MultiSemCor: an English/Italian aligned corpus word-annotated with a shared inventory of senses

Luisa BENTIVOGLI, Emanuele PIANTA, Marcello RANIERI
ITC-irst, Via Sommarive, 18
38050 Povo, Trento - Italy
{bentivo, pianta, ranieri}@itc.it

Abstract

MultiSemCor is an English/Italian parallel corpus, aligned at the word level and annotated with PoS, lemma and word sense. The parallel corpus has been created by exploiting the SemCor corpus, in which content words are lemmatized and sense-tagged with reference to the WordNet lexical database.

The main hypothesis underlying this methodology is that, given a text and its translation into another language, the semantic information is mostly preserved during the translation process. Therefore, if the texts in one language have been semantically annotated and their translations have not, annotations can be transferred from the source language to the target using word alignment as a bridge.

The strategy for creating MultiSemCor consists in (i) having SemCor texts translated into Italian by professional translators (ii) automatically aligning Italian and English texts at the word level, and (iii) automatically transferring the word sense annotations from English to the aligned Italian words. The final result of the MultiSemCor project is an Italian corpus annotated with PoS, lemma and word sense, but also an aligned parallel corpus lexically annotated with a shared inventory of word senses. More specifically, the sense inventory used is MultiWordNet (Pianta et al., 2002), a multilingual lexical database in which the Italian component is strictly aligned with the English WordNet.

The quality of the annotation of the Italian texts, resulting from the application of the transfer methodology, has been evaluated with respect to a manually annotated Gold Standard: precision amounts to 87.9% and recall to 67.2% (Bentivogli and Pianta, forthcoming).

At present MultiSemCor is composed of 116 English texts along with their corresponding 116 Italian translations. The total amount of tokens is 258,499 for English and 267,607 for Italian. The texts are fully aligned at the word level and content words are annotated with PoS, lemma, and word sense. As regards English, we have 119,802 words semantically annotated (from SemCor). As for Italian, 92,820 words are annotated with word senses that have been automatically transferred from English.

MultiSemCor can be a useful resource for a variety of tasks, both as a monolingual semantically annotated corpus and as a parallel aligned corpus. As an example, we are already using it to automatically enrich the Italian component of MultiWordNet, the reference lexicon of MultiSemCor. Moreover, the Italian component of MultiSemCor is being used as a gold standard for the evaluation of WSD systems working on Italian.

Besides NLP applications, MultiSemCor is also suitable to be consulted by humans through a Web interface (Ranieri et al., 2004). The interface we developed provides for two distinct browsing modalities. In the text-oriented modality, for each bi-text the user has access to the alignment at the sentence and word level, and to the dictionary of all the tokens of the text, with links to the sentences in which they occur. The word-oriented modality amounts to a bilingual semantic concordancer. Another important characteristic of the MultiSemCor Web interface is its integration with the on-line MultiWordNet browser. The interface and the current release of the corpus are available at: http://multisemcor.itc.it.

References

