The Preposition Project Corpora

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

Abstract

Three corpora have been developed under The Preposition Project and are available for download and use in characterizing preposition behavior. The SemEval corpus was developed for use in the SemEval 2007 task on preposition disambiguation and was drawn from FrameNet. The OEC corpus was drawn from the preposition portion of the Oxford sentence dictionary, used as example sentences for preposition definitions and based on the Oxford English Corpus. The CPA corpus was drawn from the British National Corpus following procedures used in Corpus Pattern Analysis and is intended to provide a representative sample of actual preposition usage. We describe how each of these corpora were drawn.

1 Introduction

The analysis of preposition behavior requires an understanding of a considerable number of interrelated factors. While previous work on preposition disambiguation has uncovered many of these factors, many additional needs have become more obvious. We describe three corpora that have been developed under The Preposition Project (TPP) to provide background for their use in analysis of preposition behavior.

TPP was designed to provide a well-defined framework for examining preposition behavior (Litkowski & Hargraves (2005); Litkowski & Hargraves (2006)). Recognizing that one of the most important lessons of word-sense disambiguation (WSD) studies is the need for well-defined sense inventory, we were able to obtain data from the Oxford Dictionary of English (ODE; Stevenson & Soanes (2003)) for use in TPP. In addition to an appropriate sense inventory, WSD also requires a set of instances tagged with senses from the sense inventory.

We describe three preposition corpora available for download[1]: (1) the training and test sets used in the SemEval-2007 task on preposition disambiguation, drawn from FrameNet (FN), (2) a set of sentences from the Oxford English Corpus (OEC) as examples for senses in the Oxford Dictionary of English (ODE), and (3) a set of sentences from the written portion of the British National Corpus, drawn with methodology used in the Corpus Pattern Analysis project (CPA). The first corpus covers 34 prepositions, while the latter two include all single-word prepositions and many phrasal prepositions. Each corpus consists of sentences following the SemEval format. In addition, each sentence has been lemmatized, part-of-speech tagged, and parsed with a dependency parser, in the CoNLL-X format.

In section 2, we describe the format used for all corpora. In section 3, we describe how the sentences were parsed. In section 4, we describe the SemEval corpus. In Section 5, we describe the OEC corpus. In section 6, we describe the CPA corpus.

2 The SemEval Format

The format for each corpus follows the standard lexical sample format used in Senseval and SemEval, i.e., when the objective is to disambiguate individual words. In each of the corpora, each preposition has its own XML file (e.g., underneath.xml). Each file contains a number of instances, as shown in Figure 1 (only one instance is shown in the example). The first line identifies the lexical item and its part of speech (always "prep" in these corpora). Each instance is given an identifying number and a document source. For the three

[1] These corpora are available at CL Research (http://www.clres.com) by following the links.

```
<lexelt item="underneath" pos="prep">
  <instance id="underneath.p.fn.635810" docsrc="FN">
    <answer instance="underneath.p.fn.635810" senseid="1(1)"/>
    <context>
      He always used to tuck it <head>underneath</head> the water butt .
    </context>
  </instance>
</lexelt>
```

Figure 1. Example of a Senseval/SemEval lexical sample instance.

corpora available from TPP, these are FN (FrameNet), OEC (Oxford English Corpus), and CPA (Corpus Pattern Analysis). The next line gives the answer for the instance, identifying the instance number and the TPP sense identifier. These answers are given for the SemEval and OEC corpora, but not the CPA corpus (since it has not been sense-tagged). For the test portion of the SemEval corpus, these answers are provided in an accompanying answer key file, with a KEY extension, containing the instance and sense identifiers. The next line gives the sentence (the context), with the target preposition surrounded by a "head" tag. Each sentence has been tokenized using TreeTagger,[2] that is, separated into space-separated strings, so that, for example, an apostrophe and the letter s forms a possessive token ('s) and the terminal period is separated from the preceding word.

