How Phrase Sense Disambiguation outperforms Word Sense Disambiguation for Statistical Machine Translation

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 present comparative empirical evidence arguing that a generalized *phrase sense disambiguation* approach better improves statistical machine translation than ordinary word sense disambiguation, along with a data analysis suggesting the reasons for this. Standalone word sense disambiguation, as exemplified by the Senseval series of evaluations, typically defines the target of disambiguation as a single word. But in order to be useful in statistical machine translation, our studies indicate that word sense disambiguation should be redefined to move beyond the particular case of single word targets, and instead to generalize to multi-word phrase targets. We investigate how and why the phrase sense disambiguation approach—in contrast to recent efforts to apply traditional word sense disambiguation to SMT—is able to yield statistically significant yimprovements in translation quality even under large data conditions, and consistently improve SMT across both IWSLT and NIST Chinese-English text translation tasks. We discuss architectural issues raised by this change of perspective, and consider the new model architecture necessitated by the phrase sense disambiguation approach.

*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 authors and do not necessarily reflect the views of the Defense Advanced Research Projects Agency.

1 Introduction

Until recently, attempts to apply word sense disambiguation (WSD) techniques to improve translation quality in statistical machine translation (SMT) models have met with mixed or disappointing results (e.g., Carpuat and Wu (2005), Cabezas and Resnik (2005)), suggesting that a deeper empirical exploration of the differences and consequences of the assumptions of WSD and SMT is called for.

On one hand, word sense disambiguation as a standalone task consists in identifying the correct sense of a given word among a set of predefined sense candidates. In the Senseval series of evaluations, WSD targets are typically single words, both in the lexical sample tasks, where only a predefined set of targets are considered (e.g., Kilgarriff (2001);), and in the all-words tasks, where all content word in a given corpus must be disambiguated (e.g., Kilgarriff and Rosenzweig (1999)).

This focus on single words as WSD targets might be explained by the sense inventory, which is usually derived from a manually constructed dictionary or ontology, where most entries are single words. In addition, historically, as for many other tasks, work on European languages imposed whitespace as an easy way to define convenient

word boundaries. Linguistically, however, this oversimplistic modeling approach seems rather questionable, and recalls long-held debates over the issue of what properly constitutes a "word".

In contrast, work in statistical machine translation has for some time recognized the need to segment sentences as required by the task's evaluation criteria, and today most systems use phrases or segments, and not single words, as the basic unit for lexical choice (e.g., Wu (1997); Och and Ney (2004); Koehn (2004); Chiang (2005)). Note that single-word based SMT architectures already perform a significant amount of sense disambiguation intrinsically, by virtue of combining a priori sense candidate likelihoods (from adequacy criteria as modeled by lexical translation probabilities) with contextual coherence preferences (from fluency criteria as modeled by language model probabilities). Phrasal SMT architectures, furthermore, integrate lexical collocation preferences into the disambiguation choices, raising the bar yet higher.

This suggests that to be effective at improving disambiguation accuracy within SMT architectures, sense disambiguation techniques may need to incorporate assumptions at least as strong as those already made by the SMT models. Dedicated WSD models do appear to possess traits that are promising for SMT: they employ a much broader range of features for sense selection than SMT models, and are far more sensitive to dynamic context. The question, however, is whether these advantages must be reformulated within a phrasal framework in order for the advantages to be realizable for SMT.

In this work, we empirically compare the efficacy of *phrase sense disambiguation* versus word sense disambiguation approaches toward improving translation quality of SMT models. The phrase sense disambiguation (PSD) approach generalizes word sense disambiguation to multi-word targets, aiming thereby to incorporate the crucial assumptions responsible for the success of phrasal SMT approaches into the sense disambiguation model as well. Our results and analysis show that it is indeed necessary to move away from the simplistic single-word level definition of sense disambiguation targets, in order to be useful to SMT. In effect, this argues for redefining WSD for the task of SMT. This task-driven approach to sense disambiguation requires several changes:

- Sense disambiguation targets are very different from Senseval targets.
- Sense candidates are not extracted from manually defined sense inventories, but from automatically annotated data.
- Sense disambiguation predictions require a dynamic integration architecture in SMT systems in order to be useful.

