SemEval-2007 Task 06: Word-Sense Disambiguation of Prepositions

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

Orin Hargraves
5130 Band Hall Hill Road
Westminster, MD 21158
orinhargraves@googlemail.com

Abstract

The SemEval-2007 task to disambiguate prepositions was designed as a lexical sample task. A set of over 25,000 instances was developed, covering 34 of the most frequent English prepositions, with two-thirds of the instances for training and one-third as the test set. Each instance identified a preposition to be tagged in a full sentence taken from the FrameNet corpus (mostly from the British National Corpus). Definitions from the Oxford Dictionary of English formed the sense inventories. Three teams participated, with all achieving supervised results significantly better than baselines, with a high fine-grained precision of 0.693. This level is somewhat similar to results on lexical sample tasks with open class words, indicating that significant progress has been made. The data generated in the task provides ample opportunitites for further investigations of preposition behavior.

1 Introduction

The SemEval-2007 task to disambiguate prepositions was designed as a lexical sample task to investigate the extent to which an important closed class of words could be disambiguated. In addition, because they are a closed class, with stable senses, the requisite datasets for this task are enduring and can be used as long as the problem of preposition disambiguation remains. The data used in this task was developed in The Preposition Project (TPP, Litkowski & Hargraves (2005) and Litkowski & Hargraves (2006)),[1] with further refinements to fit the requirements of a SemEval task.

[1] http://www.clres.com/prepositions.html.

In the following sections, we first describe the motivations for a preposition disambiguation task. Next, we describe the development of the datasets used for the task, i.e., the instance sets and the sense inventories. We describe how the task was performed and how it was evaluated (essentially using the same scoring methods as previous Senseval lexical sample tasks). We present the results obtained from the participating teams and provide an initial analysis of these results. Finally, we identify several further types of analyses that will provide further insights into the characterization of preposition behavior.

2 Motivation

Prepositions are a closed class, meaning that the number of prepositions remains relatively constant and that their meanings are relatively stable. Despite this, their treatment in computational linguistics has been somewhat limited. In the Penn Treebank, only two types of prepositions are recognized (IN (locative, temporal, and manner) and TO (direction)) (O'Hara, 2005). Prepositions are viewed as function words that occur with high frequency and therefore carry little meaning. A task to disambiguate prepositions would, in the first place, allow this limited treatment to be confronted more fully.

Preposition behavior has been the subject of much research, too voluminous to cite here. Three recent workshops on prepositions have been sponsored by the ACL-SIGSEM: Toulouse in 2003, Colchester in 2005, and Trento in 2006. For the most part, these workshops have focused on individual prepositions, with various investigations of more generalized behavior. The SemEval preposition disambiguation task provides a vehicle to examine whether these behaviors are substantiated with a well-defined set of corpus instances.

Prepositions assume more importance when they

Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 24–29,
Prague, June 2007. ©2007 Association for Computational Linguistics

are considered in relation to verbs. While linguistic theory focuses on subjects and objects as important verb arguments, quite frequently there is an additional oblique argument realized in a prepositional phrase. But with the focus on the verbs, the prepositional phrases do not emerge as having more than incidental importance. However, within frame semantics (Fillmore, 1976), prepositions rise to a greater prominence; frequently, two or three prepositional phrases are identified as constituting frame elements. In addition, frame semantic analyses indicate the possibility of a greater number of prepositional phrases acting as adjuncts (particularly identifying time and location frame elements). While linguistic theories may identify only one or two prepositions associated with an argument of a verb, frame semantic analyses bring in the possibility of a greater variety of prepositions introducing the same type of frame element. The preposition disambiguation task provides an opportunity to examine this type of variation.

The question of prepositional phrase attachment is another important issue. Merlo & Esteve Ferrer (2006) suggest that this problem is a four-way disambiguation task, depending on the properties of nouns and verbs and whether the prepositional phrases are arguments or adjuncts. Their analysis relied on Penn Treebank data. Further insights may be available from the finer-grained data available in the preposition disambiguation task.

