Improving Statistical Machine Translation using Word Sense Disambiguation

Marine CARPUAT marine@cs.ust.hk

Dekai WU* dekai@cs.ust.hk

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

Abstract

We show for the first time that incorporating the predictions of a word sense disambiguation system within a typical phrase-based statistical machine translation (SMT) model consistently improves translation quality across *all* three different IWSLT Chinese-English test sets, as well as producing statistically significant improvements on the larger NIST Chinese-English MT task—and moreover *never* hurts performance on any test set, according not only to BLEU but to *all eight* most commonly used automatic evaluation metrics. Recent work has challenged the assumption that word sense disambiguation (WSD) systems are useful for SMT. Yet SMT translation quality still obviously suffers from inaccurate lexical choice. In this paper, we address this problem by investigating a new strategy for integrating WSD into an SMT system, that performs *fully phrasal multi-word* disambiguation. Instead of directly incorporating a Senseval-style WSD system, we redefine the WSD task to match the exact same phrasal translation disambiguation task faced by phrase-based SMT systems. Our results provide the first known empirical evidence that lexical semantics are indeed useful for SMT, despite claims to the contrary.

1 Introduction

Common assumptions about the role and usefulness of word sense disambiguation (WSD) models in full-scale statistical machine translation (SMT) systems have recently been challenged.

On the one hand, in previous work (Carpuat and Wu, 2005b) we obtained disappointing results when using the predictions of a Senseval WSD system in conjunction with a standard word-based SMT system: we reported slightly lower BLEU scores despite trying to incorporate WSD using a number of apparently sensible methods. These results cast doubt on the assumption that sophisticated dedicated WSD systems that were developed independently from any particular NLP application can easily be integrated into a SMT system so as to improve translation quality through stronger models of context and rich linguistic information. Rather, it has been argued, SMT systems have managed to achieve significant improvements in translation quality without directly addressing translation disambiguation as a WSD task. Instead, translation disambiguation decisions are made indirectly, typically using only word surface forms and very local contextual information, forgoing the much richer linguistic information that WSD systems typically take advantage of.

On the other hand, error analysis reveals that the performance of SMT systems still suffers from inaccurate lexical choice. In subsequent empirical studies, we have shown that SMT systems perform much worse than dedicated WSD models, both supervised

*This material is based upon work supported in part by the Defense Advanced Research Projects Agency (DARPA) under GALE Contract No. HR0011-06-C-0023, and by the Hong Kong Research Grants Council (RGC) research grants RGC6083/99E, RGC6256/00E, and DAG03/04.EG09. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency.

Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 61–72, Prague, June 2007.

and unsupervised, on a Senseval WSD task (Carpuat and Wu, 2005a), and therefore suggest that WSD should have a role to play in state-of-the-art SMT systems. In addition to the Senseval shared tasks, which have provided standard sense inventories and data sets, WSD research has also turned increasingly to designing specific models for a particular application. For instance, Vickrey *et al.* (2005) and Specia (2006) proposed WSD systems designed for French to English, and Portuguese to English translation respectively, and present a more optimistic outlook for the use of WSD in MT, although these WSD systems have not yet been integrated nor evaluated in full-scale machine translation systems.

Taken together, these seemingly contradictory results suggest that improving SMT lexical choice accuracy remains a key challenge to improve current SMT quality, and that it is still unclear what is the most appropriate integration framework for the WSD models in SMT.

In this paper, we present first results with a new architecture that integrates a state-of-the-art WSD model into phrase-based SMT so as to perform *multi-word phrasal* lexical disambiguation, and show that this new WSD approach not only produces gains across *all* available Chinese-English IWSLT06 test sets for all eight commonly used automated MT evaluation metrics, but also produces statistically significant gains on the much larger NIST Chinese-English task. The main difference between this approach and several of our earlier approaches as described in Carpuat and Wu (2005b) and subsequently Carpuat *et al.* (2006) lies in the fact that we focus on repurposing the WSD system for multi-word phrase-based SMT. Rather than using a generic Senseval WSD model as we did in Carpuat and Wu (2005b), here both the WSD training and the WSD predictions are integrated into the phrase-based SMT framework. Furthermore, rather than using a single word based WSD approach to augment a phrase-based SMT model as we did in Carpuat *et al.* (2006) to improve BLEU and NIST scores, here the WSD training and predictions operate on full multi-word phrasal units, resulting in significantly more reliable and consistent gains as evaluted by many other translation accuracy metrics as well. Specifically:

- Instead of using a Senseval system, we redefine the WSD task to be exactly the same as lexical choice task faced by the multi-word phrasal translation disambiguation task faced by the phrase-based SMT system.

