Preposition Disambiguation: Still a Problem

Ken Litkowski
CL Research
9208 Gue Road
Damascus, MD 20872, USA
ken@clres.com

Abstract

Considerable recent progress has been made in preposition disambiguation using the SemEval 2007 corpus, with results reaching accuracy of over 88 percent. However, with a new corpus of tagged instances, use of the models shows a decline in performance to around 43 percent. This suggests that recent efforts suffer from an out-of-domain problem. Detailed examination of the dimensions of this problem suggests that the sampling methodology used in creating the SemEval corpus was biased to the underlying reliance on FrameNet. In view of the demonstration of increasing importance attached to prepositions in semantic role labeling, further efforts to understand preposition behavior seem warranted. Initial examination of the new data may provide a basis for future directions in preposition disambiguation.

1 Introduction

Since the SemEval 2007 task on preposition disambiguation (Litkowski & Hargraves, 2007), the corpus of 24,663 sentences has been used for the development of increasingly accurate performance in studies specifically designed for that purpose, as well as others intended for such tasks as semantic role labeling and classification. The validity of that corpus has not previously been questioned. Recently, two new corpora have been released under The Preposition Project (TPP; Litkowski & Hargraves (2005); Litkowski & Hargraves (2006)). These corpora (Litkowski (2013), along with the original SemEval corpus, are now available in both the original Senseval/SemEval lexical sample format and in CoNLL-X format where each sentence has been tokenized, lemmatized, part-of-speech tagged, and parsed using a dependency parser.[1]

One of these corpora (labeled the OEC corpus) consists of 7,650 sentences from the Oxford University Press sentence dictionary supporting various dictionaries, drawn from the Oxford English Corpus. Of these, 3,380 sentences correspond to prepositions used in the SemEval task. We use these sentences as the basis for our investigation into the representativeness of the SemEval corpus and the ability to generalize preposition disambiguation models.

In section 2, we describe the motivation for our work, namely, the emerging role of joint inference using prepositions along with predicates in semantic role labeling. In section 3, we describe our approach for analyzing the generalizability of current disambiguation models, that is, how we used the OEC corpus. In section 4, we present our results, identifying anomalous characteristics of attempting to use preposition disambiguation models on new datasets. In section 5, we explore possible reasons why these models do not generalize. In section 6, we suggest some approaches for future examination of preposition behavior. In section 7, we summarize our conclusions with some general observations on using maximum entropy or support vector machine modeling for disambiguation efforts.

2 Motivation

Several recent studies have suggested the importance of prepositions in semantic role classification. These studies move the focus away from preposition disambiguation per se into the role of prepositions in signifying semantic relations.

[1] These corpora are available at http://www.clres.com.

Zapirain et al. (2013) examined selectional preferences for semantic role classification, demonstrating that the task is better modeled using both verbs and prepositions. In addition to this primary finding, showing improved results in domain, they also reported even better performance out of domain. Srikumar & Roth (2011) defined a joint inference model that captured the interdependencies between verb semantic role labeling and relations expressed using prepositions. In addition, in a table comparing preposition sense performance between Penn Treebank and SemEval prepositions, they noted a skewed performance, with accuracies trained on one dataset performing poorly when predicting on the other dataset. They did not discuss this matter further, focusing on the development of their model using Penn Treebank data. We focus on this finding below. Srikumar & Roth (2013) focused on predicting semantic relations expressed by prepositions, defining an inventory based on collapsing related senses across prepositions. They then jointly modeled the preposition relation and its arguments. In this study, they used only the prepositions included in SemEval.

These studies are suggestive of the importance of prepositions, but have generally been limited to a relatively small set of prepositions. The new corpora from TPP provide an opportunity for extending these results from these prepositions (although the most common) to include less common prepositions and to multiword phrases that behave like prepositions. In addition, the studies suggest that the corpora upon which the findings are based may not be representative, leading to uneven performance when applied to other domains.

