IITP-MT at WAT2018: Transformer-based Multilingual Indic-English Neural Machine Translation System

Sukanta Sen, Kamal Kumar Gupta, Asif Ekbal, Pushpak Bhattacharyya
Department of Computer Science and Engineering
Indian Institute of Technology Patna
{sukanta.pcs15,kamal.pcs17,asif,pb}@iitp.ac.in

Abstract

This paper describes the systems submitted by the IITP-MT team to WAT 2018 multilingual Indic languages shared task. We submit two multilingual neural machine translation (NMT) systems (Indic-to-English and English-to-Indic) based on Transformer architecture and our approaches are similar to many-to-one and one-to-many approaches of Johnson et al. (2017). We also train separate bilingual models as baselines for all translation directions involving English. We evaluate the models using BLEU score and find that a single multilingual NMT model performs better (up to 14.81 BLEU) than separate bilingual models when the target is English. However, when English is the source language, multilingual NMT model improves only for low-resource language pairs (up to 11.60 BLEU) and degrades for relatively high-resource language pairs over separate bilingual models.

1 Introduction

In this paper, we describe our submission to multilingual Indic languages shared task at 5th Workshop on Asian Translation (WAT 2018) (Nakazawa et al., 2018). This task covers 7 Indic languages (Bengali, Hindi, Malayalam, Tamil, Telugu, Sinhalese and Urdu) and English. The objective of this shared task is to build translation models for XX-EN language pairs. By XX, we denote the set of 7 Indic languages. In this task, we submit two (single models for Indic-to-English and English-to-Indic) multilingual neural machine translation systems to translate between Indic languages and English. Unlike the European languages, most of the Indian languages do not have enough-sized parallel English translations. The parallel corpora used in this shared task have 22k to 521k parallel sentences (see Table 1), which is insufficient for NMT training. NMT is a data hungry approach and it is not possible to have sufficient amount of parallel training data for all language pairs. So building multilingual translation model by means of sharing parameters with high-resource languages is a common practice to improve the performance of low-resource language pairs. Sharing of parameters between low-resource and high-resource language pairs helps low-resource pairs to learn better model compared to model trained separately. However, it has been seen that training multiple languages together sometimes degrades the performance of some language pairs compared to a separate single bilingual model as languages may have different linguistic properties.

Recent success of end-to-end bilingual NMT systems (Kalchbrenner and Blunsom, 2013; Cho et al., 2014; Sutskever et al., 2014; Bahdanau et al., 2015) quickly gave the rise of multilingual NMT in various ways (Dong et al., 2015; Firat et al., 2016; Johnson et al., 2017). Most of the existing multilingual NMT involve non-Indic languages and are based on attentional encoder-decoder approach. We use the Transformer architecture (Vaswani et al., 2017) with subword (Sennrich et al., 2016) as basic translation unit. We develop two multilingual translation models: one is for XX→EN (7 Indic languages to English) and another is for EN→XX (English to 7 Indic languages). We also train separate bilingual

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

model as a baseline for each translation direction involving English. We evaluate the multilingual models against the bilingual models using BLEU (Papineni et al., 2002) metric. We found that multilingual NMT is better than bilingual models for all XX→EN directions, however for EN→XX directions, multilingual NMT performs better than bilingual NMT for low-resource language pairs only.

In the next section, we briefly mention some notable multilingual NMT works. We describe our submitted systems in section 3 which includes description on datasets, preprocessing, experimental setup. Results are described in section 4. Finally, the work is concluded in section 5.

2 Related Works

Dong et al. (2015) implemented a system with one-to-many mapping of languages. They translated a source language to multiple target languages where each target language decoder deals with its own attention network. Firat et al. (2016) used a single attentional network that was shared among all source-target language pairs. They used separate encoder decoder for each source and target language. Thus, the number of parameters increases as the number of language increases. Johnson et al. (2017) came up with a simple but effective approach for multilingual translation. They mixed all parallel data and trained a standard attentional encoder-decoder NMT model without any change. They used an additional token before each source sentence to specify its target language. We apply this simple approach of combining training data and then we train transformer based NMT models for building multilingual translation systems (many-to-one and one-to-many) for Indic languages.

