Evaluating FrameNet-style semantic parsing: the role of coverage gaps in FrameNet

Alexis Palmer and Caroline Sporleder
Computational Linguistics
Saarland University
{apalmer, csporled}@coli.uni-saarland.de

Abstract

Supervised semantic role labeling (SRL) systems are generally claimed to have accuracies in the range of 80% and higher (Erk and Padó, 2006). These numbers, though, are the result of highly-restricted evaluations, i.e., typically evaluating on hand-picked lemmas for which training data is available. In this paper we consider performance of such systems when we evaluate at the document level rather than on the lemma level. While it is well-known that coverage gaps exist in the resources available for training supervised SRL systems, what we have been lacking until now is an understanding of the precise nature of this coverage problem and its impact on the performance of SRL systems. We present a typology of five different types of coverage gaps in FrameNet. We then analyze the impact of the coverage gaps on performance of a supervised semantic role labeling system on full texts, showing an average oracle upper bound of 46.8%.

1 Introduction

A lot of progress has been made in semantic role labeling over the past years, but the performance of state-of-the-art systems is still relatively low, especially for deep, FrameNet-style semantic parsing. Furthermore, many of the reported performance figures are somewhat unrealistic because system performance is evaluated on hand-selected lemmas, usually under the implicit assumptions that (i) all relevant word senses (frames) of each lemma are known, and (ii) there is a suitable amount of training data for each sense. This approach to evaluation arises from the limited coverage of the available hand-coded data against which to evaluate. More realistic evaluations test systems on full text, but these same coverage limitations mean that the assumptions made in more restricted evaluations do not necessarily hold for full text. This paper provides an analysis of the extent and nature of the coverage gaps in FrameNet. A more precise understanding of the limitations of existing resources with respect to robust semantic analysis of texts is an important foundational component both for improving existing systems and for developing future systems, and it is in this spirit that we make our analysis.

Full-text semantic analysis

Automated frame-semantic analysis aims to extract from text the key event-denoting predicates and the semantic argument structure for those predicates. The semantic argument structure of a predicate describing an event encodes relationships between the participants involved in the event, e.g. who did what to whom. Knowledge of semantic argument structure is essential for language understanding and thus important for applications such as information extraction (Moschitti et al., 2003; Surdeanu et al., 2003), question answering (Shen and Lapata, 2007), or recognizing textual entailment (Burchardt et al., 2009). Evaluating an existing system for its ability to aid such tasks is unrealistic if the evaluation is lemma-based rather than text-based. Consequently, there continues to be significant interest in developing semantic role labeling (SRL) systems able to automatically compute the semantic argument structures in an input text.

Performance on the full text task, though, is typically much lower than for the more restricted evaluations. The SemEval 2007 Task on "Frame Semantic Structure Extraction," for example, required systems to identify key predicates in texts,

Coling 2008: Poster Volume, pages 928–936,
Beijing, August 2010

assign a semantic frame to the relevant predicates, identify the semantic arguments for the predicates, and finally label those arguments with their semantic roles. The systems participating in this task only obtained F-Scores between 55% and 78% for frame assignment, despite the fact that the task organizers adopted a lenient evaluation scheme which gave partial credit for near-misses (Baker et al., 2007). For the combined task of frame assigment and role labeling the performance was even lower, ranging from 35% to 54% F-Score.

Note that this distinction between evaluation schemes for SRL systems corresponds to the distinction between "lexical sample" and "all words" evaluations in word sense disambiguation, where results for the latter scheme are also typically lower (McCarthy, 2009).

The low performances are at least partly due to coverage problems. For example, Baker et al. (2007) annotated three new texts for their SemEval 2007 task. Although these new texts overlap in domain with existing FrameNet data, the task organizers had to create 40 new frames in order to complete annotation. The new frames were for word senses found in the test set but missing from FrameNet. The test set contained only 272 frames (types), meaning that nearly 15% of the frames therein were not yet defined in FrameNet. Obviously, coverage issues of this degree make full SRL a difficult task, but this is a realistic scenario that will be encountered in real applications as well.

