Statistical Machine Translation in Low Resource Settings

Ann Irvine
Center for Language and Speech Processing
Johns Hopkins University

Abstract

My thesis will explore ways to improve the performance of statistical machine translation (SMT) in low resource conditions. Specifically, it aims to reduce the dependence of modern SMT systems on expensive parallel data. We define low resource settings as having only small amounts of parallel data available, which is the case for many language pairs. All current SMT models use parallel data during training for extracting translation rules and estimating translation probabilities. The theme of our approach is the *integration of information from alternate data sources, other than parallel corpora*, into the statistical model. In particular, we focus on making use of large *monolingual* and *comparable* corpora. By augmenting components of the SMT framework, we hope to extend its applicability beyond the small handful of language pairs with large amounts of available parallel text.

1 Introduction

Statistical machine translation (SMT) systems are heavily dependent on parallel data. SMT doesn't work well when fewer than several million lines of bitext are available (Kolachina et al., 2012). When the available bitext is small, statistical models perform poorly due to the sparse word and phrase counts that define their parameters. Figure 1 gives a learning curve that shows this effect. As the amount of bitext approaches zero, performance drops drastically. In this thesis, we seek to modify the SMT model to reduce its dependence on parallel data and, thus, enable it to apply to new language pairs.

Specifically, we plan to address the following challenges that arise when using SMT systems in low resource conditions:

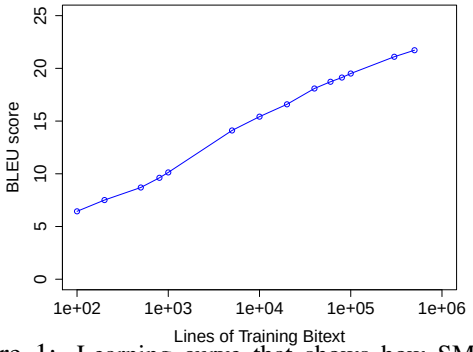

Figure 1: Learning curve that shows how SMT performance on the Spanish to English translation task increases with increasing amounts of parallel data. Performance is measured with BLEU and drops drastically as the amount of bitext approaches zero. These results use the Europarl corpus and the Moses phrase-based SMT framework, but the trend shown is typical.

- **Translating unknown words.** In the context of SMT, unknown words (or out-of-vocabulary, OOV) are defined as having never appeared in the source side of the training parallel corpus. When the training corpus is small, the percent of words which are unknown can be high.
- **Inducing phrase translations.** In high resource conditions, a word aligned bitext is used to extract a list of phrase pairs or translation rules which are used to translate new sentences. With more parallel data, this list is increasingly comprehensive. Using multi-word phrases instead of individual words as the basic translation unit has been shown to increase translation performance (Koehn et al., 2003). However, when the parallel corpus is small, so is the number of phrase pairs that can be extracted.
- **Estimating translation probabilities.** In the standard SMT pipeline, translation probabilities are estimated using relative frequency counts over the training bitext. However, when the bitext counts are sparse, probability esti-

Proceedings of the NAACL HLT 2013 Student Research Workshop, pages 54–61,
Atlanta, Georgia, 13 June 2013.

Language	#Words	Language	#Words
Nepali	0.4	Somali	0.5
Uzbek	1.4	Azeri	2.6
Tamil	3.7	Albanian	6.5
Bengali	6.6	Welsh	7.5
Bosnian	12.9	Latvian	40.2
Indonesian	21.8	Romanian	24.1
Serbian	25.8	Turkish	31.2
Ukrainian	37.6	Hindi	47.4
Bulgarian	49.5	Polish	104.5
Slovak	124.3	Urdu	287.2
Farsi	710.3	Spanish	972

Table 1: Millions of monolingual web crawl and Wikipedia word tokens

mates are likely to be noisy.

My thesis focuses on translating into English. We assume access to a small amount of parallel data, which is realistic, especially considering the recent success of crowdsourcing translations (Zaidan and Callison-Burch, 2011; Ambati, 2011; Post et al., 2012). Additionally, we assume access to larger monolingual corpora. Table 1 lists the 22 languages for which we plan to perform translation experiments, along with the total amount of monolingual data that we will use for each. We use web crawled time-stamped news articles and Wikipedia for each language. We have extracted the Wikipedia pages which are inter-lingually linked to English pages.

