Language Model Adaptation for Difficult to Translate Phrases

Behrang Mohit * **Frank Liberato** ** **Rebecca Hwa** *,**
Intelligent Systems Program *, Department of Computer Science **
University of Pittsburgh, Pittsburgh, PA 15260, USA
{behrang,liberato,hwa}@cs.pitt.edu

Abstract

This paper investigates the idea of adapting language models for phrases that have poor translation quality. We apply a selective adaptation criterion which uses a classifier to locate the most difficult phrase of each source language sentence. A special adapted language model is constructed for the highlighted phrase. Our adaptation heuristic uses lexical features of the phrase to locate the relevant parts of the parallel corpus for language model training. As we vary the experimental setup by changing the size of the SMT training data, our adaptation method consistently shows strong improvements over the baseline systems.

1 Introduction

Statistical Machine Translation (SMT) systems generally use the same setup to translate all sentences. During decoding, the SMT engine searches through a large set of model parameters. Many parameters are sparse, irrelevant and noisy with respect to the individual sentence that is being translated. The dominant solution to this problem is to train with a larger corpus. This addresses the data sparsity problem, but it creates more irrelevant model parameters. Moreover, large volumes of training data may not always be available.

In this paper, we consider ways of filtering the irrelevant and noisy parameters in order to improve translation quality. We propose a language model adaptation method for the translation of phrases. We construct one adapted language model per source language phrase by using its lexical features. Our method uses these features along with the parallel corpora sentence associations to locate the relevant target language training sentences. Furthermore, we examine the idea of adapting the language model based on the level of difficulty that a phrase presents to the SMT system. We estimate the translation difficulty of phrases in a pre-translation step with either gold standard labeling or a trained classifier. We find that only phrases that are deemed difficult for the SMT system, benefit from the adaptation. Finally, we asses the feasibility of our selective adaptation within a complete SMT pipeline.

2 Motivation

Previously we explored the estimation of the translation difficulty of phrases, using an automatic MT evaluation (BLEU) score (Mohit and Hwa, 2007). We used our estimation method to label a set of gold standard phrases with the difficulty information. Moreover, we showed that it is possible to automatically learn the translation difficulty by constructing a phrase difficulty classifier.

In this paper, our aim is to improve MT quality by focusing on what we call Difficult To Translate Phrases (DTPs) within source language sentences. We compiled a group of DTPs for a baseline SMT system. We then manually examined a group of difficult phrases to learn about the reasons that make these phrases difficult.

Among various, often overlapping reasons behind phrase difficulty, we frequently observed problems that can be reduced by modifying the language models. Not all language model problems are related to sparsity. Moreover, the focus of this paper is on problems that arise when the models are not sparse. Specifically we aim to address the following problems:

Proceedings of the 13th Annual Conference of the EAMT, pages 160–167,
Barcelona, May 2009

i. Disambiguating target language words: Target language word ambiguity can be reduced if there are distinct source language words. For example, the word *official* is ambiguous in English (person vs. feature), but it has two distinct Arabic translations for its two senses. An adapted language model trained in the right text domain, can filter in or out generation of a phrase like *Egyptian official* or *official Egyptian*.

ii. Short distance word movements: Language model can also help the decoder to decide about short term word movements. For example the ordering of adjective and nouns are reverse in Arabic and English. Generation of a phrase like *senior egyptian cleric* depends on the n-gram parameters associated with that phrase[1]. A language model that is fitted for generation of the above phrase, is likely to have a high trigram probability for the actual trigram or has high probability for the two bigrams: *senior egyptian* and *egyptian cleric* and lower probabilities for alternative bigrams such as cleric senior.

In order to solve the above problems, we bias the language model towards the domain of the translation task. Through this biasing, we filter out those parts of the training data that are irrelevant. The resulting language model is adapted for the translation of specific input (in our case, phrase). This approach is the opposite of the typical method of reducing model sparsity via data expansion. In other words, we deliberately create model sparsity in areas that are found irrelevant to the translation task.

