The IIT Bombay SMT System for ICON 2014 Tools Contest

Anoop Kunchukuttan, Ratish Puduppully, Rajen Chatterjee, Abhijit Mishra, Pushpak Bhattacharyya
Indian Institute of Technology Bombay
{anoopk, ratishp, rajen, abhijitmishra, pb}@cse.iitb.ac.in

Abstract

In this paper, we describe our submission to the ICON 2014 Tools Contest for Machine Translation. The source languages are English, Marathi, Tamil, Telugu, Bengali and the target language is Hindi. We submitted 15 systems; 5 each for the *tourism*, *health* and *general* domains. Our submission is a Phrase-based Statistical Machine Translation system with preprocessing and post-processing elements. As preprocessing, we perform source-side reordering for English-Hindi translation, and source-side word segmentation for Indian language to Hindi translation. The translation outputs were post-edited by transliterating untranslated words. Our goal was to handle key divergences between the language pairs involved by using language-independent methods which can be scaled across Indian languages without the need for expensive annotation creation. Hence, both the transliteration model and word-segmenter have been learnt using unsupervised techniques, whereas our source reordering works for any target Indian language. Our approach results in a cumulative increase of BLEU scores in range of 3-6 compared to baseline phrase based SMT systems, with source side segmentation contributing to a major chunk of improvement. This demonstrates that resources created using unsupervised methods can significantly improve SMT performance involving Indian languages.

1 Introduction

Being one of the world's most linguistically diverse countries, the need for machine translation (MT) and the research challenges it offers need not be emphasized. MT research for Indian languages is at a nascent stage as compared to translation involving English, European languages, Chinese, Arabic etc. The MT Tools Contest for ICON 2014 requires us to translate English, Tamil, Telugu, Bengali and Marathi into Hindi. The Indian languages (abbreviated as IL) exhibit shared characteristics like: (i) relatively free word order, with SOV being the canonical word order, (ii) similar orthographic systems descended from the Brahmi script based on auditory phonetic principles, (iii) vocabulary and grammatical tradition derived from Sanskrit, and (iv) morphological richness. Dravidian languages are highly agglutinative.

Morphological richness of Indian languages and structural divergence between Indian languages and English demand that we look at richer methods beyond pure phrase based techniques. To handle structural divergence for English-Hindi translation, we have used a rule-based source side reordering system which works across target Indian languages. Challenges owing to morphological richness for Indian language to Hindi translation have been addressed by segmenting the source text into its morphemes prior to translation. Our results indicate that handling these key aspects, especially morphology, yields a substantial improvement in translation quality.

Due to the unavailability of linguistic resources and tools for most Indian languages, both the transliteration model and morphanalyzer have been learnt using unsupervised techniques.

Finally, we point to the limitations of using BLEU as a metric for evaluation. Alternatively, we use METEOR for evaluation and our preliminary observations indicate that METEOR may be closer to human judgments than BLEU. This would be an interesting direction of investigation for future. We also believe that manual evaluations of the submissions should be done for any future

competitions to have a more realistic evaluation and to build a pool of human judgments which can help in the study of MT evaluation for Indian languages.

The tools and resources used in this work (transliterator [1], morphanalyzer [2] and source reordering system [3]) have been made available as open-source code for research use.

The rest of the paper is organized as follows: Section 2 describes our MT system's architecture while the system components - Source-Side Reordering, Source-Side Segmentation and Transliteration Post-Editing - are discussed in Section 3. Section 4 discusses experimental configuration, Section 5 describes the results and observations while Section 6 concludes the paper.

2 System Architecture

We have implemented two pipelined architectures: one for English-Hindi (en-hi) translation and another for Indian Language-Hindi (IL-hi) translation. This section describes both these pipelines briefly.

2.1 English-Hindi Translation Pipeline

An input sentence is processed through the following pipeline stages, illustrated by an example in Table 1.