2 Translating Unknown Words

OOV words are a major challenge in low resource SMT settings. Here, we describe several approaches to identifying translations for unknown words.

2.1 Transliteration

For non-roman script languages, in some cases, OOV words may be *transliterated* rather than *translated*. This is often true for named entities, where transliterated words are pronounced approximately the same across languages but have different spellings in the source and target language alphabets (e.g. Russian Анна translates as English *Anna*). In the case of roman script languages, of course, such words are often translated correctly without change (e.g. French *Anna* translates as English *Anna*).

In my prior work, Irvine et al. (2010a) and Irvine et al. (2010b), I have presented a language-independent approach to gathering pairs of transliterated words (specifically, names) in a pair of languages, built a module to transliterate from one language to the other, and integrated the output into an end-to-end SMT system. In my thesis, I will use this technique to hypothesize translations for OOV words. Additionally, I plan to include techniques that build upon the one described in Hermjakob et al. (2008) in order to predict when words are likely to be transliterated rather than translated. That work uses features based on an Arabic named entity tagger. In our low resource setting, we cannot assume access to such off-the-shelf tools and must adapt this existing technique accordingly.

2.2 Bilingual Lexicon Induction

Bilingual lexicon induction is the task of identifying word translation pairs in source and target language monolingual or comparable corpora. The task is well-researched, however, in prior work, Irvine and Callison-Burch (2013), we were the first to propose using *supervised* methods. Because we assume access to some small amount of parallel data, we can extract a bilingual dictionary from it to use for positive supervision. In my prior work and in the thesis, we use the following signals estimated over comparable source and target language corpora: orthographic, topic, temporal, and contextual similarity. Here, we give brief descriptions of each.

Orthographic We measure orthographic similarity between a pair of words as the normalized[1] edit distance between the two words. For non-Roman script languages, we transliterate words into the Roman script before measuring orthographic similarity.

Topic We use monolingual Wikipedia pages to estimate topical signatures for each source and target language word. Signatures contain counts of how many times a given word appears on each interlingually linked Wikipedia page, and we use cosine similarity to compare pairs of signatures.

Temporal We use time-stamped web crawl data to estimate temporal signatures, which, for a given word, contain counts of how many times that word appeared in news articles with a certain date. We expect that source and target language words which are translations of one another will appear with similar frequencies over time in monolingual data.

[1]Normalized by the average of the lengths of the two words

Contextual We score monolingual contextual similarity by first collecting context vectors for each source and target language word. The context vector for a given word contain counts of how many times words appear in its context. We use bag of words contexts in a window of size two. We gather both source and target language contextual vectors from our web crawl data and Wikipedia data (separately).

Frequency Words that are translations of one another are likely to have similar relative frequencies in monolingual corpora. We measure the frequency similarity of two words as the absolute value of the difference between the log of their relative monolingual corpus frequencies.

We propose using a supervised approach to learning how to combine the above signals into a single discriminative binary classifier which predicts whether a source and target language word are translations of one another or not. Given a classification score for each source language word paired with all English candidates, we rerank candidates and evaluate on the top-k. We give some preliminary experimental details and results here.

We have access to bilingual dictionaries for the 22 languages listed in Table 1[2]. For each language, we choose up to $8,000$ source language words among those that occur in the monolingual data at least three times and that have at least one translation in our dictionary. We randomly divide the source language words into three equally sized sets for training, development, and testing. We use the training data to train a classifier, the development data to choose the best classification settings and feature set, and the test set for evaluation.

For all experiments, we use a linear classifier trained by stochastic gradient descent to minimize squared error[3] and perform 100 passes over the training data.[4] The binary classifiers predict whether a pair of words are translations of one another or not. The translations in our training data serve as positive supervision, and the source language words in the training data paired with random English words[5] serve as negative supervision. We used our development data to tune the number of negative examples to three for each positive example. At test time, after scoring all source language words in the test set paired with all English words in our candidate set,[6] we rank the English candidates by their classification scores and evaluate accuracy in the top-k.