As mentioned above, for many tasks it is necessary to compute the semantic argument structures for the whole text, or at least for multi-sentence passages. Due to non-local relations between argument structures this is also true for tasks like question answering, where it might be possible to automatically determine a subset of lemmas which are relevant for the task. For example, in (1) it might be possible to determine that the second sentence contains the answer to the question "*Was Thomas Preston acquitted of theft?*" However, to correctly answer this question, it is necessary to resolve the null instantiation of the CHARGES role of the VERDICT frame. This null instantiation links back to the previous sentence, and resolving it might require obtaining an analysis of the word *tried*.

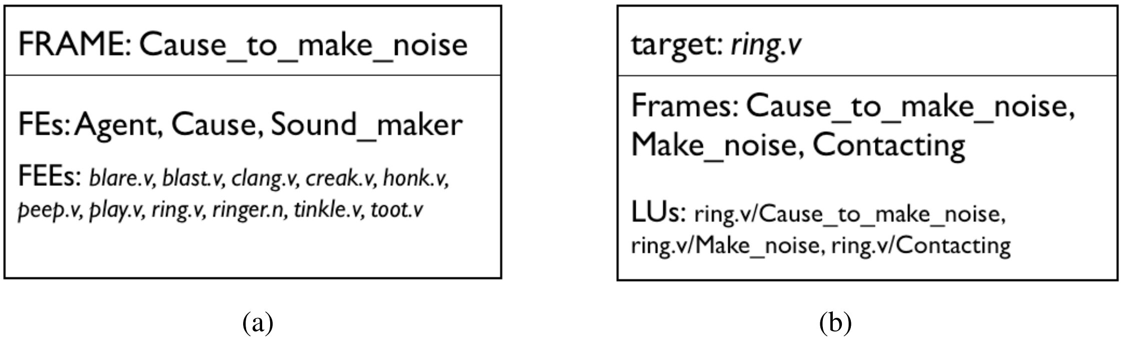
(1) [Captain Thomas Preston]$_{Defendant_i}$ was **tried**$_{Try\_defendant_i}$ for [murder]$_{Charges_{i,j}}$.

In the end [he]$_{Defendant_j}$ was **acquitted**$_{Verdict_j}$ [Ø]$_{Charges_j}$.

Performance levels obtained for full text are usually not sufficient for this kind of real-world task. FrameNet-style semantic role labeling has been shown to, in principle, be beneficial for applications that need to generalise over individual lemmas, such as recognizing textual entailment or question answering. However, studies also found that state-of-the-art FrameNet-style SRL systems perform too poorly to provide any substantial benefit to real applications (Burchardt et al., 2009; Shen and Lapata, 2007).

Extending the value of automated semantic parsing for a variety of applications requires improving the ability of systems to process unrestricted text. Several methods have been proposed to address different aspects of the coverage problem, ranging from automatic data expansion and semi-supervised semantic role labelling (Fürstenau and Lapata, 2009b; Fürstenau and Lapata, 2009a; Deschacht and Moens, 2009; Gordon and Swanson, 2007; Padó et al., 2008) to systems which can infer missing word senses (Pennacchiotti et al., 2008b; Pennacchiotti et al., 2008a; Cao et al., 2008; Burchardt et al., 2005). However, so far there has not been a detailed analysis of the problem. In this paper we provide that detailed analysis, by defining different types of coverage problems and performing analysis of both coverage and performance of an automated SRL system on three different data sets.

Section 2 of the paper provides an introduction to FrameNet and introduces the basic terminology. Section 4 describes our approach to coverage evaluation, Section 3 discusses the texts analyzed, and the analysis itself appears in Section 5. Section 6 then looks at one possibility for addressing the coverage problem. The final section presents some discussion and conclusions.



