Combining Bilingual and Comparable Corpora for Low Resource Machine Translation

Ann Irvine
Center for Language and Speech Processing
Johns Hopkins University

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

Abstract

Statistical machine translation (SMT) performance suffers when models are trained on only small amounts of parallel data. The learned models typically have both low *accuracy* (incorrect translations and feature scores) and low *coverage* (high out-of-vocabulary rates). In this work, we use an additional data resource, *comparable corpora*, to improve both. Beginning with a small bitext and corresponding phrase-based SMT model, we improve coverage by using bilingual lexicon induction techniques to learn new translations from comparable corpora. Then, we supplement the model's feature space with translation scores estimated over comparable corpora in order to improve accuracy. We observe improvements between 0.5 and 1.7 BLEU translating Tamil, Telugu, Bengali, Malayalam, Hindi, and Urdu into English.

1 Introduction

Standard statistical machine translation (SMT) models (Koehn et al., 2003) are trained using large, sentence-aligned parallel corpora. Unfortunately, parallel corpora are not always available in large enough quantities to train robust models (Kolachina et al., 2012). In this work, we consider the situation in which we have access to only a small amount of bitext for a given low resource language pair, and we wish to supplement an SMT model with additional translations and features estimated using comparable corpora in the source and target languages. Assuming access to a small amount of parallel text is realistic, especially considering the recent success of crowdsourcing translations (Zaidan and Callison-Burch, 2011; Ambati, 2011; Post et al., 2012).

We frame the shortcomings of SMT models trained on limited amounts of parallel text[1] in terms of accuracy and coverage. In this context, coverage refers to the number of words and phrases that a model has any knowledge of at all, and it is low when the training text is small, which results in a high out-of-vocabulary (OOV) rate. Accuracy refers to the correctness of the translation pairs and their corresponding probability features that make up the translation model. Because the quality of unsupervised automatic word alignments correlates with the amount of available parallel text and alignment errors result in errors in extracted translation pairs, accuracy tends to be low in low resource settings. Additionally, estimating translation probabilities[2] over sparse training sets results in inaccurate feature scores.

Given these deficiencies, we begin with a baseline SMT model learned from a small parallel corpus and supplement the model to improve its accuracy and coverage. We apply techniques presented in prior work that use *comparable corpora* to estimate similarities between word and phrases. In particular, we build on prior work in bilingual lexicon induction in order to predict translations for OOV words, improving coverage. We then use the same corpora to estimate additional translation feature scores, improving model accuracy. We see improvements in translation quality between 0.5

*Performed while faculty at Johns Hopkins University

[1] We consider low resource settings to be those with parallel datasets of fewer than 1 million words. Most standard MT datasets contain tens or hundreds of millions of words.

[2] Estimating reordering probabilities over sparse data also leads to model inaccuracies; we do not tackle that here.

and 1.7 BLEU points translating the following low resource languages into English: Tamil, Telugu, Bengali, Malayalam, Hindi, and Urdu.

2 Previous Work

Prior work shows that a variety of signals, including distributional, temporal, topic, and string similarity, may inform bilingual lexicon induction (Rapp, 1995; Fung and Yee, 1998; Rapp, 1999; Schafer and Yarowsky, 2002; Koehn and Knight, 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). Other work has used decipherment techniques to learn translations from monolingual and comparable data (Ravi and Knight, 2011; Dou and Knight, 2012; Nuhn et al., 2012). Daumé and Jagarlamudi (2011) use contextual and string similarity to mine translations for OOV words in a high resource language domain adaptation for a machine translation setting. Unlike most other prior work on bilingual lexicon induction, Daumé and Jagarlamudi (2011) use the translations in end-to-end SMT.

More recently, Irvine and Callison-Burch (2013) combine a variety of the techniques for estimating word pair similarity using source and target language comparable corpora. That work shows that only a small amount of supervision is needed to learn how to effectively combine similarity features into a single model for doing bilingual lexicon induction. In this work, because we assume access to a small amount of bilingual data, it is natural to take such a supervised approach to inducing new translations, and we directly apply that of Irvine and Callison-Burch (2013).