3 Contributions

In this paper we implement methods and experiments to answer the following questions:

i. How do we adapt the language model to overcome the translation difficulties noted in Section 2?

ii. What is a reasonable upper bound estimate of quality improvements that can be gained from adapted language models?

iii. Do all phrases have an improvement in translation from model adaptation?

iv. In a general MT test, does our proposed model adaptation framework improve translation quality?

In the following we briefly explain our approach for answering each of these questions which will be followed by experimental details and results.

[1]For this example, we are assuming that the phrase table is only providing the word to word translations

3.1 Our Adaptation Method

In SMT training, the target side of the parallel corpus is usually used for language model training. We would like to bias the language model training towards n-grams related to our translation phrase. To do so, we use the parallel corpus as a medium to locate relevant training instances. We start with the source language content words of the translation phrase. We call these source language terms, *seed* words. From the parallel corpus, we extract sentences that hold at least one of these seed words. We then include the associated target language sentence as one training sentence for the new language model. We call these training sentences, *relevant* sentences. This new relevant corpus is a much smaller subset of the original target language corpus. Some of the relevant sentences match longer n-grams with the translation task and we increase their training influence by repeating them. The repetition size is based on the length of the matched n-grams.

3.2 Estimating Upper Bounds

Estimating an upper bound for model adaptation gives us a realistic picture about the potentials of our constant resources (e.g, parallel corpus, etc.) and the expectations that we can have about language model adaptation. We present two methods which will more reliably gauge the impact of language modeling in the larger context of the decoder.

An aggressive upper bound: Given the constant phrase table, what is the closest possible decoding to the reference translation? To obtain such an upper-bound we simply train the language model with one reference translation. This ultra-overfited model is capable of generating sentences very close to the reference translation. However, the shortcomings of other translation resources such as unknown words or distortion errors are inherited when we use this language model. This upper bound tells us how much an n-gram language model, regardless of the training data, can be expected to improve translation quality.

A realistic upper bound: Unlike the aggressive upper bound scenario, in practice we train the language model on the target side of the parallel corpus. We are interested to estimate the best language model that we can build from that corpus. We still assume that we have access to the reference translation, but we no longer include it directly in the training data. Instead, we assume that we have a mechanism to choose the relevant parts of the target language corpus to train a language model. In order to train this upper bound we follow these steps:

for each n-gram in the reference translation: **do**
  if n-gram holds a content word: **then**
    Pick training sentences that hold the n-gram
  end if
  Use n-gram size to weight each training sent.
end for

In our experiments we will compare the above two upper bounds estimates against the traditional method of expanding the data for the language model training.

3.3 Model Adaptation For Phrase Translation

We would like to know if the translation of all phrases can be improved by the model adaptation. To answer this question, we apply our model adaptation method to two phrase groups: Difficult and Easy to Translate. We extract and label sets of gold standard phrases based on our variation of the procedure explained in Section 2. These sets are sentences whose most difficult or easy phrase is highlighted as the *focus* phrase. Each focus phrase is translated as part of a sentence. However, we only adapt the language model for the translation of the focus phrase, while the rest of the sentence is translated with the baseline language model.

3.4 Model Adaptation within an MT framework

Finally we would like to know if model adaptation improves the translation quality of a complete MT pipeline. We employ our adaptation method as part of a pre-translation pipeline. We still apply the adaptation to the DTP part, however we use a phrase difficulty classifier to find the most difficult phrase of each sentence. As shown in Figure 1, after the classifier finds the difficult phrase, the adapted language model is constructed for it.

After the creation of the adapted language model, the phrase is translated in the context of the full sentence, similar to the previous section.

4 Experimental Setup

We conduct our experiments on translation of Arabic to English via a phrase-based statistical translation engine. The SMT engine is the open source Phramer decoder (Olteanu, 2006) that uses the same training and decoding framework of the Pharaoh decoder (Koehn, 2004). We modify the decoder to use alternative language models for the translation of a special phrase within each sentence. We train both the baseline and the adapted language models using the SRI language modeling package (Stolcke, 2002).

