WSD for *n*-best reranking and local language modeling in SMT

Marianna Apidianaki, Guillaume Wisniewski†, Artem Sokolov, Aurélien Max†, François Yvon†

LIMSI-CNRS
† Univ. Paris Sud
BP 133, F-91403, Orsay Cedex, France
firstname.lastname@limsi.fr

Abstract

We integrate semantic information at two stages of the translation process of a state-of-the-art SMT system. A Word Sense Disambiguation (WSD) classifier produces a probability distribution over the translation candidates of source words which is exploited in two ways. First, the probabilities serve to rerank a list of *n*-best translations produced by the system. Second, the WSD predictions are used to build a supplementary language model for each sentence, aimed to favor translations that seem more adequate in this specific sentential context. Both approaches lead to significant improvements in translation performance, highlighting the usefulness of source side disambiguation for SMT.

1 Introduction

Word Sense Disambiguation (WSD) is the task of identifying the sense of words in texts by reference to some pre-existing sense inventory. The selection of the appropriate inventory and WSD method strongly depends on the goal WSD intends to serve: recent methods are increasingly oriented towards the disambiguation needs of specific end applications, and explicitly aim at improving the overall performance of complex Natural Language Processing systems (Ide and Wilks, 2007; Carpuat and Wu, 2007). This task-oriented conception of WSD is manifested in the area of multilingual semantic processing: supervised methods, which were previously shown to give the best results, are being abandoned in favor of unsupervised ones that do not rely on pre-annotated training data. Accordingly, pre-defined semantic inventories, that usually served to provide the lists of candidate word senses, are being replaced by senses relevant to the considered applications and directly identified from corpora by means of word sense induction methods.

In a multilingual setting, the sense inventories needed for disambiguation are generally built from all possible translations of words or phrases in a parallel corpus (Carpuat and Wu, 2007; Chan et al., 2007), or by using more complex representations of the semantics of translations (Apidianaki, 2009; Mihalcea et al., 2010; Lefever and Hoste, 2010). However, integrating this semantic knowledge into Statistical Machine Translation (SMT) raises several challenges: the way in which the predictions of the WSD classifier have to be taken into account; the type of context exploited for disambiguation; the target words to be disambiguated ("all-words" WSD vs. WSD restricted to target words satisfying specific criteria); the use of a single classifier versus building separate classifiers for each source word; the quantity and type of data used for training the classifier (e.g., use of raw data or of more abstract representations, such as lemmatization, allowing to deal with sparseness issues), and many others. Seemingly, the optimal way to take advantage of WSD predictions remains an open issue.

In this work, we carry out a set of experiments to investigate the impact of integrating the predictions of a cross-lingual WSD classifier into an SMT system, at two different stages of the translation process. The first approach exploits the probability distribution built by the WSD classifier over the set of translations of words found in the parallel corpus,

Proceedings of SSST-6, Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation, pages 1–9,
Jeju, Republic of Korea, 12 July 2012.

for reranking the translations in the *n*-best list generated by the SMT system. Words in the list that match one of the proposed translations are boosted and are thus more likely to appear in the final translation. Our results on the English-French IWSLT'11 task show substantial improvements in translation quality. The second approach provides a tighter integration of the WSD classifier with the rest of the system: using the WSD predictions, an additional *sentence specific* language model is estimated and used during decoding. These additional local models can be used as an external knowledge source to reinforce translation hypotheses matching the prediction of the WSD system.

In the rest of the paper, we present related work on integrating semantic information into SMT (Section 2). The WSD classifier used in the current study is described in Section 3. We then present the two approaches adopted for integrating the WSD output into SMT (Section 4). Evaluation results are presented in Section 5, before concluding and discussing some avenues for future work.

2 Related work

Word sense disambiguation systems generally work at the word level: given an input word and its context, they predict its (most likely) meaning. At the same time, state-of-the-art translation systems all consider groups of words (phrases, tuples, etc.) rather than single words in the translation process. This discrepancy between the units used in MT and those used in WSD is one of the major difficulties in integrating word predictions into the decoder. This was, for instance, one of the reasons for the somewhat disappointing results obtained by Carpuat and Wu (2005) when the output of a WSD system was directly incorporated into a Chinese-English SMT system. Because of this difficulty, other cross-lingual semantics works have considered only simplified tasks, like blank-filling, without addressing the integration of the WSD models in full-scale MT systems (Vickrey et al., 2005; Specia, 2006).

