An Evaluation and Possible Improvement Path for Current SMT Behavior on Ambiguous Nouns

Els Lefever[1,2] **and Véronique Hoste**[1,2,3]

[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

Mistranslation of an ambiguous word can have a large impact on the understandability of a given sentence. In this article, we describe a thorough evaluation of the translation quality of ambiguous nouns in three different setups. We compared two statistical Machine Translation systems and one dedicated Word Sense Disambiguation (WSD) system. Our WSD system incorporates multilingual information and is independent from external lexical resources. Word senses are derived automatically from word alignments on a parallel corpus. We show that the two WSD classifiers that were built for these experiments (English–French and English–Dutch) outperform the SMT system that was trained on the same corpus. This opens perspectives for the integration of our multilingual WSD module in a statistical Machine Translation framework, in order to improve the automated translation of ambiguous words, and by consequence make the translation output more understandable.

1 Introduction

Word Sense Disambiguation (WSD) is the NLP task that consists in assigning a correct sense to an ambiguous word in a given context. Traditionally, WSD relies on a predefined monolingual sense-inventory such as WordNet (Fellbaum, 1998) and WSD modules are trained on corpora, which are manually tagged with senses from these inventories. A number of issues arise with these monolingual supervised approaches to WSD. First of all, there is a lack of large sense-inventories and sense-tagged corpora for languages other than English. Furthermore, sense inventories such as WordNet contain very fine-grained sense distinctions that make the sense disambiguation task very challenging (even for human annotators), whereas very detailed sense distinctions are often irrelevant for practical applications. In addition to this, there is a growing feeling in the community that WSD should be used and evaluated in real application such as Machine Translation (MT) or Information Retrieval (IR) (Agirre and Edmonds, 2006).

An important line of research consists in the development of dedicated WSD modules for MT. Instead of assigning a sense label from a monolingual sense-inventory to the ambiguous words, the WSD system has to predict a correct translation for the ambiguous word in a given context. In (Vickrey et al., 2005), the problem was defined as a word translation task. The translation choices of ambiguous words are gathered from a parallel corpus by means of word alignment. The authors reported improvements on two simplified translation tasks: word translation and blank filling. The evaluation was done on an English-French parallel corpus but is confronted with the important limitation of having only one valid translation (the aligned translation in the parallel corpus) as a gold standard translation. Cabezas and Resnik (2005) tried to improve an SMT system by adding additional translations to the phrase table, but were confronted with tuning problems of this dedicated WSD feature. Specia (2006) used an inductive logic programming-based WSD system which was tested on seven ambiguous verbs in English-Portuguese translation. The latter systems already present promising results for the use of WSD in MT, but really significant improvements in terms of general machine translation qual-

Proceedings of SSST-5, Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation, pages 52–60,
ACL HLT 2011, Portland, Oregon, USA, June 2011.

ity were for the first time obtained 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.

Stroppa et al. (2007) take a completely different approach to perform some sort of implicit Word Sense Disambiguation in MT. They introduce context-information features that exploit source similarity, in addition to target similarity that is modeled by the language model, in an SMT framework. For the estimation of these features that are very similar to the typical WSD local context features (left and right context words, Part-of-Speech of the focus phrase and context words), they use a memory-based classification framework.

The work we present in this paper is different from previous research in two aspects. Firstly, we evaluate the performance of two state-of-the-art SMT systems and a dedicated WSD system on the translation of ambiguous words. The comparison is done against a manually constructed gold-standard for two language pairs, viz. English–French and English–Dutch. Although it is crucial to measure the general translation quality after integrating a dedicated WSD module in the SMT system, we think it is equally interesting to conduct a dedicated evaluation of the translation quality on ambiguous nouns. Standard SMT evaluation metrics such as BLEU (Papineni et al., 2002) or edit-distance metrics (e.g. Word Error Rate) measure the global overlap of the translation with a reference, and are thus not very sensitive to WSD errors. The mistranslation of an ambiguous word might be a subtle change compared to the reference sentence, but it often drastically affects the global understanding of the sentence.

