Investigation of Multilingual Neural Machine Translation for Indian Languages

Sahinur Rahman Laskar[1], **Riyanka Manna**[2]
Partha Pakray[1], **Sivaji Bandyopadhyay**[1]
[1]Department of Computer Science and Engineering, National Institute of Technology, Silchar, India
[2]Department of Computer Science and Engineering, Adamas University, Kolkata, India
{sahinurlaskar.nits, riyankamanna16}@gmail.com
{parthapakray, sivaji.cse.ju}@gmail.com

Abstract

In the domain of natural language processing, machine translation is a well-defined task where one natural language is automatically translated to another natural language. The deep learning-based approach of machine translation, known as neural machine translation attains remarkable translational performance. However, it requires a sufficient amount of training data which is a critical issue for low-resource pair translation. To handle the data scarcity problem, the multilingual concept has been investigated in neural machine translation in different settings like many-to-one and one-to-many translation. WAT2022 (Workshop on Asian Translation 2022) organizes (hosted by the COLING 2022) Indic tasks: English-to-Indic and Indic-to-English translation tasks where we have participated as a team named CNLP-NITS-PP. Herein, we have investigated a transliteration-based approach, where Indic languages are transliterated into English script and shared sub-word level vocabulary during the training phase. We have attained BLEU scores of 2.0 (English-to-Bengali), 1.10 (English-to-Assamese), 4.50 (Bengali-to-English), and 3.50 (Assamese-to-English) translation, respectively.

1 Introduction

Due to the advancement of deep learning techniques, neural machine translation (NMT) attains remarkable progress for single pairs translation with a large amount of bilingual corpus (Bahdanau et al., 2015; Luong et al., 2015; Vaswani et al., 2017). Moreover, NMT shows good translational performance for low-resource Indian languages (Pathak and Pakray, 2018; Pathak et al., 2018; Laskar et al., 2019a,b, 2020, 2021b,a, 2022). Recent years, researchers have been investigating multilingual NMT from various aspects, zero-shot, pivot-based, and different settings, namely, many-to-one, one-to-many, many-to-many (Johnson et al., 2017; Tan et al., 2019). In (Ramesh et al., 2022), authors developed *Samanantar*, a parallel dataset for 11 Indian languages. They converted all Indic data into a common Devanagari script and took the advantage of lexical sharing at the sub-word level for transfer learning during the training process. They explored multilingual NMT models for English-to-Indic and vice-versa by considering language tags for indicating Indic languages on the source side (Johnson et al., 2017). Similarly, we have investigated multilingual NMT in the Indic tasks of WAT2022. The difference is that instead of converting all Indic data into a common Devanagari script, we have converted all Indic data into English script and attempted to take the benefits of lexical sharing at the sub-word level for both source and target languages.

The rest of the paper is organized as follows: Section 2 presents the review of related works. The system description is briefly discussed in Section 3. Section 4 reports the results and Section 5 concludes the paper with future scope.

2 Related Work

The literature survey finds out very limited work on multilingual NMT, specifically, for English-to-Indic and Indic-to-English translation (Ramesh et al., 2022). They contributed *Samanantar* dataset which comprises parallel corpora of 11 Indic languages with English side parallel sentences and explored the multilingual NMT model for English-to-Indic and Indic-to-English. They used Fairseq (Ott et al., 2019) toolkit for transformer-based model training via multilingual settings of many-to-one and one-to-many (Johnson et al., 2017).

3 System Description

We have employed the OpenNMT-py (Klein et al., 2017) toolkit to build multilingual transformer-based NMT models for English-to-Indic and Indic-to-English translation. We have used parallel corpora provided by the

Proceedings of the 9th Workshop on Asian Translation, pages 78–81
October 17, 2022.

WAT2022 organizers (Nakazawa et al., 2022). Additionally, we have used English-Assamese parallel corpus (Laskar et al., 2020). We have maintained the equal ratio (1 : 1) for Eng-Indic (Asm/Ben/Guj/Hin/Kan/Mal/Mar/Tel/Tam/Pan/Npi/ Ory) language pairs of the dataset in the multilingual NMT settings and data statistics are presented in Table 1. We have converted all Indic data into English script using the Indic-trans, transliteration script[1] (Bhat et al., 2014). We have performed jointly byte pair encoding (sub-word level) (Sennrich et al., 2016) on the transliterated Indic sentences and English sentences with $40k$ merge operations. The sub-word level source-target vocabulary is shared during the training process of the multilingual NMT model. We have used special tokens (language tags) for Indic side languages at the one-to-many (English-to-Indic) setting (Johnson et al., 2017). We have followed the default settings of the 6 layer transformer model (Vaswani et al., 2017) in the training process. The NMT model is trained on a single GPU with early stopping criteria i.e., the model training is halted if does not converge on the validation set for more than 10 epochs. The obtained trained model is used to translate the test data provided by the WAT2022 organizers. For English-to-Indic language translation, the predicted sentences are converted into the respective Indic languages using the Indic-trans script.