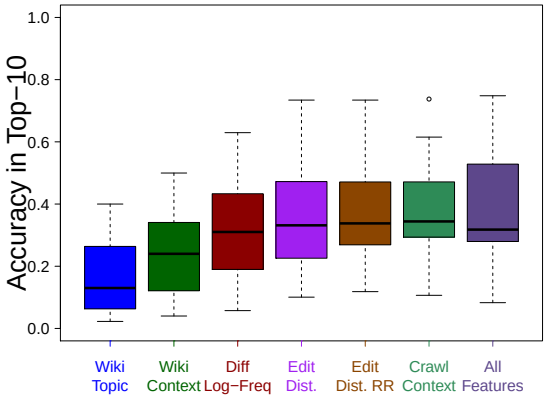

Figure 2: Performance goes up as features are greedily added to the feature space. Mean performance is slightly higher using this subset of six features (second to last bar) than using all features (last bar). Each plot represents results over our 22 languages.

We use raw similarity scores based on the signals enumerated above as features. Additionally, for each source word, we rank all English candidates with respect to each signal and include their reciprocal ranks as another set of features. Finally, we include a binary feature that indicates if a given source and target word are identical strings or not.

We train classifiers separately for each of the 22 languages listed in Table 1, and the learned weights vary based on, for example, corpora size and the relatedness of the source language and English (e.g. edit distance is informative if there are many cognates). When we use the trained classifier to predict which English words are translations of a given source word, all English words appearing at least five times in our monolingual data are candidates, and we rank them by their classification scores.

Figure 2, from left to right, shows a greedy search

[2]Details about the dictionaries in work under review.

[3]We tried using logistic rather than linear regression, but performance differences on our development set were very small and not statistically significant.

[4]We use `http://hunch.net/~vw/` version 6.1.4, and run it with the following arguments that affect how updates are made in learning: –exact adaptive norm –power t 0.5

[5]Among those that appear at least five times in our monolingual data, consistent with our candidate set.

[6]All English words appearing at least five times in our monolingual data. In practice, we further limit the set to those that occur in the top-1000 ranked list according to at least one of our signals.

Lang	MRR	Supv.	Lang	MRR	Supv.
Nepali	11.2	13.6	Somali	16.7	18.1
Uzbek	23.2	29.6	Azeri	16.1	29.4
Tamil	28.4	33.3	Albanian	32.0	45.3
Bengali	19.3	32.8	Welsh	36.1	56.4
Bosnian	32.6	52.8	Latvian	29.6	47.7
Indonesian	41.5	63.5	Romanian	53.3	71.6
Serbian	29.0	33.3	Turkish	31.4	52.1
Ukrainian	29.7	46.0	Hindi	18.2	34.6
Bulgarian	40.2	57.9	Polish	47.4	67.1
Slovak	34.6	53.5	Urdu	13.2	21.2
Farsi	10.5	21.1	Spanish	74.8	85.0

Table 2: Top-10 Accuracy on test set. Performance increases for all languages moving from the baseline (*MRR*) to discriminative training (*Supv*).

for the best subset of features. The Wikipedia topic score is the most informative stand-alone feature, and Wikipedia context is the most informative second feature. Adding features to the model beyond the six shown in the figure does not yield additional performance gains over our set of languages.

We use a model based on the six features shown in Figure 2 to score and rank English translation candidates for the test set words in each language.

Our unsupervised baseline method is based on ranked lists derived from each of the signals listed above. For each source word, we generate ranked lists of English candidates using the following six signals: Crawls Context, Crawls Time, Wikipedia Context, Wikipedia Topic, Edit distance, and Log Frequency Difference. Then, for each English candidate we compute its mean reciprocal rank[7] (MRR) based on the six ranked lists. The baseline ranks English candidates according to the MRR scores. For evaluation, we use the same test sets, accuracy metric, and correct translations.

Table 2 gives results for the baseline and our supervised technique. Across languages, the average top-10 accuracy using the baseline is 30.4, and using our technique it is 43.9, about *44% higher*.

In Section 3 we use the same features to score all *phrase pairs* in a phrase-based MT model and include them as features in tuning and decoding.

[7]The MRR of the jth English word, e_j, is $\frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank_{ij}}$, where N is the number of signals and $rank_{ij}$ is e_j's rank according to signal i.

2.3 Distributed Representations

Our third method for inducing OOV translations employs a similar intuition to that of contextual similarity. However, unlike standard contextual vectors that represent words as large vectors of counts of nearby words, we propose to use *distributed representations*. These word representations are low-dimensional and are induced iteratively using the distributed representations of nearby words, not the nearby words themselves. Using distributed representations helps to alleviate data sparsity problems.