We will begin by reviewing our phrase sense disambiguation approach for SMT and contrasting it against previous word-based models. We then describe new contrastive empirical studies aimed at directly assessing the differences. On one hand, we note that incorporating multi-word PSD into phrasal SMT reliably and consistently improves translation quality, as measured by *all* eight most commonly used evaluation metrics, on *all* four different test sets from the IWSLT and NIST Chinese-English translation tasks. On the other hand, the contrastive experiments reported here show that incorporating single-word WSD into phrasal SMT leads to unpredictable and inconsistent effects on translation quality, depending on which evaluation metric one looks at. We then turn to data analysis exploring more closely how and why the multi-word PSD approach outperforms the single-word WSD approach. The analysis shows that dynamic integration of PSD prediction is crucial to this improvement, as it allows all PSD predictions to participate in the segmention of the input sentence that yields the best translation quality.

2 Previous work

In Carpuat and Wu (2007), we proposed a novel general framework for integrating a generalized sense disambiguation method into SMT, such that phrasal lexical choice is dynamically influenced by context-dependent probabilities or scores. This Phrase Sense Disambiguation—as opposed to Word Sense Disambiguation—approach appears to be the only model to date that has been shown capable of consistently yielding improvements on translation quality across all

different test sets and automatic evaluation metrics. Other related work has all been heavily oriented toward disambiguating single words.

In perhaps the earliest study of WSD potential for SMT performance by Brown *et al.* (1991), 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 *single words* with exactly two translation candidates, and it is far from clear that the conclusions could generalize to more recent SMT architectures. In contrast with Brown *et al.*'s work, our approach incorporates the predictions of state-of-the-art WSD models (generalized to PSD models) that use rich contextual features for *any* phrase in the input vocabulary.

More recent work on WSD systems designed for the specific purpose of translation has followed the traditional word-based definition of the WSD task. Vickrey *et al.* (2005) train a logistic regression WSD model on data extracted from automatically word aligned parallel corpora, and evaluate it on a blank filling task, which is essentially an evaluation of WSD accuracy. Specia *et al.* (2007) use an inductive logic programming based WSD system to integrate expressive features for Portuguese to English translation, but this system was also only evaluated on WSD accuracy, and not integrated in a full-scale machine translation system. Even when using automatically-aligned SMT parallel corpora to define WSD tasks, as in the SemEval-2007 English Lexical Sample Task via English-Chinese Parallel Text (Ng and Chan, 2007), WSD is still defined as a word-based task.

There have been other attempts at using context information for lexical selection in SMT, but the focus was also on single words vs. multi-word phrases, and they were not evaluated in terms of translation quality. For instance, 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 do not report improved translation accuracy. Another problem in the context-sensitive SMT models of Garcia Varea *et al.* is that they strictly reside within the Bayesian source-channel model, which is word-based.

The few recent attempts at integrating *single word* based WSD models into SMT have failed to obtain clear improvements in terms of translation quality. Carpuat and Wu (2005) show that using word-based Senseval trained models does not help BLEU score when integrated in a standard word-based translation system, for a NIST Chinese-English translation task.

