ParaSense or How to Use Parallel Corpora for Word Sense Disambiguation

Els Lefever[1,2], Véronique Hoste[1,2,3] and Martine De Cock[2]

[1]LT3, Language and Translation Technology Team, University College Ghent
Groot-Brittanniëlaan 45, 9000 Gent, Belgium
[2]Dept. of Applied Mathematics and Computer Science, Ghent University
Krijgslaan 281 (S9), 9000 Gent, Belgium
[3]Dept. of Linguistics, Ghent University
Blandijnberg 2, 9000 Gent, Belgium

Abstract

This paper describes a set of exploratory experiments for a multilingual classification-based approach to Word Sense Disambiguation. Instead of using a predefined monolingual sense-inventory such as WordNet, we use a language-independent framework where the word senses are derived automatically from word alignments on a parallel corpus. We built five classifiers with English as an input language and translations in the five supported languages (viz. French, Dutch, Italian, Spanish and German) as classification output. The feature vectors incorporate both the more traditional local context features, as well as binary bag-of-words features that are extracted from the aligned translations. Our results show that the ParaSense multilingual WSD system shows very competitive results compared to the best systems that were evaluated on the SemEval-2010 Cross-Lingual Word Sense Disambiguation task for all five target languages.

1 Introduction

Word Sense Disambiguation (WSD) is the 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 (Agirre and Edmonds, 2006) . 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 (Koehn, 2005). 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). Significant improvements in terms of general MT quality were for the first time reported by Carpuat and Wu (2007) and Chan et al. (2007). Both papers describe the integration of a dedicated WSD module in a Chinese-English statistical machine translation framework and report statistically significant improvements in terms of standard MT evaluation metrics.

Several studies have already shown the validity of using parallel corpora for sense discrimination (e.g. (Ide et al., 2002)), for bilingual WSD modules (e.g. (Gale and Church, 1993; Ng et al., 2003; Diab and Resnik, 2002; Chan and Ng, 2005; Dagan and Itai, 1994)) and for WSD systems that use a combination of existing WordNets with multilingual evidence (Tufiş et al., 2004). The research described in this paper is novel as it presents a truly multilingual classification-based approach to WSD that directly incorporates evidence from four other languages. To this end, we build further on two well-known research ideas: (1) the possibility to use parallel corpora to extract translation labels and features in an automated way and (2) the assumption that incorporating evidence from multiple languages into the feature vector will be more informative than a more restricted set of monolingual or bilingual features. Furthermore, our WSD system does not use any information from external lexical resources such as WordNet (Fellbaum, 1998) or EuroWordNet (Vossen, 1998).

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 317–322,
Portland, Oregon, June 19-24, 2011.

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 (Lefever and Hoste, 2010b). 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++ (Och and Ney, 2003) 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 (Daelemans and Hoste, 2002), which has successfully been deployed in previous WSD classification tasks (Hoste et al., 2002). We performed heuristic experiments to define the parameter settings for the classifier, leading to the selection of the Jeffrey Divergence distance metric, Gain Ratio feature weighting and k = 7 as number of nearest neighbours. In future work, we plan to use an optimized word-expert approach in which a genetic algorithm performs joint feature selection and parameter optimization per ambiguous word (Daelemans et al., 2003).

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.

2.1 Training Feature Vector Construction

We created two experimental setups. The first training set incorporates the automatically generated word alignments as labels. We applied an automatic post-processing step on these word alignments in order to remove leading and trailing determiners and prepositions. In future work, we will investigate other word alignment strategies and measure the impact on the classification scores. The second training set uses manually verified word alignments as labels for the training instances. This second setup is then to be considered as the upper bound on the current experimental setup.

All English sentences were preprocessed by means of a memory-based shallow parser (MBSP) (Daelemans and van den Bosch, 2005) 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:

- features related to the **focus word itself** being the word form of the focus word, the lemma, Part-of-Speech and chunk information
- **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 extract 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 (Schmid, 1994) 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.

