English Coarse-Grained All-Words Task for SemEval

Roberto Navigli and Ken Litkowski

January 19, 2007

1 Introduction

This document contains the information about the adopted sense inventory, the format of the trial (and test) datasets and the answer files for the ENGLISH COARSE-GRAINED ALL-WORDS task at SEMEVAL.

2 Sense Inventory

The adopted sense inventory for the ENGLISH COARSE-GRAINED ALL-WORDS task is a coarse-grained version of WordNet 2.1. Sense clusters are created based on the manual validation by expert lexicographers of automatically-acquired groupings of WordNet senses. Clusters are provided in a separate file, which contains a sense cluster on each line. For instance, noun `spirit` has 8 senses in WordNet, from which we created 3 groups of senses. As a result, the file will include three lines, one for each cluster:

```
spirit%1:18:01:: spirit%1:18:00::
spirit%1:26:00:: spirit%1:07:00:: spirit%1:26:01:: spirit%1:07:02:: spirit%1:07:03::
spirit%1:10:00::
```

WordNet senses in a cluster are represented in the WordNet sense key format and are separated by spaces.

3 Format

3.1 Trial and Test Sets

The file input to systems for disambiguation will adhere to the following format:

```
<?xml version="1.0" encoding="iso-8859-1" ?>
<!DOCTYPE corpus SYSTEM "coarse-all-words.dtd">
<corpus lang="en">
<text id="d000">
<sentence id="d000.s001">
There
<instance id="d000.s001.t001" lemma="be" pos="v">was</instance>
a
<instance id="d000.s001.t002" lemma="steaming" pos="a">steaming</instance>
.
.
.

</sentence>
</text>
.
.
.

<text id="d002">
.
.
.

</text>
</corpus>
```

where each <text> tag specifies a text source whose identifier is provided by the id attribute. <sentence> tags represent single sentences within each text (again identified with an id attribute). Each <sentence> tag contains zero, one or more target words, each tagged with an <instance> element. Each instance specifies its unique identifier (id), lemma (lemma) and part of speech tag (pos). The latter can assume the values n, v, a and r for nouns, verbs, adjectives and adverbs, respectively. Instances are assumed to have an appropriate sense in the adopted sense inventory. Content words with no corresponding sense in the inventory will not be tagged.

3.2 Answer File

Systems can provide a single sense label for each instance in the test set. The format follows that of the previous SENSEVAL evaluation exercises. Systems can provide any sense in a cluster to assign the appropriate coarse sense. This will allow both systems tuned for fine-grained senses and systems exploiting the knowledge of sense groupings to participate in the evaluation exercise.