- Instead of using predefined senses drawn from manually constructed sense inventories such as HowNet (Dong, 1998), our WSD for SMT system directly disambiguates between all phrasal translation candidates seen during SMT training.

- Instead of learning from manually annotated training data, our WSD system is trained on the same corpora as the SMT system.

However, despite these adaptations to the SMT task, the core sense disambiguation task remains pure WSD:

- The rich context features are typical of WSD and almost never used in SMT.

- The dynamic integration of context-sensitive translation probabilities is not typical of SMT.

- Although it is embedded in a real SMT system, the WSD task is exactly the same as in recent and coming Senseval Multilingual Lexical Sample tasks (e.g., Chklovski *et al.* (2004)), where sense inventories represent the semantic distinctions made by another language.

We begin by presenting the WSD module and the SMT integration technique. We then show that incorporating it into a standard phrase-based SMT baseline system *consistently improves translation quality* across all three different test sets from the Chinese-English IWSLT text translation evaluation, as well as on the larger NIST Chinese-English translation task. Depending on the metric, the individual gains are sometimes modest, but remarkably, incorporating WSD *never* hurts, and helps enough to always make it a worthwile additional component in an SMT system. Finally, we analyze the reasons for the improvement.

2 Problems in context-sensitive lexical choice for SMT

To the best of our knowledge, there has been no previous attempt at integrating a state-of-the-art WSD system for fully phrasal multi-word lexical choice into phrase-based SMT, with evaluation of the resulting system on a translation task. While there are many evaluations of WSD quality, in particular the Senseval series of shared tasks (Kilgarriff and Rosenzweig (1999), Kilgarriff (2001), Mihalcea *et al.* (2004)), very little work has been done to address the actual integration of WSD in realistic SMT applications.

To fully integrate WSD into phrase-based SMT, it is necessary to perform lexical disambiguation on *multi-word phrasal* lexical units; in contrast, the model reported in Cabezas and Resnik (2005) can only perform lexical disambiguation on *single words*. Like the model proposed in this paper, Cabezas and Resnik attempted to integrate phrase-based WSD models into decoding. However, although they reported that incorporating these predictions via the Pharaoh XML markup scheme yielded a small improvement in BLEU score over a Pharaoh baseline on a single Spanish-English translation data set, we have determined empirically that applying their single-word based model to several Chinese-English datasets does *not* yield systematic improvements on most MT evaluation metrics (Carpuat and Wu, 2007). The single-word model has the disadvantage of forcing the decoder to choose between the baseline phrasal translation probabilities versus the WSD model predictions for single words. In addition, the single-word model does not generalize to WSD for phrasal lexical choice, as overlapping spans cannot be specified with the XML markup scheme. Providing WSD predictions for phrases would require committing to a phrase segmentation of the input sentence before decoding, which is likely to hurt translation quality.

It is also necessary to focus directly on translation accuracy rather than other measures such as alignment error rate, which may not actually lead to improved translation quality; in contrast, for example, Garcia-Varea *et al.* (2001) and Garcia-Varea *et al.* (2002) show improved alignment error rate with a maximum entropy based context-dependent lexical choice model, but not improved translation accuracy. In contrast, our evaluation in this paper is conducted on the actual decoding task, rather than intermediate tasks such as word alignment. Moreover, in the present work, *all* commonly available automated MT evaluation metrics are used, rather than only BLEU score, so as to maintain a more balanced perspective.

Another problem in the context-sensitive lexical choice in SMT models of Garcia Varea *et al.* is that their feature set is insufficiently rich to make much better predictions than the SMT model itself. In contrast, our WSD-based lexical choice models are designed to directly model the lexical choice in the actual translation direction, and take full advantage of not residing strictly within the Bayesian source-channel model in order to benefit from the much richer Senseval-style feature set this facilitates.

