PICT@WAT 2022: Neural Machine Translation Systems for Indic Languages

Anupam Patil
anupampatil44@gmail.com

Isha Joshi
joshiishaa@gmail.com

Dipali Kadam
ddk@pict.edu

SCTR's Pune Institute of Computer Technology, India

(Team ID: 5592)

Abstract

Translation entails more than simply translating words from one language to another. It is vitally essential for effective cross-cultural communication, thus making good translation systems an important requirement. We describe our systems in this paper, which were submitted to the WAT 2022 translation shared tasks. As part of the Multi-modal translation tasks' text-only translation sub-tasks, we submitted three Neural Machine Translation systems based on Transformer models for English to Malayalam, English to Bengali, and English to Hindi text translation. We found significant results on the leaderboard for English-Indic (en-xx) systems utilizing BLEU and RIBES scores as comparative metrics in our studies. For the respective translations of English to Malayalam, Bengali, and Hindi, we obtained BLEU scores of 19.50, 32.90, and 41.80 for the challenge subset and 30.60, 39.80, and 42.90 on the benchmark evaluation subset data.

1 Introduction

The initial approach used in machine translation was rule-based. RBMT (Rule-Based Machine Translation) models use linguistic information about both the source and the target language to generate the translation. Platforms such as Apertium [1] use this approach. Eventually, SMT (Statistical Machine Translation) models came about, which did not use a predefined set of rules but inferred the rules by analyzing the given text (Koehn and Senellart, 2010). While SMT-based models provide more natural translations, RBMT systems provide translations that are truer to the original text (Forcada et al., 2011).

Machine translation (MT) systems have struggled with ambiguity in the source language while translating text, among other challenges. With the advent of deep learning techniques, neural networks are being used for machine translation tasks. Neural machine translation (NMT) models use massive amounts of training data and computational power to correctly identify the importance of the portion of the text data to generate the output text (Popel et al., 2020).

Recent advances in neural machine translation have focused on translating a source language into a specific target language. For this job, several approaches have been offered. Early NMT architectures used a fixed length approach to generate variable length outputs. The source text's length was fixed, irrespective of the length of the text. Models such as RCTM (Kalchbrenner and Blunsom, 2013) and RNNEncdec (Cho et al., 2014) use this approach. Eventually, newer architectures began using a variable length representation for the input text. GNMT (Wu et al., 2016) and ByteNet (Kalchbrenner et al., 2016) are architectures that use layered neural networks for translation (Tan et al., 2020).

In this paper, we describe our NMT systems, which were submitted to the translation shared tasks at WAT 2022 (Nakazawa et al., 2022) .

2 Related Work

The majority of NMT research has focused on using monolingual data or parallel data that includes other language pairs. NMT systems have consistently outperformed conventional machine translation methods such as rule-based and statistical-based approaches. NMT models typically operate with a fixed vocabulary; however, the translation is an open-vocabulary problem. Several approaches have been proposed to resolve this issue. Byte-Pair-Encoding (BPE) (Sennrich et al., 2015) enables NMT model translation on open vocabulary by en-

[1] https://github.com/apertium

Proceedings of the 9th Workshop on Asian Translation, pages 106–110
October 17, 2022.

English Sentence	Malayalam Translation	Bengali Translation	Hindi Translation
Little gray metal scissors	ചാരനിറത്തിലുള്ള മെറ്റൽ കത്രിക	ছোট ধূসর ধাতব কাঁচি	छोटे ग्रे धातु के कैंची
Dog has black nose	നായയ്ക്ക് കറുത്ത മൂക്ക് ഉണ്ട്	কুকুরের নাক কালো	कुत्ते की नाक काली होती है
A person is holding a kite	ഒരു വ്യക്തി ഒരു കൈറ്റ് പിടിക്കുന്നു	একজন ব্যক্তি একটি ঘুড়ি ধরে আছেন	एक व्यक्ति पतंग पकड़ा हुआ है

Table 1: Sample translations generated by our systems for the given English inputs.

coding rare and unknown words as a sequence of subword units.

NMT models typically employ the conventional sequence-to-sequence learning architecture, made up of an encoder and a decoder. In encoder-decoder mechanisms, words are translated into word embeddings in the encoder and then transferred to the decoder, which generates the following word in the translation using an attention mechanism, encoder representations, and preceding words. Several methodologies based on deep neural networks have been proposed, such as Recurrent Neural Networks (Cho et al., 2014), LSTM (Sutskever et al., 2014), Convolutional Neural Networks (Gehring et al., 2017), and Transformers (Vaswani et al., 2017), which can serve as encoders and decoders. Several approaches have been explored for machine translation in Malayalam, Bengali and Hindi.

2.1 Malayalam

Malayalam is a Dravidian language primarily spoken in southern India. It is a low resource language with very few usable resources for the purpose of training NMT models (Premjith et al., 2019). A rule-based approach for English to Malayalam translations has been proposed by Rajan et al. (2009). A modified rule-based approach using an SMT system was introduced by Rahul et al. (2009). There is little work in English to Malayalam translation systems that are based on deep neural networks, an example being the Google NMT system (Johnson et al., 2017).