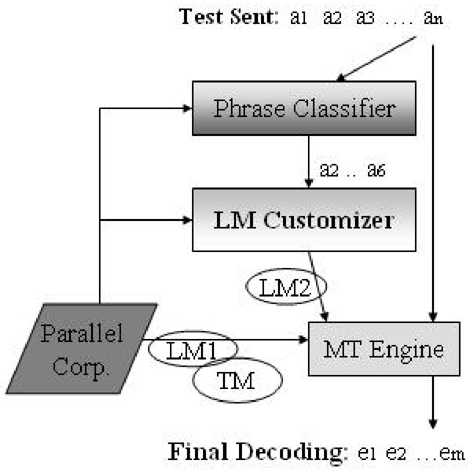

Figure 1: Translation Pipeline for Adapted LMs

4.1 Two SMT Systems

We construct two translation systems by varying the size of the training corpora. This data variation enable us to asses our approach in different translation scenarios. We use Arabic-English parallel corpora released by the Linguistic Data Consortium (LDC). The first (*small*) system is trained with one million words of parallel corpus[2]. The second *(medium)* system is cumulatively trained on an LDC corpus of 50 million words[3]. The language models for both systems are trained by the target language side of the parallel corpora. Both systems are tuned, using a development set of 500 sentences.

We use the LDC's multi-translation Arabic-English corpus [4] to extract a set of 3360 parallel phrases and label them as easy or difficult to translate. Since difficulty labels are system-specific, we label parallel phrases for both the small and medium systems. Phrase labeling and translation

[2]The corpora can be obtained from the Linguistic Data Consortium under catalog ID LDC2004T17, LDC2004T18.
[3]LDC2004E13, LDC2004E72, LDC2005E46
[4]LDC2003T18

evaluation are done based on the BLEU score (Papineni et. al., 2002).

4.2 Modified Easy-Difficult Phrase Labeling

Using the alignment-based phrase extraction tools, we automatically extract a corpus of parallel phrases. Each phrase consists of 5 to 15 source language tokens. The phrase set totally includes 32% of the sentences within the corpus. We use a held out parallel corpus to label each phrase as easy or difficult in the following round-robin fashion: Taking the translation BLEU score of the held out corpus as the referencee point, we add one phrase translation at a time and re-calculate the BLEU score. If the BLEU score is improved, the added phrase is an easy phrase (whose translation has improved the score). Otherwise, the phrase is labeled as difficult. In our labeling we use the first three reference translations to compute the BLEU scores. We keep the last reference for future MT quality evaluation. This separation reduces the bias of our labeling on our further experiments.

4.3 Building Difficulty Classifiers

When applying the language model adaptation to our full translation pipeline (Section 3.4), we use the phrase difficulty classifier to highlight the most difficult phrase of each sentence to replace our gold-standard labels. For each SMT system (small, medium), we construct a separate difficulty classifier. We compile a set of 12 language modeling features for the source and target languages to train these difficulty classifiers. We use the Support Vector Machine (SVM) as the classification model. These classifiers use the polynomial kernels which are tuned with a set of 100 development phrases.

4.4 Large Language Model Training

We also compare language model adaptation with the traditional method of using larger language models. We construct larger language models by adding up to 200 million words of the target language text to the language model training data.[5]. The larger language model's training data is cumulative with respect to the baseline models. These larger models are not in the scale of the current state of the art ultra-large models (Brant et. al., 2007). However by modifying the size of the parallel corpora along with the larger language models, we aim to simulate different data size scenarios. In this paper we report experiments where the larger language model is only used for translation of one phrase within the sentence (easy or difficult).

[5]This is a randomly chosen, continuous subset of the English Gigaword corpus.

