Experiments on Domain Adaptation for English—Hindi SMT

Rejwanul Haque, Sudip Kumar Naskar, Josef van Genabith, and Andy Way

CNGL, School of Computing,
Dublin City University, Dublin 9, Ireland
{rhaque, snaskar, josef, away}@computing.dcu.ie

Abstract. Statistical Machine Translation (SMT) systems are usually trained on large amounts of bilingual text and monolingual target language text. If a significant amount of out-of-domain data is added to the training data, the quality of translation can drop. On the other hand, training an SMT system on a small amount of training material for given in-domain data leads to narrow lexical coverage which again results in a low translation quality. In this paper, (i) we explore domain-adaptation techniques to combine large out-of-domain training data with small-scale in-domain training data for English—Hindi statistical machine translation and (ii) we cluster large out-of-domain training data to extract sentences similar to in-domain sentences and apply adaptation techniques to combine clustered sub-corpora with in-domain training data into a unified framework, achieving a 0.44 absolute corresponding to a 4.03% relative improvement in terms of BLEU over the baseline.

Keywords: statistical machine translation, domain adaptation

1 Introduction

In general, SMT models are trained on large corpora which may include quite heterogeneous topics. These topics usually define a set of terminological lexicons. Terminologies need to be translated taking into account the semantic context in which they appear. The semantic dependency problem could be overcome by learning topic-dependent translation models. There has been increased interest in incorporating data from domains with sufficient data in order to improve translation quality for small-data domains.

Several approaches have been applied to domain adaptation such as using two phrase tables jointly with a data source indicator feature added to the log-linear combination (Nakov, 2008), which has shown good results. Some researchers use multiple decoding paths of PB-SMT decoders such as Moses (Koehn *et al.*, 2007) for multi domain model adaptation (Koehn and Schroeder, 2007). Adaptations on the alignment model have been investigated where word alignments learned from a large out-of-domain corpus are used to align words for a small-scale domain (Wu *et al.*, 2005). Some researchers proposed a way to retrieve only those sentences which are most similar to the test data in order to improve the training data's match with respect to domain, topic, and style (Eck *et al.*, 2004). Recently, researchers incorporate out-of-domain data through learning phrase templates (phrase generalisation) in order to improve translation quality (Lim and Kirchhoff, 2008).

In the present work, we conduct experiments on the English—Hindi language pair. Like other Indian languages, Hindi is also a free phrase order (used with emphasis and complex structures) language. Therefore, applying adaptation techniques on such a language pair could produce interesting findings.

For adaptation purposes, previous research used similarity metrics to cluster heterogeneous corpus data into sub-corpora with homogeneous topics. In order to compute the distance

between a sentence and a cluster, different similarity metrics have been proposed. (Carter, 1994) introduced an entropy reduction based similarity metric to cluster a multi-domain monolingual corpus. A regular expression based similarity function has been defined to build class specific language models (Hasan and Ney, 2005). In our research, we explore a clustering technique based on an n-gram overlap metric to extract sentences similar to in-domain text from large out-of-domain training data.

We employ domain adaptation techniques to adapt an out-of-domain bilingual corpus to an in-domain SMT model using clustering to extract sentences similar to in-domain text from large out-of-domain training data. We apply adaptation techniques to combine sub-corpora with in-domain small-scale training data into a unified framework.

The remainder of the paper is organized as follows. In section 2 we discuss related work. Section 3 describes experimental results using our baseline SMT model. In section 4 we describe the domain adaptation techniques which are employed to combine multiple models. Section 5 presents the results obtained, together with some analysis. Section 6 concludes, and provides avenues for further work.

2 Related Work

Topic-dependent modeling was effectively applied in speech recognition to improve the quality of models (Carter, 1994). Adaptation technology has been widely used in language modeling in the same filed over the last decade (Iyer *et al.*, 1997*).*

Langlais (2002) was the first to introduce domain adaptation in SMT by integrating terminological lexicons in the translation model resulting in a significant reduction in word error rate (WER). Over the last years, many researchers have investigated the problem of combining multi-domain data. Wu and Wang (2004) and Wu *et al.* (2005) propose an alignment adaptation approach to improve domain-specific word alignment.