3 System Description

In this section, we describe datasets, preprocessing of data and experimental setup of our systems.

3.1 Datasets

We use the Indic Languages Multilingual Parallel Corpus[1] consisting of the following languages: Bengali, Hindi, Malayalam, Tamil, Telugu, Sinhalese, Urdu and English. It contains 7 parallel corpora for 7 Indic languages (translated into English), and 8 monolingual corpora. These corpora have been collected from OPUS[2] and belongs to the spoken language (OpenSubtitles) domain. For experiments, we use parallel corpora only. Training data size is presented in Table 1. For each language pair, development set and test set have 500 and 1,000 parallel sentences, respectively. Before feeding the data for training, we tokenize, truecase, subword the original corpora as preprocessing. We tokeninze English data using Moses tokenizer[3] and the Indic_NLP library[4] tool is used for tokenizing Indic language data. Tokenized English sentences are truecased using Moses truecaser script. There is no need to truecase Indic languages as they are case-insensitive.

Language Pair	#Sentences
Bengali (BN) - English	337,428
Hindi (HI) - English	84,557
Malayalam (ML) - English	359,423
Tamil (TA) - English	26,217
Telugu (TE) - English	22,165
Urdu (UR) - English	26,619
Sinhalese (SI) - English	521,726

Table 1: Training data size for each language pair.

3.2 Subword Unit

NMT works with fixed vocabulary size. To deal with large vocabulary of Indic languages, we subword each bilingual corpora independently. Sennrich et al. (2016) introduced Byte-pair-encoding (BPE) based subword unit for dealing with rare words problem in NMT. It helps to decrease the vocabulary size and to deal with unseen tokens at training and test time. With variable size of training data (see Table 1) and morphological variations among languages, vocabulary size for each language is also different. Original vocabulary size, number of BPE merge, and vocabulary size after applying BPE are shown in Table 2.

[1] http://lotus.kuee.kyoto-u.ac.jp/WAT/indic-multilingual/indic_languages_corpus.tar.gz

[2] http://opus.nlpl.eu

[3] https://github.com/moses-smt/mosesdecoder/blob/RELEASE-3.0/scripts/tokenizer/tokenizer.perl

[4] https://bitbucket.org/anoopk/indic_nlp_library

Data Pair	Source			Target		
	Original Vocab	Merge	Final Vocab	Original Vocab	Merge	Final Vocab
BN-EN	90,482	8,000	8,394	56,498	5,000	5,248
HI-EN	24,470	4,000	4,286	24,380	4,000	4,150
ML-EN	253,360	10,000	10,351	58,320	5,000	5,273
SI-EN	169,603	9,000	9,392	72,093	7,000	7,417
TA-EN	18,723	3,500	3,675	18,723	2,000	2,114
TE-EN	12,728	2,000	2,230	9,929	1,500	1,633
UR-EN	13,581	3,000	3,268	12,854	2,000	2,126

Table 2: Original vocabulary size, number of BPE merge and final vocabulary size after applying BPE for each training data pair. We decided the BPE merge values without any rigorous exploration.

