Using Parallel Corpora for Word Sense Disambiguation

Els Lefever [a,b] Véronique Hoste [a,b,c] Martine De Cock [b]

[a] *LT3, Language and Translation Technology Team, University College Ghent*
[b] *Dept. of Applied Mathematics and Computer Science, Ghent University*
[c] *Dept. of Linguistics, Ghent University*

** The full paper version of this paper is published in the Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon, June 19-24, 2011.*

1 Introduction

Word Sense Disambiguation (WSD) is the Natural Language Processing (NLP) task that consists in selecting the correct sense of a polysemous word in a given context. Most state-of-the-art WSD systems are supervised classifiers that are trained on manually sense-tagged corpora, which are very time-consuming and expensive to build. In order to overcome this acquisition bottleneck (sense-tagged corpora are scarce for languages other than English), we decided to take a multilingual approach to WSD, that builds up the sense inventory on the basis of the Europarl parallel corpus [3]. Using translations from a parallel corpus implicitly deals with the granularity problem as finer sense distinctions are only relevant as far as they are lexicalized in the target translations. It also facilitates the integration of WSD in multilingual applications such as multilingual Information Retrieval (IR) or Machine Translation (MT).

2 Experimental Setup

Starting point of the experiments was the six-lingual sentence-aligned Europarl corpus that was used in the SemEval-2010 "Cross-Lingual Word Sense Disambiguation" (CLWSD) task [4]. The task is a lexical sample task for twenty English ambiguous nouns that consists in assigning a correct translation in the five supported target languages (viz. French, Italian, Spanish, German and Dutch) for an ambiguous focus word in a given context. In order to detect the relevant translations for each of the twenty ambiguous focus words, we ran GIZA++ [5] with its default settings for all focus words. This word alignment output was then considered to be the label for the training instances for the corresponding classifier (e.g. the Dutch translation is the label that is used to train the Dutch classifier). By considering this word alignment output as oracle information, we redefined the CLWSD task as a classification task. To train our five classifiers (English as input language and French, German, Dutch, Italian and Spanish as focus languages), we used the memory-based learning (MBL) algorithm implemented in TIMBL [1], which has successfully been deployed in previous WSD classification tasks [2].

For our feature vector creation, we combined a set of English local context features and a set of binary bag-of-words features that were extracted from the aligned translations. First all English sentences were preprocessed by means of a memory-based shallow parser (MBSP) [1] that performs tokenization, Part-of-Speech tagging and text chunking. The preprocessed sentences were used as input to build a set of commonly used WSD features related to the English input sentence: (a) features related to the focus word itself being the word form of the focus word, the lemma, Part-of-Speech and chunk information and (b) local context features related to a window of three words preceding and following the focus word containing for each of these words their full form, lemma, Part-of-Speech and chunk information.

In addition to these well known monolingual features, we extracted a set of binary bag-of-words features from the aligned translation that are not the target language of the classifier (e.g. for the Dutch classifier, we extracted bag-of-words features from the Italian, Spanish, French and German aligned translations). In order to extract useful content words, we first ran Part-of-Speech tagging and lemmatisation by means of the Treetagger [6] tool. Per ambiguous focus word, a list of content words (nouns, adjectives, verbs and adverbs) was extracted that occurred in the aligned translations of the English sentences containing the focus word. One binary feature per selected content word was then created per ambiguous word: '0' in case the word does not occur in the aligned translation of this instance, and '1' in case the word does occur in the aligned translation of the training instance. For the creation of the bag-of-words features for the test sentences however, we needed to adopt a different approach as we do not have aligned translations for the English test instances at our disposal. We decided to deploy a novel strategy that uses the Google Translate API[1] to automatically generate a translation for all English test instances in the five supported languages. In a next step the automatically generated translation was preprocessed in the same way as the training translations (Part-of-Speech-tagged and lemmatized). The resulting lemmas were then used to construct the same set of binary bag-of-words features that were stored for the training instances of the ambiguous focus word. For our experiments we trained three flavors of the ParaSense system that incorporate different feature sets: (1) both the local context and translation features, (2) translation bag-of-words features and (3) English local context features.

3 Evaluation

As evaluation metrics, we used both the BEST precision metric from the SemEval "Cross-Lingual Word Sense Disambiguation" task as well as a straightforward accuracy measure. As a baseline, we selected the most frequent lemmatized translation that resulted from the automated word alignment (GIZA++). We also compared our results with the two winning SemEval-2 systems for the Cross-Lingual Word Sense Disambiguation task. The classification results show that all three flavors of the ParaSense system easily beat the baseline. Although we applied a very basic strategy for the selection of our bag-of-words translation features (we did not perform any filtering on the translations except for Part-of-Speech information), we observe that for three languages the full feature vector outperforms the classifier that uses the more traditional WSD local context features. For Dutch, the classifier that merely uses translation features even outperforms the classifier that uses the local context features. Moreover, the ParaSense system clearly outperforms the winning SemEval systems, except for Spanish where the scores are similar. As all systems, viz. the two SemEval systems as well as the three flavors of the ParaSense system, were trained on the same Europarl data, the scores illustrate the potential advantages of using a multilingual approach.

References

[1] W. Daelemans and A. van den Bosch. *Memory-based Language Processing*. Cambridge University Press, 2005.

[2] V. Hoste, I. Hendrickx, W. Daelemans, and A. van den Bosch. Parameter Optimization for Machine-Learning of Word Sense Disambiguation. *Natural Language Engineering, Special Issue on Word Sense Disambiguation Systems*, 8:311–325, 2002.

[3] P. Koehn. Europarl: a parallel corpus for statistical machine translation. In *Tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand, 2005.

[4] E. Lefever and V. Hoste. SemEval-2010 Task 3: Cross-Lingual Word Sense Disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010*, pages 15–20, Uppsala, Sweden, 2010.

[5] F. J. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

[6] H. Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on new methods in Language Processing*, Manchester, UK, 1994.

[1] `http://code.google.com/apis/language/`