Evaluation of 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

We present new direct data analysis showing that dynamically-built context-dependent phrasal translation lexicons are more useful resources for phrase-based statistical machine translation (SMT) than conventional static phrasal translation lexicons, which ignore all contextual information. After several years of surprising negative results, recent work suggests that context-dependent phrasal translation lexicons are an appropriate framework to successfully incorporate Word Sense Disambiguation (WSD) modeling into SMT. However, this approach has so far only been evaluated using automatic translation quality metrics, which are important, but aggregate many different factors. A direct analysis is still needed to understand how context-dependent phrasal translation lexicons impact translation quality, and whether the additional complexity they introduce is really necessary. In this paper, we focus on the impact of context-dependent translation lexicons on lexical choice in phrase-based SMT and show that context-dependent lexicons are more useful to a phrase-based SMT system than a conventional lexicon. A typical phrase-based SMT system makes use of more and longer phrases with context modeling, including phrases that were not seen very frequently in training. Even when the segmentation is identical, the context-dependent lexicons yields translations that match references more often than conventional lexicons.

1. Introduction

After several years of surprisingly negative results, this year has finally seen several reports of statistically significant positive improvements from novel ways of integrating Word Sense Disambiguation (WSD) methods into statistical machine translation ((Chan et al., 2007); (Giménez and Màrquez, 2007); (Carpuat and Wu, 2007b)). In particular, as we shall describe, we have introduced a generalized WSD approach called Phrase Sense Disambiguation (PSD) where the key is a reliance on an innovative kind of resource: automatically acquired fully-phrasal translation lexicons that are *fully phrasal* to the same extent as phrase-based SMT architectures.

BLEU scores and other metric are important results, but from a lexical resources standpoint, to what extent this extremely large resource is actually necessary? This type of lexical resource is orders of magnitude larger than standard translation tables in phrase-based SMT, due to the large amount of information needed for the context-dependent modeling. Obviously, our fully phrasal context-dependent translation lexicons are even larger than conventional translation lexicons. Training such translation lexicons requires enormous amounts of computation. To what extent is their contribution observably useful? In this paper, we aim to address this question by presenting new direct data analysis showing that dynamically-built context-dependent phrasal translation lexicons are indeed more useful resources for phrase-based statistical machine translation (SMT) than conventional static phrasal translation lexicons, which ignore all contextual information.

Perhaps surprisingly, most current statistical machine translation 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 standalone translation disambiguation tasks, using context-rich approaches from WSD methods in standard SMT systems surprisingly did not yield the expected improvements in translation quality (Carpuat and Wu, 2005). In recent work, we have proposed a method for designing a context-dependent lexicon specifically for a given phrase-based SMT model. The baseline SMT lexicon, which uses translation probabilities that are independent of context, is augmented with a context-dependent score, defined using insights from

*This work was supported in part by DARPA GALE contract HR0011-06-C-0023, and by the Hong Kong Research Grants Council (RGC) research grants RGC6083/99E, RGC6256/00E, and DAG03/04.EG09.

both stand-alone translation disambiguation evaluations, and standard SMT. We showed that this approach reliably helps performance on both IWSLT and NIST Chinese-English test sets, yielding consistent gains on all eight of the most commonly used automatic evaluation metrics ((Carpuat and Wu, 2007b) and (Carpuat and Wu, 2007a)).

In this paper, we focus on the impact of context-dependent translation lexicons on lexical choice in phrase-based SMT and show that context-dependent lexicons are more useful to a phrase-based SMT system than a conventional lexicon. A typical phrase-based SMT system makes use of more and longer phrases from a context-dependent lexicon than from a conventional static lexicon, including phrases that were not seen very frequently in training. Even when the segmentation is identical, the context-dependent lexicons yields translations that match references more often than conventional lexicons.

