Using Comparable Corpora to Adapt MT Models to New Domains

Ann Irvine
Center for Language and Speech Processing
Johns Hopkins University

Chris Callison-Burch
Computer and Information Science Dept.
University of Pennsylvania

Abstract

In previous work we showed that when using an SMT model trained on old-domain data to translate text in a new-domain, most errors are due to unseen source words, unseen target translations, and inaccurate translation model scores (Irvine et al., 2013a). In this work, we target errors due to inaccurate translation model scores using new-domain comparable corpora, which we mine from Wikipedia. We assume that we have access to a large old-domain parallel training corpus but only enough new-domain parallel data to tune model parameters and do evaluation. We use the new-domain comparable corpora to estimate additional feature scores over the phrase pairs in our baseline models. Augmenting models with the new features improves the quality of machine translations in the medical and science domains by up to 1.3 BLEU points over very strong baselines trained on the 150 million word Canadian Hansard dataset.

1 Introduction

Domain adaptation for machine translation is known to be a challenging research problem that has substantial real-world application. In this setting, we have access to training data in some old-domain of text but very little or no training data in the domain of the text that we wish to translate. For example, we may have a large corpus of parallel newswire training data but no training data in the medical domain, resulting in low quality translations at test time due to the mismatch.

In Irvine et al. (2013a), we introduced a taxonomy for classifying machine translation errors related to lexical choice. Our 'S4' taxonomy includes seen, sense, score, and search errors. Seen errors result when a source language word or phrase in the test set was not observed at all during training. Sense errors occur when the source language word or phrase was observed during training but not with the correct target language translation. If the source language word or phrase was observed with its correct translation during training, but an incorrect alternative outweighs the correct translation, then a score error has occurred. Search errors are due to pruning in beam search decoding. We measured the impact of each error type in a domain adaptation setting and concluded that seen and sense errors are the most frequent but that there is also room for improving errors due to inaccurate translation model scores (Irvine et al., 2013a). In this work, we target *score* errors, using comparable corpora to reduce their frequency in a domain adaptation setting.

We assume the setting where we have an old-domain parallel training corpus but no new domain training corpus.[1] We do, however, have access to a mixed-domain comparable corpus. We identify new-domain text within our comparable corpus and use that data to estimate new translation features on the translation models extracted from old-domain training data. Specifically, we focus on the French-English language pair because carefully curated datasets exist in several domains for tuning and evaluation. Following our prior work, we use the Canadian Hansard parliamentary proceedings as our old-domain and adapt models to both the medical and the science domains (Irvine et al., 2013a). At over 8 million sentence pairs,

[1] Some prior work has referred to old-domain and new-domain corpora as out-of-domain and in-domain, respectively.

the Canadian Hansard dataset is one of the largest publicly available parallel corpora and provides a very strong baseline. We give details about each dataset in Section 4.1.

We use comparable corpora to estimate several signals of translation equivalence. In particular, we estimate the contextual, topic, and orthographic similarity of each phrase pair in our baseline old-domain translation model. In Section 3, we describe each feature in detail. Using just 5 thousand comparable new-domain document pairs, which we mine from Wikipedia, and five new phrase table features, we observe performance gains of up to 1.3 BLEU points on the science and medical translation tasks over very strong baselines.

2 Related Work

Recent work on machine translation domain adaptation has focused on either the language modeling component or the translation modeling component of an SMT model. Language modeling research has explored methods for subselecting new-domain data from a large monolingual target language corpus for use as language model training data (Lin et al., 1997; Klakow, 2000; Gao et al., 2002; Moore and Lewis, 2010; Mansour et al., 2011). Translation modeling research has typically assumed that either (1) two parallel datasets are available, one in the old domain and one in the new, or (2) a large, mixed-domain parallel training corpus is available. In the first setting, the goal is to effectively make use of both the old-domain and the new-domain parallel training corpora (Civera and Juan, 2007; Koehn and Schroeder, 2007; Foster and Kuhn, 2007; Foster et al., 2010; Haddow and Koehn, 2012; Haddow, 2013). In the second setting, it has been shown that, in some cases, training a translation model on a subset of new-domain parallel training data within a larger training corpus can be more effective than using the complete dataset (Mansour et al., 2011; Axelrod et al., 2011; Sennrich, 2012; Gascó et al., 2012).