Source side reordering (reord) The input English sentence is reordered to conform to the canonical Hindi word order by applying reordering transformations on the English parse tree, and reading off the leaves of the modified tree to generate the Hindi-ordered English (HoE) sentence.

Phrase-Based Machine Translation (PB) The reordered English sentence is then translated using a standard phrase-based SMT (PB-SMT) model. The model has been trained on a parallel corpus of HoE to Hindi sentences.

Transliteration Post-Editing (translit) The untranslated words (named entities and out-of-vocabulary (OOV) words) from the previous stage are transliterated to generate top-k possible transliterations. Plugging these candidate transliterations back into the translation system can generate many potential, revised translations for the source sentence. These candidates translations are rescored using a target language model to select the best final translation for the source sentence.

2.2 Indian Language-Hindi Translation Pipeline

An input sentence is processed through the following pipeline stages, illustrated by an example in Table 2.

Source side word-segmentation (morph) The input Indian language sentence is segmented using a language independent morphanalyzer. Since Indian languages are either agglutinative and/or inflectionally rich, the segmentation of the text generates simpler morphemes. The data sparsity issue is handled by this process.

Phrase-Based Machine Translation (PB) The segmented IL sentence is then translated using a standard PB-SMT model. The model has been trained on a parallel corpus of segmented-IL to Hindi sentences.

Transliteration Post-Editing (translit) The transliteration post-editing stage works exactly as described for the English-Hindi system described in the previous section to generate the final translation for the source sentence. For Indian languages which share vocabulary, transliteration may also serve the purpose of translation of OOV words.

3 System Components

In this section, we describe various system components and design choices at length.

3.1 Source-Side Reordering

There is a significant structural divergence between Indian languages and English. Most significant among them is that Hindi is slightly free-word order with Subject-Object-Verb (SOV) being the canonical word order, whereas English has a more rigid Subject-Verb-Object (SVO) word order. Research has shown that source side reordering to conform to target side word order improves machine translation (Collins et al., 2005). The improvement is on account of two facts: (i) longer phrases can be learnt resulting in more fluent output, and (2) the decoder cannot look at long dis-

[1] https://github.com/anoopkunchukuttan/indic_nlp_resources

[2] https://github.com/anoopkunchukuttan/indic_nlp_library

[3] http://www.cfilt.iitb.ac.in/~moses/download/en_il_src_reorder/register.html

Steps	Sentence
Input Sentence	Bilirubin named colored substance is made in our body absolutely everyday .
Source side reordering	Bilirubin named colored substance in our body absolutely everyday made is .
Phrase based Translation	Bilirubin नामक रंग के पदार्थ हमारे शरीर में प्रतिदिन बनते है ।
Transliteration	वाइलीरुविन नामक रंग के पदार्थ हमारे शरीर में प्रतिदिन बनते है ।

Table 1: Example for En-Hi translation

Steps	Sentence
Input Sentence	व्यायाम आणि पदयात्रा यापैकी एखादे वर्णित प्रकाराचे दैनिक कार्य हृदय विकार आणि मधुमेह नियंत्रित करते .
Morphology Splitting	व्यायाम आणि पद यात्रा यापैकी एखाद े वर्ण ित प्रकार ाचे दैनिक कार्य हृदय विकार आणि मधुमेह नियंत्रित करते .
Phrase based Translation	व्यायाम और पदयात्रा में से किसी भी वर्णित प्रकार का दैनिक कार्य हृदय रोग और मधुमेह को नियंत्रित करता है ।
Transliteration	व्यायाम और पदयात्रा में से किसी भी वर्णित प्रकार का दैनिक कार्य हृदय रोग और मधुमेह को नियंत्रित करता है ।

Table 2: Example for Mr-Hi translation

tance reorderings due to computational considerations, but with source-side reordering, reordering in a small window is sufficient. We use a rule based approach which applies reordering transformations to the English sentence parse tree to generate the HoE sentences (Ramanathan et al., 2008; Patel et al., 2013). Although these rules were primarily developed for Hindi, they work for translation from English to any Indian languages since the the structural divergences are similar (Kunchukuttan et al., 2014).