2.2 Test Feature Vector Construction

For the creation of the feature vectors for the test instances, we follow a similar strategy as the one we used for the creation of the training instances. The first part of the feature vector contains the English

local context features that were also extracted for the training instances. For the construction of the bag-of-words features however, we need 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. Online machine translations tools have already been used before to create artificial parallel corpora that were used for NLP tasks such as for instance Named Entity Recognition (Shah et al., 2010).

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.

3 Evaluation

To evaluate our five classifiers, we used the sense inventory and test set of the SemEval "Cross-Lingual Word Sense Disambiguation" task. The sense inventory was built up on the basis of the Europarl corpus: all retrieved translations of a polysemous word were manually grouped into clusters, which constitute different senses of that given word. The test instances were selected from the JRC-ACQUIS Multilingual Parallel Corpus[2] and BNC[3]. To label the test data, native speakers provided their top three translations from the predefined clusters of Europarl translations, in order to assign frequency weights to the set of gold standard translations. A more detailed description of the construction of the data set can be found in Lefever and Hoste (2010a).

As evaluation metrics, we used both the SemEval BEST precision metric from the CLWSD task as well as a straightforward accuracy measure. The SemEval metric takes into account the frequency weights of the gold standard translations: translations that were picked by different annotators get a higher weight. For the BEST evaluation, systems can propose as many guesses as the system believes are correct, but the resulting score is divided by the number of guesses. In this way, systems that output a lot of guesses are not favoured. For a more detailed description of the SemEval scoring scheme, we refer to McCarthy and Navigli (2007). Following variables are used for the SemEval precision formula. Let H be the set of annotators, T the set of test items and h_i the set of responses for an item $i \in T$ for annotator $h \in H$. Let A be the set of items from T where the system provides at least one answer and $a_i : i \in A$ the set of guesses from the system for item i. For each i, we calculate the multiset union (H_i) for all h_i for all $h \in H$ and for each unique type (res) in H_i that has an associated frequency ($freq_{res}$).

$$Prec = \frac{\sum_{a_i:i \in A} \frac{\frac{\sum_{res \in a_i} freq_{res}}{|a_i|}}{|H_i|}}{|A|} \quad (1)$$

The second metric we use is a straightforward accuracy measure, that divides the number of correct answers by the total amount of test instances.

As a baseline, we selected the most frequent lemmatized translation that resulted from the automated word alignment (GIZA++). We also compare our results with the two winning SemEval-2 systems for the Cross-Lingual Word Sense Disambiguation task, UvT-WSD (that only participated for Dutch and Spanish) and T3-COLEUR. The UvT-WSD system (van Gompel, 2010), that also uses a k-nearest neighbor classifier and a variety of local and global context features, obtained the best scores for Spanish and Dutch in the SemEval CLWSD competition. Although we also use a memory-based learner, our method is different from this system in the way the feature vectors are constructed. Next to the incorporation of similar local context features, we also include evidence from multiple languages in our feature vector. For French, Italian and German however, the T3-COLEUR system (Guo and Diab, 2010) outperformed the other systems in the SemEval competition. This system adopts a different approach: during the training phase a monolingual WSD system processes the English input sentence and a word alignment module is used to extract the aligned translation. The English senses together with their aligned translations (and probabil-

[1] http://code.google.com/apis/language/
[2] http://wt.jrc.it/lt/Acquis/
[3] http://www.natcorp.ox.ac.uk/

ity scores) are then stored in a word sense translation table, in which look-ups are performed during the testing phase. This system also differs from the Uvt-WSD and ParaSense systems in the sense that the word senses are derived from WordNet, whereas the other systems do not use any external resources.