For many language pairs and domains, *no* new-domain parallel training data is available. Wu et al. (2008) machine translate new-domain source language monolingual corpora and use the synthetic parallel corpus as additional training data. Daumé and Jagarlamudi (2011), Zhang and Zong (2013), and Irvine et al. (2013b) use new-domain comparable corpora to mine translations for unseen words. That work follows a long line of research on bilingual lexicon induction (e.g. Rapp (1995), Schafer and Yarowsky (2002), Koehn and Knight (2002), Haghighi et al. (2008), Irvine and Callison-Burch (2013), Razmara et al. (2013)). These efforts improve S4 *seen*, and, in some instances, *sense* error types. To our knowledge, no prior work has focused on fixing errors due to inaccurate translation model *scores* in the setting where no new-domain parallel training data is available.

In Klementiev et al. (2012), we used comparable corpora to estimate several features for a given phrase pair that indicate translation equivalence, including contextual, temporal, and topical similarity. The definitions of phrasal and lexical contextual and topic similarity that we use here are taken from our prior work, where we replaced bilingually estimated phrase table features with the new features and cited applications to low resource SMT. In this work we also focus on *scoring* a phrase table using comparable corpora. However, here we work in a domain adaptation setting and seek to augment, not replace, an existing set of bilingually estimated phrase table features.

3 Phrase Table Scoring

We begin with a scored phrase table estimated using our old-domain parallel training corpus. The phrase table contains about 201 million unique source phrases up to length seven and about 479 million total phrase pairs. We use Wikipedia as a source for comparable document pairs (details are given in Section 4.1). We augment the bilingually estimated features with the following: (1) lexical and phrasal contextual similarity estimated over a comparable corpus, (2) lexical and phrasal topical similarity estimated over a comparable corpus, and (3) lexical orthographic similarity.

Contextual Similarity We estimate contextual similarity[2] by first computing a context vector for each source and target word and phrase in our phrase table using the source and target sides of our comparable corpus, respectively. We begin by collecting vectors of counts of words that appear in the context of each source and target phrase, p_s and p_t. We use a bag-of-words context consisting of the two words to the left and two words to

[2] Similar to *distributional similarity*, which is typically defined monolingually.

the right of each occurrence of each phrase. Various means of computing the component values of context vectors from raw context frequency counts have been proposed (e.g. Rapp (1999), Fung and Yee (1998)). Following Fung and Yee (1998), we compute the value of the k-th component of p_s's contextual vector, C_{p_s}, as follows:

$$C_{p_{s_k}} = n_{p_s,k} * (log(n/n_k) + 1)$$

where $n_{p_s,k}$ and n_k are the number of times the k-th source word, s_k, appears in the context of p_s and in the entire corpus, and n is the maximum number of occurrences of any word in the data. Intuitively, the more frequently s_k appears with p_s and the less common it is in the corpus in general, the higher its component value. The context vector for p_s, C_{p_s}, is M-dimensional, where M is the size of the source language vocabulary. Similarly, we compute N-dimensional context vectors for all target language words and phrases, where N is the size of the target language vocabulary.

We identify the most probable translation t for each of the M source language words, s, as the target word with the highest $p(t|s)$ under our word aligned old-domain training corpus. Given this dictionary of unigram translations, we then *project* each M-dimensional source language context vector into the N-dimensional target language context vector space. To compare a given pair of source and target context vectors, C_{p_s} and C_{p_t}, respectively, we compute their cosine similarity, or their dot product divided by the product of their magnitudes:

$$sim_{contextual}(p_s, p_t) = \frac{C_{p_s} \cdot C_{p_t}}{||C_{p_s}||||C_{p_t}||}$$

For a given phrase pair in our phrase table, we estimate *phrasal* contextual similarity by directly comparing the context vectors of the two phrases themselves. Because context vectors for phrases, which tend to be less frequent than words, can be sparse, we also compute lexical contextual similarity over phrase pairs. We define lexical contextual similarity as the average of the contextual similarity between all word pairs within the phrase pair.

