Multilingual Indian Language Translation System at WAT 2018: Many-to-one Phrase-based SMT

Tamali Banerjee[1], **Anoop Kunchukuttan**[2], **Pushpak Bhattacharya**[1]
[1] Department of Computer Science and Engineering,
Indian Institute of Technology Bombay.
{tamali,pb}@cse.iitb.ac.in
[2] Microsoft AI and Research, India.
ankunchu@microsoft.com

Abstract

This paper describes our trained models of phrase-based statistical machine translation (PBSMT) systems for Indic→English and English→Indic language-pairs, which has been submitted to the WAT 2018 shared task. In addition, we have introduced many-to-one statistical machine translation (SMT). This new approach produced comparable results in terms of translation accuracy with respect to the result of baseline SMT.

1 Introduction

India is one of the most linguistically diverse countries in the world. According to the Census of India of 2001, India has 122 major languages and 1599 other languages. These languages span four major language families. North-Indian languages such as Hindi, Bengali, Sindhi belong to the Indo-Aryan branch of the Indo-European language family, whereas South-Indian languages such as Tamil, Telugu and Malayalam belong to the Dravidian language family. These are the major language families, in addition to the Austro-Asiatic and Tibeto-Burman languages spoken by a small section of the population. In addition to the similarities between languages belonging to the same language families, there are many similarities between the four language families on account of contact over a long period of time. Hence, India is referred to as a *linguistic area* (Emeneau, 1956). This relatedness manifests itself in the form of lexical, structural and morphological similarities between these languages (Bhattacharyya et al., 2016).

In the WAT 2018 shared task (Nakazawa et al., 2018), we participated as team 'Anuvaad' and trained Indic→English and English→Indic baseline SMT systems along with our proposed many-to-one Indic→English system. In neural machine translation (NMT) (Cho et al., 2014; Sutskever et al., 2014; Bahdanau et al., 2015), multilingual transfer learning approaches (includes many-to-one, one-to-many, or many-to-many translation) have shown significant improvement in translation quality with minimal increase in network complexity, especially in the case of resource-poor languages (Johnson et al., 2017; Firat et al., 2016; Dong et al., 2015). However, SMT is still superior, when the training corpus is not big enough (Koehn and Knowles, 2017). Hence, we experimented with a multilingual many-to-one SMT system for Indic language to English translation using significantly less amount of data as compared to NMT.

2 Many-to-one SMT

Despite, the huge success of NMT, SMT can still be used to achieve comparable transnational outcome in case of data scarcity (Koehn and Knowles, 2017). In order to train a SMT system, the learning of language model and translation model requires lesser amount of data as compared to learning of NMT system (Koehn and Knowles, 2017).

Mathematically, SMT is represented as-

$$\hat{e} = \arg\max_{e}(P(e|f)) = \arg\max_{e}(P(e).P(f|e)) \quad (1)$$

where, e is a sentence of English language and f

32nd Pacific Asia Conference on Language, Information and Computation
The 5th Workshop on Asian Translation
Hong Kong, 1-3 December 2018

is a sentence of foreign language.

An SMT system selects the best translated English sentence *e* given a foreign sentence *f*. The *argmax* computation is expressed as a product of language model *P(e)* and translation model $P(f|e)$. It produces the English sentence which exhibits highest probability value for a given sentence *f*.

In many-to-one SMT model, the translation model needs to be trained on merged corpus of all Indic→English language-pair, wherein, all bilingual corpora are transliterated into a certain script-pair (Kunchukuttan et al., 2015). Further, the language model is trained on the merged form of the target language corpus.

3 Experimental Setup

We trained 14 bilingual SMT systems for 7 Indic-English language-pairs (such as Bengali, Hindi, Malayalam, Tamil, Telugu, Urdu and Sinhalese individually paired with English) in both directions. Additionally, we trained a many-to-one SMT system which was used to translate a sentence from any of these 7 Indic languages to English.

We trained our systems using training data *Indic Languages Multilingual Parallel Corpus* comprised of OpenSubtitles domain data provided by WAT 2018 organizers for shared task experiment. The data-set contains parallel corpora of 7 Indic languages as mentioned above along with their English translation. We did not use any monolingual corpus for this experiment. We extracted the data in individual files for training, tuning and testing. Table 1 shows the statistics of the data split.

.

3.1 Pre-processing

We process the corpus through appropriate filters for normalization, tokenization and truecasing using the scripts available in Moses (Koehn et al., 2007) and the Indic NLP Library[1]. Further, the training sentence length was limited to 50 words.

Following the above preprocessing steps, we generated corpora of all Indian languages transliterated in Devanagari script using the BrahmiNet transliteration system (Kunchukuttan et al., 2015), which is based on the transliteration module in Moses (Durrani et al., 2014). This data was used to train our many-to-one SMT system.

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

Language-pairs	Train	Dev	Test
bn-en	337,428	500	1000
hi-en	84,557	500	1000
ml-en	359,423	500	1000
ta-en	26,217	500	1000
te-en	22,165	500	1000
ur-en	26,619	500	1000
si-en	521,726	500	1000

Table 1: Indic-English corpora (bn- Bengali, hi- Hindi, ml- Malayalam, ta- Tamil, te- Telugu, ur- Urdu, si- Sinhalese individually paired with en-English) split statistics. The number indicates the number of sentences in the split (train, dev an test).

4 Models trained

For all our experiments, we trained the models using the Moses implementation (Koehn et al., 2007) with 3-gram language model and using the *grow-diag-final-and* heuristic for extracting phrases. We trained two types of SMT systems, baseline SMT system and many-to-one SMT system. 14 baseline SMT systems were trained using 7 parallel corpora as shown in Table 2.

