IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2024

DOI: 10.17148/IJARCCE.2024.13398

Multilingual NMT system for English to Low Resource Indic Languages - Assamese and Bengali

Kishore Kashyap[1], Shikhar Kumar Sarma[2]

Department of Information Technology, Gauhati University, Guwahati, India[1]

Department of Information Technology, Gauhati University, Guwahati, India[2]

Abstract: Neural Machine Translation has surpassed many limitations of rule-based and statistical machine translation systems and is the current state-of-the-art. Though the success of Neural Machine Translation is indisputable, still many improvements are awaited when it comes to expecting the same level of quality for translation to/from low resource languages. In this work, we tried to develop a One-To-Many Multilingual Neural Machine Translation system, which is capable of translating text from English Language to two low resource Indic languages, viz., Assamese, Bengali. We used publicly available parallel corpus. Along with the public corpus, we also used synthetic data for Assamese as the target side. We got better results in terms of BLEU, chef and TER for English to Bengali and direction English to Assamese translation direction in multilingual settings as compared to their bilingual NMT counterparts. In this paper, we have shown that both multilingualism and use of synthetic data can enhance the translation quality of languages where gold standard parallel data is very low.

Keywords: Low resource language MNMT, Multilingual Neural Machine Translation, Indian languages MT, Indic NLP, Assamese NMT, Bengali NMT.

I. INTRODUCTION

Neural Machine Translation (NMT) has revolutionized the Machine Translation (MT) domain and has witnessed impressive growth and innovation in recent years. With the introduction of Transformer architecture [1], the research on MT has shifted from statistical techniques to neural techniques. Though NMT is proven to be better in translating high resource languages, it is still a less explored area of research where many are trying to figure out a breakthrough in achieving state-of-the-art results of translation quality where the languages under study are considered as low resource language (LRL). India is a land of hundreds of languages out of which 22 languages are scheduled languages. Sadly, even most of these scheduled languages fall in the category of LRLs. The primary aim of this study is to find out whether multilingual setting could help to get better translation result as compared to their bilingual counterparts. Another secondary aim is to study the use of synthetic data to help low resource languages.

Assamese and Bengali are two of the scheduled languages of India. Spoken primarily in Assam, northeastern India, Assamese also has speakers in neighboring states and countries like Arunachal Pradesh, Meghalaya, Nagaland, Bangladesh, and Bhutan. Assamese is an agglutinative language, meaning words are formed by adding suffixes to root words. These suffixes convey grammatical information such as tense, aspect, and plurality. Due to its rich morphology, Assamese falls under the category of Morphologically Rich Languages (MRLs) [2]. This means that words can carry a lot of information, often eliminating the need for prepositions or additional words to convey grammatical meaning.

Bengali, also known by its endonym Bangla, is an Indo-Aryan language. It is a dominant language of Bangladesh and the second-most widely spoken language in India, particularly in the eastern states of West Bengal, Tripura, and parts of Assam.

Both Assamese and Bengali shares the same script for writing known as Bengali-Assamese Script (https://en.wikipedia.org/wiki/ Bengali-Assamese_script) or sometimes also called Assamese-Bengali script [3].

We did multiple experiments with both bilingual and Multilingual NMT (MNMT) involving the three languages. The results are evaluated with BLEU [4], chrF [5] and TER [6] automatic evaluation metrics using three different test sets (a) Flores-200 [7], (b) IN22-gen from IndicTrans2 [8] and (c) 500 domain specific test set created at the Department of Information Technology, Gauhati University. We got better results for Multilingual NMT as compared to its Bilingual counterparts.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2024

DOI: 10.17148/IJARCCE.2024.13398

The paper is divided into 6 sections. Section I is this introduction part. Section II and Section III outlines the review of existing works and data respectively. Experimental details are described in Section IV while Section V and Section VI shows the result and conclusion respectively.

II. REVIEW OF EXISTING WORKS

While Neural Machine Translation (NMT) has shown great success for many Indian language pairs, Assamese is indeed an under-resourced language in this domain. There is more reported research on NMT for languages like Hindi, Bengali, Marathi, Tamil, Punjabi, and Malayalam as compared to Assamese language. This is likely due to the larger speaker base and readily available training data for these languages. This limited research on Assamese NMT development can be attributed to the following factors:

Smaller Speaker Base: Assamese has a smaller speaker base compared to Hindi or Bengali, making it a "low-resource" language for NMT.