Topic Similarity Phrases and their translations are likely to appear in articles written about the same topic in two languages. We estimate topic similarity using the distribution of words and phrases across Wikipedia pages, for which we have interlingual French-English links. Specifically, we compute topical vectors by counting the number of occurrences of each word and phrase across Wikipedia pages. That is, for each source and target phrase, p_s and p_t, we collect M-dimensional topic vectors, where M is the number of Wikipedia page pairs used (in our experiments, M is typically $5,000$). We use Wikipedia's interlingual links to align the French and English topic vectors and normalize each topic vector by the total count. As with contextual similarity, we compare a pair of source and target topic vectors, T_{p_s} and T_{p_t}, respectively, using cosine similarity:

$$sim_{topic}(p_s, p_t) = \frac{T_{p_s} \cdot T_{p_t}}{||T_{p_s}||||T_{p_t}||}$$

We estimate both phrasal and lexical topic similarity for each phrase pair. As before, lexical topic similarity is estimated by taking an average topic similarity across all word pairs in a given phrase pair.

Orthographic Similarity We make use of one additional signal of translation equivalence: orthographic similarity. In this case, we do not reference comparable corpora but simply compute the edit distance between a given pair of phrases. This signal is often useful for identifying translations of technical terms, which appear frequently in our medical and science domain corpora. However, because of word order variation, we do not measure edit distance on phrase pairs directly. For example, French *embryon humain* translates as English *human embryo*; *embryon* translates as *embryo* and *humain* translates as *human*. Although both word pairs are cognates, the words appear in opposite orders in the two phrases. Therefore, directly measuring string edit distance across the phrase pair would not effectively capture the relatedness of the words. Hence, we only measure lexical orthographic similarity, not phrasal. We compute lexical orthographic similarity by first computing the edit distance between each word pair, w_s and w_t, within a given phrase pair, normalized by the lengths of the two words:

$$sim_{orth}(w_s, w_t) = \frac{ed(w_s, w_t)}{\frac{|w_s||w_t|}{2}}$$

We then compute the average normalized edit distance across all word pairs.

The above similarity metrics all allow for scores of zero, which can be problematic for our log-

Corpus	Source Words	Target Words
Training		
Canadian Hansard	161.7 m	144.5 m
Tune-1 / Tune-2 / Test		
Medical	53k / 43k / 35k	46k / 38k / 30k
Science	92k / 120k / 120k	75k / 101k / 101k
Language Modeling and Comparable Corpus Selection		
Medical	-	5.9 m
Science	-	3.6 m

Table 1: Summary of the size of each corpus of text used in this work in terms of the number of source and target word tokens.

linear translation models. We describe our experiments with different minimum score cutoffs in Section 4.2.

4 Experimental Setup

4.1 Data

We assume that the following data is available in our translation setting:

- Large old-domain parallel corpus for training
- Small new-domain parallel corpora for tuning and testing
- Large new-domain English monolingual corpus for language modeling and identifying new-domain-like comparable corpora
- Large mixed-domain comparable corpus, which includes some text from the new-domain

These data conditions are typical for many real-world uses of machine translation. A summary of the size of each corpus is given in Table 1.

Our old-domain training data is taken from the Canadian Hansard parliamentary proceedings dataset, which consists of manual transcriptions and translations of meetings of the Canadian parliament. The dataset is substantially larger than the commonly used Europarl corpus, containing over 8 million sentence pairs and about 150 million word tokens of French and English.

For tuning and evaluation, we use new-domain medical and science parallel datasets released by Irvine et al. (2013a). The medical texts consist of documents from the European Medical Agency (EMEA), originally released by Tiedemann (2009). This data is primarily taken from prescription drug label text. The science data is made up of translated scientific abstracts from the fields of physics, biology, and computer science. For both the medical and science domains, we use three held-out parallel datasets of about 40 and 100 thousand words,[3] respectively, released by Irvine et al. (2013a). We do tuning on *dev1*, additional parameter selection on *test2*, and blind testing on *test1*.

We use large new-domain monolingual English corpora for language modeling and for selecting new-domain-like comparable corpora from our mixed domain comparable corpus. Specifically, we use the English side of the medical and science training datasets released by Irvine et al. (2013a). We do not use the parallel French side of the training data at all; our data setting assumes that no new-domain parallel data is available for training.