Since the pioneering work of Carpuat and Wu (2005), several more successful ways to take WSD predictions into account have been proposed. For instance, Carpuat and Wu (2007) proposed to generalize the WSD system so that it performs a fully phrasal multiword disambiguation. However, given that the number of phrases is far larger than the number of words, this approach suffers from sparsity and computational problems, as it requires training a classifier for each entry of the phrase table.

Chan et al. (2007) introduced a way to modify the rule weights of a hierarchical translation system to reflect the predictions of their WSD system. While their approach and ours are built on the same intuition (an adaptation of a model to incorporate word predictions) their work is specific to hierarchical systems, while ours can be applied to any decoder that uses a language model. Haque et al. (2009) et Haque et al. (2010) introduce lexico-syntactic descriptions in the form of supertags as source language context-informed features in a phrase-based SMT and a state-of-the-art hierarchical model, respectively, and report significant gains in translation quality.

Closer to our work, Mauser et al. (2009) and Patry and Langlais (2011) train a global lexicon model that predicts the bag of output words from the bag of input words. As no explicit alignment between input and output words is used, words are chosen based on the (global) input context. For each input sentence, the decoder considers these word predictions as an additional feature that it uses to define a new model score which favors translation hypotheses containing words predicted by the global lexicon model. A difference between this approach and our work is that instead of using a global lexicon model, we disambiguate a subset of the words in the input sentence by employing a WSD classifier that creates a probability distribution over the translations of each word in its context.

The unsupervised cross-lingual WSD classifier used in this work is similar to the one proposed in Apidianaki (2009). The original classifier disambiguates new instances of words in context by selecting the most appropriate cluster of translations among a set of candidate clusters found in an automatically built bilingual sense inventory. The sense inventory exploited by the classifier is created by a cross-lingual word sense induction (WSI) method that reveals the senses of source words by grouping their translations into clusters according to their semantic proximity, revealed by a distributional similarity calculation. The resulting clusters represent

the source words' candidate senses. This WSD method gave good results in a word prediction task but, similarly to the work of Vickrey et al. (2005) and of Specia (2006), the predictions are not integrated into a complete MT system.

3 The WSD classifier

Our WSD classifier is a variation of the one introduced in Apidianaki (2009). The main difference is that here the classifier serves to discriminate between unclustered translations of a word and to assign a probability to each translation for new instances of the word in context. Each translation is represented by a source language feature vector that the classifier uses for disambiguation. All experiments carried out in this study are for the English (EN) - French (FR) language pair.

3.1 Source Language Feature Vectors

Preprocessing The information needed by the classifier is gathered from the EN-FR training data provided for the IWSLT'11 evaluation task.[1] The dataset consists of 107,268 parallel sentences, word-aligned in both translation directions using GIZA++ (Och and Ney, 2003). We disambiguate EN words found in the parallel corpus that satisfy the set of criteria described below.

Two bilingual lexicons are built from the alignment results and filtered to eliminate spurious alignments. First, translation correspondences with a probability lower than a threshold are discarded;[2] then translations are filtered by part-of-speech (PoS), keeping for each word only translations pertaining to the same grammatical category;[3] finally, only intersecting alignments (i.e., correspondences found in the lexicons of both directions) are retained. Given that the lexicons contain word forms, the intersection is calculated based on lemmatization information in order to perform a generalization over the contents of the lexicons. For instance, if the EN adjective *regular* is translated by *habituelle* (feminine singular form of the adjective *habituel*) in the EN-FR lexicon, but is found to translate *habituel* (masculine singular form) in the other direction, the EN-FR correspondence *regular*/*habituelle* is retained (because the two variants of the adjective are reduced to the same lemma).

All lexicon entries satisfying the above criteria are retained and used for disambiguation. In these initial experiments, we disambiguate English words having less than 20 French translations in the lexicon. Each French translation of an English word that appears more than once in the training corpus[4] is characterized by a weighted English feature vector built from the training data.