Eck *et al.* (2004) present a language model (LM) adaptation technique in SMT applying information retrieval theory following the approach of Mahajan *et al.* (1999) in speech recognition. This approach was further refined by Zhao *et al*. (2004). Hildebrand *et al.* (2005) adapt the translation model by selecting similar sentences from the available training data applying the approach of Eck *et al.* (2004). The adapted models significantly improve the translation performance compared to baseline systems.

More recently, Bulyko *et al.*, (2007) studied language model adaptation for SMT. They explored discriminative estimation of language model weights by directly optimizing machine translation evaluation metrics such as BLEU score. An improvement 0.4 BLEU score was reported.

Hasan and Ney (2005) cluster the training sentences into specific classes based on regular expressions to build class specific language models. They proposed a method of interpolating class specific and global language models following the mixture model proposed by Iyer *et al.* (1999). The results look promising in terms of perplexity reduction, as well as error rates obtained for a translation task using an *n*-best list rescoring framework. Both Yamamoto and Sumita (2007) and Foster and Kuhn (2007) extended this work to include the translation model. Yamamoto and Sumita (2007) used an unsupervised clustering technique on an unlabelled bilingual training corpus. Each cluster is regarded as a domain. Clusters are defined automatically (without human knowledge) and created by the entropy reduction based method (Carter, 1994). Civera and Juan (2007) introduce the mixture extension for HMM alignment models. This approach generates topic dependent viterbi alignments to feed a state-of-art phrase based SMT (PB-SMT).

Koehn and Schroeder (2007) investigated domain adaptations by integrating in-domain and out-of-domain language models as log-linear features in an SMT model. They also used multiple decoding paths (Birch *et al.* 2007) for combining multiple domain translation tables in the state-of-the-art PB-SMT decoder Moses (Koehn *et al.*, 2003).

Nakov (2008) combine an in-domain model (translation and reordering model) with an out-of-domain model (translation and reordering) into Moses (Koehn *et al.*, 2007). They derived log-linear features to distinguish between phrases of multiple domains by applying the data-source indicator features and showed modest improvement in translation quality.

Munteanu and Marcu (2006) automatically extract in-domain bilingual sentence pairs from large comparable corpora to enlarge the in-domain bilingual corpus. They showed a modest gain over the baseline system. Ueffing *et al.* (2007) introduced transductive semi-supervised learning for SMT, where source language corpora are used to train the models. The transductive learning can be seen as a means to adapt the SMT system to a new domain. Sentences from the devset or testset are translated repeatedly and the generated translations are added to training data to improve the performance of the SMT system. They reported a significant improvement of BLEU over the baseline.

Wu *et al.* (2008) proposed a method to perform domain adaptation for SMT, where in-domain bilingual data do not exist. The transductive learning method (Ueffing *et al.* 2007) has been used to adapt the in-domain monolingual corpus. Wu *et al.* (2008) also showed that log-linear interpolation performs better than linear interpolation to combine in-domain and out-of-domain language models as well as translation models.

Snover *et al.* (2008) describes a novel domain adaptation method for utilizing monolingual target data to improve the performance of a statistical machine translation system on news stories. For the translation of each source text, a large monolingual data set in the target language is searched for documents that might be comparable to the source text. These documents are then used to adapt the MT system to increase the probability of generating texts that resemble the comparable document. Experimental results show substantial gains.

Lim and Kirchhoff (2008) proposed a method for incorporating out-of-domain data through phrase generalization in order to improve the Italian-English translation quality. They showed a noticeable improvement in translation quality.

Finch and Sumita (2008) employed probabilistic mixture weights to combine two models for questions and declarative sentences with a general model. Foster and Kuhn (2007) used distance based weights in a mixture model. In contrast to their work, Finch and Sumita (2008) used a probabilistic classifier to determine a vector of probability representing class-membership. They performed experiments on a number of language pairs and experimental results showed the usefulness of their method.

Domain adaptation techniques can be broadly divided into two categories: (i) adaptation techniques to improve word alignment models; such as Wu *et al.* (2005) and Civera and Juan (2007) and (ii) adaptation techniques to combine multiple domain models; such as Koehn and Schroeder (2007) and Nakov (2008). The present work falls into the second type of adaptation method.

Identifying similar sentences plays an important role in domain adaptation. Experiments on detecting and clustering similar sentences from a corpus have been extensively studied in many natural language processing applications. Seno and Maria (2008) performed clustering methods to extract similar sentences from text documents.

