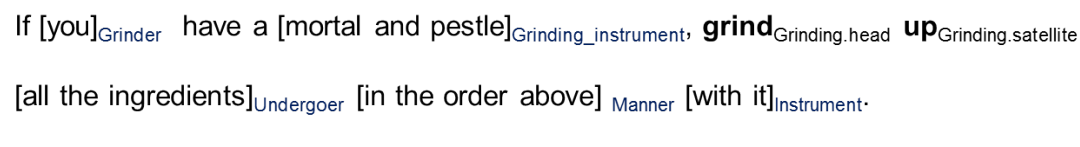

Figure 1: Example sentence from YAGS with multiword predicate and typo (*mortal* vs. *mortar*).

pus. While the FrameNet fulltext corpus contains data from various sources, we consider das-test an *in-domain* test set: all data sources of the test set are also represented in the training set.

There are two additional datasets from other domains that we use in our study on domain generalization: The **MASC** word sense sentences corpus contains FrameNet annotations for a lexical sample of roughly 100 lemmas from ANC (Passonneau et al., 2012). The Twitter-based dataset from Søgaard et al. (2015), henceforth **TW**, has some very distinctive properties: it does not provide a gold standard, but annotations by three annotators. This leads to a high variance in role annotations: the annotator TW_3 annotated only 82% of the number of roles annotated by TW_1, see Table 1. Like Søgaard et al. (2015), we report SRL results as averages over the three annotations (TW-av).

Table 1 shows statistics on these datasets. For TW, it displays the statistics for each annotator. The TW datasets are fairly small, containing only around 1,000 frame labels. The MASC dataset is of substantial size, but it constitutes a lexical sample and therefore a slightly artificial evaluation setup. There is another Twitter-based test set (Johannsen et al., 2015), which we do not use in our experiments, because it was created semi-automatically and is therefore of lower quality. We conclude that existing out-of-domain test sets for FrameNet SRL are insufficient, in particular for increasingly important domains like user-generated text, because available datasets are either small or of low quality.

YAGS: a new FrameNet test set based on user-generated text To address the need for new out-of-domain test datasets, we created **YAGS**, a new FrameNet-annotated evaluation dataset based on question-answer data from Yahoo! Answers (YA), a community-driven question-and-answer forum. The corpus is based on a random sample of 55 questions and their answers from the test split of the YA Manner Questions dataset used by Surdeanu et al. (2011) and published as part of the Yahoo! Webscope program (`https://webscope.sandbox.yahoo.com/`).

YAGS contains 1,415 sentences, 3,091 frame annotations, and 6,081 role annotations. Figure 1 shows a sentence from YAGS that demonstrates some non-standard properties of the user-generated question-answer data, such as typos (*mortal* instead of *mortar*). We publish the annotations as stand-off annotations to the original dataset.

Annotation study Each document was annotated by a two linguistically trained annotators provided with detailed guidelines and then curated by an experienced expert, all using WebAnno 2.0.0 (Yimam et al., 2014). Up to five predicates per sentence were pre-selected automatically based on lemma and POS, preferring verbal predicates to other POS, which leads to a larger proportion of verbs in YAGS. The annotation task was to identify the correct frame label for each predicate, if any, and then to identify the role spans as arguments and adjuncts of the frame, and to label them with the appropriate role. For reference, annotators accessed the FrameNet 1.5 definitions and examples with the FrameNet Explorer tool (`www.clres.com/FNExplorer.html`).

Inter-rater agreement for frame labels is Krippendorff's α=0.76; agreement for role labels given matching spans is α=0.62, and Krippendorff's α unitizing agreement for role spans is 0.7 – a good result for such a difficult task on user-generated text. Average pairwise F_1 agreement for frame labels is high at 0.96, higher than the 0.84 reported by Søgaard et al. (2015) for the TW sets. Our high frame agreement is a result of annotator experience and our elaborate annotation setup.