This analysis provides insights that complement previous evaluation results. In Senseval evaluations, we have previously evaluated lexical choice of the underlying WSD models without integration into the SMT system (Carpuat et al., 2004). In translation quality evaluations, we have reported improvements in overall translation quality when integrating PSD-augmented context-dependent lexicons into SMT. However, widely used translation quality evaluation metrics such as BLEU (Papineni et al., 2002), NIST (Doddington, 2002), METEOR (Banerjee and Lavie, 2005), etc., aggregate the impact of many different factors. These metrics compare the translation hypothesis with one or more reference translations, but ignore how the translation hypothesis was generated. In this study, we investigate how phrasal translation lexicons are actually used by the SMT decoder, focusing on comparing the usage of our new context-dependent vs. conventional static lexicons.

2. Building context-dependent lexicons for phrase-based SMT

Since the lexicons are built specifically for phrase-based SMT application, the exact same input language vocabulary and translation candidates are used. The only difference between context-dependent vs. conventional lexicons lies in the parametrization. While translation probabilities in a conventional lexicon are computed once and for all during training, and used for any occurrence of a Chinese phrase at decoding time, context-dependent translation probabilities dare computed for every occurrence of a Chinese phrase in context. In other words, while conventional lexicons are static, context-dependent lexicons are dynamic.

2.1. Defining the dynamic translation lexicon for SMT

Leveraging insights from WSD research, which has focused on accurately combining a wide range of context features into a single prediction, we have introduced PSD models that generalize WSD to phrasal translations and provide a context-dependent probability distribution over the possible translation candidates for a given Chinese phrasal lexicon entry (Carpuat and Wu, 2007b). Our 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).

Note that PSD is task-dependent and slightly differs from dedicated Senseval-style WSD.

- 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.
- The sense candidates are defined by the baseline phrasal translation lexicon, which is automatically extracted from parallel corpora, while dedicated WSD models often used 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.

2.2. Context features

The features employed are typical of WSD and are therefore far richer than those used in most SMT systems. The feature definitions are inspired by the set which yielded the best results when combined in a naive 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. Specifically, our feature set includes:

- bag-of-word context
- local collocations

- position-sensitive local POS tags
- basic dependency features

2.3. Integrating context-dependent lexicons in phrase-based SMT architectures

It is non-trivial to incorporate a context-dependent lexicon 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-dependent fashion.

For every *phrase* in a given SMT input sentence, the PSD 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.

3. Methodology

We now turn to the evaluation of actual lexical choice integrated in SMT. We conduct a comparative analysis of the usage of the context-dependent lexicon vs. the conventional lexicon by the phrase-based decoder Pharaoh on the NIST-2004 Chinese to English translation task described in previous work (Carpuat and Wu, 2007b).

3.1. Experiment set-up

The conventional SMT phrasal lexicon is learned in a standard fashion. The training data is a newswire corpus of about 2M parallel Chinese-English sentences. Phrasal translation candidates are extracted if they are consistent with the intersection of bidirectional IBM Model 4 alignments, obtained with GIZA++ (Och and Ney, 2003) and augmented to improve recall. Context-independent phrasal translation probabilities are simply maximum likelihood estimates.

The context-dependent phrasal lexicon is learned by training PSD models for each Chinese phrase using the conventional lexicon as the sense inventory, and applying those models to each occurrence of a known Chinese phrase in the NIST-2004 test set. Note that the Chinese phrases and their English phrasal translation candidates are identical in both the context-dependent and the conventional lexicons. The only difference lies in the translation probabilities, which are computed dynamically depending on the context in the first case, or computed once and for all during training, thus ignoring contextual information in the second.

Since our focus is not on a specific SMT architecture, we incorporate our lexicons in the widely-used off-the-shelf phrase-based decoder Pharaoh (Koehn, 2004), as described in (Carpuat and Wu, 2007b). Note that Pharaoh uses a log linear model of translation that combines several features in addition to the phrasal translation probabilities from the lexicon: in particular, translation probabilities in both translation directions are used, as well as lexical weights which represent the alignment of words within the phrases.