Garcia Varea *et al.* found that the best results are obtained when the training of the context-dependent translation model is fully incorporated with the EM training of the SMT system. As described below, the training of our new WSD model, though not incorporated within the EM training, is also far more closely tied to the SMT model than is the case with traditional standalone WSD models.

In contrast with Brown *et al.* (1991), our approach incorporates the predictions of state-of-the-art WSD models that use rich contextual features for any phrase in the input vocabulary. In Brown *et al.*'s early study of WSD impact on SMT performance, the authors reported improved translation quality on a French to English task, by choosing an English translation for a French word based on the single contextual feature which is reliably discriminative. However, this was a pilot study, which is limited to words with exactly two translation candidates, and it is not clear that the conclusions would generalize to more recent SMT architectures.

3 Problems in translation-oriented WSD

The close relationship between WSD and SMT has been emphasized since the emergence of WSD as an independent task. However, most of previous research has focused on using multilingual resources typically used in SMT systems to improve WSD accuracy, e.g., Dagan and Itai (1994), Li and Li (2002), Diab (2004). In contrast, this paper focuses on the converse goal of using WSD models to improve actual translation quality.

Recently, several researchers have focused on designing WSD systems for the specific purpose of translation. Vickrey *et al.* (2005) train a logistic regression WSD model on data extracted from automatically word aligned parallel corpora, but evaluate on a blank filling task, which is essentially an evaluation of WSD accuracy. Specia (2006) describes an inductive logic programming-based WSD system, which was specifically designed for the purpose of Portuguese to English translation, but this system was also only evaluated on WSD accuracy, and not integrated in a full-scale machine translation system.

Ng *et al.* (2003) show that it is possible to use automatically word aligned parallel corpora to train accurate supervised WSD models. The purpose of the study was to lower the annotation cost for supervised WSD, as suggested earlier by Resnik and Yarowsky (1999). However this result is also encouraging for the integration of WSD in SMT, since it suggests that accurate WSD can be achieved using training data of the kind needed for SMT.

4 Building WSD models for phrase-based SMT

4.1 WSD models for every phrase in the input vocabulary

Just like for the baseline phrase translation model, WSD models are defined for every phrase in the input vocabulary. Lexical choice in SMT is naturally framed as a WSD problem, so the first step of integration consists of defining a WSD model for every phrase in the SMT input vocabulary.

This differs from traditional WSD tasks, where the WSD target is a single content word. Senseval for instance has either lexical sample or all word tasks. The target words for both categories of Senseval WSD tasks are typically only content words—primarily nouns, verbs, and adjectives—while in the context of SMT, we need to translate entire sentences, and therefore have a WSD model not only for every word in the input sentences, regardless of their POS tag, but for every phrase, including tokens such as articles, prepositions and even punctuation. Further empirical studies have suggested that including WSD predictions for those longer phrases is a key factor to help the decoder produce better translations (Carpuat and Wu, 2007).

4.2 WSD uses the same sense definitions as the SMT system

Instead of using pre-defined sense inventories, the WSD models disambiguate between the SMT translation candidates. In order to closely integrate WSD predictions into the SMT system, we need to formulate WSD models so that they produce features that can directly be used in translation decisions taken by the SMT system. It is therefore necessary for the WSD and SMT systems to consider exactly the same translation candidates for a given word in the input language.

Assuming a standard phrase-based SMT system (e.g., Koehn *et al.* (2003)), WSD senses are thus either words or phrases, as learned in the SMT phrasal translation lexicon. Those "sense" candidates are very different from those typically used even in dedicated WSD tasks, even in the multilingual Senseval tasks. Each candidate is a phrase that is not necessarily a syntactic noun or verb phrase as in manually compiled dictionaries. It is quite possible that distinct "senses" in our WSD for SMT system could be considered synonyms in a traditional WSD framework, especially in monolingual WSD.

In addition to the consistency requirements for integration, this requirement is also motivated by empirical studies, which show that predefined translations derived from sense distinctions defined in monolingual ontologies do not match translation distinction made by human translators (Specia *et al.*, 2006).