FRAME: Cause_to_make_noise

FEs: Agent, Cause, Sound_maker

FEEs: *blare.v, blast.v, clang.v, creak.v, honk.v, peep.v, play.v, ring.v, ringer.n, tinkle.v, toot.v*

(a)

target: *ring.v*

Frames: Cause_to_make_noise, Make_noise, Contacting

LUs: ring.v/Cause_to_make_noise, ring.v/Make_noise, ring.v/Contacting

(b)

Figure 1: Terminology: (a) Frame with core frame elements (FEs) and frame-evoking elements (FEEs) (b) Target with possible frame assignments and resultant lexical units (LUs)

2 FrameNet

Manual annotation of corpora with semantic argument structure information has enabled the development of statistical and supervised machine learning techniques for semantic role labeling (Toutanova et al., 2008; Moschitti et al., 2008; Gildea and Jurafsky, 2002).

The two main resources are PropBank (Palmer et al., 2005) and FrameNet (Ruppenhofer et al., 2006). PropBank aims to provide a semantic role annotation for every verb in the Penn TreeBank (Marcus et al., 1994) and assigns roles on a verb-by-verb basis, without making higher-level generalizations. Whether two distinct usages of a given verb are viewed as different senses or not is thus driven by both syntax (namely, differences in syntactic argument structure) and semantics (via basic, easily-discernable differences in meaning).

FrameNet[1] is a lexicographic project whose aim it is to create a lexical resource documenting valence structures for different word senses and their possible mappings to underlying semantic argument structure (Ruppenhofer et al., 2006). In contrast to PropBank, FrameNet is primarily semantically driven; word senses (*frames*)[2] are defined mainly based on sometimes-subtle meaning differences and can thus generalise across individual lemmas, and often also across different parts-of-speech. Because FrameNet focusses on semantics it is not restricted to verbs but also provides semantic argument annotations for nouns, adjectives, adverbs, prepositions and even multi-word expressions. For example, the sentence in (2) and the NP in (3) have identical argument structures because the verb *speak* and the noun *comment* evoke the same frame STATEMENT.

(2) [The politician]$_{Speaker}$ **spoke**$_{Statement}$ [about recent developments on the labour market]$_{Topic}$.

(3) [The politician's]$_{Speaker}$ **comments**$_{Statement}$ [on recent developments on the labour market]$_{Topic}$

Since FrameNet annotations are semantically driven they are considerably more time-consuming to create than PropBank annotations. However, FrameNet also provides 'deeper' and more informative annotations than PropBank analyses (Ellsworth et al., 2004). For instance, the fact that (2) and (3) refer to the same state-of-affairs is not captured by PropBank sense distinctions.

FrameNet Terminology

The English FrameNet data consist of an inventory of frames (i.e. word senses), a set of lexical entries, and a set of annotated examples exemplifying different syntactic realizations for selected frames (known as the *lexicographic annotations*). **Frames** are conceptual structures that describe types of situations or events together with their participants. **Frame-evoking elements (FEEs)** are predicate usages which evoke a particular frame. A given lemma can evoke different

[1] `http://framenet.icsi.berkeley.edu/`

[2] We follow Erk (2005) in treating frame assignment as a word sense disambiguation task. Thus in this paper we use the terms *frame* and *sense* interchangeably.

frames in different contexts; each instance of the lemma is a separate **target** for semantic analysis. For example, (4) and (5) illustrate two different frames of the lemma *speak*.

(4) [The politician]$_{Speaker}$ **spoke**$_{Statement}$ [about recent developments on the labour market]$_{Topic}$.

(5) [She]$_{Interlocutor_1}$ doesn't **speak**$_{Chatting}$ to [anyone]$_{Interlocutor_2}$.

In this paper we follow standard use of FrameNet terminology, with the possible exception of the term *lexical unit*. Figure 1 illustrates our use of FrameNet-related terminology, focussing on (a) the CAUSE_TO_MAKE_NOISE frame and (b) the target verb lemma *ring*.

The definition of a frame determines the available roles (**frame elements** or **FEs**) of the semantic argument structure for the particular use of the predicate, as well as the status—core or peripheral—of those roles. For example, the FE TOPIC is a core role under the STATEMENT frame, but a peripheral role under the CHATTING frame.