3.2 Source-Side Word-Segmentation

Morphologically-rich/agglutinative languages have low token-to-type ratio and can generate a potentially infinite number of word types by combining morphemes. Due to this, PB-SMT systems encounter the following challenges while learning translation systems involving morphologically rich languages: (i) unreliable estimation of word translation probabilities due to data sparsity, and (ii) difficulty in generalization to test scenarios since new word forms can come up in test irrespective of the training corpus size. To address this problem, we segment the tokens in the source sentence into its constituent morphemes. The segmented corpus is used as source side of the parallel corpus, and a test sentence is also segmented before translation. We have not done any segmentation of the target language sentence, since it introduces additional complexities in ensuring morphemes of a single target word are contiguous in the translation. Since, Hindi is not highly agglutinative, this choice is reasonable; though it would clearly be insufficient for morphologically rich target languages.

Most Indian languages do not have a morphanalyzer/word-segmenter. Hence, we built unsupervised morphanalyzers using *Morfessor 2.0* [4]. The morphanalyzers are learnt from monolingual corpora using a probabilistic generative model which uses maximum-a-posteriori estimation with sparse priors inspired by the Minimum Description Length (MDL) principle (Virpioja et al., 2013). Even though these morphanalyzers do not distinguish between stem, prefix, suffix or provide any grammatical properties, the segmentation generated by this method is sufficient for preprocessing parallel corpora. We used only the word types without considering their frequencies for training since this training configuration has been shown to perform better when no annotated data is available for tuning (Virpioja et al., 2013).

3.3 Transliteration Model

Training transliteration systems requires parallel transliteration corpora. Since such corpora is generally not available for most Indian language pairs, we utilize the unsupervised approach to machine transliteration proposed by Durrani et al. (2014). This approach requires a parallel bitext corpora, which is available when building an SMT system. The transliteration model is built in two stages. First, parallel transliteration pairs are mined from a parallel corpus. Then, a transliteration model is trained from the transliteration corpus by formulating transliteration as a translation problem from

[4] http://www.cis.hut.fi/projects/morpho/morfessor2.shtml

source character strings to target character strings. Most Indian language scripts, including the ones in the competition, derive from the *Brahmi* script and are phonetic in nature. Moreover, the Unicode codepoints of these scripts are coordinated. A simple transliteration scheme is to just map the Unicode codepoints. We experimented with this approach too, but the statistical approach yields better and hence only those results are reported in the paper.

4 Experimental Configuration

We trained 15 systems over three domains - tourism, health and general. We use the MOSES toolkit (Koehn et al., 2007) for our PB-SMT system. We use the *grow-diag-final-and heuristic* for extracting phrases and the *msd-bidirectional-fe* model for lexicalized reordering. We tuned the trained models using Minimum Error Rate Training (MERT) with default parameters. For Health and Tourism, we made a training, tuning and test data split of 22800, 500, 1200 sentences respectively. For General domain, we made a training,tuning and devtest data split of 45600, 1000, 2400 sentences respectively. The official test set comprised for 500 sentences each in Health and Tourism domains and 1000 sentences in General domain for each source language.

The transliteration pairs are mined from the parallel training corpora provided for the competition. Morphology analyzers were trained using the *Leipzig Corpus* [5] and the monolingual data provided for the competition. We used a 5-gram language model built using SRILM (Stolcke and others, 2002) with Kneser-Ney smoothing (Kneser and Ney, 1995) with 1.5 million sentences from WMT monolingual corpus. The evaluation was done using the BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005) metrics. For METEOR, the synonyms were obtained from IndoWordNet (Bhattacharyya, 2010), while a variant of the IndoWordNet-assisted stemmer (Bhattacharyya et al., 2014) was used for stemming.