Vector building The feature vectors corresponding to the translations are built by exploiting information from the source contexts (Apidianaki, 2008; Grefenstette, 1994). For each translation of an EN word *w*, we extract the content words that co-occur with *w* in the corresponding source sentences of the parallel corpus (i.e. the content words that occur in the same sentence as *w* whenever it is translated by this translation). The extracted source language words constitute the features of the vector built for the translation.

For each translation T_i of *w*, let N be the number of features retained from the corresponding source context. Each feature F_j ($1 \leq j \leq N$) receives a total weight $\text{tw}(F_j, T_i)$ defined as the product of the feature's global weight, $\text{gw}(F_j)$, and its local weight with that translation, $\text{lw}(F_j, T_i)$:

$$\text{tw}(F_j, T_i) = \text{gw}(F_j) \cdot \text{lw}(F_j, T_i) \quad (1)$$

The global weight of a feature F_j is a function of the number N_i of translations (T_i's) to which F_j is related, and of the probabilities (p_{ij}) that F_j co-occurs with instances of *w* translated by each of the T_i's:

$$\text{gw}(F_j) = 1 - \frac{\sum_{T_i} p_{ij} \log(p_{ij})}{N_i} \quad (2)$$

Each of the p_{ij}'s is computed as the ratio between the co-occurrence frequency of F_j with *w* when translated as T_i, denoted as $\text{cooc\_frequency}(F_j, T_i)$,

[1] http://www.iwslt2011.org/

[2] The translation probabilities between word tokens are found in the translation table produced by GIZA++; the threshold is set to 0.01.

[3] For this filtering, we employ a PoS and lemmatization lexicon built after tagging both parts of the training corpus with TreeTagger (Schmid, 1994).

[4] We do not consider hapax translations because they often correspond to alignment errors.

and the total number of features (N) seen with T_i:

$$p_{ij} = \frac{\text{cooc\_frequency}(F_j, T_i)}{N} \quad (3)$$

Finally, the local weight $\text{lw}(F_j, T_i)$ between F_j and T_i directly depends on their co-occurrence frequency:

$$\text{lw}(F_j, T_i) = \log(\text{cooc\_frequency}(F_j, T_i)) \quad (4)$$

3.2 Cross-Lingual WSD

The weighted feature vectors corresponding to the different translations of an English word are used for disambiguation.[5] As noted in Section 3.1, we disambiguate source words satisfying a set of criteria. Disambiguation is performed by comparing the vector associated with each translation to the new context of the words in the input sentences from the IWSLT'11 test set.

More precisely, the information contained in each vector is exploited by the WSD classifier to produce a probability distribution over the translations, for each new instance of a word in context. We disambiguate word forms (not lemmas) in order to directly use the selected translations in the translated texts. However, we should note that in some cases this reduces the role of WSD to distinguishing between different forms of one word and no different senses are involved. Using more abstract representations (corresponding to senses) is one of the perspectives of this work.

The classifier assigns a score to each translation by comparing information in the corresponding source vector to information found in the new context. Given that the vector features are lemmatized, the new context is lemmatized as well and the lemmas of the content words are gathered in a bag of words. The adequacy of each translation for a new instance of a word is estimated by comparing the translation's vector with the bag of words built from the new context. If common features are found between the new context and a translation vector, an association score is calculated corresponding to the mean of the weights of the common features relatively to the translation (i.e. found in its vector). In Equation (5), $(CF_j)_{j=1}^{|CF|}$ is the set of common features between the translation vector V_i and the new context C and tw is the weight of a CF with translation T_i (cf. formula (1)).

$$\text{assoc\_score}(V_i, C) = \frac{\sum_{j=1}^{|CF|} \text{tw}(CF_j, T_i)}{|CF|} \quad (5)$$

The scores assigned to the different translations of a source word are normalized to sum up to one.