5 Experiments

We conduct two sets of experiments by varying the accuracy of easy-difficult labeling. In the first setup (Sections 5.1 and 5.2), we use gold standard difficult or easy phrases with their associated sentences[6]. In the second setup (Section 5.3), we train an easy-difficult phrase classifier and use it to find the appropriate phrase. In both experiments, we modify the language model for the translation of the highlighted phrase and translate the sentence.

5.1 Model Modification for Difficult Phrases

We start with the upper bound estimates of language model adaptations. As explained in Section 3.2, there are two upper bound language models: the realistic upper bound and the aggressive upper bound. Table 1 presents the phrase level evaluation of these upper bound estimates along with the baseline system, the larger language models, and our adaptation method. Using the realistic upper bound language models, difficult phrases get sharp improvements. This large gap is indicative of the strong potentials that the baseline training data and the idea of language model adaptation hold. Table 2 presents a sentence level evaluation of the same experiments. Since the language model modifications are applied only to one difficult phrase per sentence, the score variations are smoothed at the sentence level. However, we still observe strong score improvements for the upper bounds and our adaptation method.

LM Modif.	Small Sys	Med Sys
Baseline	16.96	18.58
100M wds LM	21.17	21.29
200M wds LM	21.92	21.83
Realistic U.B.	26.83	28.33
Aggressive U.B.	54.23	60.11
Our Adapt.	21.12	22.16

Table 1: Comparison of different LM modifications for DTPs (Phrase Level Evaluation).

[6]For the small system we use a set of 453 easy and 551 difficult phrases. For the medium system, we use a set of 471 easy and 544 difficult phrases.

LM Modif.	Small Sys	Med Sys
Baseline	21.63	25.90
100M wds LM	23.76	27.04
200M wds LM	23.85	27.55
Realistic U.B.	25.98	30.07
Aggressive U.B.	35.92	37.46
Our Adapt.	23.47	27.93

Table 2: Comparison of different LM modifications for DTPs (Sentence Level Evaluation).

When we use our language model adaptation, each sentence is translated with language models that only use 5-10% of the baseline training data. Both systems' results indicate strong improvements above the baseline system and competitive performance with the large language models. For example the results of model adaptation based on a corpus of one million words competes with the results of models trained on corpora in the scales of 100 or 200 million words.

The improvements from our adaptation method relate to the way that language model influences DTPs. We have observed that the phrase table tends to be sparse for DTPs. As a result, the DTPs are often translated word for word so that the language model has to compensate for the word order. Our adaptation method is aimed at sharpening the discriminative power of the relevant n-grams. By filtering out the irrelevant training data, we distribute the probability mass only among the n-grams that are relevant to the translation phrase. The resulting small adapted language model is tailored for discriminating between a special set of n-grams that are relevant to our translation task.

5.2 Should we modify the model for all phrases?

In the above experiments, we applied language model modifications to phrases that are difficult to translate, and observed improvements in translation quality. However, it is not clear if this pattern of improvements applies to all phrases. To find the answer, we repeat those experiments for easy to translate phrases. Tables 3 and 4 present the result of these experiments. Contrary to difficult phrases, easy phrases which are completely tuned towards the baseline language model do not gain strong improvements. In some experiments their translation quality actually deteriorates.

For example easy phrases gain a modest improvement from the realistic upper bound language model. This indicates how close the baseline language model is to our (approximately) ideal language model. In other words, the parameters of the baseline language model are tuned towards generation of sentences close to the easy phrase references. The easy phrases are so fine tuned with the baseline language model that even a much larger language model can not compete with the baseline language model.

LM Modif.	Small Sys	Med Sys
Baseline	41.81	45.45
200M wds LM	39.26	44.03
Realistic U.B.	42.91	47.19
Aggressive U.B.	73.17	74.05
Our Adapt.	41.77	45.09

Table 3: Comparison of different LM modifications for easy phrs (Phrase Level Evaluation).