We use Wikipedia as a source of comparable corpora. There are over half a million pairs of inter-lingually linked French and English Wikipedia documents.[4] We assume that we have enough monolingual new-domain data in one language to rank Wikipedia pages according to how *new-domain-like* they are. In particular, we use our new-domain English language modeling data to measure new-domain-likeness. We could have targeted our learning even more by using our new-domain French test sets to select comparable corpora. Doing so may increase the similarity between our test data and comparable corpora. However, to avoid overfitting any particular test set, we use our large English new-domain language modeling corpus instead.

For each inter-lingually linked pair of French and English Wikipedia documents, we compute the percent of English phrases up to length four that are observed in the English monolingual new-domain corpus and rank document pairs by the geometric mean of the four overlap measures. More sophisticated ways to identify new-domain-like Wikipedia pages (e.g. (Moore and Lewis, 2010)) may yield additional performance gains, but, qualitatively, the ranked Wikipedia pages seem reasonable for the purposes of generating a large set of top-k new-domain document pairs. The top-10 ranked pages for each domain are listed in Table 2. The top ranked science domain pages are primarily related to concepts from the field of physics but also include computer science and chemistry

[3] Or about 4 thousand lines each. The sentences in the medical domain text are much shorter than those in the science domain.

[4] As of January 2014.

Science	Medical
Diagnosis (artificial intelligence)	Pregabalin
Absorption spectroscopy	Cetuximab
Spectral line	Fluconazole
Chemical kinetics	Calcitonin
Mahalanobis distance	Pregnancy category
Dynamic light scattering	Trazodone
Amorphous solid	Rivaroxaban
Magnetic hyperthermia	Spironolactone
Photoelasticity	Anakinra
Galaxy rotation curve	Cladribine

Table 2: Top 10 Wikipedia articles ranked by their similarity to large new-domain English monolingual corpora.

topics. The top ranked medical domain pages are nearly all prescription drugs, which makes sense given the content of the EMEA medical corpus.

4.2 Phrase-based Machine Translation

We word align our old-domain training corpus using GIZA++ and use the Moses SMT toolkit (Koehn et al., 2007) to extract a translation grammar. In this work, we focus on phrase-based SMT models, however our approach to using new-domain comparable corpora to estimate translation scores is theoretically applicable to any type of translation grammar.

Our baseline models use a phrase limit of seven and the standard phrase-based SMT feature set, including forward and backward phrase and lexical translation probabilities. Additionally, we use the standard lexicalized reordering model. We experiment with two 5-gram language models trained using SRILM with Kneser-Ney smoothing on (1) the English side of the Hansard training corpus, and (2) the relevant new-domain monolingual English corpus. We experiment with using, first, only the old-domain language model and, then, both the old-domain and the new-domain language models.

Our first comparison system augments the standard feature set with the orthographic similarity feature, which is not based on comparable corpora. Our second comparison system uses both the orthographic feature and the contextual and topic similarity features estimated over a *random* set of comparable document pairs. The third system estimates contextual and topic similarity using new-domain-like comparable corpora. We tune our phrase table feature weights for each model separately using batch MIRA (Cherry and Foster, 2012) and new-domain tuning data. Results are averaged over three tuning runs, and we use the implementation of approximate randomization released by Clark et al. (2011) to measure the statistical significance of each feature-augmented model compared with the baseline model that uses the same language model(s).

As noted in Section 3, the features that we estimate from comparable corpora may be zero-valued. We use our second tuning sets[5] to tune a minimum threshold parameter for our new features. We measure performance in terms of BLEU score on the second tuning set as we vary the new feature threshold between $1e-07$ and 0.5 for each domain. A threshold of 0.01, for example, means that we replace all feature with values less than 0.01 with 0.01. For both new-domains, performance drops when we use thresholds lower than 0.01 and higher than 0.25. We use a minimum threshold of 0.1 for all experiments presented below for both domains.

