Context-Dependent Phrasal Translation Lexicons 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

Most current statistical machine translation (SMT) systems make very little use of contextual information to select a translation candidate for a given input language phrase. However, despite evidence that rich context features are useful in stand-alone translation disambiguation tasks, recent studies reported that incorporating context-rich approaches from Word Sense Disambiguation (WSD) methods directly into classic word-based SMT systems, surprisingly, did not yield the expected improvements in translation quality. We argue here that, instead, it is necessary to design a context-dependent lexicon that is specifically matched to a given phrase-based SMT model, rather than simply incorporating an independently built and tested WSD module. In this approach, the baseline SMT phrasal lexicon, which uses translation probabilities that are independent of context, is augmented with a context-dependent score, defined using insights from standalone translation disambiguation evaluations. This approach reliably improves performance on both IWSLT and NIST Chinese-English test sets, producing consistent gains on all eight of the most commonly used automated evaluation metrics. We analyze the behavior of the model along a number of dimensons, including an analysis confirming that the most important context features are not available in conventional phrase-based SMT models.

1 Introduction

Existing phrase-based Statistical Machine Translation (SMT) systems have very weak models of context. In a Chinese-to-English SMT system for instance, the Chinese sentence context is only directly modeled within the Chinese phrases available in the phrasal translation lexicon, and indirectly modeled by the constraints imposed by the English language model. On the English side, only very local contextual information is available from the n-gram language model.

Error analysis of phrase-based SMT systems shows lexical choice errors, suggesting that even though phrase-based SMT systems often yield competitive translation accuracy, their performance does suffer from the weakness of their models of context.

Recent trends in phrase-based SMT also suggest that these simplistic models of context are far from optimal. For instance, the use of increasingly longer n-grams in language models suggests that more output language contextual information is needed than is available with local short n-gram modeling. Example-based MT style lookup of observed phrases in the parallel corpora at run time (Vogel, 2005) seeks to make better use of training context by using translations of longer observed phrases. Such approaches, however, still do not attempt to generalize over the rich contextual information available at training time.

The problem is that most contextual features that would be expected to enable better lexical choice are ignored in conventional phrase-based SMT models: sentence context words, parts of speech, and so on. Translation lexical choice should be influenced by full sentence context, not just local n-gram effects. Structural information about the surrounding context—not merely n-gram word identities—are also informative for lexical choice.

Such considerations have led us to propose modifying existing SMT approaches such that lexical choice scoring becomes dependent on the *dynamic* context. We aim to address the weakness that conventional word-based and phrase-based SMT models employ only *static* scores for phrase translation candidates, which are precomputed at training time and do not change depending on the context of the sentence being translated.

The key idea of our new model is that probabilistic phrasal translation lexicons should be context-dependent, taking into account the dynamic *full sentence* context as registered by a battery of richer features, factoring the effect of these features into the phrase translation probabilities being used to bias lexical choice decisions. This contrasts with traditional phrase-based SMT where input sentence context is only indirectly taken into account via the coherence constraints imposed by the language model in the output language.

There are two main open issues in such an approach. First, which dynamic context features are useful to phrasal translation lexicons? Second, how can dynamic context features be incorporated into the phrasal translation lexicons?

To address these questions, we appeal to the substantial body of WSD research, which historically has occurred largely independently of the SMT community. This seems appropriate since the work in WSD has long been directly targeted at the question of how to design context features and combine all the contextual evidence into a translation

or sense prediction. In particular, the Senseval/SemEval series of workshops have extensively evaluated systems with different feature sets, as well as different machine learning models for combining contextual evidence. Recent work on WSD for SMT also provides interesting insights.

Following this approach, we propose to exploit WSD insights to build context-dependent translation lexicons for SMT. We use context features and WSD models that were designed and evaluated on several Senseval-2 lexical sample tasks (Yarowsky and Florian, 2002) and Senseval-3 tasks (Carpuat *et al.*, 2004). On the one hand, these tasks included monolingual lexical choice tasks, where word senses are defined according to some manually built ontology or semantic network such as WordNet, HowNet, or the like. More relevant to the task at hand, however, are also the *multilingual lexical choice* tasks, where word senses are directly defined as the semantic distinctions made by another language (e.g., Chklovski *et al.* (2004)).