LM Modif.	Small Sys	Med Sys
Baseline	27.75	32.70
200M wds LM	26.93	32.02
Realistic U.B.	28.12	33.66
Aggressive U.B.	37.19	39.93
Our Adapt.	27.64	32.57

Table 4: Comparison of different LM modifications for easy phrs (Sentence Level Evaluation).

The results for model modification of easy phrases show that model adaptation can be more effective if it is applied selectively to only difficult phrases. In order to apply selective model adaptation, we need a mechanism to find those difficult phrases that need special handling. Therefore we construct a translation difficulty classifier.

5.3 Selective Model Adaptation for SMT

In this experiment, we apply our model adaptation into a complete SMT pipeline. Here, we use the Figure 1 architecture to find the most difficult phrase of a sentence and selectively modify the language model. For translation of each sentence, we apply the following procedure:

i. Compile the set of all source language phrases. To reduce the scale and keep the procedure similar to our gold standard labeling, we only consider phrases that have 5 to 15 words and have a contiguous baseline translation.

ii. Extract classification features for all phrases

LM Modif.	Small Sys	Med Sys
Baseline	18.09	22.51
Our Adapt.	19.06	23.55

Table 5: An Start-to-Finish experiment with difficulty classifier in the SMT pipeline

and their translations.

iii. Classify all phrases of a sentence and use the classifier's score to choose the most difficult phrase.

iv. Construct the modified language model for the most difficult phrase.

v. Translate the sentence by using the modified language model for the difficult phrase and the baseline setup for the rest of the sentence.

For this experiment, we use a held out Arabic-English test set[7]. For both small and medium sized systems, we experiment with the baseline language model and using our model adaptation method. Table 5 compares these three variations with the baseline.

For both systems, there are steady quality improvements above the baseline. It is clear the improvements are not as strong as the case where we apply the adaptation to gold standard DTPs. This is due to classification errors where a difficult phrase is missed or an easy phrase might be classified as difficult and gets selected for model adaptation where the new model might deteriorates the phrase's translation.

6 Discussion

We worked on the problem of modifying the language model to improve translation quality of difficult phrases. From various experiments we have observed the following:

i. Language modeling plays a significant role in SMT and strongly influences the difficulty of translation.

ii. A selective adaptation of the model based on the characteristics of the translation task has a strong potential to explore.

iii. Parallel data can be heuristically used to adapt language models based on the translation task.

We modified the baseline language models in two ways: we filtered out the irrelevant training data, and highlighted the relevant part of the remaining data based on n-gram matches. In the filtering part, we aim at removing some of the target word senses that are irrelevant to the translation task. We should clarify that here we do not address the problem of data sparseness. We actually cut some portions of the (irrelevant) baseline data. However, our filtering along with the appropriate weight setting, modify the relevant parameters of the model and make them biased towards the proper domain.

Table 6 presents sample translations where language model adaptation improves translation quality. The improvement in the first sample is related to the filtering aspect of the model adaptation. The English word *official* has two senses that are mixed up in the baseline translation. Since on the Arabic side the word has two distinct meanings, given the translation phrase and the parallel corpus we are able to exclude or lower the weight of the English sentences that have the irrelevant sense. The second sample is related to case where there is trigram match (egyptian police officer) between the parallel corpus and the translation phrase. The adaptation method locates the relevant sentence and increases its weight in the new language model training.

Our adaptation method is presented as an alternative to employing a larger amount of training data. In Figure 2, we compare various sizes of language models with the language model adaptation for difficult phrases. For both small and medium systems, our proposed adaptation method is competitive with the use of larger language models. Moreover, the realistic upper bound that uses the baseline training data is well above the largest tested language model. This encourages the further exploration of our idea.

A comparison of Tables 1 and 3 shows that the large quality gap between the easy and difficult to translate phrases holds even when we use the aggressive upper bound language models. This large gap shows the limits of influence for language models, especially for the difficult phrases. Due to numerous overlapping reasons, the language model can not solely resolve all the difficulties of the translation, so the translation quality of the difficult phrases stands well bellow the easy phrases.