2.2 Bengali

Bengali is the world's seventh most widely spoken language, however, it has received less focus in NMT work due to a lack of resources and poor corpus quality. Attempts to bridge this gap, specifically with regard to machine translation have been made by proposing new corpora (Hasan et al., 2020) and the use of attention-based techniques (Dabre et al. (2021), Abujar et al. (2021)) for improving upon existing systems.

2.3 Hindi

There has been a lot of focus given to the Hindi language in NMT literature in recent years, with the availability of good quality corpora (Kunchukuttan et al. (2017), Bojar et al. (2014)) thus enabling the development of effective NMT systems.

The recent development of Multilingual Models ((Dabre et al. (2021), Kakwani et al. (2020)) primarily focused on Indic languages has helped gain traction in the research community for MT.

3 Methodology

3.1 Data Description and Preprocessing

We use the datasets provided in the WAT 2022 shared tasks for our experiments. The datasets of all the three languages comprise 28,929, 997, and 1,595 English-Indic language sentence pairs for the training, dev, and eval subsets respectively, along with their corresponding images. We train and fine-tune the models using this data. The challenge subset additionally comprises 1,400 similar instances. The Malayalam and Bengali datasets are an extension to the HindiVisualGenome dataset, thus all three sets have the same set of sentence pairs while supporting their respective language. The language pair (English-Bengali) is running for the first time as a shared task.

We perform Normalisation, which minimizes the number of unique tokens in the text, and use the SentencePiece [2] tokenizer while utilizing the Byte-Pair-Encoding (BPE) (Sennrich et al., 2015) technique on the words present in a sentence.

3.2 Models and Training

We trained models using cutting-edge transformer-based neural machine translation (NMT). The architecture is based on a standard transformer architecture with 6 self-attentive layers in both the encoder and decoder networks, each with 8 attention

[2] https://github.com/google/sentencepiece

Team	en–ml		en–bn		en–hi	
	BLEU	RIBES	BLEU	RIBES	BLEU	RIBES
nlp_novices (ours)	19.50	0.536689	32.90	0.706596	41.80	0.812483
Team 1	14.60	0.392158	22.60	0.605676	29.60	0.728801
Team 2	12.98	0.378045	22.50	0.614267	30.72	0.736262
Team 3	–	–	26.70	0.680655	37.20	0.770640

Table 2: Details of official submission results on the challenge subset of data for en–ml, en–bn and en–hi translation systems.

Team	en–ml		en–bn		en–hi	
	BLEU	RIBES	BLEU	RIBES	BLEU	RIBES
nlp_novices (ours)	30.60	0.643987	39.80	0.745190	42.90	0.816564
Team 1	30.80	0.589471	41.00	0.767212	36.20	0.785673
Team 2	30.49	0.580807	40.90	0.758246	39.78	0.776892
Team 3	–	–	40.90	0.752543	37.01	0.795302

Table 3: Details of official submission results on the evaluation subset of data for en–ml, en–bn and en–hi translation systems.

heads per layer. For all our experiments, we use transformer models that follow the strategy implemented in OPUS MT (Tiedemann and Thottingal, 2020), which utilizes the Marian-NMT (Junczys-Dowmunt et al., 2018) toolkit and finetune them on the data provided for the shared tasks.

We obtained optimal performance on the English-Malayalam and English-Hindi translation tasks using en-ml [3] and en-hi [4] bilingual NMT models respectively. For the English-Bengali translation task, we achieved competent results using a multilingual NMT model [5].

The experiments were conducted in a Linux environment using an NVIDIA Tesla P100 GPU accelerator with 16 GB RAM and CUDA 11.2 installed. We train three separate MT models for the three indic languages in our experiments. The models utilize the AdamW (Loshchilov and Hutter, 2017) optimizer for optimization of model parameters with 0.00002 as the initial learning rate.

We observed varied results based on the number of epochs of training for each Indic language we translate to. We train the English to Hindi, English to Malayalam, and English to Bengali NMT models for 30, 20, and 25 epochs respectively after observing optimal performance for the respective systems.

4 Results

The metrics used to evaluate the translations were the Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002) and Rank-based Intuitive Bilingual Evaluation Score (RIBES) (Isozaki et al., 2010) metrics. Table 2 and 3 contain the BLEU and RIBES scores[6] obtained in each translation task, i.e. English to Malayalam, English to Bengali and English to Hindi on the challenge subset and the evaluation subset. For the English to Malayalam, English to Bengali and English to Hindi translation tasks, we were able to achieve BLEU scores of 19.50, 32.90 and 41.80 respectively (on the challenge subset), as reported in Table 2. As seen in Table 3, for the evaluation set, BLEU scores of 30.60, 39.80 and 42.90 were achieved for each translation task.