Another important thread of investigation concerning preposition behavior is the task of semantic role (and perhaps semantic relation) labeling (Gildea & Jurafsky, 2002). This task has been the subject of a previous Senseval task (Automatic Semantic Role Labeling, Litkowski (2004)) and two shared tasks on semantic role labeling in the Conference on Natural Language Learning (Carreras & Marquez (2004) and Carreras & Marquez (2005)). In addition, three other tasks in SemEval-2007 (semantic relations between nominals, task 4; temporal relation labeling, task 15; and frame semantic structure extraction, task 19) address issues of semantic role labeling. Since a great proportion of these semantic roles are realized in prepositional phrases, this gives greater urgency to understanding preposition behavior.

Despite the predominant view of prepositions as function words carrying little meaning, this view is not borne out in dictionary treatment of their definitions. To all appearances, prepositions exhibit definitional behavior similar to that of open class words. There is a reasonably large number of distinct prepositions and they show a range of polysemous senses. Thus, with a suitable set of instances, they may be amenable to the same types of analyses as open class words.

3 Preparation of Datasets

The development of the datasets for the preposition disambiguation task grew directly out of TPP. This project essentially articulates the corpus selection, the lexicon choice, and the production of the gold standard. The primary objective of TPP is to characterize each of 847 preposition senses for 373 prepositions (including 220 phrasal prepositions with 309 senses)[2] with a semantic role name and the syntactic and semantic properties of its complement and attachment point. The preposition sense inventory is taken from the *Oxford Dictionary of English* (ODE, 2004).[3]

3.1 Corpus Development

For a particular preposition, a set of instances is extracted from the FrameNet database.[4] FrameNet was chosen since it provides well-studied sentences drawn from the British National Corpus (as well as a limited set of sentences from other sources). Since the sentences to be selected for frame analysis were generally chosen for some open class verb or noun, these sentences would be expected to provide no bias with respect to prepositions. In addition, the use of this resource makes available considerable information for each sentence in its identification of

[2]The number of prepositions and the number of senses is not fixed, but has changed during the course of the project, as will become clear.

[3]TPP does not include particle senses of such words as *in* or *over* (or any other particles) used with verbs to make phrasal verbs. In this context, phrasal verbs are to be distinguished from verbs that select a preposition (such as *on* in *rely on*), which may be characterized as a collocation.

[4]http://framenet.icsi.berkeley.edu/

frame elements, their phrase type, and their grammatical function. The FrameNet data was also made accessible in a form (FrameNet Explorer)[5] to facilitate a lexicographer's examination of preposition instances.

Each sentence in the FrameNet data is labeled with a subcorpus name. This name is generally intended only to capture some property of a set of instances. In particular, many of these subcorpus names include a string **pp*prep*** and this identification was used for the selection of instances. Thus, searching the FrameNet corpus for subcorpora labeled **ppof** or **ppafter** would yield sentences containing a prepositional phrase with a desired preposition. This technique was used for many common prepositions, yielding 300 to 4500 instances. The technique was modified for prepositions with fewer instances. Instead, all sentences having a phrase beginning with a desired preposition were selected.

The number of sentences eventually used in the SemEval task is shown in Table 1. More than 25,000 instances for 34 prepositions were tagged in TPP and used for the SemEval-2007 task.

3.2 Lexicon Development

As mentioned above, ODE (and its predecessor, the *New Oxford Dictionary of English* (NODE, 1997)) was used as the sense inventory for the prepositions. ODE is a corpus-based, lexicographically-drawn sense inventory, with a two-level hierarchy, consisting of a set of core senses and a set of subsenses (if any) that are semantically related to the core sense. The full set of information, both printed and in electronic form, containing additional lexicographic information, was made publicly available for TPP, and hence, the SemEval disambiguation task.

The sense inventory was not used as absolute and further information was added during TPP. The lexicographer (Hargraves) was free to add senses, particularly as the corpus evidence provided by the FrameNet data suggested. The process of refining the sense inventory was performed as the lexicographer assigned a sense to each instance. While engaged in this sense assignment, the lexicographer accumulated an understanding of the behavior of the preposition, assigning a name to each sense (characterizing its semantic type), and characterizing the syntactic and semantic properties of the preposition complement and its point of attachment or head. Each sense was also characterized by its syntactic function and its meaning, identifying the relevant paragraph(s) where it is discussed in Quirk et al (1985).