5 Results and Analysis

Tables 3-4 document the BLEU (B) and METEOR (M) scores for the systems in three domains - Tourism, Health and General. For English-Hindi, source reordering substantially improves the translation quality and transliteration contributes to a significant increase in the translation quality. For instance, in the General domain, BLEU scores with reordering show an increase of 3.8 points compared to PB-SMT and a cumulative increase of 4.8 BLEU points with transliteration enabled over the baseline. For Indian language-Hindi translation, source word-segmentation results in a sharp increase in the translation quality, whereas transliteration contributes to a modest increase. For Telugu-Hindi, source segmentation results in 5.6 BLEU points increase over the baseline and transliteration increments it by another 0.4 BLEU points. The METEOR scores also record a similar increase. Figure 1 and 2 illustrate how the preprocessing and post-processing extensions help improve the baseline translation.

Table 5 shows the BLEU and METEOR scores on the test set. The BLEU scores are lower compared to the scores obtained on the devtest dataset. On manually inspecting the translation output, we see that for many instances instead of exact n-gram match the system has chosen synonyms. BLEU does not account for such matches and hence the BLEU scores are much lower than what a manual judgment suggests. Our observations are in line with the shortcomings of BLEU as identified by Ananthakrishnan et al. (2007). Hence we also evaluated our systems using METEOR, which can account for synonym and stem matches. The METEOR scores on the devtest dataset and official test dataset are comparable, which we believe to be more representative of the quality of the translation system. Hence, we make the case for manual evaluation of the translation results on the official test dataset so that meaningful comparisons can be made. It will also result in a corpus of human evaluations which can be useful for the study of automatic evaluation metrics.

5.1 Analysis of errors

The unsupervised Morphology splitting sometimes did aggressive splitting, especially in case of named entities, hence the output translation was impacted. Example for the same is shown in Table 6. The rule-based source side reordering sometimes performs reordering which results in word order not conforming to the target Hindi word order. This impacts the fluency of the resultant translation. Table 7 gives an example.

[5] http://corpora.uni-leipzig.de/download.html

		Tourism			Health			General		
Lang Pair	**Metric**	**PB**	**PB+ reord**	**PB+ reord+ translit**	**PB**	**PB+ reord**	**PB+ reord+ translit**	**PB**	**PB+ reord**	**PB+ reord+ translit**
en-hi	B	20.87	27.22	**28.78**	24.03	28.63	**29.3**	23.55	28.34	**29.37**
	M	43.44	48.25	**50.07**	46.83	50.38	**51.22**	45.76	49.90	**51.11**

Table 3: Tourism, Health and General domain results for en-hi

		Tourism			Health			General		
Lang Pair	**Metric**	**PB**	**PB+ morph**	**PB+ morph+ translit**	**PB**	**PB+ morph**	**PB+ morph+ translit**	**PB**	**PB+ morph**	**PB+ morph+ translit**
bn-hi	B	34.38	37.1	**37.66**	36.46	38.66	**39.04**	36.24	38.61	**38.92**
	M	55.73	58.38	**58.98**	57.44	59.89	**60.37**	57.36	59.47	**59.84**
mr-hi	B	40.24	**46.86**	**46.86**	39.84	46.86	**46.86**	41.35	47.92	**47.92**
	M	60.78	**66.47**	**66.47**	60.29	**66.76**	**66.76**	61.79	**67.17**	**67.17**
ta-hi	B	17.76	22.42	**22.91**	21.55	26.05	**26.35**	20.45	25.34	**25.65**
	M	36.11	41.61	**42.31**	39.94	45.03	**45.42**	38.93	44.57	**50.00**
te-hi	B	26.99	31.77	**32.45**	29.74	35.59	**36.04**	29.88	35.43	**35.88**
	M	47.20	52.48	**53.35**	50.05	56.05	**56.68**	50.20	55.82	**56.38**