In this way, a subset of the words that occur in the input sentences from the test set are annotated with their translations and the associated scores (contextual probabilities), as shown in the example in Figure 1.[6] The WSD classifier makes predictions only for the subset of the words found in the source part of the parallel test set that were retained from the initial EN-FR lexicon after filtering. Table 1 presents the total coverage of the WSD method as well as its coverage for words of different PoS, with a focus on content words. We report the number of disambiguated words for each content PoS (cf. third column) and the corresponding percentage, calculated on the basis of the total number of words pertaining to this PoS (cf. second column). We observe that the coverage of the method on nouns and adjectives is higher than the one on verbs. Given the rich verbal morphology of French, several verbs have a very high number of translations in the bilingual lexicon (over 20) and are not handled during disambiguation. The same applies to function words (articles, prepositions, conjunctions, etc.) included in the 'all PoS' category.

4 Integrating Semantics into SMT

In this section, we present two ways to integrate WSD predictions into an SMT decoder. The first one (Section 4.1) is a simple method based on *n*-best reranking. This method, already proposed in the literature (Specia et al., 2008), allows us to easily evaluate the impact of WSD predictions on automatic translation quality. The second one (Section 4.2) builds on the idea, introduced in (Crego et al., 2010), of using an additional language model to

[5] The vectors are not used for clustering the translations as in Apidianaki (2009) but all translations are considered as candidate senses.

[6] Some source words are tagged with only one translation (e.g. *stones*_{*pierres*(1.000)}) because their other translations in the lexicon occurred only once in the training corpus and, consequently, were not considered.

PoS	# of words	# of WSD predictions	%
Nouns	5535	3472	62.72
Verbs	5336	1269	23.78
Adjs	1787	1249	69.89
Advs	2224	1098	49.37
all content PoS	14882	7088	47.62
all PoS	27596	8463	30.66

Table 1: Coverage of the WSD method

you know, one of the intense_{intenses(0.305), forte(0.306), intense(0.389)} pleasures of travel_{transport(0.334), voyage(0.332), voyager(0.334)} and one of the delights of ethnographic research_{recherche(0.225), research(0.167), études(0.218), recherches(0.222), étude(0.167)} is the opportunity_{possibilité(0.187), chance(0.185), opportunités(0.199), occasion(0.222), opportunité(0.207)} to live amongst those who have not forgotten_{oubli(0.401), oubliés(0.279), oubliée(0.321)} the old_{ancien(0.079), âge(0.089), anciennes(0.072), âgées(0.100), âgés(0.063), ancienne(0.072), vieille(0.093), ans(0.088), vieux(0.086), vieil(0.078), anciens(0.081), vieilles(0.099)} ways_{façons(0.162), manières(0.140), moyens(0.161), aspects(0.113), façon(0.139), moyen(0.124), manière(0.161)} , who still feel their past_{passée(0.269), autrefois(0.350), passé(0.381)} in the wind_{éolienne(0.305), vent(0.392), éoliennes(0.304)} , touch_{touchent(0.236), touchez(0.235), touche(0.235), toucher(0.293)} it in stones_{pierres(1.000)} polished by rain_{pluie(1.000)} , taste_{goût(0.500), goûter(0.500)} it in the bitter_{amer(0.360), amère(0.280), amertume(0.360)} leaves_{feuilles(0.500), feuillages(0.500)} of plants_{usines(0.239), centrales(0.207), plantes(0.347), végétaux(0.207)}.

Figure 1: Input sentence with WSD information

directly integrate the prediction of the WSD system into the decoder.

4.1 *N*-best List Reranking

A simple way to influence translation hypotheses selection with WSD information is to use the WSD probabilities of translation variants to produce an additional feature appended to the *n*-best list after its generation. The feature value should reflect the degree to which a particular hypothesis includes proposed WSD variants for the respective words. Rerunning the standard MERT optimization procedure on the augmented features gives a new set of model weights, that are used to rescore the *n*-best list.

We propose the following method of features construction. Given the phrase alignment information between a source sentence and a hypothesis, we verify if one or more of the proposed WSD variants for the source word occur in the corresponding phrase of the translation hypothesis. If this is the case, the corresponding probabilities are additively accumulated for the current hypothesis. At the end, two features are appended to each hypothesis in the *n*-best list: the total score accumulated for the hypothesis and the same score normalized by the number of words in the hypothesis.

Two MERT initialization schemes were considered: (1) all model weights are initialized to zero, and (2) all the weights of "standard" features are initialized to the values found by MERT and the new WSD features to zero.

