Word Sense Disambiguation vs. Statistical Machine Translation

Marine CARPUAT **Dekai WU**[1]
marine@cs.ust.hk dekai@cs.ust.hk

Human Language Technology Center
HKUST
Department of Computer Science
University of Science and Technology
Clear Water Bay, Hong Kong

Abstract

We directly investigate a subject of much recent debate: do word sense disambigation models help statistical machine translation quality? We present empirical results casting doubt on this common, but unproved, assumption. Using a state-of-the-art Chinese word sense disambiguation model to choose translation candidates for a typical IBM statistical MT system, we find that word sense disambiguation does *not* yield significantly better translation quality than the statistical machine translation system alone. Error analysis suggests several key factors behind this surprising finding, including inherent limitations of current statistical MT architectures.

1 Introduction

Word sense disambiguation or WSD, the task of determining the correct sense of a word in context, is a much studied problem area with a long and honorable history. Recent years have seen steady accuracy gains in WSD models, driven in particular by controlled evaluations such as the Senseval series of workshops. Word sense disambiguation is often assumed to be an intermediate task, which should then help higher level applications such as machine translation or information retrieval. However, WSD is usually performed and evaluated as a standalone task, and to date there have been very few efforts to integrate the learned WSD models into full statistical MT systems.

An energetically debated question at conferences over the past year is whether even the new state-of-the-art word sense disambiguation models actually have anything to offer to full statistical machine translation systems. Among WSD circles, this can sometimes elicit responses that border on implying that even asking the question is heretical. In efforts such as Senseval we tend to regard the construction of WSD models as an obviously correct, if necessarily simplified, approach that will eventually lead to essential disambiguation components within larger applications like machine translation.

There is no question that the word sense disambiguation perspective has led to numerous insights in machine translation, even of the statistical variety. It is often simply an unstated assumption that any full translation system, to achieve full performance, will sooner or later have to incorporate individual WSD components.

However, in some translation architectures and particularly in statistical machine translation (SMT), the translation engine already implicitly factors in many contextual features into lexical choice. From this standpoint, SMT models can be seen as WSD models in their own right, albeit with several major caveats.

But typical statistical machine translation models only rely on a local context to choose among lexical translation candidates, as discussed in greater detail later. It is therefore often assumed that dedicated WSD-based lexical choice models, which can incor-

[1]The authors would like to thank the Hong Kong Research Grants Council (RGC) for supporting this research in part through grants RGC6083/99E, RGC6256/00E, and DAG03/04.EG09, and several anonymous reviewers for insights and suggestions.

Proceedings of the 43rd Annual Meeting of the ACL, pages 387–394,
Ann Arbor, June 2005.

porate a wider variety of context features, can make better predictions than the "weaker" models implicit in statistical MT, and that these predictions will help the translation quality.

Nevertheless, this assumption has not been empirically verified, and we should not simply assume that WSD models can contribute more than what the SMT models perform. It may behoove us to take note of the sobering fact that, perhaps analogously, WSD has yet to be conclusively shown to help information retrieval systems after many years of attempts.

In this work, we propose to directly investigate whether word sense disambiguation—at least as it is typically currently formulated—is useful for statistical machine translation. We tackle a real Chinese to English translation task using a state-of-the-art supervised WSD system and a typical SMT model. We show that the unsupervised SMT model, trained on parallel data without any manual sense annotation, yields higher BLEU scores than the case where the SMT model makes use of the lexical choice predictions from the supervised WSD model, which are more expensive to create. The reasons for the surprising difficulty of improving over the translation quality of the SMT model are then discussed and analyzed.

2 Word sense disambiguation vs. statistical machine translation

We begin by examining the respective strengths and weaknesses of dedicated WSD models versus full SMT models, that could be expected to be relevant to improving lexical choice.

2.1 Features Unique to WSD

Dedicated WSD is typically cast as a classification task with a predefined sense inventory. Sense distinctions and granularity are often manually predefined, which means that they can be adapted to the task at hand, but also that the translation candidates are limited to an existing set.