3 Experimental Design

In attempting to extend results from joint inference studies to other prepositions, we first needed to verify the extensibility of our approach. This is where we discovered the potentially significant result that previous models may not generalize. To pursue our objectives, we locally implemented an updated version[2] of the best performing system for preposition disambiguation (Tratz & Hovy (2011)) and made modifications and extensions to this system. These alterations do not change the core of his system. They were made only for the purpose of better understanding the limitations of the underlying corpora. We have used his system in initial analyses of the other corpus mentioned in Litkowski (2013) and expect that it will provide the basis for trying to deal with the issues that are being raised in this paper.

We first implemented Tratz' system locally and developed a module to obtain and list the results of disambiguation runs (including confusion matrices). Tratz' system includes a preposition disambiguation module based on a refined sense inventory from the one used in SemEval. This sense inventory is described in Tratz (2011). We next verified that we could reproduce his SemEval disambiguation results, i.e., achieving his reported accuracy of over 88 percent, using the SemEval test corpus. After doing this, we subjected the OEC corpus to his system's part-of-speech tagging, lemmatizing, and dependency parsing. In our experiments, we were thus using the OEC corpus as the equivalent to the SemEval test corpus.

We first used Tratz' full system to disambiguate the SemEval prepositions in the OEC corpus using the refined sense inventory. In doing so, we used Tratz' mapping of TPP senses to his senses. We obtained an overall accuracy of 43.1 percent. In view of the fact that the OEC sense inventory conformed to the TPP senses, we next ran Tratz' system with the original SemEval sense inventory, obtaining an accuracy of 32.5 percent. Finally, we used the OEC corpus as the basis for training and ran the models based on this corpus against the SemEval test instances. We obtained an overall accuracy of 38.2 percent. In Table 1, we summarize the results from all these options by preposition. A detailed discussion of these results is provided in the next two sections.

4 Examination of Results

Table 1 identifies the 34 prepositions used in the SemEval 2007 task on preposition disambiguation. The table is broken down into three main columns showing the number of senses, the number of instances, and the accuracy.

[2]Available at http://sourceforge.net/projects/miacp/

	Number of Senses			Number of Instances			Accuracy				
Prep	TPP	No Data	Tratz	SE Train	SE Test	OEC	Original	Refined	OEC Using Refined	OEC Using TPP	OEC Train
about	6	0	3	710	364	36	0.956	0.984	0.694	0.694	0.871
above	9	4	4	48	23	173	0.739	0.826	0.480	0.208	0.304
across	3	0	3	319	151	40	0.967	0.993	0.550	0.600	0.609
after	11	5	6	103	53	165	0.660	0.830	0.606	0.327	0.264
against	10	4	5	195	92	197	0.902	0.978	0.645	0.193	0.456
along	4	1	3	365	173	60	0.954	0.896	0.583	0.267	0.734
among	4	1	5	100	50	70	0.820	0.780	0.343	0.386	0.580
around	6	0	5	335	155	102	0.697	0.819	0.598	0.422	0.381
as	2	1	2	174	84	11	1.000	0.952	0.364	0.545	0.286
at	12	1	10	715	367	34	0.875	0.926	0.294	0.324	0.300
before	4	1	4	47	20	44	0.900	0.900	0.477	0.545	0.600
behind	9	4	5	138	68	153	0.809	0.926	0.392	0.268	0.471
beneath	6	3	3	57	28	120	0.821	0.929	0.492	0.225	0.393
beside	3	2	3	62	29	41	1.000	1.000	0.488	0.488	0.724
between	9	2	7	211	102	165	0.961	0.990	0.382	0.370	0.529
by	22	10	16	510	248	*	0.871	0.867	*	*	*
down	5	2	4	332	153	86	0.824	0.791	0.395	0.349	0.477
during	2	0	1	81	39	40	0.846	1.000	1.000	0.650	0.538
for	15	2	17	951	478	185	0.822	0.814	0.330	0.281	0.218
from	16	0	16	1206	578	204	0.889	0.841	0.255	0.230	0.369
in	15	2	23	1397	688	101	0.783	0.786	0.347	0.366	0.315
inside	5	1	5	67	38	74	0.711	0.632	0.243	0.257	0.395
into	10	2	9	604	297	142	0.855	0.902	0.542	0.352	0.208
like	7	0	5	266	125	111	0.912	0.952	0.279	0.288	0.440
of	20	3	26	3004	1478	91	0.886	0.887	0.429	0.495	0.396
off	7	3	6	161	76	106	0.868	0.855	0.425	0.406	0.434
on	25	5	19	872	441	160	0.825	0.853	0.319	0.325	0.272
onto	3	0	3	117	58	15	0.914	0.983	0.733	0.733	0.897
over	17	5	12	200	98	235	0.755	0.878	0.255	0.255	0.418
round	8	1	5	181	82	127	0.707	0.768	*	*	0.488
through	16	1	7	441	208	208	0.490	0.962	0.380	0.202	0.298
to	17	7	14	1183	572	57	0.907	0.897	0.333	0.281	0.141
towards	6	2	7	214	102	80	0.990	0.951	*	*	0.686
with	18	3	15	1191	578	154	0.879	0.905	0.584	0.538	0.358
Overall	332	78	278	16557	8096	3380	0.857	0.881	0.431	0.325	0.382