3.3 Experimental Setup

We train 2 multilingual models namely XX→EN (Indic languages to English) and EN→XX (Englsih to Indic languages) and 14 bilingual models (7 for Indic languages to English, and 7 for English to Indic languages). All of these models are based on Transformer (Vaswani et al., 2017) network. For training the models, we use Sockeye (Hieber et al., 2017), a toolkit for NMT. Each token in training, development and test sets are split in subword units in preprocessing stage. Along with that an additional token[5] indicating which Indic language a sentence pair belong to is added at the beginning of every source[6] sentence. Then parallel data of all pairs are appended in one parallel corpus with Indic languages in one side and English on other side, for training a single multilingual model for each of EN→XX and XX→EN directions. These tokens are added with development and test sets too and likewise, development sets are also appended in a single development set. We set embedding dimension of 512, hidden dimension of 512, learning rate of 0.0002, dropout rate of 0.2. We use Adam (Kingma and Ba, 2015) optimizer. We keep mini-batch size of 2000 words[7], and maximum sentence length is restricted to 50. Rest of the hyperparameters are set to the default values of Sockeye. Training is completed on meeting early-stopping criteria (BLEU based, 10 patience) on development set. Finally, the best model is used for translating the test sets.

System			Bi	Multi	▲
BN			18.24	20.05	+1.81
HI			27.11	32.95	+5.84
ML			10.56	19.94	+9.38
SI	→	**EN**	18.22	21.35	+3.13
TA			11.58	22.42	+10.84
TE			16.15	30.96	+14.81
UR			20.02	26.56	+6.54
		BN	13.38	13.27	-0.11
		HI	24.25	26.60	+2.35
		ML	20.92	13.50	-7.42
EN	→	**SI**	12.75	10.64	-2.11
		TA	11.88	18.81	+6.93
		TE	14.21	25.81	+11.60
		UR	18.73	21.48	+2.75

Table 3: BLEU scores of our {BN, HI, ML, SI, TA, TE, UR}→EN and EN→{BN, HI, ML, SI, TA, TE, UR} systems; Bi: Bilingual Model; Multi: Multilingual Model; ▲ denotes improvement of multilingual model over bilingual model.

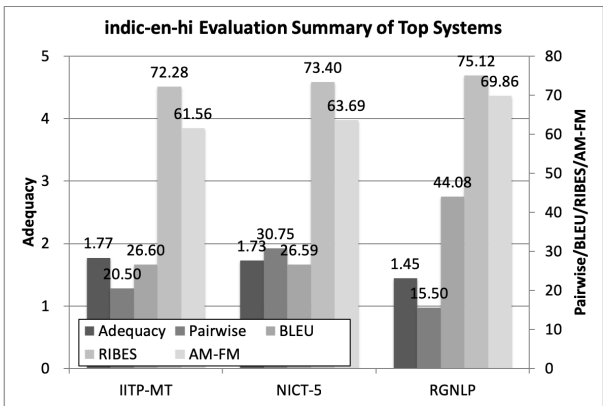
4 Results

BLEU scores of bilingual and Multilingual systems are shown in Table 3. For Indic languages to English (XX→EN), BLEU score increases in each pair of multilingual system compared to bilingual system of that pair. Here, at target side decoder has

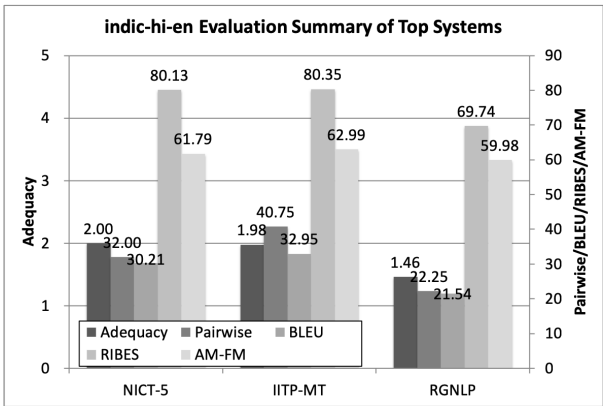
[5]We use the followings tokens: BN##, HI##, ML##, SI##, TA##, TE##, UR##

[6]Source can be either English or any Indic language depending on translation direction.

