Transformer-based Neural Machine Translation System for Hindi – Marathi: WMT20 Shared Task

Amit Kumar, Rupjyoti Baruah, Rajesh Kumar Mundotiya, Anil Kumar Singh
Department of Computer Science & Engineering
Indian Institute of Technology (B.H.U.)
Varanasi, India
{amitkumar.rs.cse17, rupjyotibaruah.rs.cse18,
rajeshkm.rs.cse16, aksingh.cse}@iitbhu.ac.in

Abstract

This paper reports the results for the Machine Translation (MT) system submitted by the NLPRL team for the Hindi – Marathi Similar Translation Task at WMT 2020. We apply the Transformer-based Neural Machine Translation (NMT) approach on both translation directions for this language pair. The trained model is evaluated on the corpus provided by shared task organizers, using BLEU, RIBES, and TER scores. There were a total of 23 systems submitted for Marathi to Hindi and 21 systems submitted for Hindi to Marathi in the shared task. Out of these, our submission ranked 6th and 9th, respectively.

1 Introduction

In the last decade and a half, neural machine translation (NMT) (Sutskever et al., 2014) has achieved great success in automatically translating human language text, outperforming statistical machine translation (SMT) (Koehn et al., 2003). Both the system require very large corpus sizes to train and evaluate the results. They, however, don't work very well for low resource data (He et al., 2016; Koehn and Knowles, 2017; Dowling et al., 2018). Translation from or to low resource languages is the major challenges faced by today's NMT systems.

Different methods have been proposed to overcome the data sparsity problem for low resource languages by researchers around the world. These include using monolingual data (Wu et al., 2019), fine-tuning (Miceli Barone et al., 2017) the high resource monolingual and parallel data on low resource data, back translation (Hoang et al., 2018), etc. They succeed up to some extent, but the success is limited, as the reported results show when compared to those for resource rich languages.

In this paper, we use the Transformer network-based NMT system (Vaswani et al., 2017) because it is among the state of the art models for machine translation. The work reported for this shared task is an extension of the work done by (Kumar and Singh, 2019) for similar languages task for 2019, which had also used a transformer based NMT system.

2 Similar Languages

Two languages are considered similar or closely related if they are close relatives in terms of the linguistic family of the linguistic family tree (or forest), or if the speakers of the two languages are in close contact over a long period of time. Contact over a long period leads to the exchange of cognates and loanwords between the speakers, sometimes even grammatical constructs.

Leveraging the close similarity of languages is one way to overcome the problem of data scarcity. Using similar features between such languages and improving translation is one of the directions for research for low resource machine translation.

For this submission, the motives behind conducting the shared task experiments are:

- To find out whether it is advantageous to use transformer-based NMT for similar languages.
- Whether using the SentencePiece [1] library without tokenization is beneficial for translation between similar languages or not.

3 Submitted System

We submitted two systems, namely, Marathi→Hindi and Hindi→Marathi. Both are the NMT systems trained on a Transformer (Vaswani et al., 2017) network. In this experiment, we did not tokenize data using any tokenizer. We directly applied SentencePiece library on the corpus. We found that directly applying

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

Proceedings of the 5th Conference on Machine Translation (WMT), pages 393–395

Parameters	Value
Encoder and decoder layers	5
Encoder embedding dimension	512
Decoder embedding dimension	512
Encoder attention heads	2
Decoder attention heads	2
Dropout	0.4
Attention dropout	0.2
Optimizer	Adam
Learning rate scheduler	inverse sqrt
Learning rate	1e-3
Minimum learning rate	1e-9
Adam-betas	(0.9, 0.98)
Number of epochs	100

Table 1: Hyperparameters used in our experiment

SentencePiece for preprocessing of data gives a better result. Since both the languages come under the category of morphologically rich and similar languages, directly applying SentencePiece on their corpus is advantageous. SentencePiece breaks the sentences into morphemes and phonemes. It extracts loanwords and cognate pairs. Breaking of sentences into subwords helps the neural translation network to learn better translations, and to generalize this knowledge to translate and produce unseen words, partly due to jointly developing the subword vocabulary.

4 Data

We trained the model on total 49434 number of Hindi - Marathi parallel corpus which belongs to three domains: News, PM India and Indic WordNet. Validation is done on total 1411 sentences. For testing, a total of 1941 sentences were used.

5 Experiment setup

We used fairseq [2] sequence to sequence encoder-decoder framework to train and evaluate the system. For hyper-parameter settings, we used the settings reported by (Guzmán et al., 2019) as these setting work well on low resource languages. Table 1 gives the hyper-parameter settings.

6 Results

Task organizers evaluate the systems using three evaluation metric: BLEU (Papineni et al., 2002), RIBES (Isozaki et al., 2010) and Translation Error Rate (TER) (Snover et al., 2006). We report the evaluation scores in table 2.

[2]https://github.com/pytorch/fairseq

system	BLEU	RIBES	TER
Marathi → Hindi	20.72	64.46	71.04
Hindi → Marathi	12.5	58.66	76.86

Table 2: Scores of our system evaluated by task organizers

7 Conclusion

In this paper, we perform experiments for translation between two similar languages: Hindi and Marathi. We submitted two systems: Marathi→Hindi and Hindi→Marathi, which were evaluated using BLEU, RIBES and TER. We found that SentencePiece works well for similar languages because it helps the Transformer in capturing the relations between two languages by providing morphemes, phonemes, cognate pairs, loanwords, etc. There were a total 23 systems submitted for Marathi → Hindi and 21 systems submitted for Hindi → Marathi in the shared task. Out of these, our system ranked 6th and 9th for Marathi → Hindi and Hindi → Marathi, respectively, considering the BLEU scores.

References

Meghan Dowling, Teresa Lynn, Alberto Poncelas, and Andy Way. 2018. Smt versus nmt: Preliminary comparisons for irish.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The flores evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english. *arXiv preprint arXiv:1902.01382*.

Wei He, Zhongjun He, Hua Wu, and Haifeng Wang. 2016. Improved neural machine translation with smt features. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 151–157. AAAI Press.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia. Association for Computational Linguistics.

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.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133.

Amit Kumar and Anil Kumar Singh. 2019. NLPRL at WAT2019: Transformer-based Tamil – English indic task neural machine translation system. In *Proceedings of the 6th Workshop on Asian Translation*, pages 171–174, Hong Kong, China. Association for Computational Linguistics.

Antonio Valerio Miceli Barone, Barry Haddow, Ulrich Germann, and Rico Sennrich. 2017. Regularization techniques for fine-tuning in neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1489–1494, Copenhagen, Denmark. 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*, pages 311–318. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200. Cambridge, MA.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

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

Lijun Wu, Yiren Wang, Yingce Xia, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2019. Exploiting monolingual data at scale for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4207–4216, Hong Kong, China. Association for Computational Linguistics.