3.2. Evaluation

Unlike evaluations of translation quality for an entire sentence or test set, focusing on lexical choice requires to know how the translation hypothesis was produced by the SMT system. More specifically, we need to know the phrasal alignment between the input sentence and the translation hypothesis (which is given by the trace option in Pharaoh).

In this study, we focus on two different but related aspects of lexical choice:

- phrasal segmentation, which can be seen as target selection for lexical choice
- translation selection for a given target

4. Context-dependent translation lexicons improve phrasal segmentation

We show that the context-dependent lexicon scores help the SMT decoder make better use of the available Chinese phrases for segmenting the input sentence. We are not interested in measuring the accuracy of a given phrasal segmentation. Phrasal segmentation is not an end in itself, but only a by-product of the full translation process, it does not make sense to define an a-priori gold standard or correct phrasal segmentation for a given sentence independently of the application. (Wu and Fung, 1994) showed that it is hard for human judges to agree on what a correct segmentation is. Instead, since our focus is on the usefulness of the translation lexicons as resources, we analyze how the Chinese phrases available in the lexicon are used.

4.1. Context-dependent lexicons encourage the decoder to use longer phrases

Context-dependent lexicons help the phrase-based SMT decoder make *truly phrasal* lexical choice: on average longer phrases are used with the context-dependent lexicons.

Figure 1 show that longer phrases are used in 1-best translations with context-dependent lexicons. With conventional phrasal translation lexicons, 63% of the phrases used are only single words, while this percentage goes down to 56% with context-dependent lexicons.

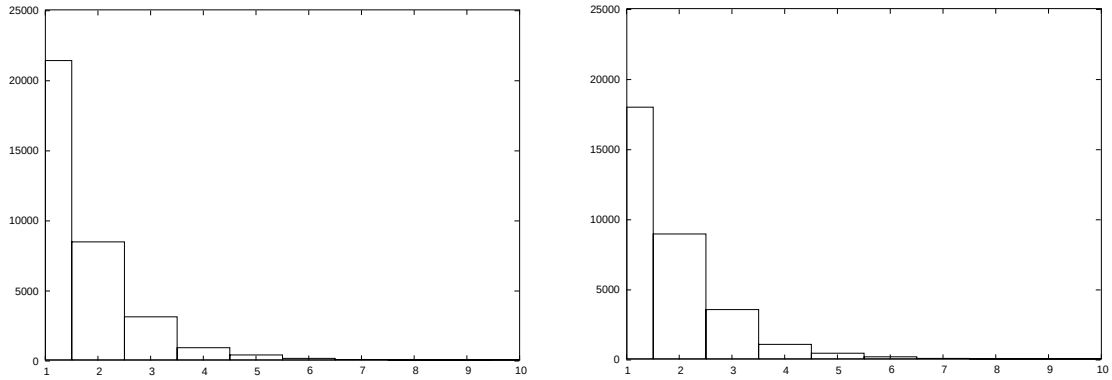

Figure 1: Frequency of phrases of length 1 to 10 used by the SMT decoder with the baseline (left plot) vs. context-dependent (right plot) lexicon on the NIST-2004 Chinese test set: fewer single words are used than with the conventional lexicon, while more longer phrases are used