Secondly, we explore the potential benefits of a real multilingual approach to WSD. The idea to use translations from parallel corpora to distinguish between word senses is based on the hypothesis that different meanings of a polysemous word are often lexicalized across languages (Resnik and Yarowsky, 2000). Many WSD studies have incorporated this cross-lingual evidence idea and have successfully applied bilingual WSD classifiers (Gale and Church, 1993; Ng et al., 2003; Diab and Resnik, 2002) or systems that use a combination of existing WordNets with multilingual evidence (Tufiş et al., 2004). Our WSD system is different in the sense that it is independent from a predefined sense-inventory (it only uses the parallel corpus at hand) and that it is truly multilingual as it incorporates information from four other languages (French, Dutch, Spanish, Italian and German depending on the target language of the classifier). Although our classifiers are still very preliminary in terms of the feature set and parameters that are used, we obtain interesting results on our test sample of ambiguous nouns. We therefore believe our system can have a real added value for SMT, as it can easily be trained for different language pairs on exactly the same corpus which is used to train the SMT system, which should make the integration a lot easier.

The remainder of this paper is organized as follows. Section 2 introduces the two machine translation systems we evaluated, while section 3 describes the feature construction and learning algorithm of our multilingual WSD system. Section 4 gives an overview of the experimental setup and results. We finally draw conclusions and present some future research in Section 5.

2 Statistical Machine Translation Systems

For our experiments, we analyzed the behavior of two phrase-based statistical machine translation (SMT) systems on the translation of ambiguous nouns. SMT generates translations on the basis of statistical models whose parameters are derived from the analysis of sentence-aligned parallel text corpora. Phrase-based SMT is considered as the dominant paradigm in MT research today. It combines a phrase translation model (which is based on the noisy channel model) and a phrase-based decoder in order to find the most probable translation e of a foreign sentence f (Koehn et al., 2003). Usually the Bayes rule is used to reformulate this translation probability:

$$argmax_e p(e|f) = argmax_e p(f|e)p(e)$$

This allows for a language model $p(e)$ that guarantees the fluency and grammatical correctness of the translation, and a separate translation model $p(f|e)$ that focusses on the quality of the transla-

tion. Training of both the language model (on monolingual data) as well as the translation model (on bilingual text corpora) requires large amounts of text data.

Research has pointed out that adding more training data, both for the translation as for the language models, results in better translation quality, (Callison-Burch et al., 2009). Therefore it is important to notice that our comparison of the two SMT systems is somewhat unfair, as we compared the Moses research system (that was trained on the Europarl corpus) with the Google commercial system that is trained on a much larger data set. It remains an interesting exercise though, as we consider the commercial system as the upper bound of how far current SMT can get in case it has unlimited access to text corpora and computational resources.

2.1 Moses

The first statistical machine translation system we used is the off-the-shelf Moses toolkit (Koehn et al., 2007). As the Moses system is open-source, well documented, supported by a very lively users forum and reaches state-of-the-art performance, it has quickly been adopted by the community and highly stimulated development in the SMT field. It also features factored translation models, which enable the integration of linguistic and other information at the word level. This makes Moses a good candidate to experiment with for example a dedicated WSD module, that requires more enhanced linguistic information (such as lemmas and Part-of-Speech tags).
We trained Moses for English–French and English–Dutch on a large subsection of the Europarl corpus (See Section 3 for more information on the corpus), and performed some standard cleaning. Table 1 lists the number of aligned sentences after cleaning the bilingual corpus, and the number of uni-, bi- and trigrams that are comprised by the language model.