In many ways, the multilingual lexical choice tasks of Senseval/SemEval embody a more empirically justifiable approach to defining the sense inventory for WSD than the monolingual ontology-based lexical choice tasks. Sense inventories constructed on the basis of manually-built ontologies inherit an enormous variety of arbitrary choices made by the ontology builder, that can damage prediction and generalization accuracy. Since ontologies are not directly observable, as the saying goes, there are as many different ontologies as there are ontology builders.

Unlike manually built ontologies, on the other hand, lexical translations are directly observable. Thus, there is far less disagreement as to the inventory of choices. The alternative translations of a word or phrase consitute an empirically validated sense inventory: a Chinese word or phrase has a different sense if it is seen to translate into a different English word or phrase. Moreover, alternative translations of a given word or phrase are extremely good at discriminating between senses in manually built ontologies, so they offer all the same advantages of any sense inventory but without the disadvantages of manually built ontologies.

Our experiments using this approach show that context-dependent translation lexicons consistently improve translation quality. The translation improvement is registered consistently by all eight most commonly used automated evaluation metrics, across four different test sets from two different tasks where contextual information available differs. No other work to our knowledge has evaluated across such a range.

A closer analysis shows that the use of our context-dependent phrasal translation lexicons directly improves phrasal lexical choice for SMT, by giving better rankings of translation candidates and more discriminative scores. This suggests that WSD features and models match the phrasal translation task defined in the SMT lexicon.

In addition, context-dependent phrasal translation lexicons encourage the weak models of context in baseline SMT to make better predictions, as can be seen in phrasal segmentation. This confirms that, unlike for instance n-best reranking models, context-dependent phrasal translation lexicons are an appropriate model of context for SMT, since their predictions help improve all stages of decoding.

We then show that rich context features are also useful and necessary for long phrases which already encode local context information and are less ambiguous than single words or very short phrases. This confirms that our models of context are much richer and powerful than those of the baseline phrase-based SMT system, and that WSD models are needed for all phrases in the translation lexicon.

2 Related work

2.1 MT-oriented WSD

Research on word sense disambiguation, which has taken place largely independently of the SMT community, has been directly targeted at the question of how to design context features and combine a wide range of contextual evidence into making a translation or sense prediction. Evaluation of WSD models is typically done on WSD accuracy only—it is implicitly assumed that better WSD models will help higher level applications such as SMT.

Recently, several researchers have focused on designing WSD systems that use rich contextual information for the specific purpose of translation, instead of any sense distinctions. 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 and allows for rich expressive context features, but this system was also only evaluated on WSD accuracy, and not integrated in a full-scale machine translation system.

2.2 Context-dependent SMT

To the best of our knowledge, our model represents the first attempt at integrating a fully phrasal context-dependent translation lexicon into SMT, where evaluation is conducted by measuring the accuracy of the resulting SMT system on a translation task (as opposed to, for example, measures of word sense disambiguation accuracy as discussed in the preceding section).

In contrast with Brown *et al.* (1991), our approach incorporates the predictions of state-of-the-art WSD models that generalize across rich contextual features for any phrase in the input vocabulary. In Brown *et al.*'s early study of contextual features 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 and full phrasal translation lexicons.

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 transla-

tion accuracy. Our evaluation in this paper is conducted on the decoding task, rather than intermediate tasks such as word alignment. Moreover, in the present work, *all* commonly available automated MT evaluation metrics are used.

Another problem in the context-dependent 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 dynamic context-dependent phrasal lexicons 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.

Finally, there have been attempts at using WSD context models for the subset of the phrasal lexicon where input phrases are single words. While this makes the WSD task identical to traditional standalone WSD, it does not seem to be an optimal modeling approach for SMT. For instance, the model reported in Cabezas and Resnik (2005) can only perform lexical disambiguation using context features 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, our experiments applying their single-word based model to several Chinese-English datasets did *not* yield systematic improvements on most MT evaluation metrics. The single-word model has the disadvantage of forcing the decoder to choose between the static context-independent phrasal translation lexicons versus the dynamic context-dependent lexicons predictions for single words. In addition, this context-dependent lexicon model for single-words does not generalize to phrasal lexicons, as overlapping spans cannot be specified with the XML markup scheme. In this framework, using a context-dependent phrasal lexicon would require committing to a phrase segmentation of the input sentence before decoding, which is likely to hurt translation quality.

Note that for languages that do not contain space characters, such as Chinese (as considered in this paper), it is not even clear what "single word" means. Any string of characters could be considered as either a word or a phrase, if we insist on forcing an analogy to European languages.

