EVALITA 2007

http://evalita.itc.it

All-Words Word Sense Disambiguation Task
Guidelines for Participants

Nicoletta Calzolari, Francesca Bertagna
Istituto di Linguistica Computazionale, Consiglio Nazionale delle Ricerche,
Via Moruzzi 1, 56100 Pisa
glottolo@ilc.cnr.it, francesca.bertagna@ilc.cnr.it

Task Description

In the all-words WSD task, systems must tag almost all of the content words in a sample of Italian corpus. Participants to the "all-words" task will be provided with a test corpus of about 5000 words (13600 tokens) where content words (nouns, verbs, adjectives and a small set of proper nouns) are semantically tagged according to the sense inventory of ItalWordNet lexicon (Roventini et al., 2003).

The Corpus (ISST)

The Italian all-words corpus consists of about 13600 word tokens, extracted from the SI-TAL [1], Italian Syntactic Semantic Treebank (ISST). The ISST (Montemagni *et al*. 2003) consists of i) a generic corpus of about 215,000 tokens, extracted from different periodicals and newspaper articles (*La Repubblica*, *Il Corriere della Sera*) and ii) a specialised corpus of about 90,000 tokens, with texts belonging to the financial domain (*Il Sole-24Ore*). The annotated corpus consists of about 5000 words and comprises a selection of Italian newspaper articles about various topics: politics, sport, news, etc. The common data format is XML.
The reference lexical resource used for the Senseval-3 sense tagging task is the lexical-semantic database IWN, developed within the framework of two different research projects: EuroWordNet (Vossen 1999) and SI-TAL, during which IWN was extended by the insertion of adjectives, adverbs and a subset of proper nouns. The IWN database

[1] SI-TAL (Integrated System for the Automatic treatment of Language) was a National Project devoted to the creation of large linguistic resources and software tools for Italian written and spoken language processing.

contains about 64,000 word senses corresponding to about 50,000 synsets.

Data Format

The DTD of the corpus and examples of the format can be download at:
http://evalita.itc.it/tasks/wsd_allWords_sample.tar.gz

The correspondence between the format of annotation and the ItalWordNet sense is:

Annotation:

```
<head id="els.morphxx.mw_xxx">noun</head>
<answer head="els.morphxx.mw_xxx" senseid="noun.S.1"/>
```

Example of the correspondent ItalWordNet entry:

```
<WORD_MEANING ID="N#13740" PART_OF_SPEECH="N">
<GLOSS></GLOSS>
<VARIANTS>
<LITERAL LEMMA="noun" SENSE="1"/>
</VARIANTS><INTERNAL_LINKS>
<RELATION TYPE="has_hyperonym" ...>
<TARGET_WM ID="5515" PART_OF_SPEECH="N" LEMMA="noun" SENSE="2"/>
</RELATION>
</INTERNAL_LINKS>
<EQ_LINKS>
<RELATION TYPE="eq_synonym" ID="1" INV_ID="1">
<TARGET_WM ID="r#9114119"/>
</RELATION>
</EQ_LINKS>
</WORD_MEANING>
```

The correpondence is:
"sense id" = [LITERAL LEMMA value + WORD_MEANING PART_OF_SPEECH value + LITERAL SENSE value]

in the corpus the POS tag for nouns is "S" while in ItalWordNet is "N";
in the corpus the POS tag for proper nouns is "SP" while in ItalWordNet is "NP";
in the corpus the POS tag for adjectives is "A" while in ItalWordNet is "AG";

The Italian all-words task also deals with a group of multiword espressions: verb phrases, adjectival phrases and noun phrases (for the representation of multiword expressions, see the All-Word DTD).

Not all the content words are tagged with a correspondent IWN sense, since they are not present in the reference resource. These are in general (i) terms with not common meanings, defined by traditional dictionaries as technical-specialistic terms, (ii) foreign words, or (iii) metaphorical uses.

During the annotation, in case of uncertainty, multiple senses have been assigned to about 90 lemmas that appeared difficult to disambiguate. In this case, the format is:

```
<head id="els.morph032.mw_xxx">verb</head>
<answer head="els.morph032.mw_xxx" senseid="verb.V.1&#38;4"/>
```

which means that the verb "verb" has been tagged with the ItalWordNet entry verb.V.1 e verb.V.4

A sensemap file (file ITAllWords_SenseMap.txt), with some words that present two or more overlapping senses, is provided to allow for the coarse-grained scoring.

Submission of system results:

- Deadline: June 1st, 2007, midnight (local time)
- System results have to be sent by e-mail to Francesca Bertagna (francesca.bertagna@ilc.cnr.it)
- Each participant has the possibility to submit a maximum of two runs (one performed by taking into account the fine-grained sense organization and one performed with the coarse-grained organization).
- System results will consist of a single data file and will have to be named as follows: evalita07_WSD_participant_run

Evaluation Metrics

Results will be evaluated by taking into account te following measures: Precision, Recall, F-Measure.

Scorer

The scorer of the Senseval-3 will be used for the analysis of the results. The scorer, with examples of format and other information can be freely downloaded from: at http://www.senseval.org/senseval3/scoring.

Contact Person

Francesca Bertagna (francesca.bertagna@ilc.cnr.it)

References

Montemagni Simonetta, Barsotti Francesco, Battista Marco Calzolari Nicoletta, Corazzari Ornella, Lenci Alessandro, Pirrelli Vito, Zampolli Antonio, Fanciulli Francesca; Massetani Maria, Raffaelli Remo, Basili Roberto, Pazienza Maria Teresa, Saracino Dario, Zanzotto Fabio, Mana Nadia, Pianesi Fabio, Delmonte Rodolfo. 2003. The syntactic-semantic *Treebank* of Italian. An Overview. *Linguistica Computazionale a Pisa vol*. I, pag.461-492.

Roventini A., Alonge A., Bertagna F., Calzolari N., Girardi C., Magnini B., Marinelli R., Speranza M., Zampolli A. (2003), *ItalWordNet*: *Building a Large Semantic Database for the Automatic Treatment of Italian,* in 'Linguistica Computazionale', Istituti Editoriali e Poligrafici Internazionali, Pisa-Roma, ISSN 0392-6907.