4.3 WSD uses the same training data as the SMT system

WSD training does not require any other resources than SMT training, nor any manual sense annotation. We employ supervised WSD systems, since Senseval results have amply demonstrated that supervised models significantly outperform unsupervised approaches (see for instance the English lexical sample tasks results described by Mihalcea *et al.* (2004)).

Training examples are annotated using the phrase alignments learned during SMT training. Every input language phrase is sense-tagged with its aligned output language phrase in the parallel corpus. The phrase alignment method used to extract the WSD training data therefore depends on the one used by the SMT system. This presents the advantage of training WSD and SMT models on exactly the same data, thus eliminating domain mismatches between Senseval data and parallel corpora. But most importantly, this allows WSD training data to be generated entirely automatically, since the parallel corpus is automatically phrase-aligned in order to learn the SMT phrase bilexicon.

4.4 The WSD system

The word sense disambiguation subsystem is modeled after the best performing WSD system in the Chinese lexical sample task at Senseval-3 (Carpuat *et al.*, 2004).

The features employed are typical of WSD and are therefore far richer than those used in most SMT systems. The feature set consists of position-sensitive, syntactic, and local collocational features, since these features yielded the best results when combined in a naïve Bayes model on several Senseval-2 lexical sample tasks (Yarowsky and Florian, 2002). These features scale easily to the bigger vocabulary and sense candidates to be considered in a SMT task.

The Senseval system consists of an ensemble of four combined WSD models:

The first model is a naïve 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 model is a maximum entropy model (Jaynes, 1978), since Klein and Manning (Klein and Manning, 2002) found that this model yielded higher accuracy than naïve Bayes in a subsequent comparison of WSD performance.

The third model is a boosting model (Freund and Schapire, 1997), since boosting has consistently turned in very competitive scores on related tasks such as named entity classification. We also use the Adaboost.MH algorithm.

The fourth model is a Kernel PCA-based model (Wu *et al.*, 2004). Kernel Principal Component Analysis or 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.

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.

4.5 Integrating WSD predictions in phrase-based SMT architectures

It is non-trivial to incorporate WSD into an existing phrase-based architecture such as Pharaoh (Koehn, 2004), since the decoder is not set up to easily accept multiple translation probabilities that are dynamically computed in context-sensitive fashion.

For every *phrase* in a given SMT input sentence, the WSD probabilities can be used as additional feature in a loglinear translation model, in combination with typical context-independent SMT bilexicon probabilities.

We overcome this obstacle by devising a calling architecture that reinitializes the decoder with dynamically generated lexicons on a per-sentence basis.

Unlike a n-best reranking approach, which is limited by the lexical choices made by the decoder using only the baseline context-independent translation probabilities, our method allows the system to make full use of WSD information for all competing phrases at all decoding stages.

5 Experimental setup

The evaluation is conducted on two standard Chinese to English translation tasks. We follow standard machine translation evaluation procedure using automatic evaluation metrics. Since our goal is to evaluate translation quality, we use standard MT evaluation methodology and do not evaluate the accuracy of the WSD model independently.

Table 1: Evaluation results on the IWSLT06 dataset: integrating the WSD translation predictions improves BLEU, NIST, METEOR, WER, PER, CDER and TER across all 3 different available test sets.

Test Set	Exper.	BLEU	NIST	METEOR	METEOR (no syn)	TER	WER	PER	CDER
Test 1	SMT	42.21	7.888	65.40	63.24	40.45	45.58	37.80	40.09
	SMT+WSD	**42.38**	**7.902**	**65.73**	**63.64**	**39.98**	**45.30**	**37.60**	**39.91**
Test 2	SMT	41.49	8.167	66.25	63.85	40.95	46.42	37.52	40.35
	SMT+WSD	**41.97**	**8.244**	**66.35**	**63.86**	**40.63**	**46.14**	**37.25**	**40.10**
Test 3	SMT	49.91	9.016	73.36	70.70	35.60	40.60	32.30	35.46
	SMT+WSD	**51.05**	**9.142**	**74.13**	**71.44**	**34.68**	**39.75**	**31.71**	**34.58**

Table 2: Evaluation results on the NIST test set: integrating the WSD translation predictions improves BLEU, NIST, METEOR, WER, PER, CDER and TER