Following this surprising result, a few attempts at integrating WSD methods into state-of-the-art SMT systems have begun to obtain slightly more encouraging results by moving away from manually-constructed sense inventories, and instead automatically defining word senses as word translation candidates, just like in SMT. Cabezas and Resnik (2005) reported that incorporating *word-based* WSD predictions via the Pharaoh XML markup scheme yielded a small improvement in BLEU score over a *phrasal* SMT baseline on a single Spanish-English translation data set. However, the result was not statistically significant, and in this paper, we will show that applying a similar single-word based model to several Chinese-English datasets does not yield systematic improvements on most MT evaluation metrics. Carpuat *et al.* (2006) also reported small improvements in BLEU score by using single-word WSD predictions in a Pharaoh baseline. However, these small improvements were obtained on a slightly weaker SMT baseline, and subsequent evaluations showed that these gains are not consistent across metrics. Giménez and Màrquez (2007) also used WSD predictions in Pharaoh for the slightly more general case of very frequent phrases, which in practice essentially limits the set of WSD targets to single words or very short phrases. However, evaluation on the single Europarl Spanish-English task did not yield consistent improvements across metrics: BLEU score did not improve, while there were small improvements in the QUEEN, METEOR and ROUGE metrics. Chan *et al.* (2007) report an improved BLEU score for a hierarchical phrase-based SMT system on a NIST Chinese-English task, by incorporating WSD predictions only for single words and short phrases of length 1 or 2. However, no results for metrics other than BLEU were reported, and no results on other tasks, so the relia-

bility of this model is not known.

What the foregoing attempts at WSD in SMT share is that (1) they focus on single words rather than full phrases, and (2) the evaluations do not show consistent improvement systematically across different tasks and metrics.

In contrast, we showed in Carpuat and Wu (2007) for the first time that generalizing WSD to exactly match *phrasal* lexical choice in SMT yields consistent improvements on 4 different test sets as measured by 8 common automatic evaluation metrics, unlike all the single-word oriented approaches. The key question left unanswered, however—which we attempt to address in the present paper—is exactly how and why it is necessary to generalize Word Sense Disambiguation to Phrase Sense Disambiguation in order to obtain this sort of consistency in translation accuracy improvement.

3 Building multi-word Phrase Sense Disambiguation models for SMT

3.1 Phrase sense disambiguation vs. word sense disambiguation

In a task-driven definition of sense disambiguation for phrase-based SMT, the PSD approach argues that disambiguation targets must be exactly the same *phrases* as in the SMT phrasal translation lexicon, so that the sense disambiguation task is identical to lexical choice for SMT. This constrasts with the standalone WSD perspective, where targets are single words, as in Senseval tasks (e.g., Kilgarriff and Rosenzweig (1999)). In SMT, phrases are typically defined as any sequence of words up to a given length. As a result, the phrasal targets for sense disambiguation need not necessarily be syntactic well-formed phrases, but rather need only be collocations defined by their surface form. This again departs from Senseval-style WSD where POS-tagging is typically decoupled from WSD, as training data is manually checked to contain instance for a single POS of the target.

In sense disambiguation for SMT, the sense candidates are those defined by the SMT translation lexicon. Sense candidates can be single words or multi-word phrases regardless of the length of the target. Note that phrasal senses do occasionally also exist in standalone WSD tasks. For instance, the Senseval English Lexical Sample tasks include WordNet phrasal senses (e.g., "polar bear" is a sense candidate for the English target word "bear".)

Given the above definitions for sense disambiguation targets and senses, annotated training data can naturally be drawn from the automatically aligned parallel corpora used to learn the SMT lexicon. Given a Chinese-English sentence pair, a WSD or PSD target in the Chinese sentence is annotated with the English phrase which is consistent with the word alignment. The definition of consistency with the word alignment should be exactly the one used for building the SMT lexicon.

Despite the differences introduced by the use of phrasal targets, the disambiguation task remains in the character and spirit of WSD. The translation lexical choice problem is exactly the same task 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. In our SMT-driven approach to PSD rather than WSD, we are only generalizing the definition of the sense disambiguation targets, and automating the sense annotation process.

3.2 Leveraging Senseval classifiers for both WSD and PSD

As in Carpuat and Wu (2007), the word sense disambiguation system is modeled after the best performing WSD system in the Chinese lexical sample task at Senseval-3 (Carpuat *et al.*, 2004). The features employed include position-sensitive, syntactic, and local collocational features, and are therefore much richer than those used in most SMT systems.