YAGS statistics and properties Table 1 presents dataset statistics for YAGS and the other test sets. Due to the predicate selection, YAGS contains a larger proportion of verbal predicates than the other sets, and has three times more frames and roles than TW, approximating the size of das-test. The proportion of core roles, roles that are obligatory for a frame and thus typically more frequent in datasets than non-core roles, in the out-of-domain test sets (TW, YAGS, MASC) is slightly smaller

data	s	f	a	n	v	r	cr
das-test	2,420	4,458	12	42	33	7,172	83
YAGS	1,415	3,091	5	18	75	6,081	74
MASC	8,444	7,226	25	42	33	11,214	78
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

Table 1: Text dataset statistics: sentences **s**; frames **f**; % of adjectives **a**, nouns **n** and verbs **v**; roles **r**, % of core roles **cr**. Subscripts for TW indicate the respective annotator.

compared to das-test. This goes along with a larger variance of roles in YAGS.

The user-generated aspect of YAGS manifests in spelling errors, and in the lack of punctuation and structure of the texts. The language is informal, but there are only few emoticons or other special words such as the hashtags typically found in tweets.

In the next section, we use the test sets from Table 1 to analyze the domain generalization capabilities of an open-source FrameNet SRL system.

4 Domain generalization capabilities of open-source FrameNet SRL

To analyze the domain generalization capabilities of contemporary open-source SRL, we ran the frame identification from `Semafor` (Das et al., 2014) with the enhanced role labeler from Kshirsagar et al. (2015), both trained on the in-domain das-train set, on the four test sets das-test, YAGS, TW, and MASC. The systems receive text annotated with predicate spans as input, which has become the standard in recent evaluations.

Evaluation script The `Semafor` evaluation script (Das et al., 2014) provides precision P, recall R, and F_1 scores for full SRL (SRL), and accuracy A for frame identification (frameId). Full SRL evaluation can be performed with and without using gold frames instead of predicted (auto) frames.

The script does not provide results on the role labeling (argument identification and labeling, roleId) alone: the scoring mechanism for *SRL/gold* also considers the by default correct gold frames. This is useful when comparing different SRL systems on the same test set, but not sufficient when 1) comparing role labeling performance on different test sets with a different ratio of frame labels to role labels (resulting from different annotation strategies), and 2) analyzing the contribution of frameId and roleId to full SRL performance across test sets.

data	frameId	roleId		SRL	
	auto	auto	gold	auto	gold
das-test	82.09	30.08	55.20	55.40	73.16
YAGS	59.62	18.60	56.99	37.22	72.58
MASC	39.52	19.46	51.74	29.05	71.08
TW-av	62.17	15.91	61.45	38.44	76.74

Table 2: `Semafor` performance on test sets in %: exact frameId A; then F_1 for roleId and SRL with system frames (auto) and gold frames (gold).

We therefore evaluate the output of the script to retain the original counts for role labels and compute scores on the role labeling proper (roleId). Moreover, there are two evaluation settings for frameId: exact frame match and partial frame match. We use the exact match setting that does not credit related frames and roles.

Results Table 2 presents scores for exact match frameId and for SRL and roleId with automatic frames (auto) and with gold frames (gold). For TW, the results are averaged over the number of annotators. According to column *SRL/auto*, we observe best `Semafor` performance for full SRL on das-test, results for the other test sets are at least 16 percentage points F_1 lower. This is mostly due to the worse frameId performance of `Semafor` on the new test sets, as shown in column *frameId*: frameId performance is at least 19 percentage points lower. This negatively affects roleId for the out-of-domain test sets (see column *roleId/auto*). *RoleId/auto* scores are also low on das-test, but higher than for the other sets.

When using gold frame labels, roleId and SRL performance improve for all test sets. As shown in columns *roleId/gold* and *SRL/gold*, the difference between in-domain and out-of-domain evaluation vanishes. Only MASC scores are still two points lower for full SRL than those for das-test. TW-av scores even surpass the in-domain scores.[2]

This shows how much FrameNet role labels are dependent on correct frame labels. Thus, it is crucial to improve the out-of-domain performance of frameId systems.

Domain dependence appears to be less of a problem for the role labeling step. The MASC dataset is the most difficult for both frameId and roleId. This is mostly a consequence of the lower training data coverage of MASC, as discussed below.