Exper.	BLEU	NIST	METEOR	METEOR (no syn)	TER	WER	PER	CDER
SMT	20.41	7.155	60.21	56.15	76.76	88.26	61.71	70.32
SMT+WSD	**20.92**	**7.468**	**60.30**	**56.79**	**71.34**	**83.87**	**57.29**	**67.38**

5.1 Data set

Preliminary experiments are conducted using training and evaluation data drawn from the multilingual BTEC corpus, which contains sentences used in conversations in the travel domain, and their translations in several languages. A subset of this data was made available for the IWSLT06 evaluation campaign (Paul, 2006); the training set consists of 40000 sentence pairs, and each test set contains around 500 sentences. We used only the pure text data, and not the speech transcriptions, so that speech-specific issues would not interfere with our primary goal of understanding the effect of integrating WSD in a full-scale phrase-based model.

A larger scale evaluation is conducted on the standard NIST Chinese-English test set (MT-04), which contains 1788 sentences drawn from newswire corpora, and therefore of a much wider domain than the IWSLT data set. The training set consists of about 1 million sentence pairs in the news domain.

Basic preprocessing was applied to the corpus. The English side was simply tokenized and case-normalized. The Chinese side was word segmented using the LDC segmenter.

5.2 Baseline SMT system

Since our focus is not on a specific SMT architecture, we use the off-the-shelf phrase-based decoder Pharaoh (Koehn, 2004) trained on the IWSLT training set. Pharaoh implements a beam search decoder for phrase-based statistical models, and presents the advantages of being freely available and widely used.

The phrase bilexicon is derived from the intersection of bidirectional IBM Model 4 alignments, obtained with GIZA++ (Och and Ney, 2003), augmented to improve recall using the grow-diag-final heuristic. The language model is trained on the English side of the corpus using the SRI language modeling toolkit (Stolcke, 2002).

The loglinear model weights are learned using Chiang's implementation of the maximum BLEU training algorithm (Och, 2003), both for the baseline, and the WSD-augmented system. Due to time constraints, this optimization was only conducted on the IWSLT task. The weights used in the WSD-augmented NIST model are based on the best IWSLT model. Given that the two tasks are quite different, we expect further improvements on the WSD-augmented system after running maximum BLEU optimization for the NIST task.

6 Results and discussion

Using WSD predictions in SMT yields better translation quality on *all* test sets, as measured by *all eight* commonly used automatic evaluation metrics.

Table 3: Translation examples with and without WSD for SMT, drawn from IWSLT data sets.

Input	请 转 乘 中央 线 。
Ref.	Please transfer to the Chuo train line.
SMT	Please turn to the Central Line.
SMT+WSD	Please transfer to Central Line.
Input	车票 在 车上 买 吗 ？
Ref.	Do I pay on the bus?
SMT	Please get on the bus?
SMT+WSD	I buy a ticket on the bus?
Input	需要 预订 吗 ？
Ref.	Do I need a reservation?
SMT	I need a reservation?
SMT+WSD	Do I need a reservation?
Input	我 想 再 确认 一下 这 张 票 的 预订 。
Ref.	I want to reconfirm this ticket.
SMT	I would like to reconfirm a flight for this ticket.
SMT+WSD	I would like to reconfirm my reservation for this ticket.
Input	步行 可以 到 那里 吗 ？
Ref.	Can I get there on foot?
SMT	Is there on foot?
SMT+WSD	Can I get there on foot?
Input	我 有 另外 一个 约会 ， 所以 请 快点 。
Ref.	I have another appointment, so please hurry.
SMT	I have an appointment for a, so please hurry.
SMT+WSD	I have another appointment, so please hurry.
Input	对不起 。 你 能 告诉 我 到 百老汇 的 路 吗 ？
Ref.	Excuse me. Could you tell me the way to Broadway?
SMT	Could you tell me the way to Broadway? I am sorry.
SMT+WSD	Excuse me, could you tell me the way to Broadway?
Input	对不起 ， 我 想 开 一个 账 户 。
Ref.	Excuse me, I want to open an account.
SMT	Excuse me, I would like to have an account.
SMT+WSD	Excuse me, I would like to open an account.