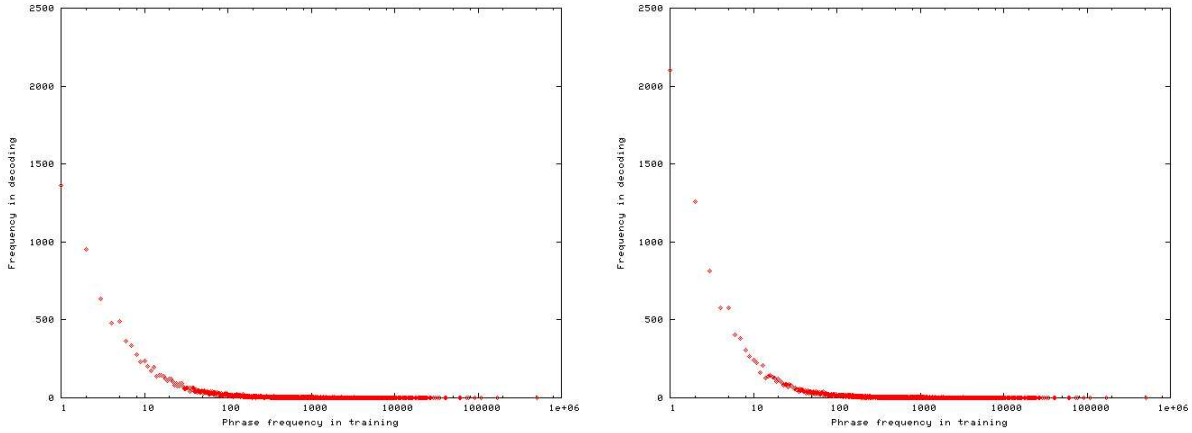

Figure 2: Distribution of training frequency of phrase types used by the SMT decoder with the baseline (left plot) vs. context-dependent (right plot) lexicon on the NIST-2004 Chinese test set: the context-dependent lexicons make more use of rare Chinese phrases that were seen only few times in the training corpus.

Using the phrasal segmentation from the baseline SMT system as a reference, we identify the segments which are a substring of the segments used with the context-dependent model: we find that the average length of those segments more than doubles, increasing from 1.25 to 2.65 with context-dependent modeling.

This analysis is consistent with previous work where several examples revealed that better phrasal segmentation is a key factor explaining the improved translation quality obtained with fully phrasal context-dependent lexicons as opposed to using context-dependent predictions for single words only (Carpuat and Wu, 2007a).

4.2. Context-dependent lexicons encourage the decoder to use more phrase types

Context-dependent lexicons help the phrase-based SMT decoder use more phrase types than with a conventional lexicon. 26% of the phrase types used with the context-dependent lexicon are not used at all with the conventional lexicon. In addition, 96% of those lexicon entries are true phrases and not single words. Note that these phrases are available in the conventional lexicon, but the conventional scoring does not encourage the decoder to make use of them.

4.3. Context-dependent lexicons encourage the decoder to use more rare phrases

Comparing the properties of the phrase types used with the context-dependent vs. conventional translation lexicons reveals that exploiting contextual information helps the SMT system pick more target types that were seen infrequently in training and have fewer training instances per translation candidate.

With the context-dependent lexicon, 38.4% more phrase types that occured 4 times or less in training are used by the SMT system. For phrases that were seen only once and twice, this figure raises respectively to 53.8% and 32.2%. Figure 1 shows the comparison of