[2]Our TW-av results are not comparable to those from Søgaard et al. (2015) because their test setup includes predicate target identification and uses different evaluation metrics.

dataset	lemmas ∉ lexicon	lemmas ∉ das-train	senses ∉ das-train	monosemous ∈ das-train
das-test	2.59	9.99	14.03	53.99
YAGS	2.79	17.33	30.36	27.07
MASC	7.45	21.72	51.25	23.51
TW_1	1.01	17.51	36.06	26.73
TW_2	1.27	17.91	51.25	27.07
TW_3	1.25	17.24	35.65	27.17

Table 3: Training data coverage of test sets in %. Sense is a combination of predicate lemma, POS and frame; lexicon refers to the `Semafor` lexicon.

Analysis In our study, it became clear that domain dependence is crucial to the frame identification step in SRL. The lower scores for the out-of-domain test sets can be a result of different domain-specific predicate-frame distributions, or a lack of coverage of the domain in the training data.

To get a better understanding of these phenomena, we compared detailed statistics of the different test sets, cf. Table 3. Das-test has the largest predicate coverage and contains a lot of monosemous predicates, which boosts the overall performance. The occurrence of fewer monosemous predicates is expected for the lexical sample dataset MASC, but might indicate a domain preference for polysemous predicates in the YAGS and TW datasets.

The percentage of unseen predicates (lemmas $\notin$ das-train) is slightly higher for the user-generated test sets than for das-test, and much higher for MASC. This is mirrored in the lower frameId performance for MASC compared to the other test sets, and the slightly higher performance of TW-av and YAGS. Not all errors can be explained by insufficient training data coverage, which indicates that domain effects occur for the out-of-domain sets.

To support this assumption, we performed a detailed error analysis on the misclassified instances for all test sets. We compute the proportion of wrongly classified instances with unseen predicates, predicates that do not occur in the training set. For MASC, the majority of the errors, 68%, are based on unseen predicates, while the number ranges between 37% and 43% for the other test sets, i.e. 37% for TW, 39% for das-test and 43% for YAGS. This shows that training data coverage is a bigger issue for MASC than for the other test sets. The proportions of in-train errors for YAGS and TW-av are similar to das-test. Together with the fact that overall proportion of errors is still much higher for the user-generated test sets YAGS and TW-av, this further supports our hypothesis of domain effects for YAGS and TW-av. Manual analysis furthermore shows that there are differences in frequently confused frames between the in-domain das-test and out-of-domain YAGS and TW-av.

In the next section, we study new methods to improve out-of-domain frame identification.

5 Frame identification with distributed word representations

Given a predicate and a set of frames associated with this predicate, a frame identification system has to choose the correct frame based on the context. In this section we introduce our frame identification method and compare it to the state of the art in both in-domain and out-of-domain settings.

Our system SimpleFrameId We developed a straightforward approach to frame identification based on distributed word representations, and were surprised to find that this simple model achieves results comparable to the state-of-the-art system, `Hermann-14`. Our initial attempts to replicate `Hermann-14`, which is not publicly available, revealed that the container-based input feature space is very sparse: there exist many syntactic paths that can connect a predicate to its arguments, but a predicate instance rarely has more than five arguments in the sentence. So by design the input representation bears no information in most of its path containers. Moreover, `Hermann-14` makes heavy use of automatically created dependency parses, which might decline in quality when applied to a new domain. We demonstrate that our simple system achieves competitive in-domain and out-of-domain performance.

Our system, called SimpleFrameId, is specified as follows: given the lexicon L, the vector space vsm and the training data, our goal is to predict the frame f given the sentence S and the predicate p. From the machine learning perspective, the lexicon and the vector space are external resources. The lexicon contains associations between predicates and frames, and we further denote the set of frames available for a predicate as $L(p)$. The vector space provides a pre-defined dense vector representation $vsm(w)$ for each word w. In our case vsm is a simple word lookup function, since we do not modify our word representations during training.

From the sentence we extract the context representation, $x_c = \frac{\sum_{w \in C} vsm(w)}{|C|}$. We experiment with two kinds of contexts: `SentBOW` includes all

the words in the sentence, i.e. $C = S$, DepBOW considers the dependency parse of the sentence and only includes direct dependents of the predicate, $C = dep(p, S)$. As for the predicate, the plain embedding from the source vector space model is used, $x_p = vsm(p)$. A simple concatenation of x_c and x_p serves as input to the disambiguation classifier D, which outputs weights $D(x_c, x_p, f)$ for each frame known to the system $f \in L$. Note that the classifier itself is agnostic to the predicate's part of speech and exact lemma and only relies on the word representations from the vsm. We experiment with two different classification methods: one is a two-layer neural network D_{NN}, the other one is D_{WSB}, which follows the line of Hermann-14 and learns representations for frames and predicates in the same latent space using the WSABIE algorithm.[3] Hyperparameters are tuned on the development sets das-dev and YAGS-dev (sampled from YAGS); we test on the remaining 2,093 instances in YAGS-test.

