Multilingual Machine Translation Systems at WAT 2021: One-to-Many and Many-to-One Transformer based NMT

Shivam Mhaskar, Aditya Jain, Aakash Banerjee, Pushpak Bhattacharyya
Department of Computer Science and Engineering
Indian Institute of Technology Bombay
Mumbai, India
{shivammhaskar, adityajainiitb, abanerjee, pb}@cse.iitb.ac.in

Abstract

In this paper, we present the details of the systems that we have submitted for the WAT 2021 MultiIndicMT: An Indic Language Multilingual Task. We have submitted two separate multilingual NMT models: one for English to 10 Indic languages and another for 10 Indic languages to English. We discuss the implementation details of two separate multilingual NMT approaches, namely one-to-many and many-to-one, that makes use of a shared decoder and a shared encoder, respectively. From our experiments, we observe that the multilingual NMT systems outperforms the bilingual baseline MT systems for each of the language pairs under consideration.

1 Introduction

In recent years, the Neural Machine Translation (NMT) systems (Vaswani et al., 2017; Bahdanau et al., 2014; Sutskever et al., 2014; Cho et al., 2014) have consistently outperformed the Statistical Machine Translation (SMT) (Koehn, 2009) systems. One of the major problems with NMT systems is that they are ***data hungry***, which means that they require a large amount of parallel data to give better performance. This becomes a very challenging task while working with low-resource language pairs for which a very less amount of parallel corpora is available. Multilingual NMT (MNMT) systems (Dong et al., 2015; Johnson et al., 2017) alleviate this issue by using the phenomenon of transfer learning among related languages, which are the languages that are related by genetic and contact relationships. (Kunchukuttan and Bhattacharyya, 2020) have shown that the lexical and orthographic similarity among languages can be utilized to improve translation quality between Indic languages when limited parallel corpora is available. Another advantage of using MNMT systems is that they support zero-shot translation, that is, translation among two languages for which no parallel corpora is available during training.

A MNMT system can also drastically reduce the total number of models required for a large scale translation system by making use of a single many-to-many MNMT model instead of having to train a separate translation system for each of the language pairs. This reduces the amount of computation and time required for training. Among various MNMT approaches, using a single shared encoder and decoder will further reduce the number of parameters and allow related languages to share vocabulary. In this paper, we describe the two MNMT systems that we have submitted for the WAT 2021 MultiIndicMT: An Indic Language Multilingual Task (Nakazawa et al., 2021) as team 'CFILT', namely one-to-many for English to Indic languages and many-to-one for Indic languages to English. This task covers 10 Indic languages which are Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Oriya, Punjabi, Tamil and Telugu.

2 Related Work

Dong et al. (2015) was the first to introduce MNMT. The authors used a one-to-many model where a separate decoder and an attention mechanism was used for each target language. Firat et al. (2016) extended this to a many-to-many setting using a shared attention mechanism. In Zoph and Knight (2016) a multi-source translation approach was proposed where multiple encoders were used, each having a separate attention mechanism. Lee et al. (2017) proposed a CNN-based character level approach where a single encoder was shared across all the source languages.

A second line of work on MNMT uses a single shared encoder and decoder (Ha et al., 2016; Johnson et al., 2017) irrespective of the number of languages on the source or the target side. An

Proceedings of the 8th Workshop on Asian Translation, pages 233–237

	en-bn	en-gu	en-hi	en-kn	en-ml	en-mr	en-or	en-pa	en-ta	en-te
ALT	20	-	20	-	-	-	-	-	-	-
Bible-uedin	-	16	62	61	61	61	-	-	-	62
CVIT-PIB	92	58	267	-	43	114	94	101	116	45
IITB 3.0	-	-	1603	-	-	-	-	-	-	-
MTEnglish2Odia	-	-	-	-	-	-	35	-	-	-
NLPC	-	-	-	-	-	-	-	-	31	-
OdiEnCorp 2.0	-	-	-	-	-	-	91	-	-	-
OpenSubtitles	411	-	92	-	383	-	-	-	32	27
PMIndia	23	41	50	29	27	29	32	28	33	33
TED2020	-	-	-	2	-	-	-	0.7	-	-
Total	546	115	2094	92	514	204	252	130	212	167

Table 1: Statistics of number of parallel sentences for each of the English-Indic language pairs across different datasets used for **training**. All the numbers are in thousands. (bn:Bengali, gu:Gujarati, hi:Hindi, kn:Kannada, ml:Malayalam, mr:Marathi, or:Oriya, pa:Punjabi, ta:Tamil, te:Telugu)

advantage of this approach is that the number of parameters are drastically reduced. Dabre et al. (2019) gives a summary of various techniques that can be used to implement MNMT systems. The MNMT systems that we have implemented are based on Johnson et al. (2017)'s approach where in one-to-many and many-to-many models a language specific token is prepended to the input sentence to indicate the target language that the model should translate to. We use transformer (Vaswani et al., 2017) architecture which has proven to give superior performance over the RNN based models (Bahdanau et al., 2014; Sutskever et al., 2014; Cho et al., 2014).