5 Results

Table 3 presents a summary of our results on the test set in each domain. Using only the old-domain language model, our baselines yield BLEU scores of 22.70 and 21.29 on the medical and science test sets, respectively. When we add the orthographic similarity feature, BLEU scores increase significantly, by about 0.4 on the medical data and 0.6 on science. Adding the contextual and topic features estimated over a random selection of comparable document pairs improves BLEU scores slightly in both domains. Finally, using the most new-domain like document pairs to estimate the contextual and topic features yields a 1.3 BLEU score improvement over the baseline in both domains. For both domains, this result is a statistically significant improvement[6] over each of the first three systems.

In both domains, the new-domain language models contribute substantially to translation quality. Baseline BLEU scores increase by about 6 and 5 BLEU score points in the medical and science domains, respectively, when we add the new-domain language models. In the medical domain, neither the orthographic feature nor the orthographic feature in combination with contextual and topic features estimated over random document pairs results in a significant BLEU score improvement. However, using the orthographic feature and the contextual and topic features estimated over new-domain document pairs yields a

[5] *test2* datasets released by Irvine et al. (2013a)

[6] p-value < 0.01

Language Model(s)	System	Medical	Science
Old	Baseline	22.70	21.29
	+ Orthographic Feature	23.09* (+0.4)	21.86* (+0.6)
	+ Orthographic & Random CC Features	23.22* (+0.5)	21.88* (+0.6)
	+ Orthographic & New-domain CC Features	23.98* (+1.3)	22.55* (+1.3)
Old+New	Baseline	28.82	26.18
	+ Orthographic Feature	29.02 (+0.2)	26.40* (+0.2)
	+ Orthographic & Random CC Features	28.86 (+0.0)	26.52* (+0.3)
	+ Orthographic & New-domain CC Features	29.16* (+0.3)	26.50* (+0.3)

Table 3: Comparison between the performance of baseline old-domain translation models and domain-adapted models in translating science and medical domain text. We experiment with two language models: *old*, trained on the English side of our Hansard old-domain training corpus and *new*, trained on the English side of the parallel training data in each new domain. We use comparable corpora of $5,000$ (1) random, and (2) the most new-domain-like document pairs to score phrase tables. All results are averaged over three tuning runs, and we perform statistical significance testing comparing each system augmented with additional features with the baseline system that uses the same language model(s). * indicates that the BLEU scores are statistically significant with $p < 0.01$.

small but significant improvement of 0.3 BLEU. In the science domain, in contrast, all three augmented models perform statistically significantly better than the baseline. Contextual and topic features yield only a slight improvement above the model that uses only the orthographic feature, but the difference is statistically significant. For the science domain, when we use the new domain language model, there is no difference between estimating the contextual and topic features over random comparable document pairs and those chosen for their similarity with new-domain data.

Differences across domains may be due to the fact that the medical domain corpora are much more homogenous, containing the often boilerplate text of prescription drug labels, than the science domain corpora. The science domain corpora, in contrast, contain abstracts from several different scientific fields; because that data is more diverse, a randomly chosen mixed-domain set of comparable corpora may still be relevant and useful for adapting a translation model.

We experimented with varying the number of comparable document pairs used for estimating contextual and topic similarity but saw no significant gains from using more than $5,000$ in either domain. In fact, performance dropped in the medical domain when we used more than a few thousand document pairs. Our proposed approach orders comparable document pairs by how new-domain-like they are and augments models with new features estimated over the top-k. As a result, using more comparable document pairs means that there is more data from which to estimate signals, but it also means that the data is less new-domain like overall. Using a domain similarity threshold to choose a subset of comparable document pairs may prove useful in future work, as the ideal amount of comparable data will depend on the type and size of the initial mixed-domain comparable corpus as well as the homogeneity of the text domain of interest.

We also experimented with using a third language model estimated over the English side of our comparable corpora. However, we did not see any significant improvements in translation quality when we used this language model in combination with the old and new domain language models.

6 Conclusion

In this work, we targeted SMT errors due to translation model *scores* using new-domain comparable corpora. Our old-domain French-English baseline model was trained on the Canadian Hansard parliamentary proceedings dataset, which, at 8 million sentence pairs, is one of the largest publicly available parallel datasets. Our task was to adapt this baseline to the medical and scientific text domains using comparable corpora. We used new-domain parallel data only to tune model parameters and do evaluation. We mined Wikipedia for new-domain-like comparable document pairs, over which we estimated several additional features scores: contextual, temporal, and orthographic similarity. Augmenting the strong baseline with our new feature set improved the quality of machine translations in the medical and science domains by up to 1.3 BLEU points.