Table 4: Tourism, Health and General domain results for bn,mr,ta,te - hi

Language Pair	**Metric**	**Health**	**Tourism**	**General**
en-hi	B	19.22	18.35	19.49
	M	43.71	42.56	43.8
bn-hi	B	28.99	29.16	28.53
	M	54.59	55.02	54.30
mr-hi	B	36.12	37.05	36.98
	M	61.69	62.17	62.16
ta-hi	B	20.65	17.81	19.31
	M	41.77	39.95	41.19
te-hi	B	20.87	27.22	28.78
	M	53.61	49.01	52.26

Table 5: Evaluation scores on the official test set

Input Sentence	गोव्यामध्ये द फीस्ट ऑफ श्री किंग्स , हॅरिटेज महोत्सव , कोंकणी नाट्य महोत्सव
Source side reordering	गोव्या मध्ये द फी स्ट ऑफ श्री किंग ्स , हॅर िटेज महोत्सव , कोंकण ी नाट्य महोत्सव

Table 6: Example for aggressive splitting with Unsupervised morphology analyzer. Here the underlined words were incorrectly split

Input Sentence	Burn on cooking 20 live scorpions in 1 litre sesame seed oil .
Source side reordering	1 in 20 live scorpions cooking on Burn sesame seed oil litre .

Table 7: Example for imprecise source side reordering

6 Conclusion

We have presented results on experiments for SMT from English, Bengali, Marathi, Tamil, Telugu into Hindi. Using source side reordering, source word segmentation and transliteration results in

improvement of upto 6 BLEU points over the baseline phrase-based SMT system. The key point is that these resources have been learnt using unsupervised methods, thus allowing them to be scaled to many Indian languages. We observe that BLEU is not a good indicator of translation quality, and more investigation is needed for better metrics of evaluation quality and as well as manual judgments of test output. In future, we plan to select outputs from various source reordering systems and investigate word segmentation of the target language.

Acknowledgements

We would like to thank the National Knowledge Network for making available computational facilities on the *Garuda Cloud* for performing computationally intensive tasks.

References

R Ananthakrishnan, Pushpak Bhattacharyya, M Sasikumar, and Ritesh M Shah. 2007. Some issues in automatic evaluation of english-hindi mt: more blues for bleu. *ICON*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65--72.

Pushpak Bhattacharyya, Ankit Bahuguna, Lavita Talukdar, and Bornali Phukan. 2014. Facilitating multi-lingual sense annotation: Human mediated lemmatizer. In *Global WordNet Conference*.

Pushpak Bhattacharyya. 2010. Indowordnet. In *In Proc. of LREC-10*.

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd annual meeting on association for computational linguistics*.

Nadir Durrani, Hieu Hoang, Philipp Koehn, and Hassan Sajjad. 2014. Integrating an unsupervised transliteration model into statistical machine translation. *EACL 2014*, page 148.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1, pages 181--184. IEEE.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177--180. Association for Computational Linguistics.

Anoop Kunchukuttan, Abhijit Mishra, Rajen Chatterjee, Ritesh Shah, and Pushpak Bhattacharyya. 2014. Sata-anuvadak: Tackling multiway translation of indian languages. In *Language Resources and Evaluation Conference*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311--318. Association for Computational Linguistics.

Raj Patel, Rohit Gupta, Prakash Pimpale, and M. Sasikumar. 2013. Reordering rules for English-Hindi SMT. In *Proceedings of the Second Workshop on Hybrid Approaches to Translation*.

Ananthakrishnan Ramanathan, Jayprasad Hegde, Ritesh Shah, Pushpak Bhattacharyya, and M. Sasikumar. 2008. Simple Syntactic and Morphological Processing Can Help English-Hindi Statistical Machine Translation. In *International Joint Conference on Natural Language Processing*.

Andreas Stolcke et al. 2002. Srilm-an extensible language modeling toolkit. In *INTERSPEECH*.

Sami Virpioja, Peter Smit, Stig-Arne Grönroos, Mikko Kurimo, et al. 2013. Morfessor 2.0: Python implementation and extensions for morfessor baseline.