[7]LDC2005T05: 606 sentences

baseline LM: the first exhibition for egyptian official of the painting on the UNK *Adapted LM:* the first official egyptian exhibition for the painting on the UNK *Reference:* the first official egyptian exhibition for painting on porcelain
baseline LM: ... that mohamed atef and is police officer former egyptian ... *Adapted LM:* ... that mohamed atef , is one of former egyptian police officer... *Reference:* ... that mohamed atif , a former egyptian police officer ...

Table 6: Sample Translation Improvements

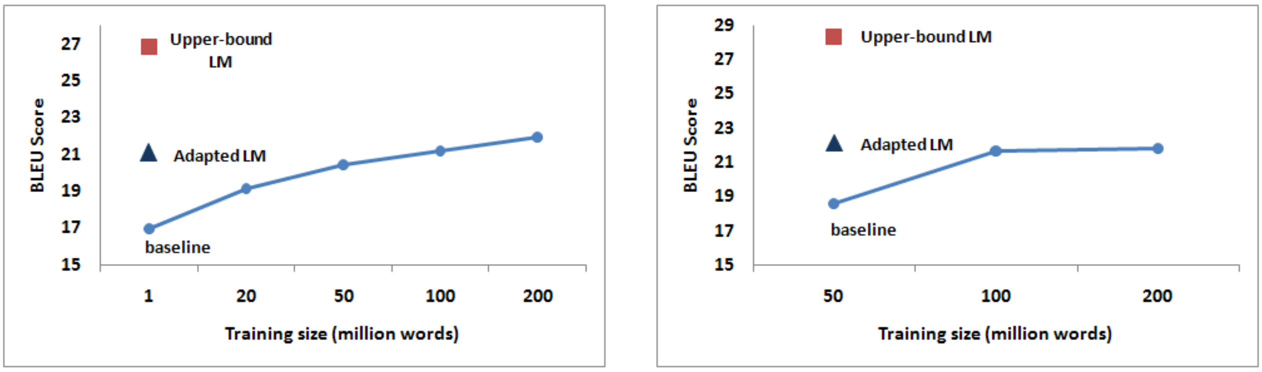

Figure 2: A comparison of model adaptation and training data expansion for systems that are trained on Small (Left) and Medium (Right) size parallel corpora

7 Related Work

Language Model Adaptation has been studied extensively in the speech processing and SMT communities (Kim and Khudanpur, 2004), (Tam et. al., 2007), (Snover et. al., 2008). The training data selection in most of the previous works involves selection of longer chunks of text. In contrast, we select training data for translation of each individual sentence. Also model adaptation has been mainly used in the translation of the entire test corpus with no special condition. Instead, we selectively use the adaptation only for translation of difficult phrases.

Our work is about improving the translation of phrases in the context of the sentence. Koehn and Knight (2003) present a frame work of isolated translation of noun phrases and re-combining the phrase translation with the rest of the sentence. Our phrase translation takes place in the context of the entire sentence but with a different language model. We still benefit from the full sentence context which reduces the translation error. The idea of decomposing the translation sentence and reattaching phrasal decodings has been also studied in the Multi Engine MT (MEMT) community. Mellebeek et. al. (2006) choose syntactically meaningful segments to decompose the sentences. In our approach we do not consider any syntactic constraint. Our major constraint is the translation difficulty of the phrase which makes us choose an alternative translation framework.

Another relevant area of work is training data subsampling. Johnson et. al. (2007) use the Fisher's exact test to validate the accuracy of the phrase table entries. They are able to reduce the size of the phrase table to 10% of its original size without a major loss of translation quality. It is not clear what percentage of the original training data is required to construct the reduced translation model. However their work confirms that training data can be used more efficiently. Ueffing et. al. (2007) also applies transduction learning to bootstrap new training sentences for SMT. New source language sentences are translated via a baseline SMT engine. Confidence estimation and model parameters are used for deciding to keep the sentence (and its decodings) in the new round of training. Our work follows a similar idea for altering the training data, but instead of adding additional data, we filter out and reweigh training data based on the relevancy to the translation task.