Lexicon-based filtering In the testing stage, the classifier outputs weights for all the frames available in the lexicon, and the best-scoring frame is selected, $f \leftarrow argmax_{f \in L} D(x_c, x_p, f)$. Since the lexicon specifies available frames for each lexical unit (i.e. lemma and POS), additional filtering can be performed, which limits the search only to the available frames, $f \leftarrow argmax_{f \in L(p)} D(x_c, x_p, f)$. If the predicate is *unknown* to the lexicon, $p \notin L$, the overall best-scoring frame is chosen. If the target has only one entry in the lexicon, it's declared unambiguous and the frame is assigned directly.

Despite being common, this setup has several flaws that can obscure the differences between systems in the testing stage. As we showed in Section 4, the FrameNet lexicon has coverage issues when applied to new domains. Neither the predicate list nor the frame associations are guaranteed to be complete, and hence the total results are highly determined by the lexicon coverage.[4] To take this into account, we also perform evaluation in the no-lexicon setting, where frames are assigned directly by the classifier and no lexicon-based filtering is performed. We find that our frame identification system performs surprisingly well in this setting, and we encourage the no-lexicon performance to be additionally reported in the future, since it better reflects the frame identification quality and smoothens the effect of lexicon coverage.

system	total	ambig	no-lex
DataBaseline	79.09	70.68	2.21
LexiconBaseline	79.05	56.62	2.21
Semafor*	83.60	69.19	-
Hermann-14* (best)	**88.41**	73.10	-
WSB+SentBOW	84.46	67.56	72.05
WSB+DepBOW	85.69	69.93	71.21
NN+SentBOW	87.63	**73.80**	**77.49**
NN+DepBOW	87.53	73.58	76.51

Table 4: In-domain system comparison on das-test, * denotes results from Hermann et al. (2014); **ambig**: evaluation on ambiguous predicates; **no-lex**: system without lexicon filter.

Baselines We employ two majority baseline models for comparison. The DataBaseline assigns frames based on how often a frame is evoked by the given predicate. This corresponds to the most frequent sense baseline in word sense disambiguation (WSD). The frames available for predicates are obtained by scanning the training data. The LexiconBaseline calculates overall frame counts first (i.e. how often a frame appears in the training data in general), and, given the predicate, selects the overall most frequent frame among the ones available for this predicate. We expect this baseline to better handle the cases when limited data is available for a given predicate sense.

Experiments In our experiments, we generate the lexicon L in the same way as in Hermann-14, by scanning the "frames" folder of the FrameNet 1.5 distribution. For the external vector space model vsm we use dependency-based word embeddings from Levy and Goldberg (2014).

In-domain performance We report the performance of our system in the in-domain setting to compare to the state-of-the-art results from Hermann-14.[5] We train our system on *das-train* and test it on *das-test* using the full FrameNet lexicon. When available, we report the no-lexicon scores as well. As Table 4 shows, our system out-

[3] In our implementation, we use the LightFM package (Kula, 2015) with the WARP option for hybrid matrix factorization.

[4] A justification for this can also be found in Hermann et al. (2014): the difference in Hermann-14 accuracy when switching from the Semafor lexicon to the full lexicon is comparable to the difference between Semafor and Hermann-14 when evaluated on the same lexicon.