7 Acknowledgements

This material is based on research sponsored by DARPA under contract HR0011-09-1-0044 and by the Johns Hopkins University Human Language Technology Center of Excellence. The views and conclusions contained in this publication are those of the authors and should not be interpreted as representing official policies or endorsements of DARPA or the U.S. Government.

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Jorge Civera and Alfons Juan. 2007. Domain adaptation in statistical machine translation with mixture modelling. In *Proceedings of the Workshop on Statistical Machine Translation (WMT)*.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: controlling for optimizer instability. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Hal Daumé, III and Jagadeesh Jagarlamudi. 2011. Domain adaptation for machine translation by mining unseen words. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

George Foster and Roland Kuhn. 2007. Mixture-model adaptation for SMT. In *Proceedings of the Workshop on Statistical Machine Translation (WMT)*.

G. Foster, C. Goutte, and R. Kuhn. 2010. Discriminative instance weighting for domain adaptation in SMT. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Pascale Fung and Lo Yuen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Jianfeng Gao, Joshua Goodman, Mingjing Li, and Kai-Fu Lee. 2002. Toward a unified approach to statistical language modeling for chinese. *ACM Transactions on Asian Language Information Processing (TALIP)*.

Guillem Gascó, Martha-Alicia Rocha, Germán Sanchis-Trilles, Jesús Andrés-Ferrer, and Francisco Casacuberta. 2012. Does more data always yield better translations? In *Proceedings of the Conference of the European Association for Computational Linguistics (EACL)*.

Barry Haddow and Philipp Koehn. 2012. Analysing the effect of out-of-domain data on SMT systems. In *Proceedings of the Workshop on Statistical Machine Translation (WMT)*.

Barry Haddow. 2013. Applying pairwise ranked optimisation to improve the interpolation of translation models. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Ann Irvine and Chris Callison-Burch. 2013. Supervised bilingual lexicon induction with multiple monolingual signals. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Ann Irvine, John Morgan, Marine Carpuat, Hal Daumé III, and Dragos Munteanu. 2013a. Measuring machine translation errors in new domains. *Transactions of the Association for Computational Linguistics*, 1(October).

Ann Irvine, Chris Quirk, and Hal Daume III. 2013b. Monolingual marginal matching for translation model adaptation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Dietrich Klakow. 2000. Selecting articles from the language model training corpus. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Alex Klementiev, Ann Irvine, Chris Callison-Burch, and David Yarowsky. 2012. Toward statistical machine translation without parallel corpora. In *Proceedings of the Conference of the European Association for Computational Linguistics (EACL)*.

Philipp Koehn and Kevin Knight. 2002. Learning a translation lexicon from monolingual corpora. In *ACL Workshop on Unsupervised Lexical Acquisition*.

Philipp Koehn and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Workshop on Statistical Machine Translation (WMT)*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Sung-Chien Lin, Chi-Lung Tsai, Lee-Feng Chien, Ker-Jiann Chen, and Lin-Shan Lee. 1997. Chinese language model adaptation based on document classification and multiple domain-specific language models. In *Fifth European Conference on Speech Communication and Technology*.

Saab Mansour, Joern Wuebker, and Hermann Ney. 2011. Combining translation and language model scoring for domain-specific data filtering. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Majid Razmara, Maryam Siahbani, Reza Haffari, and Anoop Sarkar. 2013. Graph propagation for paraphrasing out-of-vocabulary words in statistical machine translation. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Charles Schafer and David Yarowsky. 2002. Inducing translation lexicons via diverse similarity measures and bridge languages. In *Proceedings of the Conference on Natural Language Learning (CoNLL)*.

Rico Sennrich. 2012. Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of the Conference of the European Association for Computational Linguistics (EACL)*.

Jörg Tiedemann. 2009. News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, editors, *Recent Advances in Natural Language Processing (RANLP)*.

Hua Wu, Haifeng Wang, and Chengqing Zong. 2008. Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Jiajun Zhang and Chengqing Zong. 2013. Learning a phrase-based translation model from monolingual data with application to domain adaptation. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.