Recently, Klementiev et al. (2012b) induced distributed representations for the crosslingual setting. There, the induced embedding is learned jointly over multiple languages so that the representations of semantically similar words end up "close" to one another irrespective of language. They simultaneously use large monolingual corpora to induce representations for words in each language and use parallel data to bring the representations together across languages. The intuition for their approach to crosslingual representation induction comes from the multitask learning setup of Cavallanti et al. (2010). They apply this set-up to a variant of a neural probabilistic language model (Bengio et al., 2003).

In my thesis, I propose to use the distributed representations proposed by Klementiev et al. (2012b) in order to induce translations for OOV words. Additionally, I plan to learn how to compose the representations of individual words in a phrase into a single representation, allowing for the induction of *phrase* translations in addition to single words.

3 Inducing and Scoring a Phrase Table

Although by extracting OOV *word* translations we may increase the coverage of our SMT model, inducing *phrase* translations may increase performance further. In order to do so, we need to be able to score pairs of phrases to determine which have high translation probabilities. Furthermore, using alternate sources of data to score phrase pairs directly extracted from a small bitext may help distinguish good translation pairs from bad ones, which could result from incorrect word alignments, for example. In moving from words to phrases, we make use of many of the same techniques described in Section 2. Here, I present several proposals for addressing the

major additional challenges that arise for phrases, and Section 4 presents some experimental results.

3.1 Phrase translation induction

The difficulty in inducing a comprehensive set of phrase translations is that the number of phrases, on both the source and target side, is very large. In moving from the induction of word translations to phrase translations, the number of comparisons necessary to do an exhaustive search becomes infeasible. I propose to explore several ways to speed up that search in my thesis:

- Use distributed phrase representations.
- Use filters to limit the phrase pair search space. Filters should be fast and could include information such as word translations, phrase lengths, and monolingual frequencies.
- Predict when phrases should be translated as a unit, rather than compositionally. If it is possible to accurately translate a phrase compositionally from its word translations, then there is no need to induce a translation for the phrase.

3.2 Phrase translation scoring

In our prior work, Klementiev et al. (2012a), we have started to explore scoring a phrase table using comparable corpora. Given a set of phrase pairs, either induced or extracted from a small bitext, the idea is to score them using the same signals derived from comparable corpora described in the context of bilingual lexicon induction in Section 2.2. No matter the source of the phrase pairs, the hope is that such scores will help an SMT model distinguish between good and bad translations. We estimate both *phrasal* and *lexical* similarity features over phrase pairs. We estimate the first using contextual, temporal, and topical signatures over entire phrases. We estimate the latter by using the *lexical* contextual, temporal, topical, and orthographic signatures of *each word in each phrase*. We use phrasal word alignments in order to compute the *lexical* similarity between phrases. That is, we compute each similarity metric for each pair of aligned words and then, for each similarity metric, average over the word pairs. This approach is analogous to the lexical weighting feature introduced by Koehn et al. (2003).

Language	Train Words	Dev OOV Word Types	Dev OOV Word Tokens
Tamil	452k	44%	25%
Bengali	272k	37%	18%
Hindi	708k	34%	11%

Table 3: Information about datasets released by Post et al. (2012). Training data gives the number of words in the source language training set. OOV rates give the percent of development set word types and work tokens that do not appear in the training data.

4 Preliminary Results

Here we show preliminary results using our methods for translating OOV words and our methods for scoring a phrase table in end-to-end low resource machine translation. Post et al. (2012) used Amazon's Mechanical Turk to collect a small parallel corpus for several Indian languages. In our experiments, we use their Tamil, Bengali, and Hindi datasets. We use the data splits given by Post et al. (2012) and, following that work, report results on the devtest set. Table 3 shows statistics about the datasets.

In our experiments, we use the Moses phrase-based machine translation framework (Koehn et al., 2007). For each language, we extract a phrase table from the training data with a phrase limit of seven and, like Post et al. (2012), use the English side of the training data to train a language model. Throughout our experiments, we use MIRA (Chiang et al., 2009) for tuning the feature set.