The results for all five classifiers are listed in two tables. Table 1 gives an overview of the SemEval-2010 weighted precision scores, whereas Table 2 shows the more straightforward accuracy figures. Both tables list the scores averaged over all twenty test words for the baseline (most frequent word alignment), the best SemEval system (for a given language) and the two ParaSense setups: one that exclusively uses automatically generated word alignments, and one that uses the verified word alignment labels. For both setups we trained three flavors of the ParaSense system (1: local context + translation features, 2: translation features and 3: local context features).

The classification results show that for both setups all three flavors of the ParaSense system easily beat the baseline. 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. 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. In previous research (Lefever and Hoste, 2011), we showed that the classifier using evidence from all different languages was constantly better than the ones using less or no multilingual evidence. In addition, the scores also degraded relatively to the number of translation features that was used. As we used a different set of translation features for the latter pilot experiments (we only used the translations of the ambiguous words instead of the full bag-of-words features we used for the current setup), we need to confirm this trend with more experiments using the current feature sets.

Another important observation is that the classification scores degrade when using the automatically generated word alignments, but only to a minor extent. This clearly shows the viability of our setup. Further experiments with different word alignment settings and symmetrisation methods should allow us to further improve the results with the automatically generated word alignments. Using the non-validated labels makes the system very flexible and language-independent, as all steps in the feature vector creation can be run automatically.

4 Conclusion

We presented preliminary results for a multilingual classification-based approach to Word Sense Disambiguation. In addition to the commonly used monolingual local context features, we also incorporate bag-of-word features that are built from the aligned translations. Although there is still a lot of room for improvement on the feature base, our results show that the ParaSense system clearly outperforms state-of-the-art systems for all languages, except for Spanish where the results are very similar. As all steps are run automatically, this multilingual approach could be an answer for the acquisition bottleneck, as long as there are parallel corpora available for the targeted languages. Although large multilingual corpora are still rather scarce, we strongly believe there will be more parallel corpora available in the near future (large companies and organizations disposing of large quantities of parallel text, internet corpora such as the ever growing Wikipedia corpus, etc.). Another line of research could be the exploitation of comparable corpora to acquire additional training data.

In future work, we want to run additional experiments with different classifiers (SVM) and apply a genetic algorithm to perform joint feature selection, parameter optimization and instance selection. We also plan to expand our feature set by including global context features (content words from the English sentence) and to examine the relationship between the performance and the number (and nature) of languages that is added to the feature vector. In addition, we will apply semantic analysis tools (such

	French	Italian	Spanish	Dutch	German
Baseline	20.71	14.03	18.36	15.69	13.16
T3-COLEUR	21.96	15.55	19.78	10.71	13.79
UvT-WSD			23.42	17.70	
Non-verified word alignment labels					
ParaSense1 (full feature vector)	24.54	18.03	22.80	18.56	16.88
ParaSense2 (translation features)	23.92	16.77	22.58	17.70	15.98
ParaSense3 (local context features)	24.09	19.89	23.21	17.57	16.55
Verified word alignment labels					
ParaSense1 (full feature vector)	24.60	19.64	23.10	18.61	17.41
ParaSense2 (translation features)	24.29	19.15	22.94	18.25	16.90
ParaSense3 (local context features)	24.79	21.31	23.56	17.70	17.54

Table 1: SemEval precision scores averaged over all twenty test words

	French	Italian	Spanish	Dutch	German
Baseline	63.10	47.90	53.70	59.40	52.30
T3-COLEUR	66.88	50.73	59.83	40.01	54.20
UvT-WSD			70.20	64.10	
Non-verified word alignment labels					
ParaSense1 (full feature vector)	75.20	63.40	68.20	68.10	66.20
ParaSense2 (translation features)	73.20	58.30	67.60	65.90	63.60
ParaSense3 (local context features)	73.50	65.50	69.40	63.90	61.90
Verified word alignment labels					
ParaSense1 (full feature vector)	75.70	63.20	68.50	68.20	67.80
ParaSense2 (translation features)	74.70	61.30	68.30	66.80	66.20
ParaSense3 (local context features)	75.20	67.30	70.30	63.30	66.10

Table 2: Accuracy percentages averaged over all twenty test words

as LSA) on our multilingual bag-of-words sets in order to detect latent semantic topics in the multilingual feature base. Finally, we want to evaluate to which extent the integration of our WSD output helps practical applications such as Machine Translation or Information Retrieval.