3 Parsing the Corpora

The tokenized sentences of each corpus have been further processed with a lemmatizer, part-of-speech tagger, and dependency parser, using an updated version of the system described in Tratz & Hovy (2011).[3] These parses are provided in an expanded CoNLL-X format.[4] The Tratz system includes a module specifically designed to process files that use the lexical sample format shown in Figure 1.

The Tratz system creates 14 tab-separated items, compared with 10 items in the original CoNLL-X format. However, in producing these files, only 6 items are included: (1) the token counter (item 1), (2) the word form (item 2), (3) the lemma (item 3), (4) the fine-grained part of speech tag (item 5), (5), the head of the current token, i.e., the token number of its head or 0 for the ROOT of the sentence (item 7), and (6) the dependency relation of each item to its head (item 8). A key to the dependency relations as described in Tratz & Hovy (2011) is included in his distribution as well as the TPP corpora distribution.

While the Tratz system can be used to generate values for other fields, including preposition disambiguation, noun-noun compound relations, possessive relations, predicate disambiguation, and semantic role labeling, these were not created for this distribution, since these are values that may properly be the subject of investigation using these corpora.

An important note for the parsed files is that they follow the CoNLL-X format of separating sentences with a blank line. There is no identifier associated with these files. The order of the parsed files follows the order of the instances in the source files.

4 The SemEval Corpus

The SemEval corpus was taken from TPP sense-tagged instances drawn from FrameNet.[5] This is described fully in Litkowski & Hargraves (2005). Basically, the procedure involved named subcorpora used in developing the analysis for a frame. Each subcorpus is given a name which encodes syntactic properties of the subcorpus, e.g., **V-730-s20-ppacross**. TPP used these identifiers to select sentences from FrameNet, since they included instances of individual prepositions. The set of instances was then subjected to sense-tagging using the ODE inventory. Full details of this process are given at the TPP web site.[6]

Although sense tagging was performed for 57 prepositions in TPP, the SemEval corpus contains

[2] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
[3] Available at http://sourceforge.net/projects/miacp/
[4] Described in detail at http://ilk.uvt.nl/conll/,
[5] https://framenet.icsi.berkeley.edu/fndrupal/
[6] http://www.clres.com/prepositions.html

instances for only 37 of those, selected to provide sufficient sample sizes to use in SemEval. This corpus is divided into three parts: a trial set, a training set, and a test set. The trial set contains 151 instances for three prepositions (below, beyond, and near) and was intended only for use to participants in developing their systems. The training and test sets contain 24,782 instances for the remaining 34 prepositions, randomly partitioned into two-thirds of the instances for training (16,535 sentences) and one-third for the final test (8,096 sentences). The test sets do not contain the answer line shown above , but there is an accompanying answer key file for each preposition.

5 The OEC Corpus

The Oxford English Corpus was developed by Oxford University Press for use in the development of its dictionaries.[7] From this corpus, Oxford lexicographers selected a set of sentences to illustrate each sense in various dictionaries. The objective was to create a 'sentence dictionary' underlying the dictionaries, with up to 20 sentences for each sense.[8]

TPP used the sentences for preposition entries, including those directly labeled as prepositions and phrases judged to be multiword expressions displaying preposition behavior. We drew our corpus during the developmental cycle of the sentence dictionary. We used the TPP entry words to bring up and save the HTML file for each entry. We made some modifications to the files to take into account some idiosyncrasies of how the sentences for single-word prepositions and phrasal prepositions were included in these files. Many prepositions are also used as adverbs, so it was necessary to remove sentences illustrating adverb senses. The entries for many phrasal prepositions were presented in variant forms. For example, the file for **in (or out of) keeping with** included sentences showing both variants. We duplicated such files and then deleted sentences not containing the variant we wanted. After this preprocessing, we then used a script to extract the sentences for each entry, subject the sentences to TreeTagger to ensure that the sentences would be tokenized, and finally convert them to SemEval format.