Data Scarcity: Training NMT models requires vast amounts of parallel text data (sentences in both Assamese and the other language). This data is harder to find for Assamese.

In relation to Assamese MT, we found few works [16][17][18] on MT development for Assamese using SMT. In the neural domain, there are works on Assamese NMT development such as [2], [19-24]. While these works represent Assamese MT, none are addressing the issue of developing a Multilingual NMT system which includes languages other than English and Assamese.

We got English to Indic and Indic to English MNNT works in [8][15]. In [15], the authors have investigated multilingual NMT development from English to Assamese and Bengali. They have used a transliteration-based approach by converting all the Assamese and Bengali data into English script. Their reported work showed very poor result on WMT22 test sets and the authors emphasized on more research in this direction. We, thus, propose to experiment with the same language combination without using transliteration-based approach.

III. DATA

Like other experiments in neural domain, NMT system requires large amount of parallel data too. These data are used to train neural translation models which are able to compete with human translator counterparts. Though, many world languages have such large collection of parallel corpora, all languages are not so fortunate and severely lacks in required amount of data to be used to build successful NMT systems. Many Indian languages do not have a very good amount of parallel corpus, e.g., Assamese and Bodo languages.

For the purpose of these experiments, we used publicly available data from various sources. Some of these data are open sourced and some are obtained by requesting an exclusive license from the owner agency. We have listed all the sources and their links in this section.

TABLE I: DOMAIN WISE COUNT OF SENTENCES FROM NPLT ENGLISH-ASSAMESE DATA

Domain	**No of Sentences**
Agriculture	10000
Entertainment	10000
Health	25000
Tourism	25000
Total	**70000**

English-Bengali sentence level parallel data was obtained from Samananar [9] dataset which are publicly available at https://ai4bharat.iitm.ac.in/samanantar/. This is the largest publicly available parallel corpora collection for Indian languages. We extracted 2.3 million English-Bengali sentences from Samanantar. Sentences having number of tokens greater than 5 and less than 35 were extracted. These 2.3 million sentences are again filtered with the LaBSE tool [10]. A threshold of 85% LaBSE score was used to further extract the best English-Bengali Sentence pairs. After this step, the final count of number of sentences in English-Bengali pair becomes 1234566. For the English-Assamese side, we again used the Samanatar data. It has 141227 numbers of English-Assamese parallel sentences.

DOI: 10.17148/IJARCCE.2024.13398

Another set of English-Assamese parallel data are obtained from National Platform for Language Technology (NPLT). NPLT (https://nplt.in/) is a platform designed for researchers, academics, and industry professionals to access Indian language data, tools, and related web services. It contains parallel sentences from four domains, namely Agriculture, Entertainment, Health, and Tourism. The domain wise count of the data is shown in Table I.

The combined data from the above two sets for English-Assamese is now 211227 number of sentences. As this data is far less than the other pair, English-Bengali, we use the technique of synthetic data generation for English to Assamese direction. To generate target side synthetic data, we use source side English monolingual data from Samanantar. These data are then synthesised to Assamese as target side with in-house developed English-Assamese Bilingual NMT model [2]. We use 422454 sentences from the Samanantar English monolingual data to keep the ratio of collected data to synthetic data as 1:2. The final training data is shown in the Table II.

TABLE II: TOTAL NUMBER OF SENTENCES FOR BOTH THE LANGUAGE PAIRS

Language Pair	No of Sentences
English-Bengali	1234566
English-Assamese	633681

IV. EXPERIMENT

All the experiments are done using freely available toolkit - Fairseq(-py) [11]. Fairseq is a sequence modelling toolkit and uses Python, making it accessible to a wide range of researchers and developers. It's specifically designed for tasks that involve sequences of data, such as text or speech. It allows to train our own sequence models using various architectures and techniques. It also integrates seamlessly with PyTorch (https://github.com/pytorch/pytorch), a popular deep learning framework. This flexibility is valuable for researchers exploring new approaches to the problem in hand.

We used multilingual Transformer architecture, which train Transformer models for multiple language pairs simultaneously, to develop the English to Assamese and Bengali One to Many multilingual NMT system. ‘adam’ optimizer was used with adam-betas (0.9, 0.98). Learning rate was 5e-04 with warmup-updates 4000. The experiment was run with a max-epoch of 100.