3 Experimental Data

The shared task on English—Hindi SMT (Venkatapathy, 2008) released two different datasets:
(1) TIDES-IIIT Dataset: this dataset was collected for the DARPA-TIDES surprise language contest on Statistical Machine Translation in 2002. This corpus is general domain with news articles forming the greatest proportion. The training, development and test sets contain 49,504, 988 and 697 sentences respectively.
(2) EILMT-Tourism Corpus: this dataset was provided by the EILMT consortium funded by DIT, Govt. of India. This dataset is a domain-specific corpus developed purely to build machine translation systems catering for the tourism domain. The released EILMT dataset was not a

discrete set, i.e. there were some common sentences among the testset, devset and training set. Therefore, for proper evaluation we have removed the duplicate sentences among them to make it a discrete dataset. The training, development and test sets contain 6,755, 500 and 495 sentences respectively.

Table 1: Baseline results on EILMT and TIDES datasets.

Baseline	BLEU	NIST	METEOR	WER	PER
EILMT	10.93	4.56	28.59	82.06	65.67
TIDES	9.56	4.29	37.76	83.99	62.77
Concat[1]	10.78	4.65	34.18	82.28	56.48

Table 2: Experiments adding domain-specific lexicon.

TIDES Train Data+ EILMT lexicon	BLEU	NIST	METEOR	WER	PER
LM (EILMT)	8.25	4.31	37.08	84.61	58.54
LM (EILMT+TIDES)[2]	8.15	4.16	37.68	84.96	59.73

Baseline results for the two data sets are reported in Table 1. The accuracy of the EILMT baseline system (10.93 BLEU) is much higher than the TIDES baseline system (9.56 BLEU), although the size of the TIDES training data is almost eight times the EILMT training data. One possible reason may be that the sentences in TIDES are not faithful translations i.e., the target sentences convey only the meaning of the source sentences in the best possible way in the target language (Venkatapathy, 2008).

Table 1 shows that concatenating TIDES training data with EILMT training data hurts the translation quality when translating the EILMT testset (0.15 BLEU point below the baseline). In the above situation, the out-of-domain training data overwhelms the in-domain training data due to the sheer relative size. This result clearly indicates the necessity of careful adaptation of out-of-domain data in English—Hindi SMT.

3.1 Adding a Domain-Specific Lexicon

Langlais (2002) was the first to add a domain-specific lexicon into PB-SMT to improve translation quality. Following his approach, we build an SMT model adding a tourism-domain lexicon to the large TIDES training data. An SMT model is built using this combined data. The experimental results are shown in Table 2. The tourism-domain lexicon is obtained by training on the EILMT corpus using the GIZA++ toolkit[3].

The first row as well as the second row in Table 2 show that results are much lower across the all evaluation metrics except METEOR than the results obtained by concatenating the two training data sets as shown in Table 2. The system produces 2.68 BLEU points less than the concatenation model. But, surprisingly the METEOR score is much higher than for the concatenation model when translating EILMT testset. This clearly indicates that we need further study to combine multi-domain training data.

4 Domain Adaptation

In this section, we describe the domain adaptation techniques that we applied to our experiments for translation and language model adaptation. Our goal is to make use of all available training data to build language models (LM) and translation models (TM). Generally, language model and translation model adaptation in SMT are performed applying two interpolation methods: (i) linear interpolation and (ii) log-linear interpolation. Wu *et al.* (2008) pointed out that log-linear interpolation performs better than linearly interpolating multiple

[1] Concatenation of TIDES and EILMT training set (Devset, Testset and LM from EILMT data)
[2] Log-linear interpolation of EILMT and TIDES language models (This technique is described in Section 4).
[3] Available at http://www.fjoch.com/GIZA++.html

domain-specific language models and translation models. Therefore, we interpolated EILMT and TIDES language models and translation models using a log-linear combination.

4.1 Language Model Adaptation

We used the language modeling toolkit SRILM (Stolke, 2002) to build two language models from the target side of the EILMT and TIDES training data. We performed log-linear interpolation of multi-domain translation models. This results in a straight-forward combination of in-domain and out-of-domain language models. Fortunately, the PB-SMT Moses decoder supports log-linear combinations of language models. Language model weights are optimized with minimum error rate training (Och, 2003).