After sense assignments were completed, the set of instances for each preposition was analyzed against the FrameNet database. In particular, the FrameNet frames and frame elements associated with each sense was identified. The set of sentences was provided in SemEval format in an XML file with the preposition tagged as **<head>**, along with an answer key (also identifying the FrameNet frame and frame element). Finally, using the FrameNet frame and frame element of the tagged instances, syntactic alternation patterns (other syntactic forms in which the semantic role may be realized) are provided for each FrameNet target word for each sense.

All of the above information was combined into a preposition database.[6] For SemEval-2007, entries for the target prepositions were combined into an XML file as the "Definitions" to be used as the sense inventory, where each sense was given a unique identifier. All prepositions for which a set of instances had been analyzed in TPP were included. These 34 prepositions are shown in Table 1 (*below*, *beyond*, and *near* were used in the trial set).

3.3 Gold Standard Production

Unlike previous Senseval lexical sample tasks, tagging was not performed as a separate step. Rather, sense tagging was completed as an integral part of TPP. Funding was unavailable to perform additional tagging with other lexicographers and the appropriate interannotator agreement studies have not yet been completed. At this time, only qualitative assessments of the tagging can be given.

As indicated, the sense inventory for each preposition evolved as the lexicographer examined

[5]Available for the Windows operating system at http://www.clres.com for those with access to the FrameNet data.

[6]The full database is viewable in the Online TPP (http://www.clres.com/cgi-bin/onlineTPP/find_prep.cgi).

the set of FrameNet instances. Multiple sources (such as Quirk et al.) and lexicographic experience were important components of the sense tagging. The tagging was performed without any deadlines and with full adherence to standard lexicographic principles. Importantly, the availability of the FrameNet corpora facilitated the sense assignment, since many similar instances were frequently contiguous in the instance set (e.g., associated with the same target word and frame).

Another important factor suggesting higher quality in the sense assignment is the quality of the sense inventory. Unlike previous Senseval lexical sample tasks, the sense inventory was developed using lexicographic principles and was quite stable. In arriving at the sense inventory, the lexicographer was able to compare ODE with its predecessor NODE, noting in most cases that the senses had not changed or had changed in only minor ways.

Finally, the lexicographer had little difficulty in making sense assignments. The sense distinctions were well enough drawn that there was relatively little ambiguity given a sentence context. The lexicographer was not constrained to selecting one sense, but could tag a preposition with multiple senses as deemed necessary. Out of 25,000 instances, only 350 instances received multiple senses.

4 Task Organization and Evaluation

The organization followed standard SemEval (Senseval) procedures. The data were prepared in XML, using Senseval DTDs. That is, each instance was labeled with an instance identifier as an XML attribute. Within the **<instance>** tag, the FrameNet sentence was labeled as the **<context>** and included one item, the target preposition, in the **<head>** tag. The FrameNet sentence identifier was used as the instance identifier, enabling participants to make use of other FrameNet data. Unlike lexical sample tasks for open class words, only one sentence was provided as the context. Although no examination of whether this is sufficient context for prepositions, it seems likely that all information necessary for preposition disambiguation is contained in the local context.

A trial set of three prepositions was provided (the three smallest instance sets that had been developed). For each of the remaining 34 prepositions, the data was split in a ratio of two to one between training and test data. The training data included the sense identifier. Table 1 shows the total number of instances for each preposition, along with the number in the training and the test sets.

Answers were submitted in the standard Senseval format, consisting of the lexical item name, the instance identifier, the system sense assignments, and optional comments. Although participants were not restricted to selecting only one sense, all did so and did not provide either multiple senses or weighting of different senses. Because of this, a simple Perl script was used to score the results, giving precision, recall, and F-score.[7] The answers were also scored using the standard Senseval scoring program, which records a result for "attempted" rather than F-score, with precision interpreted as percent of attempted instances that are correct and recall as percent of total instances that are correct.[8] Table 1 reports the standard SemEval recall, while Tables 2 and 3 use the standard notions of precision and recall.

5 Results

Tables 2 and 3 present the overall fine-grained and coarse-grained results, respectively, for the three participating teams (University of Melbourne, Koç University, and Instituto Trentino di Cultura, IRST). The tables show the team designator, and the results over all prepositions, giving the precision, the recall, and the F-score. The table also shows the results for two baselines. The **FirstSense** baseline selects the first sense of each preposition as the answer (under the assumption that the senses are organized somewhat according to prominence). The **FreqSense** baseline selects the most frequent sense from the training set. Table 1 shows the fine-grained recall scores for each team for each preposition. Table 1 also shows the entropy and perplexity for each preposition, based on the data from the training sets.