3 Our Approach

The various types of multilingual models that we have implemented are one-to-many and many-to-one, each of which are discussed below.

3.1 One-to-Many

In a one-to-many multilingual model, the translation task involves a single source language and two or more target languages. One of the ways to achieve this is by making use of a single encoder for the source language and separate decoders for each of the target languages. The disadvantage of this method is that, as there are multiple decoders, the size of the model increases. Another way to achieve this is to use a single encoder and a single shared decoder. An advantage of this method is that the representations learnt by some language pair can further be utilized by the some other language pair. For example, the representations learnt during the training of the English-Hindi language pair can help training the English-Marathi language pair. Also, in this approach, a language specific token is prepended to the input sentence to indicate the model to which target language the input sentence should be translated.

3.2 Many-to-One

This approach is similar to the one-to-many approach. The major point of difference is that there are multiple source languages and a single target language. As a result, here we use a single shared encoder and a single decoder. Also, as the target language is same for all the source languages, it is optional to prepend a token to the input sentence unlike in the one-to-many approach which has multiple target languages for a given source language.

4 Experiments

In this section, we discuss the details of the system architecture, dataset, preprocessing, models and the training setup.

4.1 System Architecture

Table 4 lists the details of the transformer architecture used for all the experiments.

4.2 Data

The dataset provided for the shared task by WAT 2021 was used for all the experiments. We did not use any additional data to train the models. Table 1 lists the datasets used for each of the English-Indic

	Baseline		One-to-Many		
	BLEU	**RIBES**	**BLEU**	**RIBES**	**AMFM**
en → bn	12.14	0.691941	13.24	0.710664	0.777074
en → gu	18.26	0.745845	24.56	0.806649	0.817681
en → hi	33.06	0.836683	35.39	0.843969	0.821713
en → kn	11.43	0.666605	17.98	0.747233	0.816981
en → ml	10.56	0.668024	12.79	0.707437	0.805291
en → mr	-	-	18.47	0.759182	0.811499
en → or	11.19	0.644931	18.22	0.738397	0.768399
en → pa	29.00	0.810395	31.16	0.826367	0.813658
en → ta	10.97	0.662236	12.99	0.715699	0.802920
en → te	-	-	15.52	0.725496	0.789820

Table 2: Results for the one-to-many MNMT model. To obtain the baseline results, we performed the same automatic evaluation procedures as those performed in WAT 2021. The one-to-many results are the official evaluation results provided by the organizers of WAT 2021. (bn:Bengali, gu:Gujarati, hi:Hindi, kn:Kannada, ml:Malayalam, mr:Marathi, or:Oriya, pa:Punjabi, ta:Tamil, te:Telugu)

	Baseline		Many-to-One		
	BLEU	**RIBES**	**BLEU**	**RIBES**	**AMFM**
bn → en	24.38	0.772800	25.98	0.760268	0.766461
gu → en	31.92	0.799512	35.31	0.807849	0.797069
hi → en	37.72	0.847265	39.71	0.837668	0.822034
kn → en	21.30	0.738755	30.23	0.772913	0.778602
ml → en	26.80	0.786290	29.28	0.784424	0.789095
mr → en	-	-	29.71	0.786570	0.789075
or → en	-	-	30.46	0.772850	0.793769
pa → en	37.89	0.827826	38.01	0.818396	0.804561
ta → en	-	-	29.34	0.784291	0.785098
te → en	-	-	30.10	0.778981	0.783349

Table 3: Results for the many-to-one MNMT model. To obtain the baseline results, we performed the same automatic evaluation procedures as those performed in WAT 2021. The many-to-one results are the official evaluation results provided by the organizers of WAT 2021.(bn:Bengali, gu:Gujarati, hi:Hindi, kn:Kannada, ml:Malayalam, mr:Marathi, or:Oriya, pa:Punjabi, ta:Tamil, te:Telugu)

language pairs along with the number of parallel sentences. The validation and test sets have 1,000 and 2,390 sentences, respectively and are 11-way parallel.

4.3 Preprocessing

We used Byte Pair Encoding (BPE) (Sennrich et al., 2016) technique for data segmentation, that is, break up the words into sub-words. This technique is especially helpful for Indic languages as they are morphologically rich. Separate vocabularies are used for the source and target side languages. For training the one-to-many and many-to-one models, the data of all the 10 Indic languages is combined before learning the BPE codes. 48000, 48000 and 8000 merge operations are used for learning the BPE codes of the one-to-many, many-to-one and bilingual baseline models, respectively.