Table 1. Summary results showing number of senses, number of instances, and accuracy (see text for details)

4.1 Number of Senses

The column **TPP** in the number of senses is the number of senses in the original TPP sense inventory. The third column, labeled **Tratz**, is the number of senses in Tratz' refined sense inventory (as described in Tratz (2011)). The overall number of senses is 54 fewer in Tratz' inventory. While generally this might be interpreted as Tratz using more coarse-grained senses, this is not the case, as suggested by the increased number of senses for *for*, *in*, and *of*. As described in this thesis, Tratz followed a principled analysis in the development of this sense inventory. In particular, he noted that the original TPP sense inventory was not considered to be completely accurate (as discussed in Litkowski & Hargraves (2005)). For example, in tagging the SemEval corpus, the lexicographer found it necessary to increase the sense inventory by about 10 percent. No attempt was made at refining the remaining senses, many of which were indentified in TPP as problematic.

The column labeled **No Data** indicates the number of senses in the TPP sense inventory for which there were no tagged instances in the SemEval instances. The overall number, 78, indicates that there were no training (or test) instances for 25 percent of the senses in the SemEval data. This affects each of the experiments with the OEC data, either predicting the OEC senses from the SemEval training data or using the OEC corpus as training data to predict the SemEval test instances.

4.2 Number of Instances

The number of instances shows how many sentences are present in each of the corpora. The first two columns (**SE Train** and **SE Test**) show the number of training and test instances in the corpus used for the SemEval task on preposition

disambiguation. The third column (**OEC**) shows the number of instances in the OEC corpus.

As suggested in Litkowski (2013), the OEC corpus is intended to contain 20 instances for each sense; however, the table clearly shows that for some prepositions (*at, in, of, on, to,* and *with*), the number of instances per sense is much fewer. This may affect the ability of this corpus to provide sufficient data for training.

Comparison of the number of instances in the SemEval and in the OEC corpora shows considerable disparities that may be significant. There is a high number in SemEval and a low number in OEC for *about, across, as, at, in, of, to,* and *with*. An immediate question arises as to the ability of these OEC instances to serve as a training set. There is a low number in SemEval and a high number in OEC for *above*, *after*, and *beneath*. Since each of these prepositions has a large number of senses for which there was no data in the SemEval training set, an immediate question is how well these senses can be predicted in OEC.

4.3 Accuracy

The first two columns under this heading (**Original** and **Refined**) are accuracy results taken from Tratz (2011), the first using the TPP sense inventory and the second using Tratz' refined sense inventory as laid out in his thesis. Generally, the results for each preposition are quite similar, with a statistically significant difference in the overall results. Based on Tratz' development of the refined sense inventory, it is likely that his efforts represent a tightening of the sense distinctions. There are several larger discrepancies in the results for individual prepositions; further study to understand these differences might prove useful.