4 Integrating multi-word PSD vs. single-word WSD into phrasal SMT architectures

Unlike single-word WSD, it is non-trivial to incorporate the PSD predictions 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. While PSD and WSD models differ in principle only by the length of the WSD target, their integration into phrase-based SMT architectures requires significantly different strategies.

Since multi-word PSD predictions are defined for every entry in the SMT lexicon or phrase table, they can be thought of as an additional feature in the phrase table. However, unlike baseline SMT translation probabilities, these predictions are context-sensitive, and require to be updated for every new sentence. Therefore, instead of using a static phrasal translation lexicon, integration of PSD predictions require dynamically updating the phrasal translation lexicon for each sentence during decoding.

In contrast, in the single-word WSD system, since the WSD predictions only cover a subset of the phrase-table entries and the word-based targets do not have overlapping spans, it is usually possible to implement a much simpler integration architecture, by annotating the input sentence to contain the WSD predictions, as with the Pharaoh XML markup scheme.

Thus, the dynamic phrase table architecture for PSD integration necessarily generates a significant overhead. While we could in theory annotate the input sentence with phrase-based WSD predictions, just like for single-word based WSD, we argue that this approach is not optimal and would in fact hurt translation quality: annotation schemes such as the Pharaoh XML markup do not allow to annotate overlapping spans, and would thus require to commit to a phrasal segmentation of the input sentence *before* decoding. It is impossible to find an optimal phrasal segmentation before decoding, since the quality of the segmentation can only be evaluated by the translation it yields.

5 Comparative experiment setup

5.1 Data set

In order to better isolate the different effects of WSD versus PSD, comparative 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. The simpler character of these sentences faciliates clearer identification of individual factors in data analysis, compared with open domain newsire text where too many factors interfere with each other. We used a subset of this data which was made available for the IWSLT 2006 evaluation campaign; the training set consists of 40000 sentence pairs, and each test set contains around 500 sentences. We used only the pure text data, so that speech-specific issues would not interfere with our primary goal of understanding the effect of integrating WSD/PSD in a full-scale phrasal SMT model.

We also report results of the large scale evaluation of the PSD model conducted on the standard NIST Chinese-English test set (MT-04), which contains 1788 sentences drawn from newswire corpora, and is therefore of a much wider domain than the IWSLT data set.

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 in a standard fashion on the IWSLT training set, as in Carpuat and Wu (2007).

5.3 WSD and PSD models

WSD classifiers are trained for every word, while PSD classifiers are trained for every multi-word phrase in the test set vocabularies. The number of targets is therefore much higher than even in the all-words WSD tasks. For the first IWSLT test set which contains 506 sentences, we have a total of PSD 2882 targets, as opposed to only 948 WSD targets. There is on average 7.3 sense candidates and 79 training instances per PSD target.

The scale of WSD and PSD models for SMT greatly constrasts with, for instance, the Senseval-3 Chinese lexical sample task which considered only 21 single word targets, with an average of 3.95 senses and 37 training instances per target.

6 Comparative evaluation results

The comparative experiments clearly show a marked difference between single–word WSD and multi-word PSD results. Evaluation scores, summarized in Tables 2 and 3, show that multi-word PSD yields consistent improvements in

Table 1: Evaluation results on the IWSLT-07 dataset: integrating the WSD translation predictions for single words has unpredictable effects on BLEU, NIST, METEOR, WER, PER, CDER and TER across all 3 different available test sets. Using only more reliable target words, such as nouns and verbs only, or targets that have more than 30 training instances, does not yield clear improvement either.