4.2 Translation Model Adaptation

In general, translation models are built separately for each of the domain specific corpora. These models are then combined using two techniques: (i) linear interpolation (ii) log-linear interpolation. We performed the log-linear interpolation of multi-domain translation models. There are two ways of performing log-linear interpolation:

Multiple Decoding Paths: a recent feature of Moses is multiple decoding paths. This alternate decoding path model was developed by Birch *et al.* (2007). Here we use Moses' capabilities to use different decoding paths for translation model adaptation. As per our requirements, we used two and three decoding paths depending upon the number of translation models. Each decoding path is dedicated to a particular translation table. Weights for each table are optimized by minimum error rate training. In this approach, multiple lexicalised reordering tables are integrated in a log-linear combination into the Moses framework.

Data-Source Indicator Features: there is another approach introduced by Nakov (2008) to combine multiple translation tables in a log-linear combination. For each translation table, there is a data-source feature which is integrated in the log-linear combination into Moses. The data-source indicator features distinguish different domain-specific phrases. In this approach, multiple translation tables are combined together to create a single translation table by assigning a particular value for each data-source feature according to domain importance. We refer the interested reader to Nakov (2008) for the details of how these features are integrated into Moses for multi-domain translation tables. This approach is also applicable to combine multi-domain reordering tables.

4.3 Clustering Out-of-Domain Corpus Data

The idea is to extract sentences as similar as possible to the tourism domain sentences from the out-of-domain parallel corpus (TIDES). Hasan and Ney (2005) cluster sentences into specific classes based on regular expressions. They measure the perplexity of sentences as evaluation criteria. Yamamoto and Sumita (2007) applied unsupervised clustering on a bilingual training corpus. In their case, each cluster is regarded as a domain. Clusters are defined automatically (without human knowledge) and created by the entropy reduction based method (Carter, 1994). In both techniques, the total number of clusters is pre-defined by the user. Similarly, in our case, two clusters were defined, one cluster contains sentences similar to the tourism domain sentences (EILMT corpus), the other cluster contains the remaining sentences.

We build a language model on the English side of the EILMT bilingual training data. We extract high-frequency *n*-grams (upto 5-grams) and their probability distribution from the language model. Stop words are removed from the high-frequency *n*-gram list. Intuitively, the remaining high frequency *n*-gram word sequences are tourism domain-specific terminologies. For clustering purposes, we defined a similarity function that measures sentence similarity based on the occurrences of domain-specific *n*-grams in the English sentences of the TIDES bilingual training data. The similarity function is based on the *n*-gram overlap metric, which calculates the similarity score of an input sentence depending on the number of overlapping

domain-specific *n*-grams in that sentence. The sentence similarity score is derived by adding the log-probabilities of all overlapping *n*-grams. The similarity thresholds of clusters are set manually. Thus, sentences of the TIDES bilingual corpus are clustered into two sub-corpora using the above monolingual sentence clustering method. The first cluster is a sub-corpus containing parallel sentences similar to our in-domain corpus. The second cluster contains the remaining parallel sentences. The first sub-corpus (SUB1) contains 5893 sentences, (11.9%) which are closer to in-domain tourism corpus. The remaining sentences form the second sub-corpus (SUB2) containing 436111 sentences.

5 Results and Analysis

We performed two sets of experiments. The first set of experiments tests the adaptations of out-of-domain TIDES training data with in-domain EILMT training data. Experimental results are reported in Table 3. The second set of experiments is performed applying adaption techniques on clustered sub-corpora (SUB1 and SUB2) with the in-domain EILMT corpus. Experimental results are reported in Table 4. In all our adaptation experiments, we performed log-linear interpolation of in-domain and out-of-domain language models.

Table 3 shows that when data-source indicator features are applied in the adaptation technique, system performance improves 0.29 BLEU point over the EILMT baseline (Table 1). When domain adaptation using 2-decoding paths is employed, the system produces a 0.19 BLEU point improvement over the baseline.

The experimental results of clustering the out-of-domain training data are reported in Table 4. We obtained a 0.44 BLEU improvement over the baseline when applying the adaptation technique using data-source indicator features on EILMT training data and two clustered sub-corpora. This experiment improves system performance across all the evaluation metrics. A 3-decoding path adaption technique on EILMT training data and two clustered sub-corpora results in an improvement of 0.25 BLEU point over the baseline.

The experimental results in Table 3 and Table 4 clearly show that adaptation techniques improve across all the evaluation metrics over the baseline results displayed in Table 1. This indicates the usefulness and effectiveness of applying domain adaption techniques in English—Hindi Statistical machine translation.