The third column (**OEC Using Refined**) represents a direct application of Tratz' system against the OEC instances. We ran his system to part-of-speech tag, parse, and disambiguate the target prepositions to obtain his prediction of the appropriate sense using his refined sense inventory. Tratz' system includes a definition file for each preposition mapping from the TPP sense to his refined sense. These are not exact equivalencies, and provide nuanced mappings involving subset, union, and subsumption relations. In this initial analysis, we did not examine these nuances, but rather just formed unions of all the TPP senses that were related to Tratz' senses. In determining the accuracy for the OEC assignments, we merely checked whether the TPP sense indicated in the OEC corpus was a member of the set associated with the Tratz sense. As can be seen, the overall accuracy was 0.431, compared to Tratz' result of 0.881 for the SemEval test set.

The next column under accuracy (**OEC Using TPP**) involved running Tratz' system, but predicting the TPP sense. Tratz' system includes an option to use the TPP sense inventory rather than his refined inventory. In this case, we did not invoke his mapping, but used the TPP prediction directly for computing the accuracy. As can be seen, the overall result was 0.325, compared to 0.857 for Tratz' predictions with the SemEval test set. Our lower accuracy using this method seems to support the idea that Tratz' refinements provide a tighter sense inventory.

The final column under accuracy (**OEC Train**) addressed the question whether we could use the OEC instances as a training set by which to predict the SemEval test instances. We used Tratz' system to develop models for each preposition based on the OEC instances and then applied these models to the SemEval test set. As can be seen, we obtained a higher overall accuracy of 0.382, but still much lower than what Tratz obtained.

Based on our experiments in attempting to apply Tratz' state of the art system, we conclude that there is a significant mismatch in the generalizability of his system. In the next section, we describe our efforts to understand why this is the case.

5 Further Analysis of Results

In this section, we describe our efforts to understand the significant differences in our tests of applying the Tratz models to a new set of data. Unfortunately, we do not succeed, since the difficulties appear to lie much deeper than our preliminary investigations. We hope that our efforts will provide a starting point for further study in characterizing preposition behavior.

5.1 Absence of SemEval Instances

Our first hypothesis is that the SemEval instances, despite the large sample size, do not contain relevant data. As noted above, almost 25 percent of the senses identified in the OEC corpus have no instances in the SemEval data. Thus, there is no

basis on which to develop and train models for these senses.

To examine this possibility, we went through the entire OEC corpus to identify sentences that were not present in the SemEval corpus. We found 675 such sentences, about 20 percent of the corpus. If we remove these sentences, thus lowering the denominator in the accuracy calculations, we improve the accuracy in the OEC using TPP senses to 0.406 and in the OEC using Tratz' refined sense inventory to 0.539.[3] In two cases (*as* and *beside*), the accuracy improved to 100 percent. While these improvements are significant, they still do not close the gap in the overall accuracy.

5.2 Representativeness of SemEval Instances

As mentioned above, Srikumar & Roth (2011) found that SemEval instances and Penn Treebank instances did not predict each other as well as their own internal consistency. Our results support this finding. This raises the question of how representative are either set of instances. We examine this question for the SemEval instances, using the OEC instances as a benchmark.

The construction of the SemEval instances was supposedly random, with the suggestion that the sampling from the FrameNet data would provide a representative set of instances. However, FrameNet is recognized as not providing full coverage of the lexicon. We hypothesized that this lack of coverage may explain the difference in the accuracy results between the SemEval and OEC corpora.

To examine this hypothesis, we performed an analysis of the governor features identified in processing the instances using Tratz' system. We first extended his system to include a module for accessing a FrameNet dictionary.[4] The relevant governor features we used in this analysis are the lemma and the word class of the governor. Since the lemma is a single word and the word class is almost always one of the four major content word types, we excluded FrameNet lexical units consisting of multiple words (either with spaces or underscores) and those not in a principal part of speech. Our dictionary consists of a concatenation of the lexical unit and the part of speech (e.g., "**abandon.vrb**"). Our dictionary has 8631 entries with 10984 senses (e.g., **abandon.vrb** has three senses). Our dictionary also includes a list of all the frame element realizations that have been annotated for each entry. This includes a characterization of the phrase type and the grammatical function. It should be noted that many of the entries or senses have no annotations.