2.2 Google

In order to gain insights in the upper bounds for current SMT, we also analyzed the output of the Google Translate API[1] for our set of ambiguous nouns. Google Translate currently supports 57 languages. As both the amount of parallel and monolingual training data as well as the computer power are crucial for statistical MT, Google (that disposes of large computing clusters and a network of data centers for Web search) has very valuable assets at its disposal for this task. We can only speculate about the amount of resources that Google uses to train its translation engine. Part of the training data comes from transcripts of United Nations meetings (in six official languages) and those of the European Parliament (Europarl corpus). Google research papers report on a distributed infrastructure that is used to train on up to two trillion tokens, which result in language models containing up to 300 billion ngrams (Brants et al., 2007).

	French	Dutch
Number of bilingual sentence pairs		
	872.689	873.390
Number of ngrams		
unigrams	103.027	173.700
bigrams	1.940.925	2.544.554
trigrams	2.054.906	1.951.992

Table 1: Statistics resulting from the Moses training phase

3 ParaSense

This section describes the ParaSense WSD system: 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 used the sentence-aligned Europarl corpus (Koehn, 2005) for the construction of our WSD module. The following six languages were selected: English (our focus language), Dutch, French, German, Italian and Spanish. We only considered the 1-1 sentence alignments between English and the five other languages. This way we obtained a six-lingual sentence-aligned subcorpus of Europarl, that contains 884.603 sentences per language. For our experiments we used the lexical sample of twenty ambiguous nouns that was also used in the SemEval-2010 "Cross-Lingual Word Sense Disambiguation" (CLWSD) task (Lefever and Hoste, 2010b), which consists in assigning a

[1] http://code.google.com/apis/language/translate/overview.html

correct translation in five supported target languages (viz. French, Italian, Spanish, German and Dutch) for an ambiguous focus word in a given context.

In order to detect all relevant translations for the twenty ambiguous focus words, we ran GIZA++ (Och and Ney, 2003) with its default settings on our parallel corpus. The obtained word alignment output was then considered to be the classification label for the training instances for a given classifier (e.g. the French translation resulting from the word alignment is the label that is used to train the French classifier). This way we obtained all class labels (or oracle translations) for all training instances for our five classifiers (English as an input language and French, German, Dutch, Italian and Spanish as target languages). For the experiments described in this paper, we focused on the English–French and English–Dutch classifiers.

We created two experimental setups. The first training set contains the automatically generated word alignment translations as labels. A postprocessing step was applied on these translations in order to automatically filter leading and trailing determiners and prepositions from the GIZA++ output. For the creation of the second training set, we manually verified all word alignment correspondences of the ambiguous words. This second setup gives an idea of the upperbound performance in case the word alignment output could be further improved for our ambiguous nouns.

3.1 Classifier

To train our WSD classifiers, we used the memory-based learning (MBL) algorithms implemented in TIMBL (Daelemans and van den Bosch, 2005), which has successfully been deployed in previous WSD classification tasks (Hoste et al., 2002). We performed very basic 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).

3.2 Feature Construction

For the feature vector construction, we combine local context features that were extracted from the English sentence and a set of binary bag-of-words features that were extracted from the aligned translations in the four other languages (that are not the target language of the classifier).

3.2.1 Local Context Features

We extract the same set of local context features from both the English training and test instances. 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 English instances were used as input to build a set of commonly used WSD features:

- 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

These local context features are to be considered as a basic feature set. The Senseval evaluation exercises have shown that feeding additional information sources to the classifier results in better system performance (Agirre and Martinez, 2004). In future experiments we plan to integrate a.o. lemma information on the surrounding content words and semantic analysis (e.g. Singular Value Decomposition (Gliozzo et al., 2005)) in order to detect latent correlations between terms.

3.2.2 Translation Features

In addition to the commonly deployed local context features, we also extracted a set of binary bag-of-words features from the aligned translations that are not the target language of the classifier (e.g. for the French classifier, we extract bag-of-words features from the Italian, Spanish, Dutch and German aligned translations). We preprocessed all aligned translations by means of the Treetagger tool (Schmid, 1994) that outputs Part-of-Speech and

lemma information. Per ambiguous focus word, a list of all content words (nouns, adjectives, adverbs and verbs) that occurred in the aligned translations of the English sentences containing this word, was extracted. This resulted in one binary feature per selected content word per language. For the construction of the translation features for the training set, we used the Europarl aligned translations.

