NIT Rourkela Machine Translation(MT) System Submission to WAT 2022 for MultiIndicMT: An Indic Language Multilingual Shared Task

Sudhansu Bala Das
NIT Rourkela
520cs6006@nitrkl.ac.in

Atharv Biradar
PICT Pune
atharvbiradar28@gmail.com

Tapas Kumar Mishra
NIT Rourkela
mishrat@nitrkl.ac.in

Bidyut Kumar Patra
IIT BHU
bidyut.cse@iitbhu.ac.in

Abstract

Multilingual Neural Machine Translation (MNMT) exhibits incredible performance with the development of a single translation model for many languages. Previous studies on multilingual translation reveal that multilingual training is effective for languages with limited corpus. This paper presents our submission (Team Id: NITR) in the WAT 2022 for "MultiIndicMT shared task" where the objective of the task is the translation between 5 Indic languages(which are newly added in WAT 2022 corpus) into English and vice versa using the corpus provided by the organizer of WAT. Our system is based on a transformer-based NMT using fairseq modelling toolkit with ensemble techniques. Heuristic pre-processing approaches are carried out before keeping the model under training. Our multilingual NMT systems are trained with shared encoder and decoder parameters followed by assigning language embeddings to each token in both encoder and decoder. Our final multilingual system was examined by using BLEU and RIBES metric scores.

1 Introduction

This paper illustrates the submission of the MultiIndicMT shared task at the 9th Workshop on Asian Translation (WAT 2022)(Nakazawa et al., 2022) by NIT Rourkela (Team Id: NITR). Building Machine Translation (MT) model for 5 Indic languages (Assamese(as), Sindhi(sd), Sinhala(si), Urdu(ur) and Nepali (ne)) to English and vice versa is the main goal of this shared task wherein NITR has taken part. These languages are newly added in WAT 2022 corpus. The method that is most often used in machine translation is neural machine translation (Vaswani et al., 2017), (Bahdanau et al., 2014). Language pairs with fewer parallel corpora are often subject to have poor NMT performance. This happens because of a lack of translation expertise as well as overfitting, which is unavoidable in a low-resource environment. Since many Indian languages suffer from limited resources on an individual basis, creating high-quality machine translation systems for Indian languages continues to be a difficult task. However, numerous methods, including back translation (Sennrich et al., 2015), transfer learning (Zoph et al., 2016), etc., are developed to enhance the quality of low resource language translations. Additionally, training is needed for the model in each translation direction using conventional methods. So, in order to enhance the performance of language pairs with low resources, it is standard procedure to develop Multilingual Neural Machine Translation(MNMT) models by sharing parameters with languages having high resources (Firat et al., 2016), (Johnson et al., 2017), (He et al., 2016). Hence, in this regard, the shared task for WAT 2022 MultiIndicMT's goal is to verify the usefulness of MT methods for Indian languages. We have provided two MNMT models: a) one for Indic to English and the other for b) English to Indic. NITR MT System is trained on two MNMT models (Many to One and One to Many) based on Transformer Architecture using WAT 2022 MulitIndic Corpus. Our MNMT systems are based on (Johnson et al., 2017)'s method, wherein a language-specific token is appended to the input phrase in both one-to-many and many-to-many models to identify the target language to which the model needs to convert. Our training corpus are cleaned up thoroughly by using a set of heuristics techniques because the transformer model is sensitive to training noise (Liu et al., 2018). Finally, the result are presented in terms of Bilingual Evaluation Understudy (BLEU)(Papineni et al., 2002) and Rank-based Intuitive Bilingual Evaluation Score (RIBES)(Isozaki et al., 2010). In this paper, Section 2 describes the related work which is followed by the the detail description of data set in Section 3. The experimental methodology being explained in Section 4. The findings with results are discussed

in Section 5, and the paper concludes in Section 6.

2 Related Work

NMT framework can naturally include numerous languages, despite the fact that the early study on NMT focused on developing translation systems between two languages. As a result, research work on MT systems, that involves more than two languages, keeps on increasing significantly. Recently, a lot of attention is paid to multilingual neural machine translation since it allows one single model to translate between different languages. A many-to-many paradigm for multi-way translation employing shared attention and language-specific encoders and decoders is presented by (Pan et al., 2021). While transfer learning occurs implicitly in multilingualism, more explicit use of fine-tuning is an approach to accomplish the same (Zoph et al., 2016). Transliteration across scripts of related languages, as discussed in (Haddow et al., 2018) (Goyal and Sharma, 2019), may enhance the quality of multilingual models. Likewise, different methods that can be utilized to implement MNMT systems are summarised by (Dabre et al., 2020). (Sun et al., 2020) employs a fixed cross-lingual embedding, a single shared encoder, and language-specific decoders. By permitting positive transfer from the high resource languages, multiple studies on Multilingual NMT emphasizes the benefits for language pairings with low resources, enhancing the quality of the low resource ones. In terms of the BLEU score, multilingual unsupervised model tends to fare better than the bilingual unsupervised baselines. Building on earlier research by (Siddhant et al., 2022), (Bapna et al., 2022) efforts are made to combine multilingual supervised MT, zero-resource MT (Firat et al., 2016), and self-supervised learning into a single model for 1000 languages. In the next section, we give detail about the dataset which we have used.