[5] Based on the errata version of Hermann et al. (2014) in http://www.aclweb.org/anthology/P/P14/P14-1136v2.pdf

system	das-test	YAGS	MASC	TW-av
DataBaseline	79.09	52.27	43.85	47.68
LexiconBaseline	79.05	50.02	36.86	55.40
Semafor	82.09	60.01	39.52	62.17
WSB+SentBOW	84.46	59.68	54.90	66.84
WSB+DepBOW	85.69	61.50	54.56	67.14
NN+SentBOW	**87.63**	62.03	53.73	**68.67**
NN+DepBOW	87.53	**62.51**	**55.09**	67.76

Table 5: Out-of-domain frameId, total accuracy. Semafor scores calculated during our own experiments; YAGS results on YAGS-test.

performs Semafor and performs on par with the results reported for Hermann-14. One interesting observation is that our systems perform almost as well in the no-lexicon setting as the DataBaseline, which has access to the lexicon, in the total setting. To our surprise, the WSABIE-based frame identification did not yield a consistent improvement in-domain, compared to the simple NN-based approach. We also observe that in many cases the SentBOW representation performs on par with the DepBOW, while requiring significantly less data preprocessing: SentBOW only uses tokenization, whereas DepBow relies on lemmatization, POS-tagging, and dependency parsing. We attribute this effect to the fact that SentBOW provides more context information than the sparse, dependency-filtered DepBOW.

Out-of-domain performance We also investigate how well the systems perform in the out-of-domain setting. Table 5 summarizes the results. Each of the systems was trained on *das-train* and tested on a variety of test sets. As we can see, our systems outperform Semafor for all datasets. The YAGS dataset is the only dataset on which we do not strongly outperform Semafor. We attribute this to the complexity of the YAGS dataset that contains a high proportion of verbs.

Overall out-of-domain performance stays behind the F_1-agreement observed for the human annotators for TW and YAGS, which shows that there is a large margin for improvement. Corresponding scores for in-domain data are not available.

Error analysis To further investigate the performance of our system in the out-of-domain setup we analyse statistics on the errors made by the system variant NN+SentBOW.

The system's wrong predictions are affected by the lexicon in two ways. First, if the predicate is not listed in the lexicon (unknown), the system has to choose among all frames. As we have shown before, the quality of predictions for unknown predicates is generally lower. The second case is when the predicate *is* listed in lexicon (so it is not unknown), but the correct frame is not associated with this predicate. We further refer to this class of errors as *unlinked*. For unlinked predicates, the system is restricted to the set of frames provided by the lexicon, and by design has no means to select the right frame for a given predicate occurrence.

The unlinked-predicate issue points to a major design flaw in the standard frameId architecture. Although choosing among frames defined in the lexicon provides a quality boost, it also renders many instances intractable for the system, if the lexicon coverage is incomplete. As Table 6 shows, unknown and unlinked predicates are almost non-present in the in-domain case, but are a major source of errors in the out-of-domain case and even might be responsible for the majority of errors occurring due to domain shift (see MASC). It is important to point out that there is still no guarantee that these would be classified correctly once the missing linking information is available in the lexicon. However, if the correct frame is not listed among the frames available for the predicate, the misclassification is inevitable.

A more detailed analysis of the errors made by the system shows that the majority of false predictions for known and linked predicates are due to the domain differences in word usage. For example, the predicate **window** was assigned the frame *Connecting_architecture* instead of the correct frame *Time_period_of_action* in the following sentence:

"No effect of anesthetic protocol on IOP during a 12 minute measurement [**window**]."

This problem is also relevant in generic WSD (Agirre et al., 2010) and benefits from the same solutions, for instance adapting embeddings to a particular domain (Taghipour and Ng, 2015) and efficient use of embeddings (Iacobacci et al., 2016).

Another major source of errors are subtle syntactic and semantic differences between frames which are hard to resolve on the sentence level (e.g. distinguishing between *Similarity* and *Identicality* for the predicate **different**). This could be addressed by incorporating subcategorization information and document context into the disam-

dataset	% errors			accuracy loss	
	unk	unl	$\sum$	unk∪unl	total
test-das	0.83	0.66	1.49	0.18	-
YAGS-test	3.76	13.05	16.81	6.40	25.60
MASC	12.15	33.70	45.85	24.03	33.90
TW-avg	10.40	9.68	20.08	6.31	18.96

Table 6: Error sources for `NN+Dep`; **unk** is the percentage of unknown and **unl** is the percentage of unlinked predicates among misclassified instances.

biguation model, which has been proposed in recent work in FrameNet SRL, see e.g. Hermann et al. (2014) and Roth and Lapata (2015).

To further explore the impact of user-generated text, we applied word-processor spelling correction to YAGS and tested our systems on the corrected set. The results do not change significantly, which indicates that a) our distributed representations provide enough information to classify also noisy user-generated text, and b) frameId errors cannot be attributed to preprocessing problems at large scale.