For the purpose of Bilingual NMT involving English-Assamese and English-Bengali, we used standard Transformer [1] architecture and same data. All other hyper parameters are kept same and the experiment were run with a max-epoch of 100.

Tokenization of English data is done with Moses [12] tokenizer and tokenization of Assamese and Bengali data is done with tokenizer from IndicNLP Library [13].

Byte-Pair Encoding with a vocab size of 16K was used for all the experiments. BPE [14] is a subword tokenization technique commonly used in NMT. BPE breaks down words into smaller units (subwords) that might be more frequent in the training data compared to whole words. This helps to address the Out-Of-Vocabulary (OOV) problem, where the model encounters words during translation that it has not seen during training.

For all the experiments we have used a single GPU machine. The details of the hardware available are as follows: NVIDIA Quadro P1000 GPU with 4096MB of GPU memory and 640 CUDA Cores, Graphics Clock speed (min: 136MHz, max:1544MHz), Memory Transfer Rate (min:810MHz, max:5010MHz) is the only system used for the purpose of this work. The machine has a RAM of 16GB and 64 bit Intel Xeon CPU.

Train-Validation split: The total data of English-Assamese and English-Bengali pairs are split into two sets Train and Validation, which is shown in Table III.

TABLE III: TRAIN-VALIDATION SPLIT

Language Pair	Train	Validation
English-Bengali	1230566	4000
English-Assamese	629681	2000

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2024

DOI: 10.17148/IJARCCE.2024.13398

V. RESULT

We tested the system with three test sets as mentioned in Section I. For English-Bengali pair, we have not reported any score for GUIT corpus as we have not created the test set. We have compared our Multilingual NMT system with the Bilingual NMT system developed for the same language pairs. We used BLEU, chrF and TER automatic evaluation metrics for the purpose of the comparison.

The results are shown in Table IV.

TABLE IV: COMPARISON OF BILINGUAL VS MULTILINGUAL TRANSLATION SCORES FOR ENGLISH, ASSAMESE AND BENGALI

Model	**FLORES-200**			**GUIT (Own test set)**			**IN22-Gen**		
	BLEU	**chrF**	**TER**	**BLEU**	**chrF**	**TER**	**BLEU**	**chrF**	**TER**
En-As (Bilingual)	9.01	35.25	86.45	16.76	50.11	71.03	12.54	43.12	80.01
En-As (Multilingual)	**11.32**	**40.11**	**82.56**	**18.22**	**52.36**	**68.55**	**14.33**	**47.53**	**78.21**
En-Bn (Bilingual)	12.36	42.67	84.45	-	-	-	15.01	49.89	78.45
En-Bn (Multilingual)	**15.64**	**44.14**	**81.23**	-	-	-	**16.27**	**51.22**	**72.36**

All the metrics shows better result for One-to-Many Multilingual NMT as compared to Bilingual NMT system (unidirectional).

It demonstrates the effectiveness of our One-to-Many Multilingual Neural Machine Translation (NMT) system compared to a Bilingual NMT system for the language pairs we have worked with.

VI. CONCLUSION

Our findings suggest that the One-to-Many Multilingual NMT system outperforms the Bilingual NMT system based on BLEU, chrF, and TER automatic evaluation metrics. This indicates that training a single model to translate from English to multiple target languages (like Assamese, Bengali, etc.) can lead to better translation quality compared to training separate models for each language pair. When a low resource language is coupled with a language which has relatively higher number of parallel sentences, can boost the performance of the low resource language.

This is because by training a single model on multiple language pairs, the model can learn shared representations for concepts that exist across these languages. This can be particularly beneficial for LRLs where the amount of training data available in each individual language pair might be limited. By leveraging the similarities between related languages, multilingual NMT models can improve their ability to generalize and translate unseen data in the low-resource language.

Additionally, use of synthetic data is another approach we used during this experiment. And from the result, it is evident that use of both MNMT and synthetic data helped to achieve better scores for English-Assamese direction. Interestingly, with only MNMT settings, English-Bengali translation score also increased.

As a future direction, we propose to study the performance of different low resource Indic languages while coupling with languages with varying language family and script.

ACKNOWLEDGMENT

We thank the Ministry of Electronics and Information Technology (MeitY), Government of India for their assistance through the Project ISHAAN.