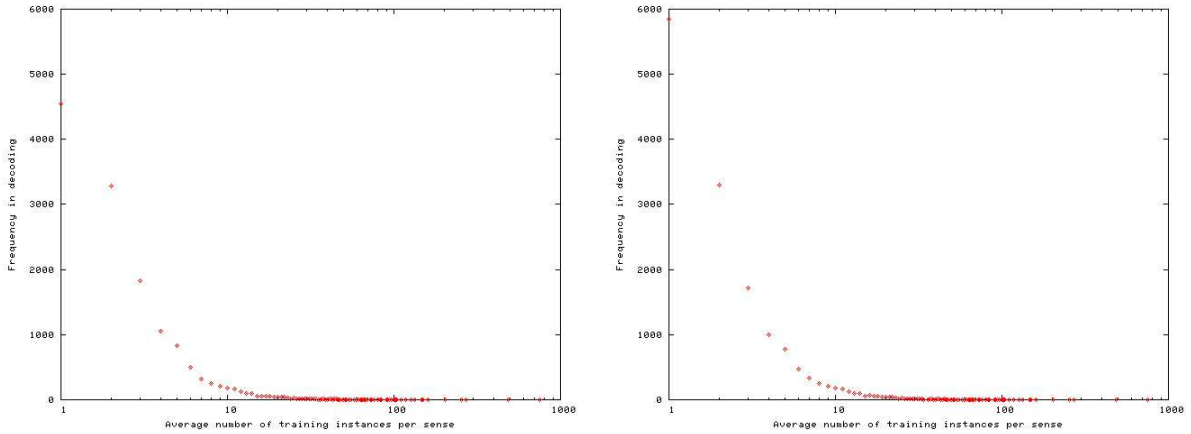

Figure 3: Distribution of average number of training instances per sense for phrase types used by the SMT decoder with the baseline (left plot) vs. context-dependent (right plot) lexicon on the NIST-2004 Chinese test set

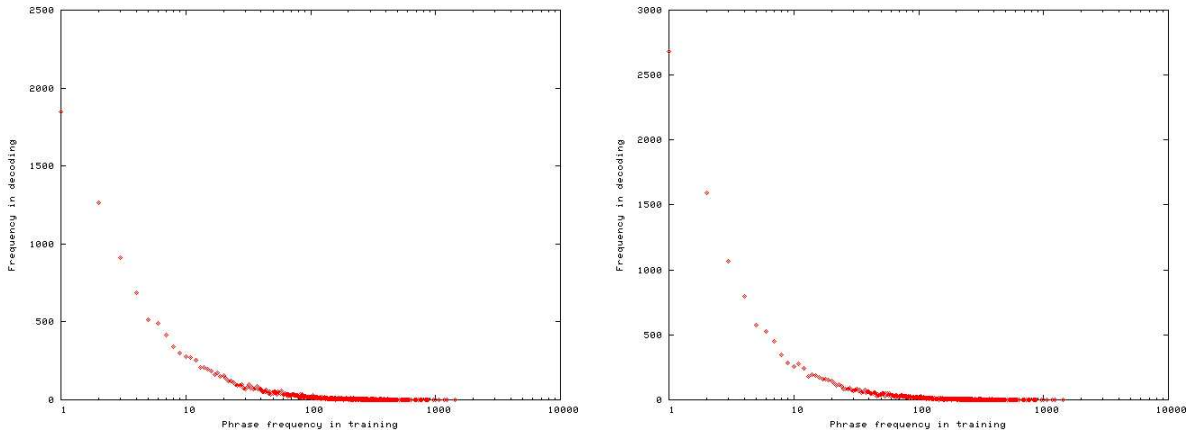

Figure 4: Distribution of average number of senses for phrase types used by the SMT decoder with the baseline (left plot) vs. context-dependent (right plot) lexicon on the NIST-2004 Chinese test set

the full distribution of phrase types used by the SMT systems according to their number of occurrences in training.

Breaking down the number of training instances per translation candidate further reveals that the same observation holds when considering the average number of training instances per translation candidate. The comparison of the actual distributions is showed in Figure 3. Figure 4 shows that context-dependent lexicons also encourages using more phrases that have only a single translation, for which we are therefore very confident.

Taken together, these observations suggest that incorporating context modeling helps SMT make better use of the phrases seen in training, and better distinguish between useful vs. noisy rare training instances.

5. Context-dependent lexicons yield better lexical choice

Second, we show that context-dependent lexicons yield better lexical choice than conventional lexicons. To some extent, the low order n-gram precisions in BLEU score and other metrics can be seen as an indirect evaluation of lexical choice accuracy. However, automatic metrics such as BLEU are only based on the translation hypothesis and ignore the segmentation and alignment between input sentence and hypothesis. Instead, we isolate lexical choice evaluation, by controlling for differences in segmentation.