Tratz' system does not attempt to identify the FrameNet frame or the kind of frame element that might be represented by a prepositional phrase. This initial analysis does not attempt to make an assignment of the FrameNet frame element that might be represented by the prepositional phrase.[5] At this time, we examined two questions: (1) Is the lemma and word class combination present in the dictionary, and (2) does the set of frame element realizations for at least one of the senses include a prepositional phrase with the target preposition as its head. We answer these questions for the two main corpora, the SemEval test set and the OEC instances, to determine the differences between them and whether these differences might explain the different accuracies.

For the first question, we found that the lemma and word class combinations of the governor was present in 93.5 percent of the instances for the SemEval set and 78.2 percent for the OEC set. This result is suggestive that the SemEval data might not be representative. For the second question, we obtained 85.6 percent for the SemEval test set and 33.0 percent for the OEC set. These results are quite similar to the overall results shown in Table 1. In Table 2, we list the results by preposition. Although these detailed results do not correspond precisely to the results shown in Table 1, they are generally in the same direction.

These results suggest that the Tratz models are capturing the FrameNet annotations. They also suggest that the lower accuracy for the OEC corpus may indicate that the governors in this corpus have not been characterized as they might be in the FrameNet fashion. More specifically, Tratz' models for individual prepositions seem to be identifying FrameNet targets (i.e., the lexical units providing the trigger for a frame) and a frame

[3] Details of which senses and how many instances for each are available upon request.

[4] https://framenet.icsi.berkeley.edu/fndrupal/. FrameNet is constantly being expanded and updated. We are using data from FrameNet 1.5.

[5] Since the SemEval instances identify the FrameNet sentence identifier, this information can be retrieved. But, this is not the purpose here.

Preposition	FrameNet Realizations SemEval	OEC
about	0.896	0.417
above	0.826	0.150
across	0.881	0.525
after	0.787	0.418
against	0.793	0.203
along	0.762	0.200
among	0.640	0.071
around	0.901	0.255
as	0.890	0.364
at	0.755	0.353
before	1.000	0.295
behind	0.742	0.157
beneath	0.786	0.025
beside	0.759	0.244
between	0.843	0.242
by	0.931	*
down	0.895	0.279
during	0.641	0.100
for	0.805	0.438
from	0.870	0.422
in	0.799	0.446
inside	0.714	0.108
into	0.879	0.620
like	0.744	0.135
of	0.919	0.549
off	0.893	0.208
on	0.855	0.444
onto	0.862	0.467
over	0.786	0.328
round	0.835	0.378
through	0.865	0.519
to	0.879	0.404
towards	0.902	0.262
with	0.848	0.552
Overall	0.856	0.330

Table 2. Proportion of Instances Where Identified Governor is a FrameNet Lexical Unit Having a Frame Element Using the Preposition

element associated with the frame (although not specifically identifying which frame element). The lower results for the OEC corpus do not depend on a FrameNet characterization, but only indicate the absence of sufficient training data.

It is important to recognize that we are not suggesting that identification of a FrameNet realization should be used as a feature in developing preposition models. The coverage of FrameNet would be problematic. It is unclear whether an attempt to map governors without FrameNet entries, such as suggested in Burchardt et al. (2005), would improve the disambiguation results for the OEC corpus. This would assume that the frame element realizations for some hypernym of the governor would be valid. Since it is well-recognized that many verbs and nouns have idiosyncratic prepositions for realizing frame elements, further investigation would be needed to support such an assumption.

In performing these analyses, we encountered a number of cases where a useful governor was not identified. There were 69 instances for the SemEval corpus (less than 1 percent) and 180 instances for the OEC corpus (5 percent). The governors identified in these cases were not content words that would likely have been analyzed in FrameNet, but which indicate other types of preposition behavior. We observed cases where the governor was another preposition (e.g., *until after*, *from among*, or *to above*), a conjunction (e.g., *or above*), verb forms for which the underlying lemma could not be identified, and general referring words (*those* or *anything*) whose semantic content was not identifiable. We did not perform an exhaustive analysis of these cases, but they provide interesting cases for further study.

5.3 Accuracy of Feature-Based Approach

For the most part, methods used for modeling preposition behavior (maximum entropy and support vector machines) depend on having a representative sample of training data from which features can be extracted. Two problems emerge: (1) absence of a representative sample, and (2) behavior that would not be identified in a feature-based approach.