The lexical entry of a lemma in FrameNet specifies a list of frames which the lemma can evoke, and the pairing of a word with a particular frame is called a **lexical unit (LU)**. Ideally there should be annotated examples for each lexical unit, exemplifying different syntactic constructions which can realize this LU. However, as we will see later (Section 5) annotated examples can be missing. Also, because FrameNet is a lexicographic project, the examples were extracted to illustrate particular usages, i.e., they are not meant to be statistically representative.

3 Data

Having introduced the basic FrameNet terminology, we now describe in more detail the data sets used in the analysis. FrameNet Release 1.3 (FN1.3), the latest release from the Berkeley FrameNet project, includes both a corpus of lexicographic annotations (FNL), which we referred to in Section 2, and a corpus of texts fully-annotated with frames and semantic role labels (FNF). Annotations in the two corpora of course cover different sets of predicates and frames, and FNL is the corpus commonly used as the basis for training supervised FrameNet-based SRL systems (Erk and Padó, 2006).

In our analysis, we look at three data sets: the lexicographic annotations from FN1.3, the full text annotations from FN1.3, and a new data set of running text that was annotated for the SemEval 2010 Task-10 (see Table 1 for details).

FrameNet Lexicographic (FNL) FrameNet started as a lexicographic project, aiming to draw up an inventory of frames and lexical units, supported by corpus evidence, to document the range of syntactic and semantic usages of each lexical unit. The annotated example sentences in this part of FN1.3 are taken from the British National Corpus (BNC). BNC is a balanced corpus, hence FNL covers, in principle, a variety of domains.

For each LU, a subset of the sentences in which it occurs was selected for annotation, and in each extracted sentence, only the target LU was annotated. The sentences were not chosen randomly but with a set of lexicographic constraints in mind. In particular the sentences should exemplify different usage. Thus ideally selected sentences would be easy to understand and not too long or complex. As a consequence of this linguistically-driven selection procedure, the annotated sentences are not statistically representative in any way. FNL provides annotations for just under 140,000 FEEs (tokens). On average, around 20 sentences are annotated for each LU. FrameNet's frame inventory contains 722 frames.[3]

FrameNet Full Texts (FNF) Starting with release 1.3, FrameNet also provides annotations of running texts. In this annotation mode, all LUs in a sentence and all sentences in a text are annotated. FN1.3 contains two subsets of full text annotations. The first of these (**PB**) contains five texts which were also annotated by the PropBank project. While all texts come from the Wall Street Journal, they are not prototypical examples of the financial domain, rather they are longer essays covering a wide variety of general interest topics

[3] Only lexical frames are included in this number. In addition to those, FrameNet 1.3 defines another 74 frames which cannot be lexicalised but are included because they provide useful generalisations in the frame hierarchy.

Data	Genre / Domain	FEEs		Frames
		Tokens	Types	Types
FNL	mixed	139,439	8370	722
PB	essays, general interest	1580	680	319
NTI	reports, foreign affairs	8271	1305	434
SE	fiction, crime	1530	680	320

Table 1: Statistics for the three data sets

(ranging from 'Bell Ringing' to 'Earthquakes'). The second subset (**NTI**) contains 12 texts from the Nuclear Threat Initiative website.[4] These texts are intelligence reports which summarize and discuss the status of various countries with regard to the development of weapons and missile systems. Statistics for both data sets are given in Table 1.

SemEval 2010 Task-10 Full Texts (SE) While the FrameNet full texts allow us to estimate coverage gaps that arise from limited training data, they do not allow us to gauge coverage problems arising from missing frames in the FN1.3 inventory. The reason for this is that the frame inventory reflects the annotations of both the lexicographic and the full text part of FN1.3, i.e., every frame annotated in one of these subsets will also be part of the inventory. To estimate the frame coverage problem on completely new texts, we therefore included a third (full text) data set that was annotated for the SemEval 2010 Task 10 on "Linking Events and Their Participants in Discourse" (Ruppenhofer et al., 2009).[5] The text is taken from Arthur Conan Doyle's "The Adventure of Wisteria Lodge". It thus comes from the fiction domain.