4.2 Local Language Models

We propose to adapt the approach introduced in Crego et al. (2010) as an alternative way to integrate the WSD predictions within the decoder: for each sentence to be translated, an additional language model (LM) is estimated and taken into account during decoding. As this additional "local" model depends on the source sentence, it can be used as an external source of knowledge to reinforce translation hypotheses complying with criteria predicted from the whole source sentence. For instance, the unigram probabilities of the additional LM can be derived from the (word) predictions of a WSD system, bigram probabilities from the prediction of phrases and so on and so forth. Although this approach was suggested in (Crego et al., 2010), this

is, to the best of our knowledge, the first time it is experimentally validated.

In practice, the predictions of the WSD system described in Section 3 can be integrated by defining, for each sentence, an additional unigram language model as follows:

- each translation predicted by the WSD classifier can be generated by the language model with the probability estimated by the WSD classifier; no information about the source word that has been disambiguated is considered;
- the probability of unknown words is set to a small arbitrary constant.

Even if most of the words composing the translation hypothesis are considered as unknown words, hypotheses that contain the words predicted by the WSD system still have a higher LM score and are therefore preferred. Note that even if we only use unigram language models in our experiments, as senses are predicted at the word level, our approach is able to handle disambiguation of phrases as well.

This approach has two main advantages over existing ways to integrate WSD predictions in an SMT system. First, no hard decisions are made: errors of the WSD can be "corrected" by the translation. Second, sense disambiguation at the word level is naturally and automatically propagated at the phrase level: the additional LM is influencing all phrase pairs using one of the predicted words.

Compared to the reranking approach introduced in the previous section, this method results in a tighter integration with the decoder. In particular, the WSD predictions are applied before search-space pruning and are therefore expected to have a more important role.

5 Evaluation

5.1 Experimental Setting

In all our experiments, we considered the TED-talk English to French data set provided by the IWSLT'11 evaluation campaign, a collection of public speeches on a variety of topics. We used the Moses decoder (Koehn et al., 2007).

The TED-talk corpus is a small data set made of a monolingual corpus ($111,431$ sentences) used to estimate a 4-gram language model with KN-smoothing, and a bilingual corpus ($107,268$ sentences) used to extract the phrase table. All data are tokenized, cleaned and converted to lowercase letters using the tools provided by the WMT organizers.[7] We then use a standard training pipeline to construct the translation model: the bitext is aligned using GIZA++, symmetrized using the grow-diag-final-and heuristic; the phrase table is extracted and scored using the tools distributed with Moses. Finally, systems are optimized using MERT on the 934 sentences of the `dev-2010` set. All evaluations are performed on the $1,664$ sentences of the `test-2010` set.

5.2 Baseline

In addition to the models introduced in Section 4, we considered two other supplementary models as baselines. The first one uses the IBM 1 model estimated during the SMT system training as a simple WSD system: for each source sentence, a unigram additional language model is defined by taking, for each source, the 20 best translations according to the IBM 1 model and their probability. Model 1 has been shown to be one of the best performing features to be added to an SMT system in a reranking step (Och et al., 2004) and can be seen as a naive WSD classifier.

To test the validity of our approach, we replicate the "oracle" experiments of Crego et al. (2010) and estimate the best gain our method can achieve. These experiments consist in using the reference to train a local n-gram language model (with n in the range 1 to 3) which amounts, in the local language model method of Section 4.2, to assuming that the WSD system correctly predicted a single translation for each source word.

5.3 Results

Table 2 reports the results of our experiments. It appears that, for the considered task, sense disambiguation improves translation performance: n-best rescoring results in a 0.37 BLEU improvement and using an additional language model brings about an improvement of up to a 0.88 BLEU. In both cases, MERT assigns a large weight to the additional fea-

[7] `http://statmt.org/wmt08/scripts.tgz`

method		BLEU	METEOR
baseline	—	29.63	53.78
rescoring	WSD (zero init)	30.00	54.26
	WSD (reinit)	29.58	53.96
additional LM	oracle 3-gram	43.56	64.64
	oracle 2-gram	39.36	62.92
	oracle 1-gram	42.92	69.39
	IBM 1	30.18	54.36
	WSD	30.51	54.38

Table 2: Evaluation results on the TED-talk task of our two methods to integrate WSD predictions.