In order to train the multilingual SMT system, first, we merged all these corpora into a single bilingual corpus, wherein, sentences of Indic languages were transliterated into Devanagari script. Transliteration is important here to leverage the lexical similarity among Indic languages (Kunchukuttan and Bhattacharyya, 2016). This was followed by the training of translation model with the mentioned 7 Indic-English bilingual corpora.

5 Postprocessing

Output translations of Indic→English language-pairs were detokenized using Moses (Koehn et al., 2007). However, for English→Indic language-pairs we did not perform any postprocessing step.

	Baseline				Many-to-one			
	BLEU	**RIBES**	**AMFM**	**Human**	**BLEU**	**RIBES**	**AMFM**	**Human**
bn→en	14.17	0.689672	0.454990	-	13.98	0.669154	0.447900	-
hi→en	25.57	0.720866	0.599880	0.750	22.45	0.709235	0.558850	5.750
ml→en	11.25	0.566812	0.376260	-	11.51	0.600102	0.365770	-
ta→en	14.34	0.671535	0.511130	29.750	14.09	0.673058	0.487250	29.250
te→en	24.05	0.729178	0.606850	-	22.13	0.714266	0.569170	-
ur→en	18.03	0.630810	0.541890	-	18.31	0.635688	0.519810	-
si→en	16.44	0.692275	0.492410	-	16.92	0.692236	0.484710	-
en→bn	11.34	0.601570	0.532680	-	-	-	-	-
en→hi	26.49	0.692385	0.657180	11.000	-	-	-	-
en→ml	14.23	0.422574	0.567090	-	-	-	-	-
en→ta	15.87	0.668548	0.756890	73.750	-	-	-	-
en→te	21.02	0.728584	0.744230	-	-	-	-	-
en→ur	21.62	0.628279	0.534550	-	-	-	-	-
en→si	11.71	0.580957	0.535950	-	-	-	-	-

Table 2: Translation accuracies of baseline and many-to-one SMT systems. bn- Bengali, hi- Hindi, ml- Malayalam, ta- Tamil, te- Telugu, ur- Urdu, si- Sinhalese individually paired with en-English.

6 Result and Discussion

Table 2 shows translation accuracies of baseline PB-SMT and many-to-one PBSMT in terms of Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002), Rank-based Intuitive Bilingual Evaluation Score (RIBES) (Group and others, 2013), Adequacy-Fluency Metrics (AMFM) (Banchs et al., 2015) and human evaluation score (HUMAN) (only for hi-en and ta-en translation models).

From the results of our experiment, we did not get any discernible improvement in translation quality by using many-to-one PBSMT compared to the baseline PBSMT systems. The many-to-one SMT approach shows minor improvement in BLEU scores for 3 Indic languages namely Malayalam, Urdu and Sindhi, and minor degradation in BLEU scores for 2 Indic languages Bengali and Tamil. A noticeable reduction was observed in BLEU scores for both Hindi and Telugu. However, only a single translation model was required for all language pairs. Thus only a single model needs to be maintained and hosted. Of course, the phrase table for the multilingual model is substantially larger than the individual models (see Table 3 for phrase table size statistics).

Language-pairs	**No. of Phrases in Phrase-table**
XX→en	15,829,552
bn→en	3,897,808
hi→en	931,612
ml→en	3,753,408
ta→en	228,846
te→en	167,668
ur→en	335,617
si→en	6,270,747

Table 3: Number of phrases in phrase-table of many-to-one (XX→en) and bilingual models.

7 Conclusion

In this paper, we have described our submissions to WAT 2018. Our multilingual PBSMT was comparable to each baseline PBSMT model, but we did not observe any major gains. However, this is only an initial study where various models have not been explored. Further investigation can help to understand if many-to-one SMT approach is useful. Many-to-one approach could also be useful for translation of code-mixed sentences.

Acknowledgments

We would like to thank Raj Dabre for his valuable inputs.

References

Dzmitry Bahdanau, KyungHyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *In Proceedings of the International Conference on Learning Representations (ICLR)*.

Rafael E Banchs, Luis F D'Haro, and Haizhou Li. 2015. Adequacy-fluency metrics: Evaluating mt in the continuous space model framework. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 23(3):472–482.

Pushpak Bhattacharyya, Mitesh Khapra, and Anoop Kunchukuttan. 2016. Statistical Machine Translation Between Related Languages. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1723–1732.

Nadir Durrani, Hassan Sajjad, Hieu Hoang, and Philipp Koehn. 2014. Integrating an unsupervised transliteration model into statistical machine translation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 148–153.

Murray B Emeneau. 1956. India as a lingustic area. *Language*, 32(1):3–16.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of NAACL-HLT*, pages 866–875.

Linguistic Intelligence Research Group et al. 2013. Ntt communication science laboratories. ribes: Rank-based intuitive bilingual evaluation score.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association of Computational Linguistics*, 5(1):339–351.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39.

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 and Pushpak Bhattacharyya. 2016. Orthographic syllable as basic unit for smt between related languages. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1912–1917.

Anoop Kunchukuttan, Ratish Puduppully, and Pushpak Bhattacharyya. 2015. Brahmi-net: A transliteration and script conversion system for languages of the indian subcontinent. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 81–85.

Toshiaki Nakazawa, Shohei Higashiyama, Chenchen Ding, Raj Dabre, Anoop Kunchukuttan, Win Pa Pa, Isao Goto, Hideya Mino, Katsuhito Sudoh, and Sadao Kurohashi. 2018. Overview of the 5th workshop on asian translation. In *Proceedings of the 5th Workshop on Asian Translation (WAT2018)*, Hong Kong, China, December.

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.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.