Klementiev et al. (2012) use comparable corpora to score an existing Spanish-English phrase table extracted from the Europarl corpus. In this work, we directly apply their technique for scoring an existing phrase table. However, unlike that work, our initial phrase tables are estimated from small parallel corpora for genuine low resource languages. Additionally, we include new translations discovered in comparable corpora.

Other prior work has mined supplemental parallel data from comparable corpora (Munteanu and Marcu, 2006; AbduI-Rauf and Schwenk, 2009; Smith et al., 2010; Uszkoreit et al., 2010; Smith et al., 2013). Such efforts are orthogonal and complementary to the approach that we take.

Language	Train Words (k)		Dev Types	Dev Tokens
	Sent	Dict	% OOV	% OOV
Tamil	335	77	44	25
Telugu	414	41	39	21
Bengali	240	7	37	18
Malayalam	263	151	6	3
Hindi	659	n/a	34	11
Urdu	616	116	23	6

Table 1: Information about datasets released by Post et al. (2012): thousands of words in the source language parallel sentences and dictionaries, and percent of development set word types (unique word tokens) and word tokens that are OOV (do not appear in either section of the training data).

Language	Web Crawls	Wikipedia
Tamil	0.1	4.4
Telugu	0.4	8.6
Bengali	2.7	3.3
Malayalam	0.1	3.7
Hindi	18.1	6.4
Urdu	285	2.5

Table 2: Millions of words of time-stamped web crawls and Wikipedia text, by language.

3 Using Comparable Corpora to Improve Accuracy and Coverage

After describing our bilingual and comparable corpora, we briefly describe the techniques proposed by Irvine and Callison-Burch (2013) and Klementiev et al. (2012). The contribution of this paper is the application and combination of these techniques in truly low resource translation conditions.

3.1 Datasets

Post et al. (2012) used Mechanical Turk to collect small parallel corpora for the following Indian languages and English: Tamil, Telugu, Bengali, Malayalam, Hindi, and Urdu. They collected both parallel sentence pairs and a dictionary of word translations.[3] We use all six datasets, which provide real low resource data conditions for six truly low resource language pairs. Table 1 shows statistics about the datasets.

Table 2 lists the amount of comparable data that we use for each language. Following both Klementiev et al. (2012) and Irvine and Callison-Burch (2013), we use time-stamped web crawls as well as interlingually linked Wikipedia documents. We use the time-stamped data to estimate temporal similarity and the interlingual Wikipedia links, which indicate documents about the same topic written in different languages, to estimate

[3] No dictionary was provided for Hindi.

topic similarity. We use both datasets in combination with a dictionary derived from the small parallel corpora to estimate contextual similarity.

3.2 Improving Coverage

In order to improve the coverage of our low resource translation models, we use bilingual lexicon induction techniques to learn translations for words which appear in our test sets but not in our training data (OOVs). Bilingual lexicon induction is the task of inducing pairs of words that are translations of one another from monolingual or comparable corpora. Irvine and Callison-Burch (2013) use a diverse set of features estimated over comparable corpora and a small set of known translations as supervision for training a discriminative classifier, which makes predictions (translation or not a translation) on test set words paired with all possible translations. Possible translations are taken from the set of all target words appearing in the comparable corpora. Candidates are ranked according to their classification scores. They achieve very good performance on the induction task itself compared with an unsupervised baseline that aggregates the same similarity features uniformly. In our setting, we have access to a small parallel corpus, which makes such a supervised approach to bilingual lexicon induction a natural choice.

We use the framework described in Irvine and Callison-Burch (2013) directly, and further details may be found there. In particular, we use the same feature set, which includes the temporal, contextual, topic, orthographic, and frequency similarity between a candidate translation pair. We derive translations to serve as positive supervision from our automatically aligned parallel text[4] and, like the prior work, use random word pairs as negative supervision. Figure 1 shows some examples of Bengali words, their correct translations, and the top-3 translations that this framework induces.

In our initial experiments, we add the highest ranked English candidate translation for each source language OOV to our phrase tables. Because all of the OOVs appear at least once in our comparable corpora,[5] we are able to mine translations for all of them. Adding these translations by definition improves the coverage of our MT models. Then, in additional sets of experiments, we also induce translations for source language words which are *low frequency* in the training data and supplement our SMT models with top-k translations, not just the highest ranked.