4 Results

The WAT2022 shared task organizer (Nakazawa et al., 2022) published the evaluation result[2] (INDIC22en-as/INDIC22as-en/INDIC22en-bn/INDIC22bn-en) at the Indic translation task for English-to-Indic and Indic-to-English and our team achieve the second position for English-to-Assamese and vice-versa translation. We have participated with a team name CNLP-NITS-PP in the English-Assamese and English-Bengali submission tracks of the same task where a total of two teams participated. The automatic evaluation metrics, BLEU (Papineni et al., 2002), RIBES (Isozaki et al., 2010) are used for evaluation of results. Table 2 presents the results of our system. The quantitative results show that our investigation of the transliteration Indic languages into English script does not provide a reasonable translation accuracy for the multilingual NMT model of English-Assamese and English-Bengali pairs translation.

Translation	BLEU	RIBES
Eng-to-Asm	1.10	0.359265
Asm-to-Eng	3.50	0.537859
Eng-to-Ben	2.00	0.503286
Ben-to-Eng	4.50	0.547407

Table 2: Our system's results (official) for Eng-Asm (English-Assamese) and Eng-Ben (English-Bengali) language pair at the Indic task.

5 Conclusion and Future Work

In this work, we have investigated multilingual NMT for Indic task of WAT2022 by taking the advantage of sub-word level source-target lexical sharing during the training. However, we need to do more experiments to improve the translational performance of low-resource pairs by utilizing pretrained multilingual models.

Acknowledgements

We want to thank the Department of Computer Science and Engineering, Center for Natural Language Processing (CNLP), Artificial Intelligence (AI) Lab at the National Institute of Technology, Silchar for providing the requisite support and infrastructure to execute this work.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Irshad Ahmad Bhat, Vandan Mujadia, Aniruddha Tammewar, Riyaz Ahmad Bhat, and Manish Shrivastava. 2014. Iiit-h system submission for fire2014 shared task on transliterated search. In *Proceedings of the Forum for Information Retrieval Evaluation*, FIRE '14, page 48–53, New York, NY, USA. Association for Computing Machinery.

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.

[1] https://github.com/libindic/indic-trans

[2] http://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/index.html

Language Pair	Type	Source	No. of Sentence
Eng-Asm	Train	Organizer	140172
		External	203315
		Total	343487
	Validation	Organizer	997
	Test	Organizer	1012
Eng-Ben/Guj/Hin/Kan/Mal/Mar/Tel/Tam/Pan/Npi/Ory	Train	Organizer	350000
	Validation	Organizer	997
	Test	Organizer	1012

Table 1: Data statistics of train, validation and test set. External: Taken permission from the organizer to use external parallel English-Assamese data (Laskar et al., 2020) (EnAsCorp1.0 has been updated and the updated version will be released at `https://github.com/cnlp-nits/EnAsCorp1.0`)

.

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.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Sahinur Rahman Laskar, Abinash Dutta, Partha Pakray, and Sivaji Bandyopadhyay. 2019a. Neural machine translation: English to hindi. In *2019 IEEE Conference on Information and Communication Technology*, pages 1–6.

Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji, Partha Pakray, and Sivaji Bandyopadhyay. 2020. EnAsCorp1.0: English-Assamese corpus. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 62–68, Suzhou, China. Association for Computational Linguistics.

Sahinur Rahman Laskar, Partha Pakray, and Sivaji Bandyopadhyay. 2019b. Neural machine translation: Hindi-Nepali. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 202–207, Florence, Italy. Association for Computational Linguistics.

Sahinur Rahman Laskar, Partha Pakray, and Sivaji Bandyopadhyay. 2021a. Neural machine translation: Assamese–bengali. In *Modeling, Simulation and Optimization: Proceedings of CoMSO 2020*, pages 571–579. Springer Singapore.

Sahinur Rahman Laskar, Partha Pakray, and Sivaji Bandyopadhyay. 2021b. Neural machine translation for low resource assamese–english. In *Proceedings of the International Conference on Computing and Communication Systems: I3CS 2020, NEHU, Shillong, India*, volume 170, page 35. Springer.

Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji, Partha Pakray, and Sivaji Bandyopadhyay. 2022. Improved neural machine translation for low-resource english–assamese pair. *Journal of Intelligent and Fuzzy Systems*, 42(5):4727–4738.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Toshiaki Nakazawa, Hideya Mino, Isao Goto, Raj Dabre, Shohei Higashiyama, Shantipriya Parida, Anoop Kunchukuttan, Makoto Morishita, Ondřej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, Yusuke Oda, and Sadao Kurohashi. 2022. Overview of the 9th workshop on Asian translation. In *Proceedings of the 9th Workshop on Asian Translation (WAT2022)*, Gyeongju, Republic of Korea. 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 the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

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*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Amarnath Pathak and Partha Pakray. 2018. Neural machine translation for indian languages. *Journal of Intelligent Systems*, pages 1–13.

Amarnath Pathak, Partha Pakray, and Jereemi Bentham. 2018. English–mizo machine translation using neural and statistical approaches. *Neural Computing and Applications*, 30:1–17.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Deepak Kumar, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *Trans. Assoc. Comput. Linguistics*, 10:145–162.

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.

Xu Tan, Yi Ren, Di He, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2019. Multilingual neural machine translation with knowledge distillation. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.