Acknowledgments

We thank the anonymous reviewers for their valuable remarks. This research was funded by the University College Research Fund.

References

E. Agirre and P. Edmonds, editors. 2006. *Word Sense Disambiguation. Algorithms and Applications*. Text, Speech and Language Technology. Springer, Dordrecht.

M. Carpuat and D. Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 61–72, Prague, Czech Republic.

Y.S. Chan and H.T. Ng. 2005. Scaling Up Word Sense Disambiguation via Parallel Texts. In *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI 2005)*, pages 1037–1042, Pittsburgh, Pennsylvania, USA.

Y.S. Chan, H.T. Ng, and D. Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 33–40, Prague, Czech Republic.

W. Daelemans and V. Hoste. 2002. Evaluation of Machine Learning Methods for Natural Language Processing Tasks. In *Proceedings of the third International Conference on Language Resources and Evaluation (LREC'02)*, pages 755–760.

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

W. Daelemans, V. Hoste, F. De Meulder, and B. Naudts. 2003. Combined optimization of feature selection and

algorithm parameters in machine learning of language. *Machine Learning*, pages 84–95.

I. Dagan and A. Itai. 1994. Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, 20(4):563–596.

M. Diab and P. Resnik. 2002. An Unsupervised Method for Word Sense Tagging Using Parallel Corpora. In *Proceedings of ACL*, pages 255–262.

C. Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

W.A. Gale and K.W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

W. Guo and M. Diab. 2010. COLEPL and COLSLM: An Unsupervised WSD Approach to Multilingual Lexical Substitution, Tasks 2 and 3 SemEval 2010. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 129–133, Uppsala, Sweden. Association for Computational Linguistics.

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

N. Ide, T. Erjavec, and D. Tufiş. 2002. Sense discrimination with parallel corpora. . In *ACL-2002 Workhop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 54–60, Philadelphia.

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

E. Lefever and V. Hoste. 2010a. Construction of a Benchmark Data Set for Cross-Lingual Word Sense Disambiguation. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

E. Lefever and V. Hoste. 2010b. 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.

E. Lefever and V. Hoste. 2011. Examining the Validity of Cross-Lingual Word Sense Disambiguation. In *Proceedings of the Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2011)*, Tokyo, Japan.

D. McCarthy and R. Navigli. 2007. SemEval-2007 Task 10: English Lexical Substitution Task. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53, Prague, Czech Republic.

H.T. Ng, B. Wang, and Y.S. Chan. 2003. Exploiting parallel texts for word sense disambiguation: An empirical study. In *41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 455–462, Sapporo, Japan.

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

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

R. Shah, B. Lin, A. Gershman, and R. Frederking. 2010. SYNERGY: A Named Entity Recognition System for Resource-scarce Languages such as Swahili using Online Machine Translation. In *Proceedings of the Second Workshop on African Language Technology (AFLAT 2010)*, Valletta, Malt.

D. Tufiş, R. Ion, and N. Ide. 2004. Fine-Grained Word Sense Disambiguation Based on Parallel Corpora, Word Alignment, Word Clustering and Aligned Wordnets. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, pages 1312–1318, Geneva, Switzerland, August. Association for Computational Linguistics.

M. van Gompel. 2010. UvT-WSD1: A Cross-Lingual Word Sense Disambiguation System. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 238–241, Uppsala, Sweden. Association for Computational Linguistics.

P. Vossen, editor. 1998. *EuroWordNet: a multilingual database with lexical semantic networks*. Kluwer Academic Publishers, Norwell, MA, USA.