Source	Induced Translations	Correct Translation
গাণিতিকভাবে	mathematical equal ganitikovabe	mathematically
ফাংশন	function functions variables	function
অভিষেক	made goal earned	inauguration

Figure 1: Examples of OOV Bengali words, our top-3 ranked induced translations, and their correct translations.

3.3 Improving Accuracy

In order to improve the accuracy of our models, we use comparable corpora to estimate additional features over the translation pairs in our phrase tables and include those features in tuning and decoding. This approach follows that of Klementiev et al. (2012). We compute both phrasal features and lexically smoothed features (using word alignments, like the Moses lexical translation probabilities) for all of the following except orthographic similarity, for which we only use lexically smoothed features,[6] resulting in nine additional features: temporal similarity based on time-stamped web crawls, contextual similarity based on web crawls and Wikipedia (separately), orthographic similarity using normalized edit distance, and topic similarity based on inter-lingually linked Wikipedia pages. Our hope is that by adding a diverse set of similarity features to the phrase tables, our models will better distinguish between good and bad translation pairs, improving accuracy.

4 Experiments

4.1 Experimental setup

We use the data splits given by Post et al. (2012) and, following that work, include the dictionaries in the training data and report results on the devtest set using case-insensitive BLEU and four references. We use the Moses phrase-based MT framework (Koehn et al., 2007). For each language, we extract a phrase table with a phrase limit of seven. In order to make our results comparable to those of Post et al. (2012), we follow that work and use

[4]GIZA++ intersection alignments over all training data.

[5]The Post et al. (2012) datasets are crowdsourced English translations of source Wikipedia text. Using Wikipedia as comparable corpora, we observe all OOVs at least once.

[6]Because the words within a phrase pair are often reordered, phrase-level orthographic similarity is unreliable.

Language	Top-1 Acc.	Top-10 Acc.
Tamil	4.5	10.2
Telugu	32.8	47.9
Bengali	17.9	29.8
Malayalam	12.9	23.0
Hindi	44.3	57.6
Urdu	16.1	33.8

Table 3: Percent of word types in a held out portion of the training data which are translated correctly by our bilingual lexicon induction technique. Evaluation is over the top-1 and top-10 outputs in the ranked lists for each source word.

the English side of the training data to train a language model. Using a language model trained on a larger corpus (e.g. the English side of our comparable corpora) may yield better results, but such an improvement is orthogonal to the focus of this work. Throughout our experiments, we use the batch version of MIRA (Cherry and Foster, 2012) for tuning the feature set.[7] We rerun tuning for all experimental conditions and report results averaged over three tuning runs (Clark et al., 2011).

Our baseline uses the bilingually extracted phrase pairs and standard translation probability features. We supplement it with the top ranked translation for each OOV to improve coverage (+ OOV Trans) and with additional features to improve accuracy (+Features). In Section 4.2, we make each modification separately and then together. Then we present additional experiments where we induce translations for low frequency words, in addition to OOVs (4.3), append top-k translations (4.4), vary the amount of training data used to induce the baseline model (4.5), and vary the amount of comparable corpora used to estimate features and induce translations (4.6).

4.2 Results

Before presenting end-to-end MT results, we examine the performance of the supervised bilingual lexicon induction technique that we use for translating OOVs. In Table 3, top-1 accuracy is the percent of source language words in a held out portion of the training data[8] for which the highest ranked English candidate is a correct translation.[9] Performance is lowest for Tamil and highest for Hindi. For all languages, top-10 accuracy is much higher than the top-1 accuracy. In Section 4.4, we explore appending the top-k translations for OOV words to our model instead of just the top-1.

Table 4 shows our results adding OOV translations, adding features, and then both. Additional translation features alone, which improve our models' accuracy, increase BLEU scores between 0.18 (Bengali) and 0.60 (Malayalam) points.

Adding OOV translations makes a big difference for some languages, such as Bengali and Urdu, and almost no difference for others, like Malayalam and Tamil. The OOV rate (Table 1) is low in the Malayalam dataset and high in the Tamil dataset. However, as Table 3 shows, the translation induction accuracy is low for both. Since few of the supplemental translations are correct, we don't observe BLEU gains. In contrast, induction accuracies for the other languages are higher, OOV rates are substantial, and we do observe moderate BLEU improvements by supplementing phrase tables with OOV translations.