As we do not dispose of similar aligned translations for our test instances (where we only have the English test sentences at our disposal), we had to adopt a different strategy. We decided to use the Google Translate API to automatically generate translations for all English test instances in the five target languages. This automatic translation process can be done using whatever machine translation tool, but we chose the Google API because of its easy integration. Online machine translation 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). Similarly, Navigli and Ponzetto (2010) used the Google Translate API to enrich BabelNet, a wide-coverage multilingual semantic network, with lexical information for all languages.

Once the automatic aligned translations were generated, we preprocessed them in the same way as we did for the aligned training translations. In a next step, we again selected all content words from these translations and constructed the binary bag-of-words features.

4 Evaluation

To evaluate the two machine translation systems as well as the ParaSense system on their performance on the lexical sample of twenty ambiguous words, 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]. There were in total 50 test instances for each of the twenty ambiguous words in the sample. To label the test data, native speakers assigned three valid translations from the predefined clusters of Europarl translations to each test instance. A more detailed description of the construction of the data set can be found in (Lefever and Hoste, 2010a). As evaluation metric, we used 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++).

The output of the ParaSense WSD module consists of a lemmatized translation of the ambiguous focus word in the target language. The output of the two statistical machine translation systems, however, is a translation of the full English input sentence. Therefore we manually selected the translation of the ambiguous focus word from the full translation, and made sure the translation was put in its base form (masculine singular form for nouns and adjectives, infinitive form for verbs).

Table 2 lists the accuracy figures for the baseline, two flavors of the ParaSense system (with and without correction of the word alignment output), Moses and Google for English–French and English–Dutch.

A first conclusion is that all systems beat the most frequent sense baseline. As expected, the Google system (where there was no limitation on the training data) achieves the best results, but for French the considerable difference in training size only leads to modest performance gains compared to the ParaSense System. Another interesting observation is that the ParaSense system that uses manually verified translation labels hardly beats the system that uses automatically generated class labels. This is promising as it makes the manual interventions on the data superfluous and leads to a fully automatic system development process.

Figure 1 illustrates the accuracy figures for French for all three systems (for the ParaSense system we used the flavor that incorporates the non-validated translation labels) on all individual test words.

The three curves follow a similar pattern, except for some words where Moses (*mood, scene, side*) or both Moses and ParaSense (*figure*) perform worse. As the curves show, some words (e.g. *coach, figure,*