This corpus contains 7,650 sentences covering 635 senses for 259 prepositions. This corpus was intended to provide a suitable set of instances for those prepositions not covered in the SemEval corpus. These files all follow the format shown above, providing the TPP sense identifier for each preposition. There is no split into training and test sets. Since the OEC-based sentence dictionary was not designed to be representative, this corpus does not provide any indication of the relative frequency of preposition sense usage, usually considered to be essential for the kinds of methods employed in WSD studies.

6 The CPA Corpus

The CPA corpus of prepositions was created following the principles developed for the Corpus Pattern Analysis for verbs. Hanks (2004a) describes methods for a Corpus Pattern Analysis of verbs, based on the Theory of Norms and Exploitations (Hanks (2004b) and Hanks (2013)). [9] The basic technique developed in CPA is to draw a sample of sentences for a verb under analysis, using the Word Sketch Engine (WSE),[10] and then to develop patterns of usage that exhaustively covers all the instances.

In general, an initial sample of 250 instances is drawn for each verb from the written portion of the British National Corpus. When there are no more than 250 instances, all instances are analyzed. When there are more than 250 instances, the corpus is sampled, selecting 250 instances. If the behavior for a particular verb is especially complex, requiring the development of a large number of patterns, the sample size is increased, maintaining the requirement that all instances be covered by a pattern.

We developed the CPA corpus with the idea of performing a corpus pattern analysis of prepositions (Litkowski, 2012). We have followed the principle of selecting 250 instances for each preposition. To do this, we used WSE to select the instances, specifying the option to obtain full

[7] http://oxforddictionaries.com/us/words/the-oxford-english-corpus

[8] http://oxforddictionaries.com/us/words/example-sentences

[9] CPA was developed as a pilot project, designed to develop a Pattern Dictionary of English. A description of this project can be found at http://nlp.fi.muni.cz/projects/cpa/. The initial patterns can be viewed at http://deb.fi.muni.cz/pdev/. CPA is now being followed by a full project, known as Disambiguation of Verbs by Collocation, at http://clg.wlv.ac.uk/projects/DVC/.

[10] http://www.sketchengine.co.uk/

sentences, rather than the keyword-in-context format. For single-word prepositions, we used the WSE specification to search for instances tagged as prepositions. For phrasal prepositions, we used the WSE technique for looking for phrases, rather than specifying a part of speech. After submitting the query, we saved the results in a text file. The results include header information showing the number of hits in the full corpus, the query, whether we had a random sample (if the number of hits was greater than 250), and two lines for each instance. The first line contained all the text up to and including the target preposition or phrase; the second line contained the remaining text of the sentence.

We next converted the text files for each preposition into XML files adhering to the SemEval format. In this conversion, we first subjected each sentence to TreeTagger to ensure that the sentence would be fully tokenized.

The CPA corpus consists of 48,105 sentences for 271 single-word and phrasal prepositions, with 250 or more instances for 140 of these items. In most cases, we have 250 instances when the total number of hits was greater than 250. For prepositions likely to have a large number of senses, we drew larger samples. We have 500 sentences for nine prepositions: *at, by, for, from, in, into, like, over,* and *through*. We have 750 sentences for three prepositions: *of, on, to,* and *with*.

This corpus is not tagged. In other words, it conforms to the SemEval format shown in Figure 1, but does not include an instance line. This corpus is intended for further examination of preposition behavior. We believe this corpus is representative, unlike the SemEval and OEC corpora.

7 Summary

We have described three corpora available under The Preposition Project. Combined, these corpora contain 80,537 sentences. The raw sentences are provided in SemEval format. The parsed sentences are provided in CoNLL-X format.

References

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

Patrick Hanks. 2004b. The Syntagmatics of Metaphor and Idioms. *International Journal of Lexicography*, 17(3):245-74.

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

Ken Litkowski. 2013. *Corpus Pattern Analysis of Prepositions*. Technical Report 12-02. 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.

Angus Stevenson and Catherine Soanes (Eds.). 2003. *The Oxford Dictionary of English*. Oxford: Clarendon Press.

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.