To improve accuracy, dedicated WSD models typically employ features that are not limited to the local context, and that include more linguistic information than the surface form of words. This often requires several stages of preprocessing, such as part-of-speech tagging and/or parsing. (Preprocessor domain can be an issue, since WSD accuracy may suffer from domain mismatches between the data the preprocessors were trained on, and the data they are applied to.) For example, a typical dedicated WSD model might employ features as described by Yarowsky and Florian (2002) in their "feature-enhanced naive Bayes model", with position-sensitive, syntactic, and local collocational features. The feature set made available to the WSD model to predict lexical choices is therefore much richer than that used by a statistical MT model.

Also, dedicated WSD models can be supervised, which yields significantly higher accuracies than unsupervised. For the experiments described in this study we employed supervised training, exploiting the annotated corpus that was produced for the Senseval-3 evaluation.

2.2 Features Unique to SMT

Unlike lexical sample WSD models, SMT models simultaneously translate complete sentences rather than isolated target words. The lexical choices are made in a way that heavily prefers *phrasal cohesion* in the output target sentence, as scored by the language model. That is, the predictions benefit from the sentential context of the *target* language. This has the general effect of improving translation fluency.

The WSD accuracy of the SMT model depends critically on the phrasal cohesion of the target language. As we shall see, this phrasal cohesion property has strong implications for the utility of WSD.

In other work (forthcoming), we investigated the inverse question of evaluating the Chinese-to-English SMT model on word sense disambiguation performance, using standard WSD evaluation methodology and datasets from the Senseval-3 Chinese lexical sample task. We showed the accuracy of the SMT model to be significantly lower than that of all the dedicated WSD models considered, even after adding the lexical sample data to the training set for SMT to allow for a fair comparison. These results highlight the relative strength, and the potential hoped-for advantage of dedicated supervised WSD models.

3 The WSD system

The WSD system used for the experiments is based on the model that achieved the best performance, by a large margin, on the Senseval-3 Chinese lexical sample task (Carpuat *et al.*, 2004).

3.1 Classification model

The model consists of an ensemble of four voting models combined by majority vote.

The first voting model is a naive Bayes model, since Yarowsky and Florian (2002) found this model to be the most accurate classifier in a comparative study on a subset of Senseval-2 English lexical sample data.

The second voting model is a maximum entropy model (Jaynes, 1978), since Klein and Manning (2002) found that this model yielded higher accuracy than naive Bayes in a subsequent comparison of WSD performance. (Note, however, that a different subset of either Senseval-1 or Senseval-2 English lexical sample data was used for their comparison.)

The third voting model is a boosting model (Freund and Schapire, 1997), since has consistently turned in very competitive scores on related tasks such as named entity classification (Carreras *et al.*, 2002) . Specifically, an AdaBoost.MH model was used (Schapire and Singer, 2000), which is a multiclass generalization of the original boosting algorithm, with boosting on top of decision stump classifiers (i.e., decision trees of depth one).

The fourth voting model is a Kernel PCA-based model (Wu *et al.*, 2004). Kernel Principal Component Analysis (*KPCA*) is a nonlinear kernel method for extracting nonlinear principal components from vector sets where, conceptually, the n-dimensional input vectors are nonlinearly mapped from their original space R^n to a high-dimensional feature space F where linear PCA is performed, yielding a transform by which the input vectors can be mapped nonlinearly to a new set of vectors (Schölkopf *et al.*, 1998). WSD can be performed by a Nearest Neighbor Classifier in the high-dimensional KPCA feature space. (Carpuat *et al.*, 2004) showed that KPCA-based WSD models achieve close accuracies to the best individual WSD models, while having a significantly different bias.

All these classifiers have the ability to handle large numbers of sparse features, many of which may be irrelevant. Moreover, the maximum entropy and boosting models are known to be well suited to handling features that are highly interdependent.

The feature set used consists of position-sensitive, syntactic, and local collocational features, as described by Yarowsky and Florian (2002).

3.2 Lexical choice mapping model

Ideally, we would like the WSD model to predict English translations given Chinese target words in context. Such a model requires Chinese training data annotated with English senses, but such data is not available. Instead, the WSD system was trained using the Senseval-3 Chinese lexical sample task data. (This is suboptimal, but reflects the difficulties that arise when considering a real translation task; we cannot assume that sense-annotated data will always be available for all language pairs.)