[7] Precision is the percent of total correct instances and recall is the percent of instances attempted, so that an F-score can be computed.

[8] The standard SemEval (Senseval) scoring program, **scorer2**, does not work to compute a coarse-grained score for the preposition instances, since senses are numbers such as "4(2a)" and not alphabetic.

Table 2. Fine-Grained Scores (All Prepositions - 8096 Instances)			
Team	**Prec**	**Rec**	**F**
MELB-YB	0.693	1.000	0.818
KU	0.547	1.000	0.707
IRST-BP	0.496	0.864	0.630
FirstSense	0.289	1.000	0.449
FreqSense	0.396	1.000	0.568

Table 3. Coarse-Grained Scores (All Prepositions - 8096 Instances)			
Team	**Prec**	**Rec**	**F**
MELB-YB	0.755	1.000	0.861
KU	0.642	1.000	0.782
IRST-BP	0.610	0.864	0.715
FirstSense	0.441	1.000	0.612
FreqSense	0.480	1.000	0.649

As can be seen, all participating teams performed significantly better than the baselines. Additional improvements occurred at the coarse grain, although the differences are not dramatically higher.

All participating teams used supervised systems, using the training data for their submissions. The University of Melbourne used a maximum entropy system using a wide variety of syntactic and semantic features. Koç University used a statistical language model (based on Google ngram data) to measure the likelihood of various substitutes for various senses. IRST-BP used Chain Clarifying Relationships, in which contextual lexical and syntactic features of representative contexts are used for learning sense discriminative patterns. Further details on their methods are available in their respective papers.

6 Discussion

Examination of the detailed results by preposition in Table 1 shows that performance is inversely related to polysemy. The greater number of senses leads to reduced performance. The first sense heuristic has a correlation of -0.64; the most frequent sense heuristic has a correlation of -0.67. the correlations for MELB, KU, and IRST are -0.40, -0.70, and -0.56, respectively. The scores are also negatively correlated with the number of test instances. The correlations are -0.34 and -0.44 for the first sense and the most frequent sense heuristics. For the systems, the scores are -0.17, -0.48, and -0.39 for Melb, KU, and IRST.

The scores for each preposition are strongly negatively correlated with entropy and perplexity, as frequently observed in lexical sample disambiguation. For MELB-YB and IRST-BP, the correlation with entropy is about -0.67, while for KU, the correlation is -0.885. For perplexity, the correlation is -0.55 for MELB-YB, -0.62 for IRST-ESP , and -0.82 for KU.

More detailed analysis is required to examine the performance for each preposition, particularly for the most frequent prepositions (*of*, *in*, *from*, *with*, *to*, *for*, *on*, *at*, *into*, and *by*). Performance on these prepositions ranged from fairly good to mediocre to relatively poor. In addition, a comparison of the various attributes of the TPP sense information with the different performances might be fruitful. Little of this information was used by the various systems.

7 Conclusions

The SemEval-2007 preposition disambiguation task can be considered successful, with results that can be exploited in general NLP tasks. In addition, the task has generated considerable information for further examination of preposition behavior.

References

Xavier Carreras and Lluis Marquez. 2004. Introduction to the CoNLL-2004 Shared Task: Semantic Role Labeling. In: *Proceedings of CoNLL-2004*.

Xavier Carreras and Lluis Marquez. 2005. Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling. In: *Proceedings of CoNLL-2005*.

Charles Fillmore. 1976. Frame Semantics and the Nature of Language. *Annals of the New York Academy of Sciences*, 280: 20-32.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic Labeling of Semantic Roles. *Computational Linguistics, 28* (3), 245-288.

Kenneth C. Litkowski. 2004. Senseval-3 Task: Automatic Labeling of Semantic Roles. In *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*. ACL. 9-12.

Kenneth C. Litkowski & Orin Hargraves. 2005. The Preposition Project. In: *ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistic Formalisms*

and Applications, University of Essex - Colchester, United Kingdom. 171-179.

Kenneth C. Litkowski.& 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.