The text was manually annotated with frame-semantic argument structure by two experienced annotators. Similar to the FNF texts, the annotators aimed to annotate all LUs in the text. To do so, some new frames had to be created for previously un-encountered LUs. These new frames are not part of FN1.3 and we can thus use them to estimate coverage problems arising from missing frames. Details for the data set can be found in Table 1. This data set is very similar to the PB set in terms of size, FEE type-token ratio and number of frames (types).

4 Types of Coverage Gaps

Semantic role labelling systems have to perform two sub-tasks: (i) identifying the correct frame for a given lemma and context, and (ii) identifying and labeling the frame elements. The most severe coverage problems typically arise with the first subtask. Furthermore, coverage problems related to frame identification have a knock-on effect on role identification and labeling because the choice of the correct frame determines which roles are available. Therefore, we focus on the frame identification task in this paper.

Attempts to do automated frame assignment on unrestricted text invariably encounter problems associated with limited coverage of frame-evoking elements in FrameNet. However, not every coverage gap is the same, and the precise nature of a coverage gap influences potential strategies for addressing it. In this section we describe the different types of coverage gaps. We proceed from less problematic coverage gaps to more problematic ones, in the sense that the former can be addressed more straighforwardly by automated systems than can the latter.

4.1 NOTR gaps

Some coverage gaps occur when lexical units (LUs) defined in FrameNet lack corresponding annotated examples; these gaps are the result of lacking training data, hence we call them **NOTR** gaps. To give a sense of the abundance of such gaps, of the 10,191 LUs defined in FN1.3, annotated examples are available for only 6727.

NOTR-LU: lexical unit with no training data. In many cases, an LU — a specific pairing of a target lemma with one frame — may be defined in FrameNet, thus potentially accessible to an automated system, but lacking labeled training material. For example, FrameNet defines two LUs for the noun *ringer*: with the frames CAUSE TO MAKE NOISE and SIMILARITY. It is clear that the occurrence of *ringer* in (6) belongs to the former LU, even given a very limited context. The lexicographic annotations, though, provide training material only for the SIMILARITY frame.

[4] http://www.nti.org

[5] The data set is available from http://semeval2.fbk.eu/semeval2.php?location=data.

(6) Then, at a signal, the **ringers** begin varying the order in which the bells sound without altering the steady rhythm of the striking.

NOTR-LU gaps pose particular problems to a fully-supervised SRL system, because such a system cannot learn anything about the context in which the CAUSE TO MAKE NOISE frame is more appropriate. A NOTR-LU gap is identified for an LU even if training data is available for other senses (i.e. other LUs) of the target lemma.

NOTR-TGT: target with no training data. In other cases, a target lemma may be defined as participating in one or more LUs, but with no training data available for *any* of them. In other words, a supervised automated system trained only on the available annotated examples will fail to learn any potential frame assignments for the target lemma. Such is the case for *art*, which in FrameNet is assigned the single frame CRAFT, but for which FNL contains no training data.

(7) The **art** of change-ringing is peculiar to the English, and, like most English peculiarities, unintelligible to the rest of the world.

Whereas a NOTR-LU gap obscures a particular frame assignment for a target lemma, a NOTR-TGT gap indicates a complete absence in the lexicographic corpus of annotated data for the lemma.

4.2 UNDEF gaps

The previous coverage problems arise from a lack of annotated data, an issue which conceivably could be addressed through further annotation. More serious problems arise when a text contains word senses, words, or frames not contained in FrameNet. We call such elements 'undefined'; specifically, they receive no treatment in FN1.3.

UNDEF-LU: lexical unit not defined. Coverage gaps of this sort occur when the frame inventory for a given lemma is not complete. In other words, at least one LU for the lemma exists in FrameNet, but one or more other LUs are missing. For example, the noun *installation* occurs in FrameNet with the frames LOCALE BY USE and INSTALLING. The sense of an art installation, which is an instance of the frame PHYSICAL ARTWORKS, is missing.

UNDEF-TGT: target not addressed. In the worst case, all LUs for a target lemma might be missing, i.e., the lemma does not occur in the FrameNet lexicon at all. The noun *fabric* is an example. Though it has at least two distinct senses—that of cloth or material and that of a framework (e.g. *the fabric of society*)—FrameNet provides no help for determining appropriate frames for instances of this lemma.