3 Dataset

We used the dataset given by the organisers for generating the parallel corpus for Assamese, Nepali, Sindhi, Sinhala and Urdu language. The organizer have shared the MultiIndicMT WAT 2022 corpora, which is made up of roughly 15 million parallel sentences for 15 language pairs. From that corpus, we have used the OPUS corpus (Tiedemann, 2012) for the language pairs of Assamese, Nepali, Sindhi, Urdu, Sinhala and English. No additional data is used from any other sources. Table 1 shows the data statistics of parallel corpus provided by WAT 2022 organizers. Urdu is having the largest number of sentences whereas Assamese and Nepali are relatively low in corpus size.

Table 1: Parallel corpora statistics

EN to Indic	Sentences
en-as	140000
en-ne	700000
en-ur	6100000
en-sd	1700000
en-si	3300000

4 Methodology

In this section, we give details about the system those are submitted to the WAT2022 for Multi-IndicMT Shared Task (Nakazawa et al., 2022). We present findings for two categories of models: a) Many-En: Multilingual many-to-one system trained with all parallel data of five language pairs that are provided in WAT 2022, and b) En-Many: Multilingual many-to-one system trained with parallel data using the same corpus but in opposite direction. In this method, a shared encoder-decoder transformer architecture is employed to train our multilingual models.

4.1 Data Preprocessing

MultiIndicMT WAT 2022 corpora contains noisy sentences in many languages. So, filtering and pre-processing are carefully done to remove those. According to earlier research (Junczys-Dowmunt, 2018), a strict data filtering strategy is essential to keep quality of data. Out of many pre-processing techniques used by us, some of them are mentioned as inspired by (Li et al., 2019).

- Remove the sentence pair if either the source or the target sentence contains words longer than 35.
- If the source sentence has at least 10 characters in a different language, remove the sentence pair.
- Remove the sentence pair if the source sentence contains at least 60 % characters from a different language (UTF-8 ranges are utilised for this purpose).

- Remove sentences in which the language on the source and target sides is the same.
- Remove any sentences that have redundant translations or HTML elements.

Table 2: Filtered Parallel corpora statistics

EN to Indic	Filtered	Filtered Sentences
en-as	3.60%	134960
en-ne	5.80%	659400
en-ur	13.74%	5261860
en-sd	7.62%	1570460
en-si	11.65%	2915550

With implementation of the above techniques, We filtered the bilingual corpus accounting to approximate 8.48% sentence being filtered from the complete corpus as shown in Table 2. Then, we tokenize data for both Indian languages and English using the IndicNLP library and the Moses Tokenizer (Koehn et al., 2007) respectively.

4.2 Tokenization

Indic Languages do not share many terms at the non-root level despite having many cognates. Utilizing Indian languages at the sub-word level, which assures greater vocabulary overlap, is therefore the more effective strategy. As a result, we have used the widely accepted method of byte pair encoding (BPE) to break down each word into its sub-word parts (Sennrich et al., 2015). BPE units, which come in a variety of lengths, offer the proper context for translation systems involving related languages. Data sparsity is not an issue because their vocabularies are significantly smaller than those of the morpheme and word-level models. Learning BPE merging rules helps in a situation where numerous languages are involved. It not only helps in identifying common sub-words among them, but also ensures that each language pair is segmented properly.

5 Experimental Setup

This section describes the complete pipeline used to produce the translation systems for the WAT MultiIndic shared task submission.

5.1 Finetuning and Training

A multilingual model makes it possible to translate between several languages using a common word piece vocabulary. This is much easier than training separate models for each language pair. The Transformer(Vaswani et al., 2017) model (with 6 layers of encoder and decoder, 8 heads, 512 embedding size, and 2048 feed-forward size for each of them) is applied to implement our work. NITR MT System was trained on NVIDIA Quadro RTX 5000 machine having one GPU card. Further, for the implementation of the multilingual system, the advantage of Fairseq (Ott et al., 2019) library is considered. The method adopted by us is put forth by (Johnson et al., 2017) towards provisioning of a "language-specific token" driven technique that shares the attention mechanism and a single encoder-decoder network to create multilingual models. The input sequence includes a language token to indicate the direction of translation. Given this input, the decoder learns to produce the goal. This method, which is proven to be easy and efficient, compels the model to generalize across linguistic boundaries during training. To optimize model parameters, we have employed the Adam optimizer (Kingma and Ba, 2015).
Irrespective of time and resource constraints in order to experiment with several models, the last five checkpoints (360000–400000 iterations) are combined. Based on the correctness of the validation set, all our models are trained with early stopping criteria. After reassembling translated BPE segments during testing, the sentences translated are reverted to the previous language scripts. Lastly, the precision of our translation models is evaluated through BLEU and RIBES.