In order to compute the *potential* BLEU gains that we could realize by correctly translating all OOV words (achieving 100% accuracy in Table 3), we perform an oracle experiment. We use automatic word alignments over the test sets to identify correct translations and append those to the phrase tables.[10] The results, in Table 4, show possible gains between 4.3 (Telugu and Bengali) and 0 (Malayalam) BLEU points above the baseline. Not surprisingly, the possible gain for Malayalam, which has a very low OOV rate, is very low. Our *+OOV Trans.* model gains between 0% (Tamil) and 38% (Urdu) of the potential improvement.

Using comparable corpora to improve both accuracy (+Features) and coverage (+OOV Trans.) results in translations that are better than applying either technique alone for five of the six languages. BLEU gains range from 0.48 (Bengali) to 1.39 (Urdu). We attribute the particularly good Urdu performance to the relatively large comparable corpora (Table 2). As a result, we have already begun to expand our web crawls for all languages. In Section 4.6, we present results varying the amount of Urdu-English comparable corpora used to induce translations and estimate additional features.

Table 4 also shows the Hiero (Chiang, 2005) and SAMT (Zollmann and Venugopal, 2006) results that Post et al. (2012) report for the same

[7] We experimented with MERT and PRO as well but saw consistently better baseline performance using batch MIRA.

[8] Described in Section 3.2. We retrain with all training data for MT experiments.

[9] Post et al. (2012) gathered up to six translations for each source word, so some have multiple correct translations

[10] Because the automatic word alignments are noisy, this oracle is conservative.

Experiment	Tamil		Telugu		Bengali		Malayalam		Hindi		Urdu	
	BLEU	Diff.	BLEU	Diff.	BLEU	Diff.	BLEU	Diff.	BLEU	Diff.	BLEU	Diff.
Baseline	9.45		11.72		12.07		13.55		15.01		20.39	
+Features	9.77	+0.32	11.96	+0.24	12.25	+0.18	14.15	+0.60	15.34	+0.33	20.97	+0.58
+OOV Trans.	9.45	0.00	12.20	+0.48	**12.74**	+0.67	13.65	+0.10	15.59	+0.58	21.30	+0.91
+Feats & OOV	**9.98**	+0.53	**12.25**	+0.53	12.55	+0.48	**14.18**	+0.63	**16.08**	+1.07	**21.78**	+1.39
OOV Oracle	12.32	+2.87	16.04	+4.32	16.41	+4.34	13.55	0.00	17.72	+2.71	22.80	2.41
Hiero	9.81		12.46		12.72		13.72		15.53		19.53	
SAMT	9.85		12.61		13.53		14.28		17.29		20.99	

Table 4: BLEU performance gains that target coverage (+OOV Trans.) and accuracy (+Features), and both (+Feats & OOV). OOV oracle uses OOV translations from automatic word alignments. Hiero and SAMT results are reported in Post et al. (2012).

datasets. Both syntax-based models outperform the phrase-based MT baseline for each language except Urdu, where the phrase-based model outperforms Hiero. Here, we extend a phrase-based rather than a syntax-based system because it is simpler. However, our improvements may also apply to syntactic models (future work). Because our efforts have focused on the accuracy and coverage of translation pairs and have not addressed reordering or syntax, we expect that combining them with an SAMT grammar will result in state-of-the art performance.

4.3 Translations of Low Frequency Words

Given the positive results in Section 4.2, we hypothesize that mining translations for low frequency words, in addition to OOV words, may improve accuracy. For source words which only appear a few times in the parallel training text, the bilingually extracted translations in the standard phrase table are likely to be inaccurate. Therefore, we perform additional experiments varying the minimum source word training data frequency for which we induce additional translations. That is, if $freq(w_{src}) \leq M$, we induce a new translation for it and include that translation in our phrase table. Note that in the results presented in Table 4, $M = 0$. In these experiments, we include our additional phrase table features estimated over comparable corpora and hope that these scores will assist the model in choosing among multiple translation options for low frequency words, one or more of which is extracted bilingually and one of which is induced using comparable corpora. Table 5 shows the results when we vary M. As before, we average BLEU scores over three tuning runs.