[2] http://wt.jrc.it/lt/Acquis/

[3] http://www.natcorp.ox.ac.uk/

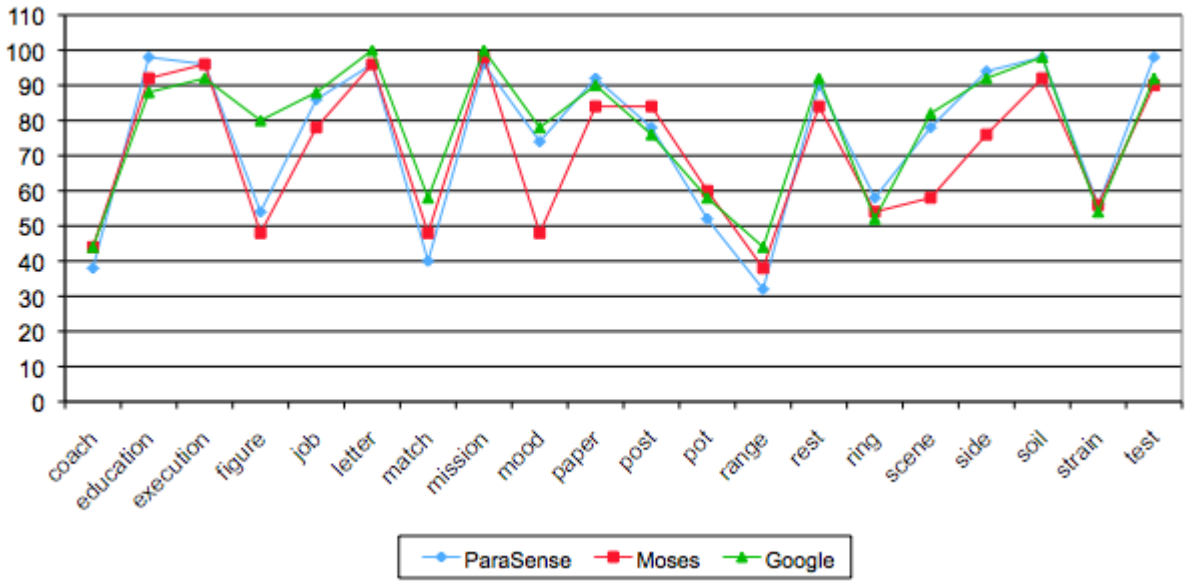

Figure 1: Accuracy figures per system for all 20 test words

	French	Dutch
Baseline	63%	59%
ParaSense system		
Non Corrected word alignment labels	75%	68%
Corrected word alignment labels	76%	68%
SMT Systems		
Moses	71%	63%
Google	78%	74%

Table 2: Accuracy figures averaged over all twenty test words

	Number of Instances	Number of Translations
coach	66	11
education	4380	55
execution	489	26
figure	2298	167
job	7531	184
letter	1822	75
match	109	21
mission	1390	46
mood	100	26
paper	3650	94
post	998	68
pot	63	27
range	1428	145
rest	1739	80
ring	143	46
scene	284	50
side	3533	261
soil	287	16
strain	134	40
test	1368	92

Table 3: Number of instances and classes for all twenty test words in French

match, range) are particularly hard to disambiguate, while others obtain very high scores (e.g. *letter, mission, soil*). The almost perfect scores for the latter can be explained by the fact that these words all have a very generic translation in French (respectively *lettre, mission, sol*) that can be used for all senses of the word, although there might be more suited translations for each of the senses depending on the context. As the manual annotators could pick three good translations for each test instance, the most generic translation often figures between the gold standard translations.

The low scores for some other words can often be explained through the relationship with the number of training instances (corresponding to the frequency of the word in the training corpus) and the ambiguity (number of translations) per word. As is shown in Table 3, both for *coach* and *match* there are very few examples in the corpus, while *figure* and *range*

are very ambiguous (respectively 167 and 145 translations to choose from).

The main novelty of our ParaSense system lies in the application of a multilingual approach to perform WSD, as opposed to the more classical approach that only uses monolingual local context features. Consequently we also ran a set of additional experiments to examine the contribution of the different translation features to the WSD performance. Table 4 shows the accuracy figures for French and Dutch for a varying number of translation features including the other four languages: Italian, Spanish, French and Dutch for the French classifier or French for the Dutch classifier. The scores clearly confirm the validity of our hypothesis: the classifiers using translation features are constantly better than the one that merely uses English local context features. For French, the other two romance languages seem to contribute most: the classifier that uses Italian and Spanish bag-of-words features achieves the best performance (75.50%), whereas the classifier that incorporates German and Dutch translations obtains the worst scores (71.90%). For Dutch, the interpretation of the scores is less straightforward: the Italian-German combination achieves the best result (69%), but the difference with the other classifiers that use two romance languages (Italian-Spanish: 67.70% and Italian-French: 67.20%) is less salient than for French. In order to draw final conclusions on the contribution of the different languages, we probably first need to optimize our feature base and classification parameters. For the current experiments, we use very sparse bag-of-words features that can be optimized in different ways (e.g. feature selection, reduction of the bag-of-words features by applying semantic analysis such as Singular Value Decomposition, etc.).