DOI: 10.17148/IJARCCE.2024.13398

REFERENCES

[1]. A. Vaswani et al., 'Attention is all you need', in *Advances in neural information processing systems*, 2017, pp. 5998–6008.
[2]. K. Kashyap, S. K. Sarma, and M. A. Ahmed, 'Improving translation between English, Assamese bilingual pair with monolingual data, length penalty and model averaging', *International Journal of Information Technology*, vol. 16, no. 3, pp. 1539–1549, Mar. 2024.
[3]. S. Mahanta, 'Assamese', *Journal of the International Phonetic Association*, vol. 42, no. 2, pp. 217–224, 2012.
[4]. K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, 'Bleu: a Method for Automatic Evaluation of Machine Translation', in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.
[5]. M. Popović, 'chrF++: words helping character n-grams', in *Proceedings of the Second Conference on Machine Translation*, 2017, pp. 612–618.
[6]. M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, 'A Study of Translation Edit Rate with Targeted Human Annotation', in *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, 2006, pp. 223–231.
[7]. N. Team et al., 'No Language Left Behind: Scaling Human-Centered Machine Translation', *arXiv [cs.CL]*. 2022.
[8]. J. Gala et al., 'IndicTrans2: Towards High-Quality and Accessible Machine Translation Models for all 22 Scheduled Indian Languages', *Transactions on Machine Learning Research*, 2023.
[9]. G. Ramesh et al., 'Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages', *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 145–162, 02 2022.
[10]. F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, 'Language-agnostic BERT Sentence Embedding', *arXiv [cs.CL]*. 2020.
[11]. M. Ott et al., 'fairseq: A Fast, Extensible Toolkit for Sequence Modeling', in *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.
[12]. P. Koehn et al., 'Moses: Open Source Toolkit for Statistical Machine Translation', in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, 2007, pp. 177–180.
[13]. A. Kunchukuttan, 'The IndicNLP Library', 2020. [Online]. *Available: https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf*.
[14]. R. Sennrich, B. Haddow, and A. Birch, '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), 2016, pp. 1715–1725.
[15]. S. R. Laskar, R. Manna, P. Pakray, and S. Bandyopadhyay, 'Investigation of Multilingual Neural Machine Translation for Indian Languages', in *Proceedings of the 9th Workshop on Asian Translation*, 2022, pp. 78–81.
[16]. K. K. Baruah, P. Das, A. Hannan, and S. K. Sarma, 'Assamese-English Bilingual Machine Translation', *CoRR*, vol. abs/1407.2019, 2014.
[17]. P. Das and K. K. Baruah, 'Assamese to English statistical machine translation integrated with a transliteration module', *International Journal of Computer Applications*, vol. 100, no. 5, 2014.
[18]. M. T. Singh, R. Borgohain, and S. Gohain, 'An English-assamese machine translation system', *International Journal of Computer Applications*, vol. 93, no. 4, 2014.
[19]. R. Baruah, R. K. Mundotiya, and A. K. Singh, 'Low resource neural machine translation: Assamese to/from other indo-aryan (indic) languages', *Transactions on Asian and Low-Resource Language Information Processing*, vol. 21, no. 1, pp. 1–32, 2021.
[20]. M. Ahmed, K. Talukdar, P. Boruah, P. S. K. Sarma, and K. Kashyap, 'GUIT-NLP's Submission to Shared Task: Low Resource Indic Language Translation', in *Proceedings of the Eighth Conference on Machine Translation*, 2023, pp. 935–940.
[21]. M. A. Ahmed, K. Kashyap, and S. K. Sarma, 'Tokenization effect on neural machine translation: an experimental investigation for English-Assamese', in *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 2023, pp. 1–7.
[22]. R. Dutt, T. A. Kusupati, A. Srivastava, and B. Nath, 'Neural Machine Translation for English-Assamese Language Pair using Transformer', in *2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT)*, 2022, pp. 1–5.
[23]. B. Nath, S. Sarkar, and S. Das, 'Development of Neural Machine Translator for English-Assamese Language Pair', in *Advanced Techniques for IoT Applications: Proceedings of EAIT 2020*, 2022, pp. 279–288.
[24]. S. R. Laskar, P. Pakray, and S. Bandyopadhyay, 'Neural machine translation for low resource assamese--english', in *Proceedings of the International Conference on Computing and Communication Systems: I3CS 2020*, NEHU, Shillong, India, 2021, pp. 35–44.