The results are shown in Table 1 for IWSLT and Table 2 for the NIST task. Paired bootstrap resampling shows that the improvements on the NIST test set are statistically significant at the 95% level.

Remarkably, integrating WSD predictions helps all the very different metrics. In addition to the widely used BLEU (Papineni *et al.*, 2002) and NIST (Doddington, 2002) scores, we also evaluate translation quality with the recently proposed Meteor (Banerjee and Lavie, 2005) and four edit-distance style metrics, Word Error Rate (WER), Position-independent word Error Rate (PER) (Tillmann *et al.*, 1997), CDER, which allows block reordering (Leusch *et al.*, 2006), and Translation Edit Rate (TER) (Snover *et al.*, 2006). Note that we report Meteor scores computed both with and without using WordNet synonyms to match translation candidates and references, showing that the improvement is not due to context-independent synonym matches at evaluation time.

Comparison of the 1-Best decoder output with and without the WSD feature shows that the sentences differ by one or more token respectively for 25.49%, 30.40% and 29.25% of IWSLT test sets 1,

Table 4: Translation examples with and without WSD for SMT, drawn from the NIST test set.

Input	没有 任何 议员 投票 反对 他 。
SMT	Without any congressmen voted against him.
SMT+WSD	No congressmen voted against him.
Input	俄 在 车臣 实行 的 政策 以及 对 独 联 体 邻国 的 态度 更 是 令 美国 担忧 。
SMT	Russia's policy in Chechnya and CIS neighbors attitude is even more worried that the United States.
SMT+WSD	Russia's policy in Chechnya and its attitude toward its CIS neighbors cause the United States still more anxiety.
Input	至于 美国 的 人权 状况 呢 ？
SMT	As for the U.S. human rights conditions?
SMT+WSD	As for the human rights situation in the U.S.?
Input	我 参 拜 是 为了 祈求 日本 的 和平 与 繁荣 。
SMT	The purpose of my visit to Japan is pray for peace and prosperity.
SMT+WSD	The purpose of my visit is to pray for peace and prosperity for Japan.
Input	为 防范 恐怖活动 ， 洛杉矶 警方 采取 了 前所未有 的 严密 保安 措施 。
SMT	In order to prevent terrorist activities Los Angeles, the police have taken unprecedented tight security measures.
SMT+WSD	In order to prevent terrorist activities Los Angeles, the police to an unprecedented tight security measures.

2 and 3, and 95.74% of the NIST test set.

Tables 3 and 4 show examples of translations drawn from the IWSLT and NIST test sets respectively.

A more detailed analysis reveals WSD predictions give better rankings and are more discriminative than baseline translation probabilities, which helps the final translation in three different ways.

- The rich context features help rank the correct translation first with WSD while it is ranked lower according to baseline translation probability scores .
- Even when WSD and baseline translation probabilities agree on the top translation candidate, the stronger WSD scores help override wrong language model predictions.
- The strong WSD scores for phrases help the decoder pick longer phrase translations, while using baseline translation probabilities often translate those phrases in smaller chunks that include a frequent (and incorrect) translation candidate.

For instance, the top 4 Chinese sentences in Table 4, are better translated by the WSD-augmented system because the WSD scores help the decoder to choose longer phrases. In the first example, the phrase "没有 任何" is correctly translated as a whole as "No" by the WSD-augmented system, while the baseline translates each word separately yielding an incorrect translation. In the following three examples, the WSD system encourages the decoder to translate the long phrases "更 是 令 美国 担忧", "美国 的 人权 状况", and "祈求 日本 的 和平 与 繁荣" as single units, while the baseline introduces errors by breaking them down into shorter phrases.

The last sentence in the table shows an example where the WSD predictions do not help the baseline system. The translation quality is actually much worse, since the verb "采取" is incorrectly translated as "to", despite the fact that the top candidate predicted by the WSD system alone is the much better translation "has taken", but with a relatively low probability of 0.509.

7 Conclusion

We have shown for the first time that integrating multi-word phrasal WSD models into phrase-based

SMT consistently helps on all commonly available automated translation quality evaluation metrics on all three different test sets from the Chinese-English IWSLT06 text translation task, and yields statistically significant gains on the larger NIST Chinese-English task. It is important to note that the WSD models *never* hurt translation quality, and always yield individual gains of a level that makes their integration always worthwile.