Test Set	Experiment	BLEU	NIST	METEOR	METEOR (no syn)	TER	WER	PER	CDER
#1	Baseline	42.21	7.888	65.40	63.24	40.45	45.58	37.80	40.09
	+WSD (all words)	41.94	7.911	65.55	63.52	40.59	45.61	37.75	40.09
	+WSD (nouns and verbs)	42.19	7.920	65.97	63.88	40.64	45.88	37.58	40.14
	+WSD (>30)	42.08	7.902	65.43	63.30	40.52	45.57	37.80	40.06
#2	Baseline	41.49	8.167	66.25	63.85	40.95	46.42	37.52	40.35
	+WSD (all words)	41.31	8.161	66.23	63.72	41.34	46.82	37.98	40.69
	+WSD (nouns and verbs)	41.25	8.135	66.08	63.40	41.30	46.76	37.85	40.65
	+WSD (>30)	41.56	8.186	66.44	63.89	40.87	46.36	37.57	40.35
#3	Baseline	49.91	9.016	73.36	70.70	35.60	40.60	32.30	35.46
	+WSD (all words)	49.73	9.017	73.32	70.82	35.72	40.61	32.10	35.30
	+WSD (nouns and verbs)	49.58	9.003	73.07	70.46	35.94	40.84	32.40	35.62
	+WSD (>30)	50.11	9.043	73.60	70.98	35.41	40.38	32.23	35.30

Table 2: Evaluation results on the IWSLT-06 dataset: integrating the multi-word PSD translation predictions for all phrases improves BLEU, NIST, METEOR, WER, PER, CDER and TER across all 3 different available test sets. In contrast, using the traditional single-word WSD approach has an unreliable impact on translation quality.

Test Set	Experiment	BLEU	NIST	METEOR	METEOR (no syn)	TER	WER	PER	CDER
#1	Baseline	42.21	7.888	65.40	63.24	40.45	45.58	37.80	40.09
	+WSD (all words)	41.94	7.911	65.55	63.52	40.59	45.61	37.75	40.09
	+PSD (all phrases)	**42.38**	**7.902**	**65.73**	**63.64**	**39.98**	**45.30**	**37.60**	**39.91**
#2	Baseline	41.49	8.167	66.25	63.85	40.95	46.42	37.52	40.35
	+WSD (all words)	41.31	8.161	66.23	63.72	41.34	46.82	37.98	40.69
	+PSD (all phrases)	**41.97**	**8.244**	**66.35**	**63.86**	**40.63**	**46.14**	**37.25**	**40.10**
#3	Baseline	49.91	9.016	73.36	70.70	35.60	40.60	32.30	35.46
	+WSD (all words)	49.73	9.017	73.32	70.82	35.72	40.61	32.10	35.30
	+PSD (all phrases)	**51.05**	**9.142**	**74.13**	**71.44**	**34.68**	**39.75**	**31.71**	**34.58**

translation quality, across *all* metrics and on *all* test sets, including statistically significant improvements on the large NIST task, while in contrast, the impact of single-word WSD on translation quality is highly unpredictable. In particular, the single-word WSD results are inconsistent across different test sets, and depend on which evaluation metric is chosen.

In order to measure the impact of WSD on translation quality, the translation results were evaluated using *all eight* of the most commonly used automatic evaluation metrics. In addition to the widely used BLEU (Papineni *et al.*, 2002) and NIST (Doddington, 2002) scores, we also evaluate translation quality with METEOR (Banerjee and Lavie, 2005), Word Error Rate (WER), Position-independent word Error Rate (PER) (Tillmann *et al.*, 1997), CDER (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.

In the sections that follow, we investigate various reasons that PSD outperforms WSD, drawing from data analysis on these comparative experiments.

7 Single-word WSD yields unreliable results

Using WSD predictions for all the single words in a given test set has an unreliable impact on translation quality, as can be seen in Table 1. While it yields a very small, non-significant gain on NIST and METEOR on Test Set 1, it yields worse BLEU, NIST and METEOR scores for all the other test sets.

In order to check that this disappointing result cannot be simply explained by the effect of unusual target words, we perform two sets of additional experiments. We attempt to consider only target words that are closer to those used in Senseval evaluations for which these WSD models were initially designed, and demonstrated good performance.

Instead of using WSD predictions for all the whitespace separated tokens that were seen during training, we restrict our set of WSD targets to nouns and verbs. This is slightly closer to the definition of targets in Senseval tasks, which typically include nouns, verbs and sometimes adjectives, but never punctuation or any function word. Table 1 shows that this does not help translation quality compared to the baseline system, and actually underperforms using WSD predictions for all words.