PoS	baseline	WSD
Nouns	67.57	69.06
Verbs	45.97	47.76
Adjectives	51.79	53.94
Adverbs	52.17	56.25

Table 3: Contrastive lexical evaluation: % of words correctly translated within each PoS class

tures during tuning. When rescoring *n*-best, an improvement is observed only when the weights are initialized to zero and not to the weights resulting from the previous optimization, maybe because of the difficulty to exit the local minimum MERT had found earlier.

As expected, integrating the WSD predictions with an additional language model results in a larger improvement than simple rescoring, which shows the importance of applying this new source of information early in the translation pipeline, before search space pruning. Also note that the system using the IBM 1 predictions is outperformed by the system using the WSD classifier introduced in Section 3, showing the quality of its predictions.

Oracle experiments stress the high potential of the method introduced in (Crego et al., 2010) as a way to integrate external sources of knowledge: all three conditions result in large improvements over the baseline and the proposed methods. It must, however, be noted that contrary to the WSD method introduced in Section 3, these oracle experiments rely on sense predictions for all source words and not only content words. Surprisingly enough, predicting phrases instead of words results only in a small improvement. Additional experiments are required to explain why 2-gram oracle achieved such a low performance.

5.4 Contrastive lexical evaluation

All the measures used for evaluating the impact of WSD information on translation show improvements, as discussed in the previous section. We complement these results with another measure of translation performance, proposed by Max et al. (2010), which allows for a more fine-grained contrastive evaluation of the translations produced by different systems. The method permits to compare the results produced by the systems on different word classes and to take into account the source words that were actually translated. We focus this evaluation on the classes of content words (nouns, adjectives, verbs and adverbs) on which WSD had an important coverage. Our aim is, first, to explore how these words are handled by a WSD-informed SMT system (the system using the local language models) compared to the baseline system that does not exploit any semantic information; and, second, to investigate whether their disambiguation influences the translation of surrounding non-disambiguated words.

Table 3 reports the percentage of words correctly translated by the semantically-informed system within each content word class: consistent gains in translation quality are observed for all parts-of-speech compared to the baseline, and the best results are obtained for nouns.

	baseline				WSD			
	w_{-2}	w_{-1}	w_{+1}	w_{+2}	w_{-2}	w_{-1}	w_{+1}	w_{+2}
Nouns	64.01	68.69	75.17	64.6	65.47	70.46	76.3	66.6
Verbs	68.67	67.58	63	62.19	69.98	68.89	64.85	64.25
Adjectives	63.1	64.39	64.28	66.55	64.09	65.65	64.76	69.33
Adverbs	70.8	69.44	68.67	66.38	71	71.21	70	67.22

Table 4: Impact of WSD prediction on the surrounding words

Table 4 shows how the words surrounding a disambiguated word *w* (noun, verb, adjective or adverb) in the text are handled by the two systems. More precisely, we look at the translation of words in the immediate context of *w*, i.e. at positions w_{-2}, w_{-1}, w_{+1} and w_{+2}. The left column reports the percentage of correct translations produced by the baseline system (without disambiguation) for words in these positions; the right column shows the positive impact that the disambiguation of a word has on the translation of its neighbors. Note that this time we look at disambiguated words and their context without evaluating the correctness of the WSD predictions. Nevertheless, even in this case, consistent gains are observed when WSD information is exploited. For instance, when a noun is disambiguated, 70.46% and 76.3% of the immediately preceding (w_{-1}) and following (w_{+1}) words, respectively, are correctly translated, versus 68.69% and 75.17% of correct translations produced by the baseline system.

6 Conclusion and future work

The preliminary results presented in this paper on integrating cross-lingual WSD into a state-of-the-art SMT system are encouraging. Both adopted approaches (*n*-best rescoring and local language modeling) benefit from the predictions of the proposed cross-lingual WSD classifier. The contrastive evaluation results further show that WSD improves not only the translation of disambiguated words, but also the translation of neighboring words in the input texts.