Our experiments compare the following:

- A baseline phrase-based model, using phrase pairs extracted from the training data and the standard phrasal and lexical translation probabilities based on the bitext.
- Baseline supplemented with word translations induced by our baseline unsupervised bilingual lexicon induction method (Section 2.2)
- Baseline supplemented with word translations induced by our supervised bilingual lexicon induction methods (Section 2.2).
- Baseline model supplemented with additional features, estimated over comparable corpora (Section 3.2).
- Baseline model supplemented with induced word translations and also additional features.

Table 4 shows our results. Adding additional phrase table features increased BLEU scores from

Experiment	K	Tamil BLEU	Tamil Diff.	Bengali BLEU	Bengali Diff.	Hindi BLEU	Hindi Diff.
Baseline		9.16		12.14		14.85	
+ Mono. Features		9.70	+0.54	12.54	+0.40	15.16	+0.31
+ Unsupervised Word Translations	1	9.33	+0.17	12.11	-0.03	15.37	+0.52
+ Supervised Word Translations	1	9.76	+0.60	12.38	+0.24	15.64	+0.79
+ Mono. Feats. & Sup. Trans.	1	10.20	+1.04	**13.01**	**+0.87**	15.84	+0.99
+ Mono. Feats. & Sup. Trans.	5	**10.41**	**+1.25**	12.64	+0.50	**16.02**	**+1.17**
+ Mono. Feats. & Sup. Trans.	10	10.12	+0.96	12.57	+0.43	15.86	+1.01

Table 4: BLEU performance gains that target coverage and accuracy separately and together. We add the top-K ranked translations for each OOV source word.

0.31 BLEU points for Hindi to 0.54 for Tamil.

Next, we monolingually induced translations for all development and test set source words. We experimented with adding translations for source words with low training data frequencies in addition to OOV words but did not observe BLEU improvements beyond what was gained by translating OOVs alone. Our BLEU score gains that result from improving OOV coverage, *+Supervised Word Translations*, range from 0.24 for Bengali to 0.79 for Hindi and outperform the unsupervised lexicon induction baseline for all three languages.

Using comparable corpora to supplement both the feature space and the coverage of OOVs results in translations that are better than applying either technique alone. For all languages, the BLEU improvements are approximately additive. For Tamil, the total BLEU point gain is 1.25, and it is 1.17 for Hindi and 0.87 for Bengali. Table 4 shows results as we add the top-k ranked translation for each OOV word and vary k. For Tamil and Hindi, we get a slight boost by adding the top-5 translations instead of the single best but get no further gains with the top-10.

5 Previous Work

Prior work on bilingual lexicon induction has shown that a variety of signals derived from monolingual data, including distributional, temporal, topic, and string similarity, are informative (Rapp, 1995; Fung and Yee, 1998; Koehn and Knight, 2002; Schafer and Yarowsky, 2002; Monz and Dorr, 2005; Huang et al., 2005; Schafer, 2006; Klementiev and Roth, 2006; Haghighi et al., 2008; Mimno et al., 2009; Mausam et al., 2010; Daumé and Jagarlamudi, 2011). This thesis builds upon this work and uses a diverse set of signals for translating full sentences, not just words. Recently, Ravi and Knight (2011), Dou and Knight (2012), and Nuhn et al. (2012) have worked toward learning a phrase-based translation model from monolingual corpora, relying on *decipherment* techniques. In contrast to that research thread, we make the realistic assumption that a small parallel corpus is available for our low resource languages. With a small parallel corpus, we are able to take advantage of supervised techniques, changing the problem setting dramatically.

Since the early 2000s, the AVENUE (Carbonell et al., 2002; Probst et al., 2002; Lavie et al., 2003) project has researched ways to rapidly develop MT systems for low-resource languages. In contrast to that work, my thesis will focus on a language-independent approach as well as integrating techniques into current state-of-the-art SMT frameworks. In her thesis, Gangadharaiah (2011) tackles several data sparsity issues within the example-based machine translation (EBMT) framework. Her work attempts to tackle some of the same data sparsity issues that we do including, in particular, phrase table coverage. However, our models for doing so are quite different and focus much more on the use of a variety of new non-parallel data resources.

Other approaches to low resource machine translation include extracting parallel sentences from comparable corpora (e.g. Smith et al. (2010)) and translation crowdsourcing. Our efforts are orthogonal and complementary to these.

6 Conclusion

My thesis will explore using alternative data sources, other than parallel text, to inform statistical machine translation models. In particular, I will build upon a long thread of research on bilingual lexicon induction from comparable corpora. The result of my thesis will be broadening the applicability of current SMT frameworks to language pairs and domains for which parallel data is limited.