5.1. Methodology

Since the Chinese phrases that are translated change with the phrasal segmentation of the input sentence, we choose to compare lexical choice for identical segments. This allows to separate all other factors of variation and to perform a direct comparison of the lexical choice accuracy for the context-dependent vs. conventional lexicons. Again, there is no available gold standard that exactly give correct translations for a given Chinese phrase. To solve this problem we use an approximation and check instead whether the phrasal translation match any of the reference transla-

Table 1: Comparing lexical choice accuracy with dynamic context-dependent vs. static context-independent translation lexicons for identical Chinese phrases: the confusion matrix shows that dynamic predictions correct the static predictions more frequently than vice-versa.

Lexicon	Static	match	no match
Dynamic			
match		1435	**2139**
no match		**683**	2272

tions available.

5.2. Context-dependent lexicon predictions match reference translations better for identical segments

For identical segments, when the context-dependent lexicon and the conventional lexicon yield different lexical choice, the context-dependent lexicon selects translations that match the references more often than the conventional lexicon. Drilling further, we find that 48% of the lexical choices that do not match the references with the conventional lexicon are corrected by the context-dependent lexicon. Table 1 shows the confusion matrix for the predictions of the context-dependent lexicon vs. the conventional lexicon.

6. Related work

Until recently, there has been surprisingly little interest in integrating context information in SMT lexicons. Early attempts were not clearly sucessful and more recent proposals focused on the primary goal of improving translation quality. None of these studies directly focus on the usage of the translation lexicons as we did here.

6.1. Early attempts at integrating context are limited by strictly residing within the IBM translation models

There have been a few early attempts at defining context-dependent translation lexicons, but their usefulness in SMT is unclear. (Berger et al., 1996) first used maximum entropy modeling to integrate local context information into IBM translation models, but they do not perform any significant evaluation of the impact on translation quality: only two example sentences are showed to be better translated with context-dependent modeling. In addition, their context-dependent translation models reside strictly within the Bayesian source-channel IBM models, which imposes restrictions on context modeling:

- Only a restricted feature set can be used: The probability distribution is estimated using maximum entropy based on position-sensitive local collocational features in a window of 3 words around the English word. Unlike in this work, syntactic information and wider sentential context are ignored.

- Contextual information is extracted from the *output* language: it is harder to extract reliable context features from noisy decoding hypotheses than from clean input sentences, as it is done in our context-dependent lexicons.

(Garcia-Varea et al., 2001) have extended the model proposed by (Berger et al., 1996) to include context features from both input and output language, but the resulting feature set is still insufficiently rich to make much better predictions than the SMT model itself. In addition to the output language context features defined by (Berger et al., 1996), words in a window of three words to the left in the input language are also considered. In contrast, our context-dependent lexicons are designed to directly model the lexical choice in the actual translation direction, and benefit from the much richer Senseval-style feature set.
In addition, unlike in the present work, Garcia Varea et al. did not fully integrate their context-dependent models were in decoding. (Garcia-Varea et al., 2001) and (Garcia-Varea et al., 2002) only report improved alignment error rates over IBM models 4 and 5 on the German-English Verbmobil corpus, and omit to evaluate the impact on translation quality. This is an issue since alignment is only an intermediate task, and improved alignments does not necessarily imply better translation quality. In later work, the context-dependent translation models are found to yield small but not statistical significant improvements in WER when used to rescore n-best lists (Garcia-Varea and Casacuberta, 2005). Full integration within decoding search was not attempted.

6.2. Recent work exclusively focused on translation quality

Our first attempt at using context-rich approaches from Senseval WSD in standard SMT systems surprisingly did not yield the expected improvements in translation quality (Carpuat and Wu, 2005). Following this disappointing results, several alternatives to strict Senseval-style WSD have been proposed, but all these proposals only evaluated the impact on automatic metrics of translation quality and did not directly study the use of the translation lexicons.