UNDEF-FR: frame not defined. Finally, it may be not only that the LU is missing, but that there is no definition in FrameNet for the correct frame given the context. For example, in the sports domain the lemma *ringer* can have the sense of (*a horseshoe thrown so that it encircles the peg*); to our knowledge, this sense is not available in FrameNet.

5 Coverage gaps and automated processing

With the exception of work on extending coverage, most FrameNet-style semantic role labeling studies draw both training and evaluation data from FNL. This is an unrealistic evaluation scenario for full-text semantic analysis, as such evaluation limits the domain for which prediction can occur to those lexical entries treated in FNL. For systems which do not attempt any generalization beyond those lexical entries with training data, this limits the system to 5864 lemmas for which it can make predictions regarding frame assignment and role labeling.

Disregarding whether annotations have yet been provided for the lexical units in FNL still limits us to 8370 frame-evoking elements (targets). To better understand the potential of current frame-semantic resources for semantic analysis of unrestricted text, we evaluate coverage of the FNL annotations against the texts in FNF, as well as against the SemEval text. We then analyze the performance of an off-the-shelf, supervised SRL system, Shalmaneser (Erk and Padó, 2006), on the same texts, with a focus on the types of

Dataset	TR-LU	NOTR-LU	NOTR-TGT	UNDEF-LU	UNDEF-FR
PB	42.66	9.56	47.78	–	–
NTI	46.77	7.77	45.46	–	–
SE	51.64	6.86	26.01	3.40	12.09

Table 2: FrameNet coverage for analyzed texts

errors made and the upper bound on performance for this system.

5.1 FrameNet coverage

As described in Section 4, in many cases a lexical unit, a frame-evoking element, or a frame may simply not be represented in FrameNet. In other cases, the entity may be in FN1.3 but lacking training data. Of the 722 frames defined in FN1.3, for example, annotations exist for 502.

For the three data sets analyzed, Table 2 shows the degree of coverage provided by FNL for the gold-standard frame annotations. First, the TR-LU column shows the non-problematic cases, for which the correct frame annotation is available in FrameNet, with training data. The next two columns represent training gaps related to lack of training data: NOTR-LU are cases for which training data exists for the target, but not for the correct sense of the target, and NOTR-TGT instances are those for which no training data at all exists for the target.

Because all targets annotated in the FNF texts (i.e. PB and NTI above) are incorporated in FN1.3, gaps due to missing LUs, targets, or frames do not exist for those texts. The same does not hold for the SemEval (SE) text. For 3.4% of the annotated SemEval targets, an LU is entirely missing from the lemma's frame inventory in FrameNet, and in just over 12% of cases both the lemma and the frame are missing. In total, more than 15% of LUs appearing in the gold-standard SemEval annotations are not defined at all within FrameNet. This figure accords with that found by Baker et al. (2007).

5.2 Error analysis of full-text frame assignment

Here we examine the errors made by Shalmaneser for frame assignment on the three data sets. The upper bound on apparent performance is fixed by the number of targets for which Shalmaneser has seen training data, namely the sum of TR-LU and NOTR-LU in Table 2.[6]

Dataset	Correct	Type(i)	Type(ii)	Type(iii)
PB	36.71	5.95	9.56	47.78
NTI	41.22	5.55	7.77	45.46
SE	46.67	4.97	6.86	41.50

Table 3: Shalmaneser performance on texts

We consider three categories of errors: (i) *normal or true errors* are misclassifications when the correct label has been seen in the training data. In this category we also count errors resulting from incorrect lemmatization. (ii) *label-not-seen errors* are misclassifications when the correct label does not appear in the training data and thus is unavailable to the classifier. Finally, (iii) *no-chance errors* occur when the system has no information for either a given target or a given frame. Table 3 shows the prevalence of each error type for each data set, given as the percentage of all frame-assignment targets.