In general, modest BLEU score gains are made as we supplement our phrase-based models with induced translations of low frequency words. The highest performance is achieved when M is between 5 and 50, depending on language. The largest gains are 0.5 and 0.3 BLEU points for Bengali and Urdu, respectively, at $M = 25$. This is not surprising; we also saw the largest relative gains for those two languages when we added OOV translations to our baseline model. With the addition of low frequency translations, our highest performing Urdu model achieves a BLEU score that is 1.7 points higher than the baseline.

Language	Base.	M: trans added for $freq(w_{src}) \leq M$					
		0	1	5	10	25	50
Tamil	9.5	10.0	9.9	10.2	10.2	9.9	**10.2**
Telugu	11.7	12.3	12.2	12.3	**12.4**	12.3	11.9
Bengali	12.1	12.6	12.8	13.0	12.9	**13.1**	13.0
Malayalam	13.6	14.2	14.1	**14.2**	14.2	13.9	13.9
Hindi	15.0	16.1	16.1	16.2	**16.2**	16.0	15.8
Urdu	20.4	21.8	21.8	21.8	21.9	**22.1**	21.8

Table 5: Varying minimum training data frequency of source words for which new translations are induced and included in the phrase-based model. In all cases, the top-1 induced translation is added to the phrase table and features estimated over comparable corpora are included (i.e. +Feats & Trans model).

In different data conditions, inducing translations for low frequency words may result in better or worse performance. For example, the size of the training set impacts the quality of automatic word alignments, which in turn impacts the reliability of translations of low frequency words. However, the experiments detailed here suggest that including induced translations of low frequency words will not hurt performance and may improve it.

4.4 Appending Top-K Translations

So far we have only added the top-1 induced translation for OOV and low frequency source words to our phrase-based model. However, the bilingual lexicon induction results in Table 3 show that accuracies in the top-10 ranked translations are, on average, nearly twice the top-1 accuracies. Here, we explore adding the top-k induced translations. We hope that our additional phrase table features estimated over comparable corpora will enable the

Language	Base.	k: top-k translations added				
		1	3	5	10	25
Tamil	9.5	10.0	**10.0**	9.8	10.0	10.0
Telugu	11.7	**12.3**	11.7	11.9	11.7	11.6
Bengali	12.1	12.6	12.6	12.6	12.7	**12.8**
Malayalam	13.6	14.2	14.2	14.2	**14.2**	14.1
Hindi	15.0	**16.1**	16.0	15.9	15.9	15.9
Urdu	20.4	21.8	**21.8**	21.7	21.5	21.6

Table 6: Adding top-k induced translations for source language OOV words, varying k. Features estimated over comparable corpora are included (i.e. +Feats & Trans model). The highest BLEU score for each language is highlighted. In many cases differences are less than 0.1 BLEU.

decoder to correctly choose between the k translation options. We induce translations for OOV words only ($M = 0$) and include all comparable corpora features.

Table 6 shows performance as we append the top-k ranked translations for each OOV word and vary k. With the exception of Bengali, using a k greater than 1 does not increase performance. In the case of Bengali, and additional 0.2 BLEU is observed when the top-25 translations are appended. In contrast, we see performance decrease substantially for other languages (0.7 BLEU for Telugu and 0.2 for Urdu) when the top-25 translations are used. Therefore, we conclude that, in general, the models do not sufficiently distinguish good from bad translations when we append more than just the top-1. Although using a k greater than 1 means that more correct translations are in the phrase table, it also increases the number of possible outputs over which the decoder must search.

4.5 Learning Curves over Parallel Data

In the experiments above, we only evaluated our methods for improving the accuracy and coverage of models trained on small amounts of bitext using the full parallel training corpora released by Post et al. (2012). Here, we apply the same techniques but vary the amount of parallel data in order to generate learning curves. Figure 2 shows learning cures for all six languages. In all cases, results are averaged over three tuning runs. We sample both parallel sentences and dictionary entries.