As mentioned above, there were significant disparities in the number of instances for the SemEval and OEC corpora. When we made the predictions for the OEC corpus, we also generated confusion matrices for each preposition. For nearly every preposition, one sense predominated as the prediction, usually with much more than 50 percent of the values. For example, for *above*, 141 of the 173 instances were predicted to be the TPP sense 4(2), with virtually all of the instances in each of the nine OEC senses predicted to have this sense. Some of the more polysemous prepositions (*for*, *from*, *of*, *on*, *over*, and *with*) had more of a spread, but still with dominant predicted senses. Combined with the results of the last section, showing the close ties with FrameNet frame realizations, it is possible that the features for such instances dominate the frequency distributions and the resultant predictions.

Most of the features generated in Tratz' system are properties of individual words, identified with various rules and then characterized principally

with WordNet properties (e.g., synonyms, hypernyms, file number, and gloss words). We have suggested that the overall success of this system is based on the ability to identify the governor and the complement. However, this does not capture some of the lexicographic behavior associated with many senses.

Each preposition is likely to be used in some idiomatic expressions, such as *for example*. In TPP, many of these have been identified in treatments associated with the more common prepositions. Several senses have peculiar characteristics, such as *day after day* (repetition), *at fourteen* (requiring an age), *from ... to ...* (expressing a range), *three into twelve* (expressing division), *to the 4th power* (expressing exponentiation), *five to ten* (expressing time), and *a boy of 15* (expressing age). These are not captured in the feature analysis and should perhaps act as options in a decision tree.

More generally, as Srikumar & Roth (2013) point out, some senses may be disambiguated with the use of ontological knowledge. In their example, the sense of *at* in a phrase *arrive at ...* depends on the object of the preposition, either a location or a time. When the dictionary definitions of prepositions are examined, many of them characterize the object in some way. For example, one sense of *on* is "regularly taking (a drug or medicine)", as in *he is on morphine*. TPP likewise provides characterizations of the objects and the governors. For example, for the sense of *above*, "at a higher volume or pitch than," the object is specified as "the quieter of two noises or sounds" and the governor as "nouns and verbs denoting sound or aural perception."

Tratz' system uses features such as hypernyms, either immediate or extended, from the WordNet hierarchy. Srikumar & Roth do likewise, with a maximum depth of four, as well as adding type categories defined by word-similarity driven clusters. The TPP characterizations do not lend themselves to precise formulations. In general, this is an area for further investigation, particularly since, given the association with FrameNet entries, which are likely to be mostly verbs, where the development of verb classes might prove most useful.

6 Areas for Further Investigation

In the previous sections, we have identified the existence of a continuing problem with preposition disambiguation and we have examined several dimensions as the possible sources of this problem. We have not presented a solution to the problem, but hopefully we have provided a starting point from which further investigation may proceed. At this point, we can only speculate on some areas that might be pursued.

We believe the chief problem is the lack of a representative corpus with which the major methods (maximum entropy and support vector machines) can provide generalizable predictions. Litkowski (2013) identifies a third corpus developed under The Preposition Project, the CPA corpus, drawn from the British National Corpus using the Corpus Pattern Analysis system of Hanks (2004). This corpus is being analyzed using principles described in Hanks (2013), with the intention of developing a pattern dictionary of English prepositions.[6] This corpus consists of over 48,000 sentences for 304 prepositions, including 170 phrasal prepositions. This corpus has been pre-processed with Tratz' system. This data will provide the basis for examining such factors as the extent of coverage of FrameNet triggers among the governors identified by Tratz' system.

Another main problem is the difficulty in providing suitable ontological categories for the governors and complements. We have provided an initial examination of FrameNet and suggested further investigation of the transitivity of frame element realizations. With the prevalence of verb governors, another resource that should be examined is VerbNet (Kipper, et al., (2006)). While the type categories of Srikumar & Roth were applied to modeling semantic relations, they also suggested that they were useful in preposition disambiguation. On the other hand, many entries for the complement or attachment in TPP were annotated as having "no pattern"; such cases may continue to be problematic.