2.3 WSD vs. SMT

In previous work, we have obtained seemingly conflicting empirical evidence on the usefulness of WSD in SMT. When we integrated the WSD predictions of Senseval-style WSD models into a word-based SMT system in a number of ways, for the first time, we surprisingly obtained a *decrease* in BLEU score (Carpuat and Wu, 2005b). However, we also showed that SMT systems alone perform much worse that WSD systems on a WSD task (Carpuat and Wu, 2005a), which suggests that WSD should have something to offer to SMT. Taken together, these results suggest that a better framework for integrating contextual evidence in SMT is needed. In this paper, we argue that such a framework is provided by context-dependent phrasal translation lexicons.

3 Context-dependent phrasal translation lexicons

As mentioned earlier, there are two main open issues in moving toward context-dependent phrasal translation lexicons. First, which dynamic context features are useful? Second, how can dynamic context features be incorporated into a probabilistic phrasal translation lexicon? Our approach to these two questions is described here.

3.1 Rich context features

The first key issue is how to define a rich set of dynamic context features. Our approach is to directly use the feature sets that have evolved in the course of extensive work in WSD shared task evaluations. Our feature definitions are inspired by the set which yielded the best results when combined in a naïve Bayes model on several Senseval-2 lexical sample tasks (Yarowsky and Florian, 2002). The dynamic context features we employ are typical of WSD models, and are therefore far richer than those used in most SMT systems. These features scale easily to the bigger vocabulary and sense candidates to be considered in a SMT task. Specifically, our feature set includes:

- bag-of-word context
- local collocations
- position-sensitive local POS tags
- basic dependency features

3.2 Defining the dynamic translation lexicon for SMT as WSD

The second key issue is how to incorporate the above dynamic context features into the probability estimates given by the phrasal translation lexicons, such that the translation probability depends on the context of the particular sentence being translated.

Our approach is to use a state-of-the-art WSD model to provide a context-dependent probability distribution over the possible English translation candidates for a given Chinese phrasal lexicon entry. This approach leverages experience from WSD research, which has focused on accurately combining a wide range of context features into a single sense categorization prediction. The word sense disambiguation subsystem we use is modeled after the best performing WSD system in the Chinese lexical sample task at Senseval-3 (Carpuat *et al.*, 2004).

Note that the WSD task definition is now task-dependent, and thus, differs slightly from dedicated Senseval-style WSD in the following respects:

- The basic unit to disambiguate is any Chinese entry in the phrasal translation lexicon. It can be any single word or multi-word phrase, unlike in Senseval-style WSD models were typically only single content words are disambiguated.

Table 1: Evaluation results on the IWSLT-06 dataset: Integrating the WSD-based context-dependent phrasal translation lexicon improves BLEU, NIST, METEOR, WER, PER, CDER and TER across all 3 different available test sets.

Test Set	Exp.	BLEU	NIST	METEOR	METEOR (no syn)	TER	WER	PER	CDER
Test 1	Baseline	42.21	7.888	65.40	63.24	40.45	45.58	37.80	40.09
	+ WSD	**42.38**	**7.902**	**65.73**	**63.64**	**39.98**	**45.30**	**37.60**	**39.91**
Test 2	Baseline	41.49	8.167	66.25	63.85	40.95	46.42	37.52	40.35
	+ WSD	**41.97**	**8.244**	**66.35**	**63.86**	**40.63**	**46.14**	**37.25**	**40.10**
Test 3	Baseline	49.91	9.016	73.36	70.70	35.60	40.60	32.30	35.46
	+ 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-based context-dependent phrasal translation lexicon improves BLEU, NIST, METEOR, WER, PER, CDER and TER.

Exp.	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
+ WSD	**20.62**	**7.538**	**59.99**	**56.38**	**72.53**	**85.09**	**58.62**	**68.54**

- The sense candidates are defined by the baseline phrasal translation lexicon, which is automatically extracted from parallel corpora, while dedicated WSD models for the monolingual lexical sample tasks use manually built sense inventories.

- To be consistent with the sense definitions, the training samples are also automatically extracted from the phrase-aligned parallel corpus: for every sentence pair where a consistent phrasal alignment is found for a phrasal lexicon entry, we can extract a Chinese sentence where the Chinese phrase is sense-annotated with its aligned phrasal translation. This presents the advantage of not requiring any manual annotation effort, while keeping the training data of the context-dependent phrasal translation lexicon consistent with that of the baseline lexicon.