We have proposed to consistently integrate WSD models both during training, where sense definitions and sense-annotated data are automatically extracted from the word-aligned parallel corpora from SMT training, and during testing, where the phrasal WSD probabilities are used by the SMT system just like all the other lexical choice features.

Context features are derived from state-of-the-art WSD models, and the evaluation is conducted on the actual translation task, rather than intermediate tasks such as word alignment.

It is to be emphasized that this approach does not merely consist of adding a source sentence feature in the log linear model for translation. On the contrary, it remains a real WSD task, defined just as in the Senseval Multilingual Lexical Sample tasks (e.g., Chklovski *et al.* (2004)). Our model makes use of typical WSD features that are almost never used in SMT systems, and requires a dynamically created translation lexicon on a per-sentence basis.

To our knowledge this constitues the first attempt at fully integrating state-of-the-art WSD with conventional phrase-based SMT. Unlike previous approaches, the WSD targets are not only single words, but *multi-word phrases*, just as in the SMT system. This means that WSD senses are unusually predicted not only for a limited set of single words or very short phrases, but for all phrases of arbitrarily length that are in the SMT translation lexicon. The single word approach, as we reported in Carpuat *et al.* (2006), improved BLEU and NIST scores for phrase-based SMT, but subsequent detailed empirical studies we have performed since then suggest that single word WSD approaches are less successful when evaluated under all other MT metrics (Carpuat and Wu, 2007). Thus, fully phrasal WSD predictions for longer phrases, as reported in this paper, are particularly important to improve translation quality.

The results reported in this paper cast new light on the WSD vs. SMT debate, suggesting that a close integration of WSD and SMT decisions should be incorporated in a SMT model that successfully uses WSD predictions. Our objective here is to demonstrate that this technique works for the widest possible class of models, so we have chosen as the baseline the most widely used phrase-based SMT model. Our positive results suggest that our experiments could be tried on other current statistical MT models, especially the growing family of tree-structured SMT models employing stochastic transduction grammars of various sorts (Wu and Chiang, 2007). For instance, incorporating WSD predictions into an MT decoder based on inversion transduction grammars (Wu, 1997)—such as the Bracketing ITG based models of Wu (1996), Zens *et al.* (2004), or Cherry and Lin (2007)—would present an intriguing comparison with the present work. It would also be interesting to assess whether a more grammatically structured statistical MT model that is less reliant on an n-gram language model, such as the syntactic ITG based "grammatical channel" translation model of (Wu and Wong, 1998), could make more effective use of WSD predictions.

References

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgement. In *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguistics (ACL-2005)*, Ann Arbor, Michigan, June 2005.

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.

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

Marine Carpuat and Dekai Wu. Evaluating the word sense disambiguation performance of statistical machine translation. In *Proceedings of the Second International Joint Conference on Natural Language Processing (IJCNLP)*, pages 122–127, Jeju Island, Republic of Korea, 2005.

Marine Carpuat and Dekai Wu. Word sense disambiguation vs. statistical machine translation. In *Proceedings of the annual meeting of the association for computational linguistics (ACL-05)*, Ann Arbor, Michigan, 2005.

Marine Carpuat and Dekai Wu. How phrase sense disambiguation outperforms word sense disambiguation for statistical machine translation. Forthcoming, 2007.

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.

Marine Carpuat, Yihai Shen, Xiaofeng Yu, and Dekai Wu. Toward integrating word sense and entity disambiguation into statistical machine translation. In *Third International Workshop on Spoken Language Translation (IWSLT 2006)*, Kyoto, November 2006.

Colin Cherry and Dekang Lin. Inversion Transduction Grammar for joint phrasal translation modeling. In Dekai Wu and David Chiang, editors, *NAACL-HLT 2007 / AMTA Workshop on Syntax and Structure in Statistical Translation (SSST)*, pages 17–24, Rochester, NY, April 2007.

Timothy Chklovski, Rada Mihalcea, Ted Pedersen, and Amruta Purandare. The Senseval-3 multilingual English-Hindi lexical sample task. In *Proceedings of Senseval-3, Third International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 5–8, Barcelona, Spain, July 2004. SIGLEX, Association for Computational Linguistics.