Finally, an attempt should be made to transform the statistical results into lexicographic characterizations. We believe that certain long distance dependencies (i.e., spanning a few words as in idiomatic senses) are not presently captured

[6] See http://www.clres.com/pdep.html and http://www.clres.com/db/TPPEditor.html.

by these models. While these are not likely to have a significant impact on disambiguation results, they may capture some phenomena that do occur with prepositions. In addition, a better characterization of ontological components may assist in refining TPP specifications for the governors and complements that can be used in creating a refined sense inventory, much like what Tratz did in his thesis.

7 Conclusions

In attempting to validate and then exploit (for semantic role labeling) apparent progress in preposition disambiguation, we found that applying a state-of-the-art system to a new tagged corpus achieved an accuracy of only 40 percent, compared to reported levels as high as 88 percent. This result suggested that preposition behavior can also be subject to problems of domain adaptation. We examined the details of the predictions for the new corpus, looking at differences in sense inventories, number of instances in respective test sets, and different ways of making the predictions.

We examined the details of the predictions for the new corpus, looking at differences in sense inventories, number of instances in respective test sets, and different ways of making the predictions. We attempted to understand where the shortfalls occurred, identifying lack of suitable training data for many senses and the apparent non-representativeness of the standard SemEval corpus used to investigate preposition disambiguation. Our results suggest that the SemEval corpus is heavily skewed to the frames that have been analyzed in FrameNet.

We have identified several areas where future investigations might focus. We suggest that these areas need to include the kind of feature analysis used in prior disambiguation studies, but may also require methodological approaches that do not rely as much on frequency-based data and that provide additional lexicographic analysis.

Acknowledgments

We are grateful for comments on a draft of this paper from Orin Hargraves, Stephen Tratz, and Vivek Srikumar. This paper was also submitted to the *Transactions of the Association for Computational Linguistics*. We are grateful for the comments from three anonymous reviewers, who characterized the paper as a work in progress.

References

Aljoscha Burchardt, Katrin Erk, and Anette Frank. 2005. A WordNet Detour to FrameNet. In *Proceedings of the GLDV 2005 Workshop GermaNet II*. Bonn.

Patrick Hanks. 2004. Corpus Pattern Analysis. In *EURALEX Proceedings.* Vol. I, pp. 87-98. Lorient, France: Université de Bretagne-Sud

Patrick Hanks. 2013. *Lexical Analysis: Norms and Exploitations*. MIT Press.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2006. A Large-Scale Extension of VerbNet with Novel Verb Classes. In *Proceedings of EURALEX*. Torino, Italy. 173-184.

Ken Litkowski. 2013. *The Preposition Project Corpora*. Technical Report 13-01. Damascus, MD: CL Research.

Ken Litkowski and Orin Hargraves. 2005. The preposition project. *ACL-SIGSEM Workshop on "The Linguistic Dimensions of Prepositions and Their Use in Computational Linguistic Formalisms and Applications"*, pages 171–179.

Ken Litkowski and Orin Hargraves. 2006. Coverage and Inheritance in The Preposition Project. In: *Proceedings of the Third ACL-SIGSEM Workshop on Prepositions*. Trento, Italy.ACL. 89-94.

Ken Litkowski and Orin Hargraves. 2007. SemEval-2007 Task 06: Word-Sense Disambiguation of Prepositions. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic.

Vivek Srikumar and Dan Roth. 2011. A Joint Model for Extended Semantic Role Labeling. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. ACL, 129-139.

Vivek Srikumar and Dan Roth. 2013. Modeling Semantic Relations Expressed by Prepositions. *Transactions of the Association for Computational Linguistics*, 1.

Stephen Tratz. 2011. *Semantically-Enriched Parsing for Natural Language Understanding.* PhD Thesis, University of Southern California.

Stephen Tratz and Eduard Hovy. 2011. A Fast, Accurate, Non-Projective, Semantically-Enriched Parser. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, Scotland, UK.

Zapirain, B., E. Agirre, L. Marquez, and M. Surdeanu. 2013. Selectional Preferences for Semantic Role Classification. *Computational Linguistics*, 39:3.