Despite these differences, the WSD model supporting our context-dependent phrasal translation lexicon is still a pure WSD model. Just as in any Senseval/SemEval multilingual lexical sample task (e.g., Chklovski *et al.* (2004)), the task consists of disambiguating between semantic distinctions made by another language.

4 Evaluation on full-scale translation

We have conducted a comprehensive evaluation on two standard Chinese to English translation tasks, using all eight of the most commonly employed automated evaluation metrics. For every task, we evaluate translation quality with both BLEU (Papineni *et al.*, 2002) and NIST (Doddington, 2002) scores along with the recently proposed METEOR (Banerjee and Lavie, 2005) with and without WordNet synonyms. In addition, we report for each task 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).

Since our goal is to evaluate actual translation quality, we restrict ourselves to standard MT evaluation methodology. We do not evaluate the disambiguation accuracy of the embedded WSD models independently, as we cannot safely assume that higher WSD evaluation scores necessarily lead to higher translation accuracies.

4.1 Two very different tasks with 4 test sets

One set of experiments was 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 IWSLT 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, 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 experiment was 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.

4.2 A standard baseline SMT system

Our aim is to lay out an approach that can be expected to work in any reasonably common phrase-based SMT implementation. Since our focus is not on a specific SMT architecture, we chose the widely-used off-the-shelf phrase-based decoder Pharaoh (Koehn, 2004). Pharaoh implements a beam search decoder for phrase-based statistical models, and presents the advantages of being freely available and widely used.

Table 3: Translation examples with and without the WSD-based dynamic context-dependent phrasal translation lexicon, drawn from IWSLT data sets.

Input	请 转 乘 中央 线 。
Ref.	Please transfer to the Central train line .
SMT	Please turn to the Central Line .
+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 ?
+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 ?
+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 .
+WSD	I would like to reconfirm my reservation for this ticket .
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 .
+WSD	Excuse me , could you tell me the way to Broadway ?
Input	我 想 开 一个 账 户 。
Ref.	I want to open an account .
SMT	I would like to have an account .
+WSD	I would like to open an account .

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

The loglinear model weights were learned using Chiang's implementation of the maximum BLEU training algorithm (Och, 2003), both for the baseline and for the WSD-augmented system.

In the remaining sections, we discuss a number of different analyses of the experimental results.

5 Context-dependent modeling consistently improves translation

The most obvious observation on the experimental results, as shown in Table 1 for the IWSLT task and Table 2 for the NIST task, is that making the phrasal translation lexicons context-dependent produces higher translation quality on *all* test sets, as measured by *all eight* commonly used automated evaluation metrics. Paired bootstrap resampling shows that the improvements on the much larger NIST test set are statistically significant at the 95% level.

These are the only results to date we are aware of to show such consistent improvement across the entire range of metrics. Few previous attempts at integrating WSD predictions have shown even marginal improvement on any metric, much less on all metrics. The fact that the METEOR scores rise both with and without using WordNet synonyms to match translation candidates and references strongly suggests, moreover, that the improvement is not merely due to context-independent synonym matches at evaluation time.

6 Context-dependent modeling helps even on small-scale single-domain IWSLT tasks

Since the IWSLT task exclusively uses a relatively small set of short sentences from the travel domain, one might argue that lexical choice is easier for IWSLT than for NIST where the training data consists of larger, more heterogeneous parallel texts from the broader news domain.

However, even for the simpler IWSLT task, the weak models of context of the baseline phrase-based SMT system fail to capture sufficient context information, and augmenting phrasal translation lexicons with rich context features proves to be useful as can be seen in the examples from Table 3. These examples illustrate that, even in a single domain, there are genuine sense ambiguities in SMT phrasal translation lexicons (e.g. "turn" vs. "transfer" in the first example, "get" vs. "buy" in the second example, "open" vs. "have" in the last example).

Across all the IWSLT test sets, an average of 19 features per occurrence of a Chinese phrase are observed and used to build the dynamic context-dependent lexicon. This confirms that the rich WSD-style context features are indeed used for SMT translation lexicons, even in the single domain IWSLT corpus where sentences are quite short.