The Chinese lexical sample task includes 20 target words. For each word, several senses are defined using the HowNet knowledge base. There are an average of 3.95 senses per target word type, ranging from 2 to 8. Only about 37 training instances per target word are available.

For the purpose of Chinese to English translation, the WSD model should predict English translations instead of HowNet senses. Fortunately, HowNet provides English glosses. This allows us to map each HowNet sense candidate to a set of English translations, converting the monolingual Chinese WSD system into a translation lexical choice model. We further extended the mapping to include any significant translation choice considered by the SMT system but not in HowNet.

4 The SMT system

To build a representative baseline statistical machine translation system, we restricted ourselves to making use of freely available tools, since the potential contribution of WSD should be easier to see against this baseline. Note that our focus here is not on the SMT model itself; our aim is to evaluate the impact of WSD on a real Chinese to English statistical machine translation task.

Table 1: Example of the translation candidates before and after mapping for the target word "路" (*lu*)

HowNet Sense ID	HowNet glosses	HowNet glosses + improved translations
56520	distance	distance
56521	sort	sort
56524	Lu	Lu
56525, 56526, 56527, 56528	path, road, route, way	path, road, route, way, circuit, roads
56530, 56531, 56532	line, means, sequence	line, means, sequence, lines
56533, 56534	district, region	district, region

4.1 Alignment model

The alignment model was trained with GIZA++ (Och and Ney, 2003), which implements the most typical IBM and HMM alignment models. Translation quality could be improved using more advanced hybrid phrasal or tree models, but this would interfere with the questions being investigated here. The alignment model used is IBM-4, as required by our decoder. The training scheme consists of IBM-1, HMM, IBM-3 and IBM-4, following (Och and Ney, 2003).

The training corpus consists of about 1 million sentences from the United Nations Chinese-English parallel corpus from LDC. This corpus was automatically sentence-aligned, so the training data does not require as much manual annotation as for the WSD model.

4.2 Language model

The English language model is a trigram model trained on the Gigaword newswire data and on the English side of the UN and Xinhua parallel corpora. The language model is also trained using a publicly available software, the CMU-Cambridge Statistical Language Modeling Toolkit (Clarkson and Rosenfeld, 1997).

4.3 Decoding

The ISI ReWrite decoder (Germann, 2003), which implements an efficient greedy decoding algorithm, is used to translate the Chinese sentences, using the alignment model and language model previously described.

Notice that very little contextual information is available to the SMT models. Lexical choice during decoding essentially depends on the translation probabilities learned for the target word, and on the English language model scores.

5 Experimental method

5.1 Test set selection

We extracted the Chinese sentences from the NIST MTEval-04 test set that contain any of the 20 target words from the Senseval-3 Chinese lexical sample target set. For a couple of targets, no instances were available from the test set. The resulting test set contains a total of 175 sentences, which is smaller than typical MT evaluation test sets, but slightly larger than the one used for the Senseval Chinese lexical sample task.

5.2 Integrating the WSD system predictions with the SMT model

There are numerous possible ways to integrate the WSD system predictions with the SMT model. We choose two different straightforward approaches, which will help analyze the effect of the different components of the SMT system, as we will see in Section 6.5.

5.2.1 Using WSD predictions for decoding

In the first approach, we use the WSD sense predictions to constrain the set of English sense candidates considered by the decoder for each of the target words. Instead of allowing all the word translation candidates from the translation model, when we use the WSD predictions we override the translation model and force the decoder to choose the best translation from the predefined set of glosses that maps to the HowNet sense predicted by the WSD model.

Table 2: Translation quality with and without the WSD model

Translation System	BLEU score
SMT	0.1310
SMT + WSD for postprocessing	0.1253
SMT + WSD for decoding	0.1239
SMT + WSD for decoding with improved translation candidates	0.1232

5.2.2 Using WSD predictions for postprocessing

In the second approach, we use the WSD predictions to postprocess the output of the SMT system: in each output sentence, the translation of the target word chosen by the SMT model is directly replaced by the WSD prediction. When the WSD system predicts more than one candidate, a unique translation is randomly chosen among them. As discussed later, this approach can be used to analyze the effect of the language model on the output.