We consider various ways for extending this work. First, future experiments will involve the use of more abstract representations of senses than individual translations, by applying a cross-lingual word sense induction method to the training corpus prior to disambiguation. We will also experiment with disambiguation at the level of lemmas, to reduce sparseness issues, and with different ways for handling lemmatized predictions by the SMT systems. Furthermore, we intend to extend the coverage of the WSD method by exploring other filtering methods for cleaning the alignment lexicons, and by addressing the disambiguation of words of all PoS.

Acknowledgments

This work was partly funded by the European Union under the FP7 project META-NET (T4ME), Contract No. 249119, and by OSEO, the French agency for innovation, as part of the Quaero Program.

References

Marianna Apidianaki. 2008. Translation-oriented Word Sense Induction Based on Parallel Corpora. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

Marianna Apidianaki. 2009. Data-driven Semantic Analysis for Multilingual WSD and Lexical Selection in Translation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-09)*, pages 77–85, Athens, Greece.

Marine Carpuat and Dekai Wu. 2005. Word Sense Disambiguation vs. Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 387–394, Ann Arbor, Michigan.

Marine Carpuat and Dekai Wu. 2007. Improving Statistical Machine Translation using Word Sense Disambiguation. In *Proceedings of the Joint EMNLP-CoNLL Conference*, pages 61–72, Prague, Czech Republic.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word Sense Disambiguation Improves Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-07)*, pages 33–40, Prague, Czech Republic.

Josep Maria Crego, Aurélien Max, and François Yvon. 2010. Local lexical adaptation in Machine Translation through triangulation: SMT helping SMT. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 232–240, Beijing, China.

Gregory Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Norwell, MA.

Rejwanual Haque, Sudip Naskar, Yanjun Ma, and Andy Way. 2009. Using supertags as source language context in SMT. In *Proceedings of the 13th Annual Meeting of the European Association for Machine Translation (EAMT 2009)*, pages 234–241, Barcelona, Spain.

Rejwanul Haque, Sudip Kumar Naskar, Antal Van Den Bosch, and Andy Way. 2010. Supertags as source language context in hierarchical phrase-based SMT. In *Proceedings of AMTA 2010: The Ninth Conference of the Association for Machine Translation in the Americas*, pages 210–219, Denver, CO.

N. Ide and Y. Wilks. 2007. Making Sense About Sense. In E. Agirre and P. Edmonds, editors, *Word Sense Disambiguation, Algorithms and Applications*, pages 47–73. Springer.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual ACL Meeting, Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

Els Lefever and Veronique Hoste. 2010. SemEval-2010 Task 3: Cross-lingual Word Sense Disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2), ACL 2010*, pages 15–20, Uppsala, Sweden.

Arne Mauser, Saša Hasan, and Hermann Ney. 2009. Extending statistical machine translation with discriminative and trigger-based lexicon models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 210–217, Singapore, August.

Aurélien Max, Josep Maria Crego, and François Yvon. 2010. Contrastive Lexical Evaluation of Machine Translation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

Rada Mihalcea, Ravi Sinha, and Diana McCarthy. 2010. SemEval-2010 Task 2: Cross-Lingual Lexical Substitution. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2), ACL 2010*, pages 9–14, Uppsala, Sweden.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, and Dragomir Radev. 2004. A smorgasbord of features for statistical machine translation. In *Proceedings of HLT-NAACL 2004*, pages 161–168, Boston, Massachusetts, USA.

Alexandre Patry and Philippe Langlais. 2011. Going beyond word cooccurrences in global lexical selection for statistical machine translation using a multilayer perceptron. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 658–666, Chiang Mai, Thailand, November.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Lucia Specia, Baskaran Sankaran, and Maria Das Graças Volpe Nunes. 2008. n-Best Reranking for the Efficient Integration of Word Sense Disambiguation and Statistical Machine Translation. In *Proceedings of the 9th international conference on Computational linguistics and intelligent text processing*, CICLing'08, pages 399–410, Berlin, Heidelberg. Springer-Verlag.

Lucia 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.

David Vickrey, Luke Biewald, Marc Teyssier, and Daphne Koller. 2005. Word-Sense Disambiguation for Machine Translation. In *Proceedings of the Joint Conference on Human Language Technology / Empirical Methods in Natural Language Processing (HLT-EMNLP)*, pages 771–778, Vancouver, Canada.