6 Results

The quality of our translation files are evaluated by the organisers using BLEU and RIBES, based on metrics on the official WAT 2022 MultiIndicMT test set (Nakazawa et al., 2022). To determine the BLEU scores of baseline models, multi-bleu.perl script is availed. When evaluating the Multi-IndicMT task, organizers prefer to tokenized reference and hypothesis files to find out the BLEU score. Moses-tokenizer is used for tokenization. We present results provided by the organizers for English to Indic and Indic to English language pairs which are based on the translation files that we have submitted (Nakazawa et al., 2022). Table 3 and 4 displays the multilingual models official BLEU and RIBES scores. In terms of scores, we notice that Urdu is having more than 15 BLEU score for both

Table 3: Result of One to Many(EN -> Indic) languages considering the evaluation Metrics.

	Baseline		**Our System**	
en->Indic	**BLEU**	**RIBES**	**BLEU**	**RIBES**
en->as	-	-	10.20	0.634631
en->ur	-	-	19.60	0.718763
en->sd	-	-	6.30	0.579323
en->si	-	-	9.50	0.647028

Table 4: Result of Many to One (Indic->English) languages considering the evaluation Metrics.

	Baseline		**Our System**	
Indic->en	**BLEU**	**RIBES**	BLEU	RIBES
as->en	-	-	15.50	0.706743
ne->en	-	-	8.00	0.546125
ur->en	-	-	20.50	0.744934
sd->en	-	-	15.40	0.709039
si->en	-	-	8.20	0.632228

the directions (En->Indic and Indic->En). Because of the time and resource constraints, we were not able to work with other indic languages.

7 Conclusion

In this paper, we highlight the MultiIndicMT shared task as submitted by us to WAT 2022. Through provisioning of two multilingual NMT models, one-to-many (English to 5 Indic languages) and many-to-one (4 Indic languages to English) we get competitive outcomes. In our process, test-runs are executed combining with several preprocessing and training strategies sequentially. Although we have used sufficient data filtering techniques, still it is observed that the training data gets contaminated with noise. Therefore, investigating more efficient data filtering methods as well as their effects on MT performance is another promising future area. In future, we look forward to extend our research that will help in fine-tuning of both encoder and decoder during the monolingual unsupervised training in order to improve the quality of the synthetic data generated during the process.

References

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

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, et al. 2022. Building machine translation systems for the next thousand languages. *arXiv preprint arXiv:2205.03983*.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Computing Surveys*, pages 1–38.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. *arXiv preprint arXiv:1601.01073*.

Vikrant Goyal and Dipti Misra Sharma. 2019. The iiit-h gujarati-english machine translation system for wmt19. In *Proceedings of the Fourth Conference on Machine Translation*, pages 191–195.

Barry Haddow, Nikolay Bogoychev, Denis Emelin, Ulrich Germann, Roman Grundkiewicz, Kenneth Heafield, Antonio Valerio Miceli-Barone, and Rico Sennrich. 2018. The university of edinburgh's submissions to the wmt18 news translation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 399–409.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. *Advances in neural information processing systems*, pages 1–9.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *In Proceedings of conference on empirical methods in natural language processing*, pages 944–952.

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 for Computational Linguistics, pages 339–351.

Marcin Junczys-Dowmunt. 2018. Microsoft's submission to the wmt2018 news translation task: How i learned to stop worrying and love the data. *arXiv preprint arXiv:1809.00196*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the Fourty fifth Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180.

Bei Li, Yinqiao Li, Chen Xu, Ye Lin, Jiqiang Liu, Hui Liu, Ziyang Wang, Yuhao Zhang, Nuo Xu, and Zeyang Wang. 2019. The niutrans machine translation systems for wmt. In *Proceedings of the Fourth Conference on Machine Translation*, pages 257–266.

Hairong Liu, Mingbo Ma, Liang Huang, Hao Xiong, and Zhongjun He. 2018. Robust neural machine translation with joint textual and phonetic embedding. *arXiv preprint arXiv:1810.06729*.

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. *arXiv preprint arXiv:1904.01038*.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. *arXiv preprint arXiv:2105.09501*.

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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Aditya Siddhant, Ankur Bapna, Orhan Firat, Yuan Cao, Mia Xu Chen, Isaac Caswell, and Xavier Garcia. 2022. Towards the next 1000 languages in multilingual machine translation: Exploring the synergy between supervised and self-supervised learning. *arXiv preprint arXiv:2201.03110*.

Haipeng Sun, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2020. Unsupervised neural machine translation with cross-lingual language representation agreement. *ACM Transactions on Audio, Speech, and Language Processing*, pages 1170–1182.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218. Citeseer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.