7 Acknowledgements

The research presented in this paper was done in collaboration with my advisor, Chris Callison-Burch. 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

Vamshi Ambati. 2011. *Active Learning for Machine Translation in Scarce Data Scenarios*. Ph.D. thesis, Carnegie Mellon University.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research (JMLR)*, 3:1137–1155.

Jaime G. Carbonell, Katharina Probst, Erik Peterson, Christian Monson, Alon Lavie, Ralf D. Brown, and Lori S. Levin. 2002. Automatic rule learning for resource-limited mt. In *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA)*.

Giovanni Cavallanti, Nicoló Cesa-bianchi, and Claudio Gentile. 2010. Linear algorithms for online multitask classification. *Journal of Machine Learning Research (JMLR)*, 11:2901–2934.

David Chiang, Kevin Knight, and Wei Wang. 2009. 11,001 new features for statistical machine translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

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)*.

Qing Dou and Kevin Knight. 2012. Large scale decipherment for out-of-domain machine translation. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*.

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)*.

Rashmi Gangadharaiah. 2011. *Coping with Data-sparsity in Example-based Machine Translation*. Ph.D. thesis, Carnegie Mellon University.

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)*.

Ulf Hermjakob, Kevin Knight, and Hal DaumŐ Iii. 2008. Name translation in statistical machine translation learning when to transliterate. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Fei Huang, Ying Zhang, and Stephan Vogel. 2005. Mining key phrase translations from web corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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, Chris Callison-Burch, and Alexandre Klementiev. 2010a. Transliterating from all languages. In *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA)*.

Ann Irvine, Mike Kayser, Zhifei Li, Wren Thornton, and Chris Callison-Burch. 2010b. Integrating output from specialized modules in machine translation: transliterations in joshua. *Prague Bulletin of Mathematical Linguistics*, pages 107–116.

Alexandre Klementiev and Dan Roth. 2006. Weakly supervised named entity transliteration and discovery from multilingual comparable corpora. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

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

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012b. Inducing crosslingual distributed representations of words. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

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

Philipp Koehn, Franz Joseph Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

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)*.

Prasanth Kolachina, Nicola Cancedda, Marc Dymetman, and Sriram Venkatapathy. 2012. Prediction of learning curves in machine translation. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Alon Lavie, Stephan Vogel, Lori Levin, Erik Peterson, Katharina Probst, Ariadna Font, Rachel Reynolds, Jaime Carbonelle, and Richard Cohen. 2003. Experiments with a Hindi-to-English transfer-based MT system under a miserly data scenario. *ACM Transactions on Asian Language Information Processing (TALIP)*, 2.

Mausam, Stephen Soderland, Oren Etzioni, Daniel S. Weld, Kobi Reiter, Michael Skinner, Marcus Sammer, and Jeff Bilmes. 2010. Panlingual lexical translation via probabilistic inference. *Artificial Intelligence*, 174:619–637, June.

David Mimno, Hanna Wallach, Jason Naradowsky, David Smith, and Andrew McCallum. 2009. Polylingual topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Christof Monz and Bonnie J. Dorr. 2005. Iterative translation disambiguation for cross-language information retrieval. In *Proceedings of the Conference on Research and Developments in Information Retrieval (SIGIR)*.

Malte Nuhn, Arne Mauser, and Hermann Ney. 2012. Deciphering foreign language by combining language models and context vectors. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Matt Post, Chris Callison-Burch, and Miles Osborne. 2012. Constructing parallel corpora for six indian languages via crowdsourcing. In *Proceedings of the Workshop on Statistical Machine Translation (WMT)*.

Katharina Probst, Lori Levin, Erik Peterson, Alon Lavie, and Jaime Carbonell. 2002. MT for minority languages using elicitation-based learning of syntactic transfer rules. *Machine Translation*, 17:245–270, December.

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)*.

Sujith Ravi and Kevin Knight. 2011. Deciphering foreign language. 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)*.

Charles Schafer. 2006. *Translation Discovery Using Diverse Similarity Measures*. Ph.D. thesis, Johns Hopkins University.

Jason R. Smith, Chris Quirk, and Kristina Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Omar F. Zaidan and Chris Callison-Burch. 2011. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.