It would also be interesting to use the gold standard or correct sense of the target words instead of the WSD model predictions in these experiments. This would give an upper-bound on performance and would quantify the effect of WSD errors. However, we do not have a corpus which contains both sense annotation and multiple reference translations: the MT evaluation corpus is not annotated with the correct senses of Senseval target words, and the Senseval corpus does not include English translations of the sentences.

6 Results

6.1 Even state-of-the-art WSD does not help BLEU score

Table 2 summarizes the translation quality scores obtained with and without the WSD model. Using our WSD model to constrain the translation candidates given to the decoder hurts translation quality, as measured by the automated BLEU metric (Papineni *et al.*, 2002).

Note that we are evaluating on only difficult sentences containing the problematic target words from the lexical sample task, so BLEU scores can be expected to be on the low side.

6.2 WSD still does not help BLEU score with improved translation candidates

One could argue that the translation candidates chosen by the WSD models do not help because they are only glosses obtained from the HowNet dictionary. They consist of the root form of words only, while the SMT model can learn many more translations for each target word, including inflected forms and synonyms.

In order to avoid artificially penalizing the WSD system by limiting its translation candidates to the HowNet glosses, we expand the translation set using the bilexicon learned during translation model training. For each target word, we consider the English words that are given a high translation probability, and manually map each of these English words to the sense categories defined for the Senseval model. At decoding time, the set of translation candidates considered by the language model is therefore larger, and closer to that considered by the pure SMT system.

The results in Table 2 show that the improved translation candidates do not help BLEU score. The translation quality obtained with SMT alone is still better than when the improved WSD Model is used. The simpler approach of using WSD predictions in postprocessing yields better BLEU score than the decoding approach, but still does not outperform the SMT model.

6.3 WSD helps translation quality for very few target words

If we break down the test set and evaluate the effect of the WSD per target word, we find that for all but two of the target words WSD either hurts the BLEU score or does not help it, which shows that the decrease in BLEU is not only due to a few isolated target words for which the Senseval sense distinctions

are not helpful.

6.4 The "language model effect"

Error analysis revealed some surprising effects. One particularly dismaying effect is that even in cases where the WSD model is able to predict a better target word translation than the SMT model, to use the better target word translation surprisingly often still leads to a lower BLEU score.

The phrasal coherence property can help explain this surprising effect we observed. The translation chosen by the SMT model will tend to be more likely than the WSD prediction according to the language model; otherwise, it would also have been predicted by SMT. The translation with the higher language model probability influences the translation of its neighbors, thus potentially improving BLEU score, while the WSD prediction may not have been seen occurring within phrases often enough, thereby lowering BLEU score.

For example, we observe that the WSD model sometimes correctly predicts "impact" as a better translation for "冲击" (*chongji*), where the SMT model selects "shock". In these cases, some of the reference translations also use "impact". However, even when the WSD model constrains the decoder to select "impact" rather than "shock", the resulting sentence translation yields a lower BLEU score. This happens because the SMT model does not know how to use "impact" correctly (if it did, it would likely have chosen "impact" itself). Forcing the lexical choice "impact" simply causes the SMT model to generate phrases such as "against Japan for peace constitution impact" instead of "against Japan for peace constitution shocks". This actually lowers BLEU score, because of the n-gram effects.

6.5 Using WSD predictions in postprocessing does not help BLEU score either

In the postprocessing approach, decoding is done before knowing the WSD predictions, which eliminates the "language model effect". Even in these conditions, the SMT model alone is still the best performing system.

The postprocessing approach also outperforms the integrated decoding approach, which shows that the language model is not able to make use of the WSD predictions. One could expect that letting the decoder choose among the WSD translations also yields a better translation of the context. This is indeed the case, but for very few examples only: for instance the target word "地方" (*difang*) is better used in the integrated decoding ouput "the place of local employment" , than in the postprocessing output "the place employment situation". Instead, the majority of cases follow the pattern illustrated by the following example where the target word is "老" (*lao*): the SMT system produces the best output ("the newly elected President will still face old problems"), the postprocessed output uses the fluent sentence with a different translation ("the newly elected President will still face outdated problems"), while the translation is not used correctly with the decoding approach ("the newly elected President will face problems still to be outdated").