Comparison of different language models in SMT is one of our challenges. In this paper, we used the end result (translation) quality to evaluate the language models. An alternative method to consider is the *gold-in-sands* framework (Zhang, 2008). The idea is that a better language model should be able to rank the reference translations higher than alternative translations. It is implied that such a language model is more capable of generating sentences close to the reference translation.

8 Conclusion

In this paper, we presented a heuristics for training language models that are adapted for specific phrases deemed difficult to translate. When applied to a complete SMT pipeline, our model adaptation method improves the translation quality and competes with larger language models. For many target languages where large volumes of monolingual data is not available, usage of a larger language model is not an option and model adaptation is even more helpful.

We are working in several directions: We would like to extend our model adaptation by considering the interaction between the translation and the language models. Also, we are interested to expand our comparisons of the adapted language models vs. other language models (eg. baseline), outside the MT decoding. One area to explore is the usage of language model-based re-ranking of the reference translations.

Acknowledgments

This work has been supported in parts by US National Science Foundation Grants IIS-0832381 and IIS-0745914. We would like to thank our reviewers, the NLP group at the University of Pittsburgh and the MT group at the Carnegie Mellon University for their valuable comments and suggestions.

References

Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, Jeffrey Dean 2007. Large Language Models in Machine Translation In Proceedings of *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning(EMNLP-CoNLL-07).*

J Howard Johnson, Joel Martin, George Foster and Roland Kuhn 2007. Improving Translation Quality by Discarding most of Phrase Table In Proceedings of *EMNLP-CoNLL-2007.*

W. Kim and S. Khudanpur. Cross lingual latent semantic analysis for language model. In Erhard Hinrichs and Dan Roth, editors, *Proceedings of the International Conferences on Acoustic Speech and Signal Processing, ICASSP 2004.*

Philip Koehn and Kevin Knight 2003. Feature-Rich Statistical Translation of Noun Phrases In Proceedings of *43rd Annual Meeting of Association for Computational Linguistics (ACL-03).*

Philip Koehn 2004. Pharaoh: A Beam Search Decoder for Phrase-based Statistical Machine Translation Models Technical Report, USC Information Sciences Institute.

Bart Mellebeek, Karolina Owczarzak, Josef Van Genabith and Andy Way. 2006. Multi-Engine Machine Translation by Recursive Sentence Decomposition In Proceedings of *American Machine Translation Association Conference (AMTA-06).*

Behrang Mohit and Rebecca Hwa 2007. Localization of Difficult-to-Translate Phrases. In Proceedings of *ACL Workshop on Statistical Machine Translation.*

Marian Olteanu 2006. Phramer: An open-source statistical phrase-based mt decoder In Proceedings of *NAACL Workshop on Statistical Machine Translation.*

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 Association for Computational Linguistics (ACL-2002).*

Matthew Snover, Bonnie Dorr and Richard Schwartz. 2008. Language and translation model adaptation using comparable corpora In Proceedings of *Conference on Empirical Methods in Natural Language Processing (2008).*

Andreas Stolcke 2008. SRILM – an extensible language modeling toolkit. In Proceedings of *Conference on Empirical Methods in Natural Language Processing (2002).*

Yik-Cheung Tam, Ian Lane, and Tania Schultz. Bilingual-lsa based lm adaptation for spoken language translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics.*

Nicola Ueffing, Gholamreza Haffari and Anoop Sarkar 2007. Transductive Learning for Statistical Machine Translation In Proceedings of *EMNLP-CoNLL-2007.*

Ying Zhang 2008. Structured Language Model for Statistical Machine Translation PhD Thesis, Carnegie Mellon University. Pittsburgh, USA.