6 Discussion and outlook

Our analysis in Section 4 shows that domain adaptation is mainly required for the frameId step of FrameNet SRL. Unlike in PropBank SRL, in FrameNet SRL there is no significant performance drop for roleId once correct frames are available. The number of available roles given the correct frame is lower, on average 10, which reduces the complexity of the roleId task.

In Section 5 we introduced a simple, yet efficient frame identification method and evaluated it on in-domain and out-of-domain data. The method achieves competitive in-domain results, and outperforms the best available open-source system in out-of-domain accuracy. We also observe that our system performs well in the newly introduced `no-lexicon` evaluation setting, where no lexicon-based filtering is applied.

We identified a major issue in the standard frameId architecture: shifting to a new domain might render the predicate-frame associations in the FrameNet lexicon incomplete, which leads to errors for a standard classifier trained on in-domain data. One could optimize a frameId system to work in the `no-lexicon` setting which does not rely on the lexicon knowledge at all. However, in this setting the classification results are currently lower. Manually or automatically increasing both predicate and predicate-frame association coverage of the FrameNet lexicon could help, and we suggest investigating this line of research in future work.

While our method achieves state-of-the-art results on out-of-domain data, overall results are still significantly lower than the human performance observed for YAGS and TW, which shows that there is large room for improvement. Some further benefits could be gained from combining the WSABIE and NN-based classification, using advanced context representations, e.g. *context2vec* (Melamud et al., 2016) and incorporating syntactic information into the model. The out-of-domain performance could be further improved by adapting word representations to a new domain.

A direct comparison to the `Hermann-14` system in the out-of-domain setup would shed some more light on the properties of the task affecting the out-of-domain performance. On the one hand, we expect `Hermann-14` to perform worse due to its heavy reliance on syntactic information, which might decline in quality when moved to a new domain; on the other hand, the WSABIE-based classification might smoothen this effect. We make our dataset publicly available to enable comparison to related work.[6]

7 Conclusion

Domain dependence is a well-known issue for supervised NLP tasks such as FrameNet SRL. To the best of our knowledge, there is no recent study of the domain dependence of FrameNet SRL, also prohibited by a lack of appropriate datasets.

To address this problem, we 1) present the first comprehensive study of the domain generalization performance of the open-source `Semafor` system on several diverse benchmark sets. As a prerequisite, we introduce YAGS, a new, substantially sized test set in the domain of user-generated question-and-answer text. We find that the major bottleneck for out-of-domain FrameNet SRL is the frame identification step; we 2) explore a promising way to improve out-of-domain frame identification, i.e. using distributed word representations. Our simple frame identification system based on distributed word representations achieves higher scores for out-of-domain frame identification than previous systems and approaches state-of-the-art results in-domain. To support reproducibility of our results, we publish the YAGS test set annotations and our frame identification system for research purposes.

[6] `www.ukp.tu-darmstadt.de/ood-fn-srl`

Acknowledgements

This work was supported by FAZIT-Stiftung and by the German Research Foundation (DFG) through grant GU 798/18-1 (QAEduInf) and the research training group "Adaptive Preparation of Information form Heterogeneous Sources" (AIPHES, GRK 1994/1). We thank Orin Hargraves and our annotators for their excellent work on the annotation study, Dr. Richard Eckart de Castilho for support regarding WebAnno, as well as Dr. Judith Eckle-Kohler and the anonymous reviewers for their comments on earlier versions of this paper.

References

Eneko Agirre, Oier López de Lacalle, Christiane Fellbaum, Shu-Kai Hsieh, Maurizio Tesconi, Monica Monachini, Piek Vossen, and Roxanne Segers. 2010. SemEval-2010 Task 17: All-Words Word Sense Disambiguation on a Specific Domain. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 75–80. Association for Computational Linguistics.

Collin Baker, Michael Ellsworth, and Katrin Erk. 2007. SemEval-2007 Task 19: Frame Semantic Structure Extraction. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 99–104, Prague, Czech Republic, June. Association for Computational Linguistics.

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. Association for Computational Linguistics.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 120–128, Sydney, Australia, July. Association for Computational Linguistics.