Table 3: BLEU scores per target word: WSD helps for very few target words

Target word	**SMT**	**SMT + WSD**
把握 bawo	0.1482	0.1484
包 bao	0.1891	0.1891
材料 cailiao	0.0863	0.0863
冲击 chongji	0.1396	0.1491
地方 difang	0.1233	0.1083
分子 fengzi	0.1404	0.1402
活动 huodong	0.1365	0.1465
老 lao	0.1153	0.1136
路 lu	0.1322	0.1208
起来 qilai	0.1104	0.1082
钱 qian	0.1948	0.1814
突出 tuchu	0.0975	0.0989
研究 yanjiu	0.1089	0.1089
运动 zhengdong	0.1267	0.1251
走 zhou	0.0825	0.0808

6.6 BLEU score bias

The "language model effect" highlights one of the potential weaknesses of the BLEU score. BLEU penalizes for phrasal incoherence, which in the present study means that it can sometimes sacrifice adequacy for fluency.

However, the characteristics of BLEU are by

no means solely responsible for the problems with WSD that we observed. To doublecheck that n-gram effects were not unduly impacting our study, we also evaluated using BLEU-1, which gave largely similar results as the standard BLEU-4 scores reported above.

7 Related work

Most translation disambiguation tasks are defined similarly to the Senseval Multilingual lexical sample tasks. In Senseval-3, the English to Hindi translation disambigation task was defined identically to the English lexical sample task, except that the WSD models are expected to predict Hindi translations instead of WordNet senses. This differs from our approach which consists of producing the translation of complete sentences, and not only of a predefined set of target words.

Brown *et al.* (1991) proposed a WSD algorithm to disambiguate English translations of French target words based on the single most informative context feature. In a pilot study, they found that using this WSD method in their French-English SMT system helped translation quality, manually evaluated using the number of acceptable translations. However, this study is limited to the unrealistic case of words that have exactly two senses in the other language.

Most previous work has focused on the distinct problem of exploiting various bilingual resources (e.g., parallel or comparable corpora, or even MT systems) to help WSD. The goal is to achieve accurate WSD with minimum amounts of annotated data. Again, this differs from our objective which consists of using WSD to improve performance on a full machine translation task, and is measured in terms of translation quality.

For instance, Ng *et al.* (2003) showed that it is possible to use word aligned parallel corpora to train accurate supervised WSD models. The objective is different; it is not possible for us to use this method to train our WSD model without undermining the question we aim to investigate: we would need to use the SMT model to word-align the parallel sentences, which could too strongly bias the predictions of the WSD model towards those of the SMT model, instead of combining predictive information from independent sources as we aim to study here.

Other work includes Li and Li (2002) who propose a bilingual bootstrapping method to learn a translation disambiguation WSD model, and Diab (2004) who exploited large amounts of automatically generated noisy parallel data to learn WSD models in an unsupervised bootstrapping scheme.

8 Conclusion

The empirical study presented here argues that we can expect that it will be quite difficult, at the least, to use standard WSD models to obtain significant improvements to statistical MT systems, even when supervised WSD models are used. This casts significant doubt on a commonly-held, but unproven, assumption to the contrary. We have presented empirically based analysis of the reasons for this surprising finding.

We have seen that one major factor is that the statistical MT model is sufficiently accurate so that within the training domain, even the state-of-the-art dedicated WSD model is only able to improve on its lexical choice predictions in a relatively small proportion of cases.

A second major factor is that even when the dedicated WSD model makes better predictions, current statistical MT models are unable to exploit this. Under this interpretation of our results, the dependence on the language model in current SMT architectures is excessive. One could of course argue that drastically increasing the amount of training data for the language model might overcome the problems from the language model effect. Given combinatorial problems, however, there is no way at present of telling whether the amount of data needed to achieve this is realistic, particularly for translation across many different domains. On the other hand, if the SMT architecture cannot make use of WSD predictions, even when they are in fact better than the SMT's lexical choices, then perhaps some alternative model striking a different balance of adequacy and fluency is called for. Ultimately, after all, WSD is a method of compensating for sparse data. Thus it may be that the present inability of WSD models to help improve accuracy of SMT systems stems not from an inherent weakness of dedicated WSD models, but rather from limitations of present-day SMT architectures.