In contrast with Senseval target words, which are picked so that representative training data can be obtained, we are using every target word in the vocabulary, whatever the available training data. In order to check that the target words with few training instances are not hurting the contribution of other targets, we try to restrict our set of target words to those for which at least 30 instances were seen during training. Table 1 shows that this does not have a reliable effect on translation quality either, yielding small gains in BLEU, NIST and METEOR scores over the baseline for Test Sets 2 and 3, but hurting BLEU on Test Set 1. While the results are overall slightly better than when using all WSD predictions for all words, there is no clear trend for improvement.

These results show that considering only single words as sense disambiguation targets does not allow the SMT system to reliably exploit WSD predictions. This holds even when only targets that meet conditions that are closer to Senseval evaluations, where our WSD models are known to achieve good performance.

8 Multi-word PSD consistently improves translation quality

In contrast with the unreliable single-word WSD results, using phrasal multi-word PSD predictions in SMT remarkably yields better translation quality on *all* test sets, as measured by *all eight* commonly used automatic evaluation metrics. The results are shown in Table 2 for IWSLT and Table 3 for the NIST task. Paired bootstrap resampling shows that the improvements on the NIST test set are statistically significant at the 95% level.

Comparison of the 1-Best decoder output with and without the PSD 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, 2 and 3, and 95.74% of the NIST test set.

9 Multi-word PSD helps the decoder find a more useful segmentation of the input sentence

Analysis reveals that integrating PSD into SMT helps the decoder select a phrase segmentation of the input sentence which allows to find better translations than word-based WSD. We sampled translation examples from the IWSLT test sets, so that both word-based and phrase-based results are available for comparison. In addition, the relatively short sentence length of this corpus helps give a clearer understanding of the impact of WSD. Consider the following example:

Input 我 想 再 确认 一下 这 张 票 的 预订 。

Reference I want to reconfirm this ticket.

Table 3: Evaluation results on the NIST test set: integrating the PSD translation predictions improves BLEU, NIST, METEOR, WER, PER, CDER and TER.

Experiment	BLEU	NIST	METEOR	METEOR (no syn)	TER	WER	PER	CDER
Baseline	20.20	7.198	59.45	56.05	75.59	87.61	60.86	72.06
+PSD	**20.62**	**7.538**	**59.99**	**56.38**	**72.53**	**85.09**	**58.62**	**68.54**

WSD I would like to reconfirm a flight for this ticket.

PSD I want to reconfirm my reservation for this ticket.

Here, in the input segment “这张票的预订”, the particle “的” is in the same segment as the preceding word when using multi-word PSD predictions (“ 票的 ”), while the single-word WSD prefers to use “ 的预订”. This results in an incorrect translation of the phrase “的预订” as “flight for”. In contrast, PSD prefers to use the target “预订”, which ranks the correct “reservation” as the top translation candidate with a very confident probability of 0.94, as opposed to 0.28 only for the baseline context-independent translation probability used in the single-word WSD-augmented model. Similarly, consider:

Input 请 转 乘 中央 线 。

Reference You should transfer to the Central Line.

WSD Please turn to the Central Line.

PSD Please transfer to Central Line.

Here, PSD translates the segment “转乘” as a single unit and selects the correct translation “transfer to”, while WSD separately translates the words “转” and “乘” into the incorrect “turn to”. The multi-word PSD model correctly ranks “transfer to” as its translation candidate, but it is interesting to note that all other translation candidates (e.g., “have a connection to”) are better than “turn to”, because the sense disambiguation target phrase itself contains disambiguating information, and is therefore a better lexical choice unit. Consider a further example:

Input 我 想 打电话 到 日本 的 东京 ， 现在 东京 是 几 点 ？

Reference I’d like to call Tokyo, Japan. What time is it now in Tokyo?