Table 3: Domain Adaptation combining EILMT and TIDES corpus.

Adaptation	Experiments	BLEU	NIST	METEOR	WER	PER
Data Source Indicator Features	TM_{EILMT}+TM_{TIDES}	**11.22**	4.65	33.1	82.56	56.46
Multiple Decoding Paths	TM_{EILMT}+TM_{TIDES}	11.12	**4.69**	**34.46**	**80.44**	**55.63**

Table 4: Domain Adaptation combining EILMT and clustered sub-corpora.

Adaptation	Experiments	BLEU	NIST	METEOR	WER	PER
Data Source Indicator Features	TM_{EILMT}+TM_{SUB1}	11.07	4.59	31.04	81.93	56.72
	TM_{EILMT}+TM_{SUB1}+TM_{SUB2}	**11.37**	**4.68**	**34.29**	**81.72**	**55.94**
Multiple Decoding Paths	TM_{EILMT}+TM_{SUB1}	11.00	4.56	31.56	82.04	56.71
	TM_{EILMT}+TM_{SUB1}+TM_{SUB2}	11.16	4.63	34.26	81.87	56.28

Furthermore, the system developed by adaptation techniques on clustered sub-corpora and the in-domain training corpus produces the highest improvements (0.44 BLEU absolute; corresponding to 4.03% relative) over the baseline. This improvement clearly indicates the effectiveness of our approach to cluster out-of-domain corpus data.

6 Conclusion and Future Work

Lexical coverage is an important issue when an SMT system is built, particularly with languages with small-scale training data. But, this problem will worsen if out-of-domain training data is added in an improper way. Applying domain adaptation techniques to combine

multiple domain corpora, we have successfully increased the lexical coverage in an SMT model with improved system performance. The highest improvement in terms of BLEU (4.03% relative) is achieved when a domain adaption technique by data-source indicator feature is applied on clustered sub-corpora and in-domain training data. Therefore, clustering a large out-of-domain corpus to extract sentences similar to an in-domain corpus is an interesting finding of this research. Domain adaptation techniques using multiple domain data have been successfully employed on a new language pair English—Hindi. To the best of our knowledge, no research has been published to date on domain adaptation in SMT on Indian languages.

The improvement in translation quality is definitely due to the better lexical coverage of phrases. In future, we want to perform a manual evaluation on the output to see how the adapted system's translation differs from baseline translation. For clustering purposes, we defined a simple similarity function which is based on an *n*-gram overlap metric. In future, we want to improve the similarity function by substituting the *n*-gram overlap metric with a weighted *n*-gram overlap metric (setting larger weights for the longer *n*-gram word sequences). In the present work, a bilingual out-of-domain corpus is clustered based on the occurrences of domain-specific *n*-grams in the source language. In future, we want to consider the target language as well as the source language. We also intend to investigate translating Hindi—English to see whether similar improvements are achieved. We also want to automatically tune the threshold value for clustering based on machine translation evaluation metric.

References

Bulyko, I., S. Matsoukas, R. Schwartz, L. Nguyen and J. Makhoul. 2007. Language Model Adaptation in Machine Translation from Speech. In *Proceedings of the 32nd International Conference on Acoustics, Speech, and Signal Processing (ICASSP-2007)*, Honolulu, Hawaii, USA, pp.117–120.

Carter, D. 1994. Improving Language Models by Clustering Training Sentences. In *Proceedings of ACL-1994*, New Mexico State University, Las Cruces, New Mexico, USA. pp.59-64.

Civera, J. and A. Juan. 2007. Domain adaptation in statistical machine translation with mixture modelling. *Proceedings of ACL-2007.* Prague, Czech Republic, pp. 177–180.

Eck M., S. Vogel and A. Waibel. 2004. Language model adaptation for statistical machine translation based on information retrieval. In *Proceedings of the 4th International Conference on language resources and evaluation (LREC-2004).* Lisbon, Portugal, pp.327–330.

Finch, A. and E. Sumita. 2008. Dynamic model interpolation for statistical machine translation. In *Proceedings of ACL-08: HLT. Third Workshop on Statistical Machine Translation*. Columbus, OH, pp.208-215.

Foster, G. and R. Kuhn. 2007. Mixture-model adaptation for SMT. *ACL 2007: Proceedings of the Second Workshop on Statistical Machine Translation*, Prague, Czech Republic, pp.128–135.