All six learning curves show similar trends. In all experimental conditions, BLEU performance increases approximately linearly with the log of the amount of training data. Additionally, supplementing the baseline with OOV translations improves performance more than supplementing the baseline with additional phrase table scores based

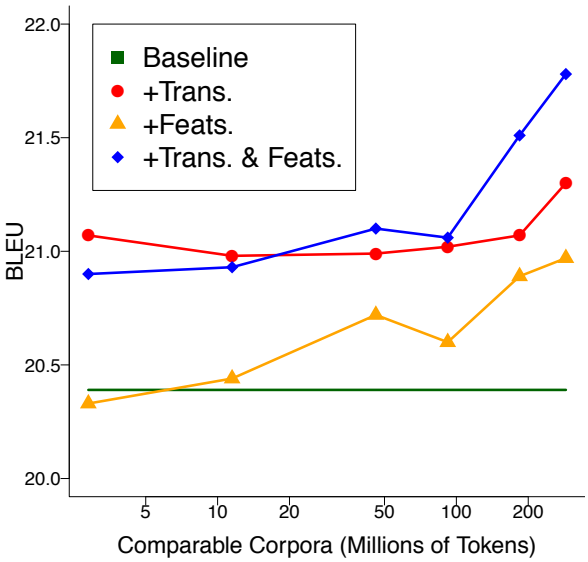

Figure 3: English to Urdu translation results using varying amounts of comparable corpora to estimate features and induce translations.

on comparable corpora. However, in most cases, supplementing the baseline with both translations and features improves performance more than either alone. Performance gains are greatest when very little training data is used. The Urdu learning curve shows the most gains as well as the cleanest trends across training data amounts. As before, we attribute this to the relatively large comparable corpora available for Urdu.

4.6 Learning Curves over Comparable Corpora

In our final experiment, we consider the effect of the amount of *comparable corpora* that we use to estimate features and induce translations. We present learning curves for Urdu-English because we have the largest amount of comparable corpora for that pair. We use the full amount of parallel data to train a baseline model, and then we randomly sample varying amounts of our Urdu-English comparable corpora. Sampling is done separately for the web crawl and Wikipedia comparable corpora. Figure 3 shows the results. As before, results are averaged over three tuning runs.

The phrase table features estimated over comparable corpora improve end-to-end MT performance more with increasing amounts of comparable corpora. In contrast, the amount of comparable corpora used to induce OOV translations does not impact the performance of the resulting MT system as much. The difference may be due