We have provided a comparative analysis between the effects of using fine-tuned pre-trained models and models trained from scratch. The optimal results on the leaderboard were obtained using the fine-tuned models. Table 4 depicts the difference in performance of the models with and without pre-training under similar training methods. To obtain comparable results using non pre-trained models, additional training and data resources would be required.

[3] https://github.com/Helsinki-NLP/OPUS-MT-train/tree/master/models/en-ml

[4] https://github.com/Helsinki-NLP/Tatoeba-Challenge/tree/master/models/eng-hin

[5] https://github.com/Helsinki-NLP/Tatoeba-Challenge/tree/master/models/eng-mul

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

Language Pair	With pre-training		Without pre-training	
	Test	Challenge	Test	Challenge
en-ml	30.60	19.50	0.65342	0.21455
en-bn	39.80	32.90	0.00057	0.14980
en-hi	42.90	41.80	2.03818	0.92631

Table 4: Effect of using pre-trained models on the performance by comparative analysis using BLEU scores.

The disparity between the performance with respect to the challenge set and the evaluation set (reported in Table 2 and 3) can be attributed to the following reasons:

- Firstly, the challenge set had 1232 unique English words (ignoring stopwords), while the evaluation set had 1256; the number of common words between the two of them being only 552.

- Additionally, the number of intersecting terms (ignoring stopwords) between the train set + the challenge set and the train set + the evaluation set is 976 and 1109 respectively.

The above reasons may explain the ambiguity that arises when it comes to the translation of some unique words.

Table 1 illustrates sample translations of three common English sentences taken from the shared task data. The table reports translations of the given English inputs in Malayalam, Bengali, and Hindi.

5 Conclusion

In this paper, we discuss the submissions made to three tasks at WAT 2022: Neural Machine Translation Systems for Indic Languages. We participated in the text-only subtask of the multimodal translation tasks of English to Malayalam, English to Bengali, and English to Hindi translations. In the future, we would like to experiment with multimodal MT models and incorporate multimodal aspects for the facilitation of better translation systems.

References

Sheikh Abujar, Abu Kaisar Mohammad Masum, Abhishek Bhattacharya, Soumi Dutta, and Syed Akhter Hossain. 2021. English to bengali neural machine translation using global attention mechanism. In *Emerging Technologies in Data Mining and Information Security*, pages 359–369. Springer.

Ondřej Bojar, Vojtěch Diatka, Pavel Rychlỳ, Pavel Straňák, Vít Suchomel, Aleš Tamchyna, and Daniel Zeman. 2014. Hindencorp-hindi-english and hindi-only corpus for machine translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3550–3555.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh M Khapra, and Pratyush Kumar. 2021. Indicbart: A pre-trained model for natural language generation of indic languages. *arXiv preprint arXiv:2109.02903*.

Mikel L Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25(2):127–144.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *International conference on machine learning*, pages 1243–1252. PMLR.

Tahmid Hasan, Abhik Bhattacharjee, Kazi Samin, Masum Hasan, Madhusudan Basak, M Sohel Rahman, and Rifat Shahriyar. 2020. Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for bengali-english machine translation. *arXiv preprint arXiv:2009.09359*.

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.

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*, 5:339–351.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, et al. 2018. Marian: Fast neural machine translation in c++. *arXiv preprint arXiv:1804.00344*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961.

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.

Nal Kalchbrenner, Lasse Espeholt, Karen Simonyan, Aaron van den Oord, Alex Graves, and Koray Kavukcuoglu. 2016. Neural machine translation in linear time. *arXiv preprint arXiv:1610.10099*.

Philipp Koehn and Jean Senellart. 2010. Convergence of translation memory and statistical machine translation. In *Proceedings of AMTA Workshop on MT Research and the Translation Industry*, pages 21–31.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2017. The iit bombay english-hindi parallel corpus. *arXiv preprint arXiv:1710.02855*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

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.

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.

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtskỳ. 2020. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature communications*, 11(1):1–15.

B Premjith, M Anand Kumar, and KP Soman. 2019. Neural machine translation system for english to indian language translation using mtil parallel corpus. *Journal of Intelligent Systems*, 28(3):387–398.

C Rahul, K Dinunath, Remya Ravindran, Soman, and KP. 2009. Rule based reordering and morphological processing for english-malayalam statistical machine translation. In *2009 International Conference on Advances in Computing, Control, and Telecommunication Technologies*, pages 458–460. IEEE.

Remya Rajan, Remya Sivan, Remya Ravindran, and KP Soman. 2009. Rule based machine translation from english to malayalam. In *2009 International Conference on Advances in Computing, Control, and Telecommunication Technologies*, pages 439–441. IEEE.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Zhixing Tan, Shuo Wang, Zonghan Yang, Gang Chen, Xuancheng Huang, Maosong Sun, and Yang Liu. 2020. Neural machine translation: A review of methods, resources, and tools. *AI Open*, 1:5–21.

Jörg Tiedemann and Santhosh Thottingal. 2020. Opus-mt–building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation.

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.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.