Most attempts limit the use of context-dependents models to the subset of the translation 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. (Cabezas and Resnik, 2005) used word-based Senseval WSD predictions to augment a Spanish-English phrase-based translation system and report small but not statistically significant improvements in BLEU score. (Giménez and Màrquez, 2007) also used WSD predictions in a phease-based SMT system 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 reliability of this model is not known.

In contrast, the context-dependent lexicons used in this work are defined for the entire phrasal vocabulary considered by the SMT system. 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 (Carpuat and Wu, 2007b). In direct contrastive experiments, we showed that it is necessary to use context-dependent translation probabilities for the entire phrasal lexicon in order to obtain those reliable improvements in translation quality(Carpuat and Wu, 2007a). (Stroppa et al., 2007) obtained also statistically significant improvements on NIST but not on BLEU score on Italian-English and Chinese-English IWSLT tasks, by augmenting the phrase-based Pharaoh SMT system with context-dependent phrasal translation probabilities learned using decision trees. However, their WSD models are weaker than our Senseval inspired PSD models, since their feature set is limited to co-occurring words and POS tags in a context window of only two words around the target.

These evaluations definitely show that fully phrasal context-dependent translation lexicons help translation quality in SMT, but they do not directly address the usage of the context-dependent lexicons as resources, as we have done in this work.

7. Conclusion

Our study reveals some of the reasons why 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.

Interestingly, improvements in translation quality are not only due to better lexical choice for a given target, which could be expected given improvements in metrics such as BLEU, NIST and METEOR, but also due to better phrasal segmentation of the input sentences and better use of the input phrases available in the lexicon. Specifically:

- For the exact same parallel training data, more and longer phrases are used in decoding, including phrases that were seen only few times in training: this suggests that context modeling help better exploit the available vocabulary.
- After compensating for differences in phrasal segmentation, the decoder selects better translations with context-dependent lexicons than with conventional lexicons.

This is consistent with previous contrastive studies which showed that using fully phrasal, as opposed to single-word, context-dependent lexicons is crucial to obtain reliable improvements in translation quality (Carpuat and Wu, 2007a).

This study therefore suggests that despite the additional complexity of extracting context features, training and applying WSD models, context-dependent phrasal translation lexicons are worth integrating into SMT. In this work, 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)).

8. References

Satanjeev Banerjee and Alon Lavie. 2005. 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.

Adam Berger, Stephen Della Pietra, and Vincent Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–68.

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

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

Marine Carpuat and Dekai Wu. 2007a. How phrase sense disambiguation outperforms word sense disambiguation for statistical machine translation. In *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2007)*, pages 43–52, Skovde, Sweden, September.

Marine Carpuat and Dekai Wu. 2007b. Improving statistical machine translation using word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*, pages 61–72, Prague, June.

Marine Carpuat, Weifeng Su, and Dekai Wu. 2004. 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. SIGLEX, Association for Computational Linguistics.

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

George Doddington. 2002. 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.

Ismael Garcia-Varea and Francisco Casacuberta. 2005. Maximum entropy modeling: A suitable framework to learn context-dependent lexicon models for statistical machine translation. *Machine Learning*, 60:135–158.

Ismael Garcia-Varea, Franz Och, Hermann Ney, and Francisco Casacuberta. 2001. 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.

Ismael Garcia-Varea, Franz Och, Hermann Ney, and Francisco Casacuberta. 2002. 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.

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

Philipp Koehn. 2004. 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.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–52. This is the citation to use for GIZA++.

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

Nivolas Stroppa, Antal van den Bosch, and Andy Way. 2007. Exploiting source similarity for smt using context-informed features. In *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2007)*, Skovde, Sweden, September.

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

Dekai Wu and Pascale Fung. 1994. Improving Chinese tokenization with linguistic filters on statistical lexical acquisition. In *ANLP-94: 4th Conference on Applied Natural Language Processing*, Stuttgart, October.

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

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