WSD I want to make a call to Tokyo, Japan is Tokyo time now?

PSD I want to make a call to Tokyo, Japan what time is it now in Tokyo?

The PSD system translates the phrase “是 几 点” as a single target into “what time is”, with a confident PSD probability of 0.90. This prediction is not used by the WSD-augmented system, because the context-independent baseline translation probabilities prefers the incorrect translation “what time does it” higher than “what time is”, with much less confident scores (0.167 vs. 0.004). As a result, using only WSD predictions leads the words “是” and ”几 点” to be translated separately, and incorrectly.

In contrast, the following example demonstrates how multi-word PSD helps in selecting a mix of both longer and shorter phrases where appropriate:

Input 请 给 我 修理 一下 或 给 我 换 一下 。

Reference Please fix it or exchange it.

WSD Please fix it or I change it for me.

PSD Please give me fix it or exchange it for me.

In particular, by translating the phrase “请给我” as a whole, multi-word PSD avoids the problem caused by the incorrect reordering of the pronoun “I” in single-word WSD. The phrase translation is not optimal, but it is better than the single-word WSD translation, which does not make much sense because of the incorrect reordering. At the same time, the multi-word PSD predictions do not translate the phrase “东京 是 几 点” as a single target, which helps pick the better translation “exchange”.

It is worth noting that using multi-word PSD sometimes yields better lexical choice than single-word WSD even in cases when the same phrase segmentation of the input sentence is arrived at. This is the case in the following examples:

Input 全 是 个人 物品 。

Reference This is all my personal luggage.

WSD Is it all personal effects.

PSD They are all personal effects.

Input 咖啡 和 红茶 ， 您 要 哪个 ？

Reference Which would you like, coffee or tea?

WSD Which would you like, and coffee black tea?

PSD Which would you like, black tea or coffee?

The targets that are translated differently are single words in both sentences, which means that the WSD/PSD predictions are identical in the WSD-augmented SMT and PSD-augmented SMT experiments. However, the translation candidate selected by the decoder differs. In the first example, the WSD/PSD scores incorrectly prefer "and" with a probability of 0.967 to the better "or" translation, which is only given a probability of 0.002. However, the PSD-based translation for the whole sentence is correct, while the WSD-based translation is incorrectly ordered, perhaps letting the language model prefer the phrase "and coffee" which was seen 10 times more in the training set than the correctly ordered "and tea". Although this phenomenon requires more analysis, we suspect that having WSD predictions for every phrase in the SMT lexicon allows to learn better log linear model weights than for word-based WSD predictions.

10 When WSD/PSD predictions go wrong

The following examples show that for some sentences using sense disambiguation, whether single-word WSD or multi-word PSD, occasionally does not help or even hurts translation quality. Consider the following example:

Input 我 要 送 餐 服务 。

Reference Room service, please.

WSD I will take meal service.

PSD I want to eat service.

Here, the single word target "送" is incorrectly translated as "eat" and "meal", while a better translation candidate, "order", is given a lower WSD score. Another problem with this sentence is that the word "服务" is not seen alone during training, but in the collocation "房间 服务", so that "服务" was aligned to "service" only during training, and "room service" is not a translation candidate for "服务" in the SMT phrasal translation lexicon. WSD/PSD can only help to rank the given candidates, and there is nothing they can do when the correct translation is not in the original SMT phrasal translation lexicon.

Similarly, consider the following example:

Input 啊 。 给 我 帐单 。

Reference Uhh. Give me a Tab.

WSD Oh. I have the bill.

PSD Well, let me check.

The incorrect translation of "帐 单 。" as "check ." by the multi-word PSD model inappropriately influences the translation of the context, resulting in a sentence translation whose meaning has nothing in common with the reference.

This, of course, highlights the fact that for extremely short sentences containing only neutral words or extremely polysemous function words, WSD/PSD is not a good idea. In Example 7, there is actually no solid contextual evidence upon which the sense disambiguation model can decide whether "帐 单" should be translated as "bill", "check", or "tab". "给" is the highly polysemous verb "give", and "我" is the neutral word "I". In fact, without document level context, it would be hard even for a human translator to pick the right translation.