Xavier Carreras and Lluís Màrquez. 2005. Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 152–164, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Danilo Croce, Cristina Giannone, Paolo Annesi, and Roberto Basili. 2010. Towards open-domain semantic role labeling. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 237–246, Uppsala, Sweden, July. Association for Computational Linguistics.

Dipanjan Das and Noah A. Smith. 2011. Semi-Supervised Frame-Semantic Parsing for Unknown Predicates. In *Proc. 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, California. Association for Computational Linguistics.

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

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

Katrin Erk and Sebastian Padó. 2006. SHALMANESER – A Toolchain For Shallow Semantic Parsing. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, volume 6, pages 527–532, Genoa, Italy. ELRA.

Charles J. Fillmore, Christopher R. Johnson, and Miriam R.L. Petruck. 2003. Background to FrameNet. *International journal of lexicography*, 16(3):235–250.

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, September. Association for Computational Linguistics.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 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 2009): Shared Task*, pages 1–18, Boulder, Colorado, June. Association for Computational Linguistics.

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, June. Association for Computational Linguistics.

Fei Huang and Alexander Yates. 2010. Open-domain semantic role labeling by modeling word spans. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 968–978, Uppsala, Sweden, July. Association for Computational Linguistics.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Embeddings for Word Sense Disambiguation: An Evaluation Study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 897–907, Berlin, Germany, August. Association for Computational Linguistics.

Anders Johannsen, Héctor Martínez Alonso, and Anders Søgaard. 2015. Any-language frame-semantic parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2062–2066, Lisbon, Portugal, September. Association for Computational Linguistics.

Richard Johansson and Pierre Nugues. 2008. 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, August. Coling 2008 Organizing Committee.

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, July. Association for Computational Linguistics.

Maciej Kula. 2015. Metadata embeddings for user and item cold-start recommendations. In Toine Bogers and Marijn Koolen, editors, *Proceedings of the 2nd Workshop on New Trends on Content-Based Recommender Systems co-located with 9th ACM Conference on Recommender Systems (RecSys 2015)*, volume 1448 of *CEUR Workshop Proceedings*, pages 14–21, Vienna, Austria, September. CEUR-WS.org.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*, pages 302–308. The Association for Computer Linguistics.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning generic context embedding with bidirectional LSTM. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*, pages 51–61.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS '13)*, pages 3111–3119, Lake Tahoe, Nevada, USA.

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

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.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 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, October. Association for Computational Linguistics.

Michael Roth and Mirella Lapata. 2015. Context-aware frame-semantic role labeling. *Transactions of the Association for Computational Linguistics*, 3:449–460.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2010. FrameNet II: Extended Theory and Practice. Technical report, ICSI, University of California, Berkeley.

Anders Søgaard, Barbara Plank, and Héctor Martínez Alonso. 2015. Using Frame Semantics for Knowledge Extraction from Twitter. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 2447–2452, Austin, Texas, USA.

Anders Søgaard. 2013. Semi-supervised learning and domain adaptation in natural language processing. *Synthesis Lectures on Human Language Technologies*, 6(2):1–103.

Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. 2008. The conll 2008 shared task on joint parsing of syntactic and semantic dependencies. In *CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 159–177, Manchester, England, August. Coling 2008 Organizing Committee.

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

Kaveh Taghipour and Hwee Tou Ng. 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*, pages 314–323, Denver, Colorado, May–June. Association for Computational Linguistics.

Jason Weston, Samy Bengio, and Nicolas Usunier. 2011. WSABIE: Scaling Up to Large Vocabulary Image Annotation. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three*, IJCAI'11, pages 2764–2770, Barcelona, Catalonia, Spain. AAAI Press.

Haitong Yang, Tao Zhuang, and Chengqing Zong. 2015. Domain adaptation for syntactic and semantic dependency parsing using deep belief networks. *Transactions of the Association for Computational Linguistics*, 3:271–282.

Seid Muhie Yimam, Richard Eckart de Castilho, Iryna Gurevych, and Chris Biemann. 2014. Automatic Annotation Suggestions and Custom Annotation Layers in WebAnno. In Kalina Bontcheva and Zhu Jingbo, editors, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. System Demonstrations*, pages 91–96, Stroudsburg, PA 18360, USA. Association for Computational Linguistics.