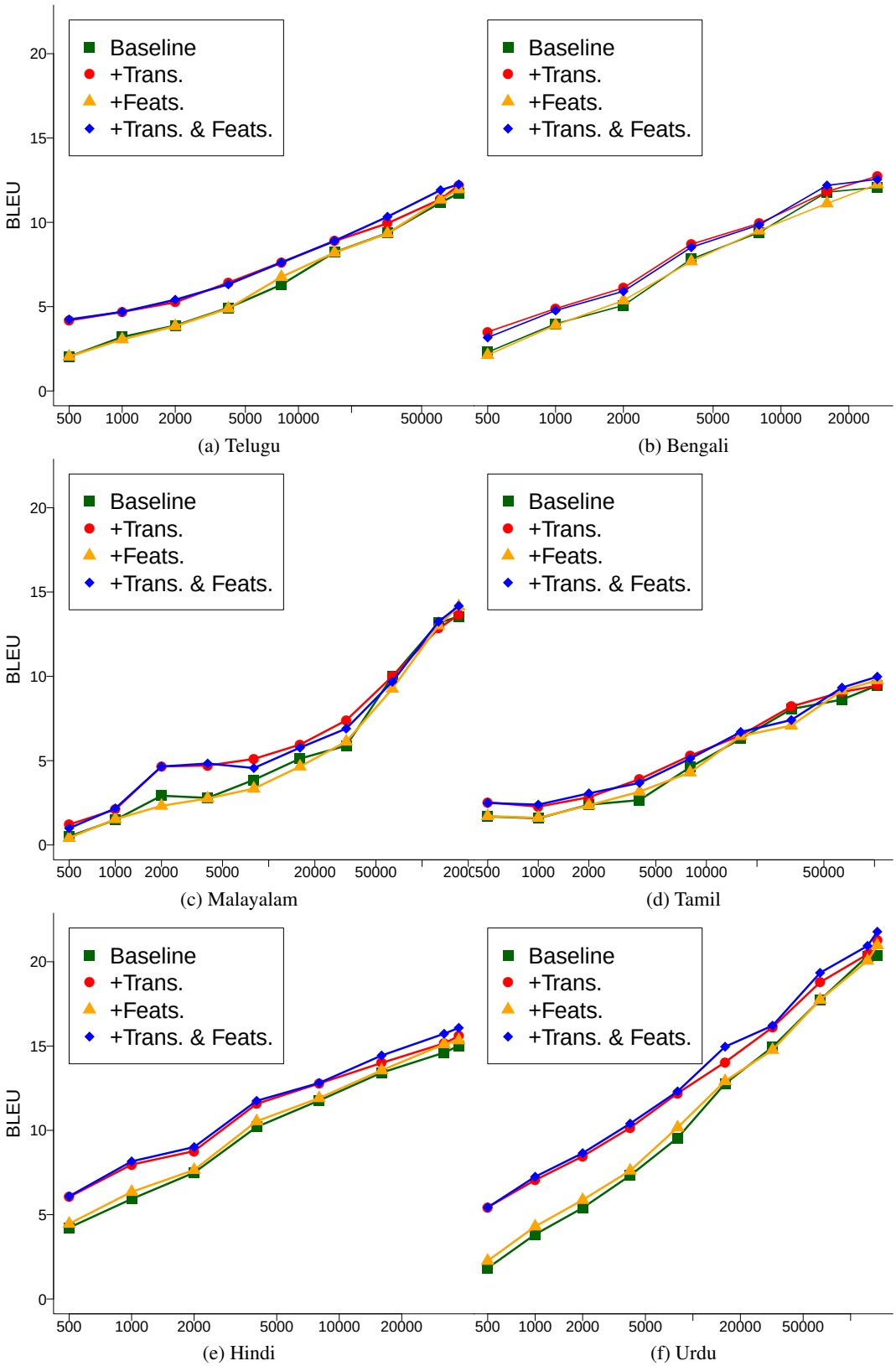

Figure 2: Comparison of learning curves over lines of parallel training data for four SMT systems: our baseline phrase-based model (baseline), model that supplements the baseline with translations of OOV words induced using our supervised bilingual lexicon induction framework (+Trans), model that supplements the baseline with additional phrase table features estimated over comparable corpora (+Feats), and a system that supplements the baseline with both OOV translations and additional features (+Trans & Feats).

to the fact that data sparsity is always more of an issue when estimating features over *phrase pairs* than when estimating features over *word pairs* because phrases appear less frequently than words in monolingual corpora. Our comparable corpora features are estimated over phrase pairs while translations are only induced for OOV words, not phrases. So, it makes sense that the former would benefit more from larger comparable corpora.

5 Conclusion

As Post et al. (2012) showed, it is reasonable to assume a small parallel corpus for training an SMT model even in a low resource setting. We have used comparable corpora to improve the accuracy and coverage of phrase-based MT models built using small bilingual corpora for six low resource languages. We have shown that our methods improve BLEU score performance independently and that their combined impact is nearly additive. Additionally, our results show that adding induced translations of low frequency words improves performance beyond what is achieved by inducing translations for OOVs alone. Finally, our results show that our techniques improve relative performance most when very little parallel training data is available.

6 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

Sadaf AbduI-Rauf and Holger Schwenk. 2009. On the use of comparable corpora to improve smt performance. In *Proceedings of the Conference of the European Association for Computational Linguistics (EACL)*.

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

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)*.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

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)*.

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)*.

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)*.

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)*.

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. 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, 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)*.

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)*.

Dragos Munteanu and Daniel Marcu. 2006. Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

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)*.

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)*.

Jason Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. 2013. Dirt cheap web-scale parallel text from the common crawl. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Jakob Uszkoreit, Jay M. Ponte, Ashok C. Popat, and Moshe Dubiner. 2010. Large scale parallel document mining for machine translation. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

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)*.

Andreas Zollmann and Ashish Venugopal. 2006. Syntax augmented machine translation via chart parsing. In *Proceedings of the Workshop on Statistical Machine Translation (WMT)*.