These observations suggest that in future evolutions of these directions, we might want to trigger PSD based on a cursory examination of sentence properties, in order to avoid hurting translation quality when there is simply no context information for PSD to exploit.

11 Conclusion

We have presented new comparative empirical evidence and data analysis strongly indicating that in order to be useful for improving the translation quality of current phrasal SMT performance levels, we will need *phrase sense disambiguation* models that are generalized to disambiguate phrasal target words, rather than traditional single-word sense diambiguation models. On one hand, the experimental results conducted on both the IWSLT-06 and NIST Chinese-English translation tasks, using eight different automatic evaluation metrics, have shown that—remarkably—incorporating phrase sense disambiguation *consistently* improves translation quality on *all* test sets for *all* evaluation metrics. But on the other hand, contrastive results where traditional single-word oriented WSD is incorporated into SMT leads to unpredictable effects on translation quality depending on the metric used, thus tending to confirm that the generalization from word sense disambiguation to phrase sense disambiguation is indeed necessary.

Analysis suggests that this very different behavior is made possible by the dynamic integration of phrase-based WSD predictions into SMT, which allow all phrase targets to compete during decoding, instead of forcing the SMT system to use a particular segmentation of its input sentence.

References

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for mt evaluation with improved correlation with human judgement. In *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 *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. Word sense disambiguation vs. statistical machine translation. In *the annual meeting of the association for computational linguistics (ACL-05)*, Ann Arbor, Michigan, 2005.

Marine Carpuat and Dekai Wu. Improving statistical machine translation using word sense disambiguation. In *The 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*, Prague, June 2007.

Marine Carpuat, Weifeng Su, and Dekai Wu. Augmenting ensemble classifi cation for word sense disambiguation with a Kernel PCA model. In *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.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. Word sense disambiguation improves statistical machine translation. In *45th Annual Meeting of the Association for Computational Linguistics (ACL-07)*, Prague, June 2007.

David Chiang. A hierarchical phrase-based model for statistical machine translation. In *the 43th Annual Meeting of Computational Linguistics (ACL-2005)*, Ann Arbor, Michigan, June 2005.

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

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

Ismael Garcia-Varea, Franz Och, Hermann Ney, and Francisco Casacuberta. Refi ned lexicon models for statistical machine translation using a maximum entropy approach. In *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. Effi cient integration of maximum entropy lexicon models within the training of statistical alignment models. In *AMTA-2002*, pages 54–63, Tiburon, California, October 2002.

Jesús Giménez and Lluís Màrquez. Context-aware discriminative phrase selection for statistical machine translation. In *Workshop on Statistical Machine Translation*, Prague, June 2007.

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 *Senseval-2, Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 17–20, Toulouse, France, July 2001. SIGLEX, Association for Computational Linguistics.

Philipp Koehn. Pharaoh: a beam search decoder forphrase-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 Ueffi ng, and Hermann Ney. Effi cient mt evaluation using block movements. In *EACL-2006 (11th Conference of the European Chapter of the Association for Computational Linguistics)*, pages 241–248, Trento, Italy, April 2006.

Hwee Tou Ng and Yee Seng Chan. English lexical sample task via english-chinese parallel text. In *4th International Workshop on Semantic Evaluation (SemEval-2007)*, Prague, June 2007.

Franz Josef Och and Hermann Ney. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449, 2004.

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

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

Lucia Specia, Maria das Gračas Volpe Nunes, and Mark Stevenson. Learning expressive models for word sense disambiguation. In *45th Annual Meeting of the Association for Computational Linguistics (ACL-2007)*, Prague, June 2007.

Christoph Tillmann, Stefan Vogel, Hermann Ney, A. Zubiaga, and H. Sawaf. Accelerated dp-based search for statistical translation. In *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. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–404, 1997.