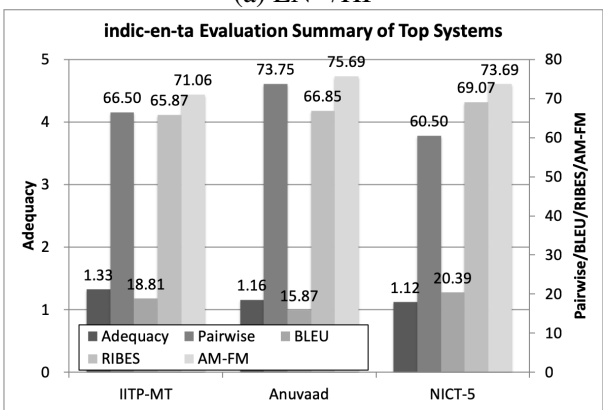
[7]Sockeye supports word based batching too.



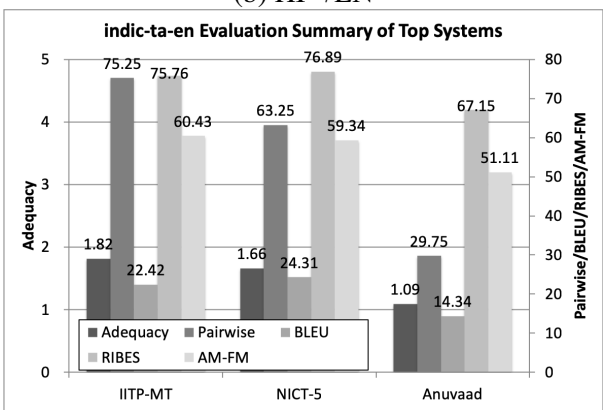
(a) EN→HI



(b) HI→EN



(c) EN→TA



(d) TA→EN

Figure 1: Official bar charts showing Adequacy, Pairwise Human Evaluation, BLEU, RIBES and AM-FM scores of top systems for multilingual Indic languages shared task at WAT 2018.

to deal with only one language i.e. English (EN). It shows sharing of parameters improves the performance of high-resource language pairs ({BN, ML, SI}→EN) as well as low-resource language pairs ({HI, TA, TE, UR}→EN). Unlike XX→EN, BLEU scores of EN→XX improve only for low-resource language pairs (EN→{HI, TA, TE, UR}) at the cost of BLEU scores of high-resource language pairs (EN→{BN, ML, SI}). For EN→XX multilingual system, a single decoder has to deal with multiple languages with different vocabulary and different linguistic features; that is why it is difficult for a single decoder to handle information of each target language.

For multilingual shared task, the official Pairwise Human evaluation and Adequacy scores (Nakazawa et al., 2018) were released for four translation directions only: EN→HI, HI→EN, EN→TA and TA→EN. Figure 1 shows the comparison of Pairwise Human evaluation and Adequacy scores along with BLEU, RIBES (Isozaki et al., 2010), AM-FM (Banchs et al., 2015) scores of top three systems for each of the four translation directions.

5 Conclusion

In this paper, we described our submission to WAT 2018 multilingual Indic languages shared task. We submitted two multilingual NMT models: many-to-one (7 Indic languages to English) and one-to-many (English to 7 Indic languages). Our multilingual NMT is based on Transformer architecture. We evaluated our models using BLEU score and found that multilingual NMT performs better than separately trained bilingual NMT models when the target side has only one language (English) and the improvement is higher for low-resource languages (up to 14.81 BLEU points). However, performance of multilingual NMT degrades compared to bilingual mod-

els for the relatively high-resource languages when the target has many languages.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *International Conference on Learning Representation(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.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111.

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)*, pages 1723–1732. Association for Computational Linguistics.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875. Association for Computational Linguistics.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: A toolkit for neural machine translation. *arXiv preprint arXiv:1712.05690*.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952. Association for Computational Linguistics.

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

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *International Conference on Learning Representation(ICLR)*.

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, Philadelphia, Pennsylvania.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725. 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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.