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

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.

George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Human Language Technology conference (HLT-2002)*, San Diego, CA, 2002.

Zhendong Dong. Knowledge description: what, how and who? In *Proceedings of International Symposium on Electronic Dictionary*, Tokyo, Japan, 1998.

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.

Ismael Garcia-Varea, Franz Och, Hermann Ney, and Francisco Casacuberta. Refined lexicon models for statistical machine translation using a maximum entropy approach. In *Proceedings of the 39th annual meeting of the association for computational linguistics (ACL-01)*, Toulouse, France, 2001.

Ismael Garcia-Varea, Franz Och, Hermann Ney, and Francisco Casacuberta. Efficient integration of maximum entropy lexicon models within the training of statistical alignment models. In *Proceedings of AMTA-2002*, pages 54–63, Tiburon, California, October 2002.

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

Adam Kilgarriff and Joseph Rosenzweig. Framework and results for English Senseval. *Computers and the Humanities*, 34(1):15–48, 1999. Special issue on SENSEVAL.

Adam Kilgarriff. English lexical sample task description. In *Proceedings of Senseval-2, Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 17–20, Toulouse, France, July 2001. SIGLEX, Association for Computational Linguistics.

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.

Philipp Koehn, Franz Och, and Daniel Marcu. Statistical phrase-based translation. In *Proceedings of HLT/NAACL-2003*, Edmonton, Canada, May 2003.

Philipp Koehn. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *6th Conference of the Association for Machine Translation in the Americas (AMTA)*, Washington, DC, September 2004.

Gregor Leusch, Nicola Ueffing, and Hermann Ney. Efficient MT evaluation using block movements. In *Proceedings of EACL-2006 (11th Conference of the European Chapter of the Association for Computational Linguistics)*, pages 241–248, Trento, Italy, April 2006.

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.

Rada Mihalcea, Timothy Chklovski, and Adam Kilgariff. The Senseval-3 English lexical sample task. In *Proceedings of Senseval-3, Third International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 25–28, Barcelona, Spain, July 2004. SIGLEX, Association for Computational Linguistics.

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

Franz Josef Och. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, 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.

Michael Paul. Overview of the IWSLT06 evaluation campaign. In *Third International Workshop on Spoken Language Translation (IWSLT 2006)*, Kyoto, November 2006.

Philip Resnik and David Yarowsky. Distinguising systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(2):113–133, 1999.

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

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of AMTA*, pages 223–231, Boston, MA, 2006. Association for Machine Translation in the Americas.

Lucia Specia, Maria das Gračas Volpe Nunes, Gabriela Castelo Branco Ribeiro, and Mark Stevenson. Multilingual versus monolingual WSD. In *EACL-2006 Workshop on Making Sense of Sense: Bringing Psycholinguistics and Computational Linguistics Together*, pages 33–40, Trento, Italy, April 2006.

Lucia Specia. A hybrid relational approach for WSD—first results. In *Proceedings of the COLING/ACL 06 Student Research Workshop*, pages 55–60, Sydney, July 2006. ACL.

Andreas Stolcke. SRILM—an extensible language modeling toolkit. In *International Conference on Spoken Language Processing*, Denver, Colorado, September 2002.

Christoph Tillmann, Stefan Vogel, Hermann Ney, A. Zubiaga, and H. Sawaf. Accelerated DP-based search for statistical translation. In *Proceedings of Eurospeech'97*, pages 2667–2670, Rhodes, Greece, 1997.

David Vickrey, Luke Biewald, Marc Teyssier, and Daphne Koller. Word-sense disambiguation for machine translation. In *Joint Human Language Technology conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, Vancouver, 2005.

Dekai Wu and David Chiang, editors. *NAACL-HLT 2007 / AMTA Workshop on Syntax and Structure in Statistical Translation (SSST)*. Association for Computational Linguistics, Rochester, NY, USA, April 2007.

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, Spain, 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.

Dekai Wu. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–404, 1997.

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 *20th International Conference on Computational Linguistics (COLING-2004)*, Geneva, August 2004.