Paola Merlo and Eva Esteve Ferrer. 2006. The Notion of Argument in Prepositional Phrase Attachment. *Computational Linguistics, 32* (3), 341-377.

The New Oxford Dictionary of English. 1998. (J. Pearsall, Ed.). Oxford: Clarendon Press.

Thomas P. O'Hara. 2005. Empirical Acquisition of Conceptual Distinctions via Dictionary Definitions. Ph.D. Thesis. New Mexico State .

The Oxford Dictionary of English. 2003. (A. Stevension and C. Soanes, Eds.). Oxford: Clarendon Press.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, & Jan Svartik. (1985). *A comprehensive grammar of the English language.* London: Longman.

Table 1. SemEval-2007 Preposition Disambiguation

				Number of Instances			Fine-Grained Recall				
							Participating Teams			Baselines	
Prepostition	Senses	Ent	Perp	Total	Training	Test	Melb	KU	IRST	First Sense	Freq Sense
about	6	0.63	1.54	1074	710	364	0.885	0.934	0.780	0.885	0.885
above	9	1.80	3.49	71	48	23	0.652	0.522	0.565	0.043	0.609
across	3	0.23	1.17	470	319	151	0.960	0.960	0.914	0.960	0.960
after	11	2.15	4.44	156	103	53	0.472	0.585	0.585	0.434	0.434
against	10	1.89	3.69	287	195	92	0.880	0.793	0.826	0.446	0.435
along	4	0.30	1.23	538	365	173	0.954	0.954	0.936	0.954	0.954
among	4	1.55	2.93	150	100	50	0.660	0.680	0.620	0.300	0.300
around	6	2.05	4.13	490	335	155	0.561	0.535	0.381	0.155	0.452
as	2	0.00	1.00	258	174	84	1.000	1.000	0.988	1.000	1.000
at	12	2.38	5.21	1082	715	367	0.790	0.662	0.646	0.425	0.425
before	4	1.33	2.51	67	47	20	0.600	0.850	0.800	0.450	0.450
behind	9	1.31	2.47	206	138	68	0.662	0.676	0.471	0.662	0.662
beneath	6	1.22	2.33	85	57	28	0.714	0.679	0.750	0.571	0.571
beside	3	0.00	1.00	91	62	29	1.000	1.000	1.000	1.000	1.000
between	9	2.11	4.31	313	211	102	0.814	0.765	0.892	0.422	0.422
by	22	2.53	5.77	758	510	248	0.730	0.556	0.391	0.000	0.371
down	5	1.18	2.26	485	332	153	0.654	0.647	0.680	0.438	0.438
during	2	1.00	2.00	120	81	39	0.769	0.564	0.667	0.615	0.385
for	15	2.84	7.17	1429	951	478	0.573	0.395	0.456	0.036	0.238
from	16	2.85	7.21	1784	1206	578	0.642	0.415	0.512	0.279	0.279
in	15	2.81	7.01	2085	1397	688	0.561	0.436	0.494	0.362	0.362
inside	5	1.63	3.10	105	67	38	0.579	0.579	0.605	0.368	0.526
into	10	2.14	4.41	901	604	297	0.616	0.539	0.586	0.290	0.451
like	7	1.26	2.40	391	266	125	0.856	0.808	0.592	0.120	0.768
of	20	3.14	8.80	4482	3004	1478	0.681	0.374	0.144	0.000	0.205
off	7	1.16	2.23	237	161	76	0.658	0.776	0.408	0.171	0.763
on	25	3.42	10.68	1313	872	441	0.624	0.469	0.351	0.218	0.206
onto	3	0.60	1.52	175	117	58	0.879	0.879	0.776	0.879	0.879
over	17	2.52	5.73	298	200	98	0.510	0.510	0.480	0.010	0.327
round	8	2.31	4.95	263	181	82	0.610	0.512	0.000	0.037	0.378
through	16	2.71	6.54	649	441	208	0.524	0.538	0.481	0.322	0.495
to	17	2.43	5.38	1755	1183	572	0.745	0.579	0.558	0.322	0.322
towards	6	0.71	1.63	316	214	102	0.931	0.873	0.833	0.873	0.873
with	18	3.05	8.27	1769	1191	578	0.699	0.455	0.635	0.149	0.249
Total	332			24653	16557	8096	0.693	0.547	0.496	0.289	0.396