It can be seen that the frame assignment accuracy is relatively low for all three texts (between 37% and 47%). However, only a relatively small proportion of the misclassifications are due to true errors made by the system. Furthermore, a large amount of errors (41% to 48%, with an average of 46.8%) is due to cases where important information is missing from FrameNet (Type (iii) errors). Consequently, improving the semantic role labeller by optimising the feature space or the machine learning framework is going to have very little effect. A much more promising path would be to investigate methods which might enable the SRL system to deal gracefully with unseen data. One possible strategy is discussed in the next section.

[6] By 'apparent performance' we mean the system's own evaluation of its accuracy on frame assignment.

6 Frame and lemma overlap

One potential strategy for improving full-text semantic analysis without performing additional annotation is to take advantage of semantic overlap as it is represented in FrameNet. We can look at two different types of overlap in FrameNet: **lemma overlap** and **frame overlap**.

6.1 Lemma overlap

The approach of treating frame assignment as a word sense disambiguation task (as, e.g., by Shalmaneser) relies on the overlap of LUs with the same lemma and trains lemma-based classifiers on all training instances for all LUs involving that lemma. One way to consider using labeled material in FrameNet to improve performance on targets for which we have no labeled material is to generalize over lemmas associated with the same frame. The idea is to use training instances from related lemmas to build a larger training set for lemmas with little or no annotated data.

Of the 8370 lemmas in FN, 8358 share a single frame with at least one other lemma. 890 overlap on two frames with at least one other lemma, and 111 have 3-frame overlap with at least one other lemma. Only 16 lemmas show an overlap of four or more frames. These groupings are:

1. clang.v, clatter.v, click.v, thump.v
2. hit.v, smack.v, swing.v, turn.v
3. drop.v, rise.v
4. remember.v, forget.v
5. examine.v, examination.n
6. withdraw.v, withdrawal.n

The first two groupings are sets of words that are closely semantically related, the second two are opposite pairs, and the third two are verb-nominalization pairs.

The lemma overlap groups differ with respect to how much training data they make accessible.

6.2 Frame overlap

Another possibility to be considered is generalization over all instances of a given frame. For the 502 frames with annotated examples, the number of annotated instances ranges from one (SAFE SITUATION, BOARD VEHICLE, and ACTIVITY START to 6233 (SELF MOTION), with an average of 278 training instances per frame.

In future work we will examine the effectiveness of binary frame-based classifiers, abstracting away from individual predicates to predict whether a given lemma belongs to the frame in question (for a related study see Johansson and Nugues (2007)). A potential drawback to this approach is the loss of predicate-specific information. We know, for example, about verbs that they tend to have typical argument structures and typical syntactic realizations of those argument structures.

In addition to this frame-overlap approach, we will consider the impact on coverage of using coarser-grained versions of FrameNet in which frames have been merged according to frame relations defined over the FrameNet hierarchy, using the FrameNet Transformer tool described in (Ruppenhofer et al., 2010).

7 Conclusions

Although it is clear that the capability to do shallow semantic analysis on unrestricted text, and on complete documents or text passages, would help performance on a number of key tasks, currently-available resources seriously limit our potential for achieving this with supervised systems. The analysis in this paper aims for a better understanding of the precise nature of these limitations in order to address them more deliberately and with a principled understanding of the coverage problems faced by current systems.

To this end, we outline a typology of coverage gaps and analyze both coverage of FrameNet and performance of a supervised semantic role labeling system on three different full-text data sets, totaling over 150,000 frame-assignment targets. We find that, on average, 46.8% of targets are not covered under straight supervised-classification approaches to frame assignment.

Acknowledgments

This research has been funded by the German Research Foundation DFG under the MMCI Cluster of Excellence. Thanks to the anonymous reviewers, Josef Ruppenhofer, Ines Rehbein, and Hagen Fürstenau for interesting and helpful comments and discussions, and to Collin Baker for assistance with data.

References

C. Baker, M. Ellsworth, K. Erk. 2007. Semeval-2007 task 19: Frame semantic structure extraction. In *Proceedings of SemEval-2007*.

A. Burchardt, K. Erk, A. Frank. 2005. A WordNet Detour to FrameNet. In *Proceedings of the GLDV-05 Workshop GermaNet II*.