7 Context-dependent modeling improves phrasal lexical choice

Making the phrasal translation lexicons context-dependent instead of context-independent causes different phrasal translations to be chosen in a large proportion of the cases. The output sentence translations change for 25.49%, 30.40% and 29.25% of IWSLT test sets 1, 2 and 3, respectively. In contrast, 95.74% of the sentence translations change for the NIST test set. The fact that this percentage is much higher for the NIST test set can be explained by sentence length. Since the Chinese sentences are much longer in the NIST test set (examples are shown in Table 4), there are many more opportunities for the context-dependent phrasal translation lexicon to change the decoder's decisions.

The rich context features employed by the dynamic context-dependent approach tends to produce more accurate rankings of the alternative phrasal translation candidates, frequently correctly giving the top rank to the best translation. In contrast, the baseline context-independent translation probabilities do not correctly pick the top-ranked phrase translation as frequently.

Moreover, the scores given by the context-dependent phrasal translation lexicon tend to be more discriminative

Table 4: Translation examples with and without the WSD-based dynamic context-dependent phrasal translation lexicon, 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.

than baseline context-independent translation probabilities. This helps because the language and distortion models in SMT architectures can inadvertently override a top-ranked translation candidate. But because the context-dependent translation scores for the top-ranked candidate tend to be stronger (since the richer context features allow the model to be more confident), it becomes more difficult to inadvertently override the top-ranked phrase translation candidate.

8 Conventional SMT lacks the most useful context features

Further investigation reveals that the most useful context features for the context-dependent phrasal translation lexicon are dynamic context features not available in conventional phrase-based SMT. In order to get a sense of the usefulness of each class of context features, we performed an analysis of the feature weights learned by our maximum entropy WSD predictor. The higher the feature weight, the more useful that particular feature can be assumed to be. We normalized the feature weights for each WSD model, and then computed the average weight of features in each feature class over all Chinese phrase occurrences.

This analysis showed the top two most useful feature classes learned on the IWSLT data to be the POS tag preceding the ambiguous Chinese phrase and the POS tag following the ambiguous Chinese phrase. POS tags allow to generalize over all training examples seen, and are never used in conventional phrase-based SMT.

After the POS features, the third most useful class of features is the full sentence context as represented using bag-of-words. This, again, is information that is not available to phrase-based SMT, since all translation decisions only use local context. In phrase-based SMT, the full sentence context can only be memorized as a long phrase in the translation lexicon. In contrast, the bag-of-words models generalize over all the sentential contexts observed during training for all Chinese phrases.

9 Context-dependent modeling improves phrasal segmentation

Our dynamic context-dependent features do not only improve lexical choice, interestingly, but also the segmentation of the input sentence.

In phrase-based SMT systems, the segmentation of the input sentence indirectly makes use of input language local context by using translations for overlapping input phrases to build competing hypotheses during decoding. Analysis shows that the context-rich features incorporated in our predictions of the dynamic lexicon help the decoder to use longer input phrases on average.

This shows that our dynamic context-dependent lexicon is an appropriate model for integrating rich context features into phrase-based SMT, since its predictions propagate to all stages of decoding, even improving the implicit use of context through segmentation in traditional phrase-based SMT.

10 Rich WSD-style context features are necessary for the entire phrasal lexicon

One could argue that the phrase-based SMT systems do not need sophisticated WSD scores for the entire phrasal lexicon. In particular, it is unclear whether WSD-style rich context features, initially designed for single word disambiguation, are necessary to translate long phrases which intrinsically encode local context. If overlapping phrases of different length occur in the context-independent lexicon, the entries for the longer phrases can be seen as translations of single words or short phrases in their local context, which are less ambiguous than single words, and therefore might not require sophisticated WSD models.

Table 5: Evaluation results on the IWSLT-06 dataset: Integrating the full dynamic context-dependent lexicon for all phrases improves BLEU, NIST, METEOR, WER, PER, CDER and TER across all 3 different available test sets. In contrast, using the context-dependent WSD predictions only for single words has an unreliable impact on translation quality.

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

However, limiting feature-rich WSD predictions to single words has an unreliable effect on translation quality. Table 5 shows that, unlike using the full dynamic context-dependent lexicon, restricting it to single input words does not reliably improve translation quality across all metrics and all test sets when compared to the context-independent baseline phrase-based SMT system. For instance, using only dynamic context-dependent predictions for single words improves both versions of the METEOR metric on test set 1 compared to the baseline. However, on test set 3, METEOR strangely decreases when similarity matching is used, and increases when similarity matching is removed. This shows that the results are not consistent across different test sets for a given metric. For a fixed test set, say test set 1, the NIST and METEOR metrics are slightly improved while the BLEU and WER score gets worse, which shows that the results are not consistent across the most widely used automated evaluation metrics.