4.4 Baseline Models

The baseline MT models are bilingual MT models based on the vanilla transformer architecture. We have trained 20 separate bilingual MT models, 10 for English to each Indic language and 10 for each Indic language to English.

4.5 Models and Training

For this task, we built two separate MNMT systems, a one (English) to many (10 Indic languages) model and a many (10 Indic languages) to one (English)

	Encoder	Decoder
No. of layers	6	6
No. of attention heads	8	8
Embedding dimensions	512	512
FFNN hidden layer dim	2048	2048

Table 4: System architecture details

model. In our one-to-many model, we used the transformer architecture with a single encoder and a single shared decoder. The encoder used the English vocabulary and the decoder used a shared vocabulary of all the Indic languages. In our many-to-one model, we used the transformer architecture with a single shared encoder and a single decoder. Here the encoder used a shared vocabulary of all the Indic languages and English vocabulary is used for the decoder. In both of these MNMT models, we prepended a language specific token to the input sentence.

We used the fairseq (Ott et al., 2019) library for implementing the multilingual systems. For training, we used Adam optimizer with betas '(0.9,0.98)'. The initial learning rate used was 0.0005 and the inverse square root learning rate scheduler was used with 4000 warm-up updates. The dropout probability value used was 0.3 and the criterion used was label smoothed cross entropy with label smoothing of 0.1. We used an update frequency, that is, after how many batches the backward pass is performed, of 8 for the multilingual models and 4 for the bilingual baseline models.

During decoding we used the beam search algorithm with a beam length of 5 and length penalty of 1. The many-to-one model was trained for 160 epochs and the one-to-many model was trained for 145 epochs. The model with the best average BLEU score was chosen as the best model. The average BLEU score for a MNMT model was calculated by taking the average of the BLEU scores obtained across all the language pairs.

5 Results and Analysis

The Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002) metric, the Rank-based Intuitive Bilingual Evaluation Score (RIBES) (Isozaki et al., 2010) metric and Adequacy-Fluency Metrics (AMFM) (Banchs et al., 2015) are used to report the results. Table 2 and 3 lists the results for all our experiments.

The baseline results are obtained by training bilingual models and then we have used automatic evaluation procedures same as those performed in WAT 2021. The one-to-many and many-to-one results are those reported by WAT 2021 on our submitted translation files.

We observe that for all language pairs in both the translation directions, the MNMT models give superior performance as compared to the bilingual NMT models. For relatively high resource language pairs like English-Hindi and English-Bengali the increase in BLEU score is less while for relatively low resource language pairs like English-Kannada and English-Oriya the increase in BLEU score is substantial. From the above observation it follows that low resource language pairs benefit much more from multilingual training than high resource language pairs. An increase of up to 8.93 BLEU scores (for Kannada to English) is observed using MNMT systems over the bilingual baseline NMT systems.

6 Conclusion

In this paper, we have discussed our submission to the WAT 2021 MultiIndicMT: An Indic Language Multilingual Task. We have submitted two separate MNMT models: a one-to-many (English to 10 Indic languages) model and a many-to-one (10 Indic languages to English) model. We evaluated our models using BLEU and RIBES scores and observed that the MNMT models outperform the separately trained bilingual NMT models across all the language pairs. We also observe that for the lower resource language pairs the improvement in performance is much more as compared to that for the higher resource language pairs.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

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*, 23(3):472–482.

KyungHyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *CoRR*, abs/1409.1259.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2019. A survey of multilingual neural machine translation. *CoRR*, abs/1905.05395.

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, Beijing, China. 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, San Diego, California. Association for Computational Linguistics.

Thanh-Le Ha, Jan Niehues, and Alexander H. Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. *CoRR*, abs/1611.04798.

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, Cambridge, MA. 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.

Philipp Koehn. 2009. *Statistical machine translation*. Cambridge University Press.

Anoop Kunchukuttan and Pushpak Bhattacharyya. 2020. Utilizing language relatedness to improve machine translation: A case study on languages of the indian subcontinent. *arXiv preprint arXiv:2003.08925*.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2017. Fully character-level neural machine translation without explicit segmentation. *Transactions of the Association for Computational Linguistics*, 5:365–378.

Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, and Sadao Oda, Yusuke Kurohashi. 2021. Overview of the 8th workshop on Asian translation. In *Proceedings of the 8th Workshop on Asian Translation*, Bangkok, Thailand. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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, Berlin, Germany. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 30–34, San Diego, California. Association for Computational Linguistics.