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

Datasets and Formats

The set of prepositions included in the SemEval-2007 task is drawn from The Preposition Project (TPP). The prepositions include the 17 most common prepositions, as well as 20 additional very common prepositions, as shown in the accompanying table. This table shows the number of senses for each preposition and the total number of instances that have been tagged (disambiguated) in The Preposition Project. All these prepositions, except for three prepositions used for the trial set, are included in the SemEval-2007 test set, with approximately two-thirds of the instances used for training and the remaining one-third as the test set. The training and test sets were generated randomly from the available instances.

Each preposition included in the task will have a training set and a test set. Each instance is a sentence drawn from the FrameNet sentences and is included exactly as transcribed there, except that the target preposition to be disambiguated is tagged with a head element. The set of instances for each preposition will be contained in an XML-valid file, with the top element lexelt (e.g., <lexelt item="for.p">). Each instance will consist of a sentence identified as the context; the training set will also contain an answer element, as in the following example.

```
<instance id="for.p.fn.331141" docsrc="FN">
        <answer instance="for.p.fn.331141" senseid="2(2)"/>
        <context>
                He used to work as a policeman and , on the few occasions when he was in
                extreme danger , automatically did whatever had to be done without any
                thought <head>for</head> his own safety .
        </context>
</instance>
```

The sense identifier is a number assigned in The Preposition Project using the sense inventory from the Oxford Dictionary of English (ODE). The full sense inventory for each preposition will be provided and will include a mnemonic semantic relation/role name, and properties of the preposition complement and the head or attachment point of the preposition. The answer key for the training set will include not only the sense identifier, but also a FrameNet frame and frame element name, as assigned by the FrameNet lexicographers. Participants may use the FrameNet sentence identifier to make use of other information generated by the FrameNet lexicographers. (Note that in the sense descriptions in the dictionary files, the numbers in parentheses provide an indication of the granularity of the senses. If a sense number in parentheses contains a letter or the full sense number is followed by a dash and a number, e.g., 3(2a)-1, the sense is a fine-grained sense. Results will be scored by both coarse and fine grains. Surprisingly, only in rare instances did the lexicographer assign multiple senses to instances.)

A trial dataset has been prepared for the instances and senses of the prepositions **below**, **beyond**, and **near**. This dataset consists of sense inventories (*.defs.xml), instance files (*.sents.xml), and

answer keys (*.sents,key). The sense inventories were generated from data available in the The Preposition Project. The help file from the Online TPP is included in the trial dataset.

The sense inventories include some extraneous information from the data files used in TPP. Essentially, the relevant information for each sense is contained in **S** (sections) tags; each sense has a sense identifier (**senseid**). ODE data are contained primarily is definition fields (**df**), examples (**ex** and **exg**), and grammar groups (**gg)**. TPP generated data are contained in the following fields (which are described in detail in the accompanying help file): **srtype** (semantic relation or role name), **qsyn** (Quirk syntax), **qpar** (Quirk paragraph), **cprop** (complement properties), **aprop** (attachment properties), **frfes** (Frame::Element pairs), **opreps** (other prepositions - short), **fepreps** (other prepositions - long), **srel** (sense relations), and **com** (comments). Some of these fields may be empty for some senses; minimally, there will be information in **srtype**, **cprop**, and **aprop** fields. **An important consideration for the SemEval task is that senses in the sense inventory that do not have an accompanying set of *frfes* will have no instances in the sentences.** In many cases, but not all, the TPP lexicographer provided a *treatment* of a preposition's senses (including **near** and **beyond**, but not **below**); these treatments are included in the trial dataset. In the case of more common prepositions, these treatments may be somewhat elaborate.

Submissions

Results should be submitted in tar or zip files using a short identifier for the name of the participating team. Individual answer files should be submitted for each preposition that the participating team wishes to analyze, i.e., for **below**, there should be a file **below.p.results**. Each preposition file should contain one line for each instance that is attempted. Each line should appear as follows:

below.p below.p.fn.194575 3(1b)

where the items are separated by white space (e.g., either a space or a tab), consisting of the preposition name (the preposition followed by "p"), the instance as given in the instance id, and the sense(s), using the identifier(s) as identified in the preposition's dictionary. More than one answer is permitted for a given instance, although higher scores will be achieved when only an exact sense is identified.