To further test this, our experiments could be tried on other statistical MT models. For example, the WSD model's predictions could be employed in a Bracketing ITG translation model such as Wu (1996) or Zens *et al.* (2004), or alternatively they could be incorporated as features for reranking in a maximum-entropy SMT model (Och and Ney, 2002), instead of using them to constrain the sentence translation hypotheses as done here. However, the preceding discussion argues that it is doubtful that this would produce significantly different results, since the inherent problem from the "language model effect" would largely remain, causing sentence translations that include the WSD's preferred lexical choices to be discounted. For similar reasons, we suspect our findings may also hold even for more sophisticated statistical MT models that rely heavily on n-gram language models. A more grammatically structured statistical MT model that less n-gram oriented, such as the ITG based "grammatical channel" translation model (Wu and Wong, 1998), might make more effective use of the WSD model's predictions.

References

Peter Brown, Stephen Della Pietra, Vincent Della Pietra, and Robert Mercer. Word-sense disambiguation using statistical methods. In *Proceedings of 29th meeting of the Association for Computational Linguistics*, pages 264–270, Berkeley, California, 1991.

Marine Carpuat, Weifeng Su, and Dekai Wu. Augmenting ensemble classification for word sense disambiguation with a Kernel PCA model. In *Proceedings of Senseval-3, Third International Workshop on Evaluating Word Sense Disambiguation Systems*, Barcelona, July 2004. SIGLEX, Association for Computational Linguistics.

Xavier Carreras, Lluís Màrques, and Lluís Padró. Named entity extraction using AdaBoost. In Dan Roth and Antal van den Bosch, editors, *Proceedings of CoNLL-2002*, pages 167–170, Taipei, Taiwan, 2002.

Philip Clarkson and Ronald Rosenfeld. Statistical language modeling using the CMU-Cambridge toolkit. In *Proceedings of Eurospeech '97*, pages 2707–2710, Rhodes, Greece, 1997.

Mona Diab. Relieving the data acquisition bottleneck in word sense disambiguation. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, 2004.

Yoram Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *Journal of Computer and System Sciences, 55(1)*, pages 119–139, 1997.

Ulrich Germann. Greeedy decoding for statistical machine translation in almost linear time. In *Proceedings of HLT-NAACL-2003. Edmonton, AB, Canada*, 2003.

E.T. Jaynes. *Where do we Stand on Maximum Entropy?* MIT Press, Cambridge MA, 1978.

Dan Klein and Christopher D. Manning. Conditional structure versus conditional estimation in NLP models. In *Proceedings of EMNLP-2002, Conference on Empirical Methods in Natural Language Processing*, pages 9–16, Philadelphia, July 2002. SIGDAT, Association for Computational Linguistics.

Cong Li and Hang Li. Word translation disambiguation using bilingual bootstrapping. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 343–351, 2002.

Hwee Tou Ng, Bin Wang, and Yee Seng Chan. Exploiting parallel texts for word sense disambiguation: An empirical study. In *Proceedings of ACL-03, Sapporo, Japan*, pages 455–462, 2003.

Franz Och and Hermann Ney. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of ACL-02, Philadelphia*, 2002.

Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–52, 2003.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002.

Robert E. Schapire and Yoram Singer. BoosTexter: A boosting-based system for text categorization. *Machine Learning*, 39(2):135–168, 2000.

Bernhard Schölkopf, Alexander Smola, and Klaus-Rober Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5), 1998.

Dekai Wu and Hongsing Wong. Machine translation with a stochastic grammatical channel. In *Proceedings of COLING-ACL'98*, Montreal,Canada, August 1998.

Dekai Wu, Weifeng Su, and Marine Carpuat. A Kernel PCA method for superior word sense disambiguation. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, Barcelona, July 2004.

Dekai Wu. A polynomial-time algorithm for statistical machine translation. In *Proceedings of 34th Annual Meeting of the Association for Computational Linguistics*, Santa Cruz, California, June 1996.

David Yarowsky and Radu Florian. Evaluating sense disambiguation across diverse parameter spaces. *Natural Language Engineering*, 8(4):293–310, 2002.

Richard Zens, Hermann Ney, Taro Watanabe, and Eiichiro Sumita. Reordering constraints for phrase-based statistical machine translation. In *Proceedings of COLING-2004*, Geneva,Switzerland, August 2004.