This confirms that the full sentential context and the syntactic features used by WSD models are necessary to translate long phrases as well as single words, and therefore that WSD is an appropriate framework for integrating contextual information into traditional phrase-based SMT.

Note that we also reported small improvements in BLEU score by using single-word WSD predictions in a Pharaoh baseline in Carpuat *et al.* (2006). These small improvements were obtained on a slightly weaker SMT baseline. On the contrary, Table 5 shows that BLEU scores now actually slightly decrease with our stronger baseline.

This restricted lexicon approach is similar to the proposal by Cabezas and Resnik (2005) who used the XML input scheme to provide word-based WSD predictions in the Pharaoh decoder. They obtained small gains in BLEU score on the Spanish-English Europarl task. However, their report does not check consistency of this improvement using other evaluation metrics and other data sets.

11 Conclusion

We have described a new SMT approach—dynamic context-dependent phrasal translation lexicon modeling—that draws insights from WSD-inspired translation tasks utilizing far richer *full sentence* context features than found in conventional SMT, leading to translation quality improvement in remarkably consistent fashion across varied data sets and all eight commonly used automated evaluation metrics. The improvements hold across three different test sets from the Chinese-English IWSLT 2006 test translation evaluation, as well as on a larger-scale NIST Chinese-English translation task at statistically significant levels. Even for small-scale single-domain IWSLT tasks where the individual gains are relatively small, incorporating the context-dependent phrasal lexicon never hurts, and helps enough to make it a worthwile additional component in a traditional SMT system.

Our study indicates that context-dependent phrasal translation lexicon modeling provides an appropriate modeling framework for successfully integrating the kind of predictions made by WSD-style modules into SMT architectures. Unlike in previous work where using WSD scores did not help translation quality (Carpuat and Wu, 2005b), our context-dependent phrasal translation lexicon allows combining the strengths of WSD and SMT models, by using the rich contextual features and machine learning models from WSD, while allowing the SMT system to make use of the WSD scores at all stages of decoding, since the context-dependent WSD scores are defined for every phrase in the bilexicon, just like regular context-independent probabilities.

Since our aim was to study this approach for the broadest possible class of models, we chose one of the most widely used SMT models as the baseline, namely flat phrase-based SMT. In light of the encouraging results, dynamic context-dependent phrasal translation lexicons might also be integrated into other current SMT models such as tree-structured SMT models employing various kinds of stochastic transduction grammars (e.g., Wu (1997), Wu and Chiang (2007)). For example, the context-dependent predictions might be utilized by a Bracketing ITG based decoder such as that of Wu (1996), Zens *et al.* (2004), or Cherry and Lin (2007), or alternatively a more grammatically structured statistical MT model that is less reliant on n-gram language modeling, such as the syntactic ITG based "grammatical channel" translation model of (Wu and Wong, 1998). The question remains open as to which type of SMT model could make most effective use of context-dependent phrasal translation lexicons.

Acknowledgements

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.

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. Evaluating the word sense disambiguation performance of statistical machine translation. In *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 *annual meeting of the association for computational linguistics (ACL-05)*, Ann Arbor, Michigan, 2005.

Marine Carpuat, Weifeng Su, and Dekai Wu. Augmenting ensemble classification 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.

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 *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 *Human Language Technology conference (HLT-2002)*, San Diego, CA, 2002.

Ismael Garcia-Varea, Franz Och, Hermann Ney, and Francisco Casacuberta. Refined lexicon models for statistical machine translation using a maximum entropy approach. In *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 *AMTA-2002*, pages 54–63, Tiburon, California, October 2002.

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 Ueffing, and Hermann Ney. Efficient 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.

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 *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 *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. A hybrid relational approach for WSD - first results. In *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 *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.

Stephan Vogel. PESA: Phrase pair extraction as sentence splitting. In *Machine Translation Summit X*. Phuket, Thailand, 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 *COLING-ACL'98*, Montreal,Canada, August 1998.

Dekai Wu. A polynomial-time algorithm for statistical machine translation. In *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,Switzerland, August 2004.