A. Burchardt, M. Pennacchiotti, S. Thater, M. Pinkal. 2009. Assessing the impact of frame semantics on textual entailment. *Journal of Natural Language Engineering, Special Issue on Textual Entailment*, 15(4):527–550.

D. D. Cao, D. Croce, M. Pennacchiotti, R. Basili. 2008. Combining word sense and usage for modeling frame semantics. In *Proceedings of STEP-08*.

K. Deschacht, M.-F. Moens. 2009. Semi-supervised Semantic Role Labeling Using the Latent Words Language Model. In *Proceedings of EMNLP-09*.

M. Ellsworth, K. Erk, P. Kingsbury, S. Padó. 2004. PropBank, SALSA, and FrameNet: How Design Determines Product. In *Proceedings LREC 2004 Workshop on Building Lexical Resources from Semantically Annotated Corpora*.

K. Erk, S. Padó. 2006. Shalmaneser – a toolchain for shallow semantic parsing. In *Proceedings of LREC-06*.

K. Erk. 2005. Frame assignment as word sense disambiguation. In *Proceedings of IWCS 6*.

H. Fürstenau, M. Lapata. 2009a. Graph alignment for semi-supervised semantic role labeling. In *Proceedings of EMNLP 2009*.

H. Fürstenau, M. Lapata. 2009b. Semi-supervised semantic role labeling. In *Proceedings of EACL 2009*.

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

A. Gordon, R. Swanson. 2007. Generalizing semantic role annotations across syntactically similar verbs. In *Proceedings of ACL 2007*.

R. Johansson, P. Nugues. 2007. Using WordNet to extend FrameNet coverage. In *Proceedings of the Workshop on Building Frame-semantic Resources for Scandinavian and Baltic Languages, NODALIDA*.

M. Marcus, G. Kim, M. A. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz, B. Schasberger. 1994. The Penn Treebank: Annotating predicate argument structure. In *ARPA Human Language Technology Workshop*.

D. McCarthy. 2009. Word Sense Disambiguation: An Overview. *Language and Linguistics Compass*, 3(2):537–558.

A. Moschitti, P. Morarescu, S. Harabagiu. 2003. Open-domain information extraction via automatic semantic labeling. In *Proceedings of FLAIRS*.

A. Moschitti, D. Pighin, R. Basili. 2008. Tree Kernels for Semantic Role Labeling. *Computational Linguistics*, 34(2).

S. Padó, M. Pennacchiotti, C. Sporleder. 2008. Semantic role assignment for event nominalisations by leveraging verbal data. In *Proceedings of Coling 2008*.

M. Palmer, D. Gildea, P. Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–105.

M. Pennacchiotti, D. D. Cao, R. Basili, D. Croce, M. Roth. 2008a. Automatic induction of FrameNet lexical units. In *Proceedings of EMNLP-08*.

M. Pennacchiotti, D. D. Cao, P. Marocco, R. Basili. 2008b. Towards a Vector Space Model for FrameNet-like Resources. In *Proceedings of LREC-08*.

J. Ruppenhofer, M. Ellsworth, M. R. L. Petruck, C. R. Johnson, J. Scheffczyk. 2006. FrameNet II: Extended Theory and Practice.

J. Ruppenhofer, C. Sporleder, R. Morante, C. Baker, M. Palmer. 2009. SemEval-2010 Task 10: Linking Events and Their Participants in Discourse. In *Proceedings of SEW-2009*.

J. Ruppenhofer, M. Pinkal, J. Sunde. 2010. Generating FrameNets of various granularities. In *Proceedings of LREC 2010*.

D. Shen, M. Lapata. 2007. Using semantic roles to improve question answering. In *Proceedings of EMNLP-2007*.

M. Surdeanu, S. Harabagiu, J. Williams, P. Aarseth. 2003. Using predicate-argument structures for information extraction. In *Proceedings of ACL 2003*.

K. Toutanova, A. Haghighi, C. D. Manning. 2008. A Global Joint Model for Semantic Role Labeling. *Computational Linguistics*, 34(2).