Hasan, S. and H. Ney. 2005. Clustered language models based on regular expressions for SMT. In *Proceedings of 10th EAMT conference*, Budapest, pp.119–125.

Hildebrand, A.S., M. Eck, S. Vogel and A. Waibel. 2005. Adaptation of the translation model for statistical machine translation based on information retrieval. In *Proceedings of 10th EAMT conference "Practical applications of machine translation"*, Budapest, pp.133–142.

Iyer, R.M., M. Ostendorf and H. Gish. 1997. Using Out-of-Domain Data to Improve In-Domain Language Models. *IEEE Signal Processing Letters*, pp.221–223.

Iyer, R.M. and M. Ostendorf. 1999. Modeling long distance dependence in language: Topic mixtures versus dynamic cache models. *IEEE Transactions on Speech and Audio Processing*, 7(1):30–39.

Koehn, P., F.J. Och and D. Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT/NAACL 2003*, Edmonton, Canada, pp.48–54.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin and E. Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *ACL-2007: Proceedings of demo and poster sessions*, Prague, Czech Republic, pp.177-180.

Koehn, P. and J. Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Proceedings of ACL-2007*. Prague, Czech Republic, pp.224–227.

Langlais, P. 2002. Improving a general-purpose statistical translation engine by terminological lexicons. In *Proceedings of Coling-2002: Second international workshop on computational terminology (COMPUTERM 2002)*, Taipei, Taiwan, pp.1–7.

Lim, C. and K. Kirchhoff. 2008. Domain Adaptation Through Phrase Generalization for Improved Statistical Machine Translation Quality. *UWEE Technical Report*.

Mahajan, M., D. Beeferman and X.D. Huang. 1999. Improved topic-dependent language modeling using information retrieval techniques. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Phoenix, AZ.

Munteanu, D.S. and D. Marcu. 2006. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics* 31 (4), pp.477–504.

Nakov, P. 2008. Improving English-Spanish statistical machine translation: experiments in domain adaptation, sentence paraphrasing, tokenization, and recasing. In *Proceedings of ACL-08: HLT. Third Workshop on Statistical Machine Translation*, The Ohio State University, Columbus, OH, USA. pp.147–150.

Och, F. J. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of ACL-2003*, Sapporo, Japan, pp.160–167.

Stolcke, A. 2002. SRILM – an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing, vol. 2*. Denver, CO, pp. 901-904.

Seno, E.R.M. and M. das G.V. Nunes. 2008. Some Experiments on Clustering Similar Sentences of Texts in Portuguese. In *Proceedings of the International Conference on Computational Processing of Portuguese* (*PROPOR 2008*). Aveiro, Portugal, pp.133-142.

Snover, M., B. Dorr and R. Schwartz. 2008. Language and translation model adaptation using comparable corpora. In *Proceedings of EMNLP-2008*, Honolulu, Hawaii, USA, pp.857–866.

Ueffing, N., G. Haffari and A. Sarkar. 2007. Transductive learning for statistical machine translation. In *Proceedings of ACL-2007*, Prague, Czech Republic, pp.25–32.

Venkatapathy, S. 2008. NLP Tools Contest – 2008: Summary. *In ICON-2008: Proceedings of the NLP Tools Contest: Statistical Machine Translation (English to Hindi)*, Pune, India.

Wu, H., H. Wang and Z. Liu. 2005. Alignment model adaptation for domain-specific word alignment. In *Proceedings of ACL-05*, Ann Arbor, MI. pp.467–474.

Wu, H. and H. Wang. 2004. Improving domain-specific word alignment with a general bilingual corpus. Machine translation: from real users to research. In *Proceedings of AMTA 2004*, Washington, DC, pp.262–271.

Wu, H., H. Wang and C. Zong. 2008. Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In *Proceedings of the 22nd International Conference on Computational Linguistics* (*Coling 2008*), Manchester, UK, pp.993–1000.

Yamamoto, H. and E. Sumita. 2007. Bilingual cluster based models for statistical machine translation. In *Proceedings of EMNLP-CoNLL'07*, Prague, Czech Republic, 2007; pp.514–523.

Zhao, B., M. Eck and S. Vogel. 2004. Language model adaptation for statistical machine translation with structured query models. In *Proceedings of 20th International Conference on Computational Linguistics* (*Coling 2004*), University of Geneva, Switzerland, pp.1–7.