	French	**Dutch**
Baseline	63.10	59.40
All four translation features		
It, Es, De, Nl/Fr	75.20	68.10
Three translation features		
It, Es, De	75.00	67.80
Es, De, Nl/Fr	74.70	66.30
It, De, Nl/Fr	75.20	68.20
It, Es, Nl/Fr	75.30	67.90
Average	75.05	67.55
Two translation features		
Es, De	74.70	67.80
It, De	75.10	69.00
De, Nl/Fr	71.90	68.00
It, Es	75.50	67.70
Es, Nl/Fr	74.20	68.10
It, Nl/Fr	75.30	67.20
Average	74.45	67.96
One translation feature		
De	74.50	66.50
Es	75.20	68.40
It	74.90	66.70
Nl/Fr	73.80	66.20
Average	74.60	66.95
No translation features		
None	73.50	63.90

Table 4: Accuracy figures for French and Dutch for a varying number of translation features including the other four languages viz. Italian (It), Spanish (Es), German (De) and French (Fr) or Dutch (Nl)

5 Conclusion

We presented a thorough evaluation of two statistical Machine Translation systems and one dedicated WSD system on a lexical sample of English ambiguous nouns. Our WSD system incorporates both monolingual local context features and bag-of-words features that are built from aligned translations in four additional languages. The best results are obtained by Google, the SMT system that is built with no constraints on data size or computational resources. Although there is still a lot of room for improvement on the feature base and optimization of the WSD classifiers, our results show that the ParaSense system outperforms Moses that is built with the same training corpus.

We also noticed large differences among the test words, often related to the number of training instances and the number of translations the classifier (or decoder) has to choose from.

Additional experiments with the ParaSense system incorporating a number of varying translations features allow us to confirm the validity of our hypothesis. The classifiers that use the multilingual bag-of-words features clearly outperform the classifier that only uses local context features.

In future work, we want to expand our feature set and apply a genetic algorithm to perform joint feature selection, parameter optimization and instance

selection. In addition, we will apply semantic analysis tools (such as SVD or 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 the decoder to pick the correct translation in a real SMT framework.

References

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

E. Agirre and D. Martinez. 2004. Smoothing and Word Sense Disambiguation. In *Proceedings of EsTAL - España for Natural Language Processing*, Alicante, Spain.

Th. Brants, A.C. Popat, P. Xu, F.J. Och, and J. Dean. 2007. Large Language Models in Machine Translation. In *Proceedings of the 2007 Joint Conference on Empirical methods in Natural Language Processing and Computational Natural Language Learning*, pages 858–867.

C. Cabezas and P. Resnik. 2005. Using wsd techniques for lexical selection in statistical machine translation. Technical report, Institute for Advanced Computer Studies, University of Maryland.

C. Callison-Burch, Ph. Koehn, Ch. Monz, and J. Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the 4th EACL Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece.

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, 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 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.

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.

A.M. Gliozzo, C. Giuliano, and C. Strapparava. 2005. Domain Kernels for Word Sense Disambiguation. In

43nd Annual Meeting of the Association for Computational Linguistics. (ACL-05).

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.

P. Koehn, F.J. Och, and D. Marcu. 2003. Statistical Phrase-based translation. In *HLT-NAACL 2003: conference combining Human Language Technology conference series and the North American Chapter of the Association for Computational Linguistics conference series*, pages 48–54, Edmonton, Canada.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

P. 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.

R. Navigli and S.P. Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden.

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.

K. Papineni, S. Roukos, T. Ward, and Zhu W.-J. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

Ph. Resnik and D. Yarowsky. 2000. Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(3):113–133.

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.

L. Specia. 2006. A Hybrid Relational Approach for WSD - First Results. In *Proceedings of the COLING/ACL 2006 Student Research Workshop*, pages 55–60, Sydney, Australia.

N. Stroppa, A. van den Bosch, and A. Way. 2007. Exploiting source similarity for smt using context-informed features. In *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2007)*.

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.

D. Vickrey, L. Biewald, M. Teyssier, and D. Koller. 2005. Word-sense disambiguation for machine translation. In *Proceedings of EMNLP05*, pages 771–778.