Neural Machine Translation for Tamil–Telugu Pair

Sahinur Rahman Laskar, Bishwaraj Paul, Prottay Kumar Adhikary
Partha Pakray, Sivaji Bandyopadhyay
Department of Computer Science and Engineering
National Institute of Technology Silchar
Assam, India
{sahinur_rs, bishwaraj_ug, prottay_ug, partha}@cse.nits.ac.in
sivaji.cse.ju@gmail.com

Abstract

The neural machine translation approach has gained popularity in machine translation because of its context analysing ability and its handling of long-term dependency issues. We have participated in the WMT21 shared task of similar language translation on a Tamil-Telugu pair with the team name: CNLP-NITS. In this task, we utilized monolingual data via pre-train word embeddings in transformer model based neural machine translation to tackle the limitation of parallel corpus. Our model has achieved a bilingual evaluation understudy (BLEU) score of 4.05, rank-based intuitive bilingual evaluation score (RIBES) score of 24.80 and translation edit rate (TER) score of 97.24 for both Tamil-to-Telugu and Telugu-to-Tamil translations respectively.

1 Introduction

Machine translation (MT) works as an interface that handles language ambiguity concerns via automatic translation between two different languages. Neural machine translation (NMT) attains state-of-the-art results for both high and low-resource language pairs translation (Cho et al., 2014; Sutskever et al., 2014; Bahdanau et al., 2015; Luong et al., 2015; Pathak et al., 2018; Pathak and Pakray, 2018; Laskar et al., 2019; Laskar et al., 2020a). The NMT utilizes an artificial neural network to predicts the likelihood of a sequence of words. But NMT requires a sizeable parallel corpus to get effective MT output, challenging for low-resource pair translation. In this WMT21 shared task, we have participated on a similar language pair translation task of Tamil–Telugu pair using NMT. We aim to utilize similarity features among such a similar language pair and monolingual data to overcome the less availability of parallel corpus. The transformer model (Vaswani et al., 2017) based NMT is considered in this work, since it outperforms RNN based NMT. Moreover, NMT performance can be enhanced utilizing monolingual data (Weng et al., 2019; Wu et al., 2019; Ramachandran et al., 2017; Variš and Bojar, 2019; Qi et al., 2018). To evaluate the performance of our system's output, WMT21 organizer used standard evaluation metrics, namely, BLEU (Papineni et al., 2002), RIBES (Isozaki et al., 2010) and TER (Snover et al., 2006) which are reported in Section 4.

2 Related Work

There are limited works on the Tamil–Telugu pair (Chakravarthi et al., 2021). The literature survey found similar works on Indian similar language pairs, such as Hindi–Nepali (Laskar et al., 2019) and Hindi–Marathi (Laskar et al., 2020b) at WMT19 and WMT20. Both (Laskar et al., 2020b, 2019) used transformer model based NMT. Moreover, Ramachandran et al. (2017); Variš and Bojar (2019); Qi et al. (2018) pre-trained methods are incorporated in NMT to utilize advantage of monolingual data for low-resource pairs translation. In this work, GloVe (Pennington et al., 2014) pre-trained word embeddings are used in transformer model (Vaswani et al., 2017) based NMT for both Tamil-to-Telugu and Telugu-to-Tamil translation.

3 System Description

Our system mainly consists of the following parts: data preprossessing, model training and testing. These have been described in the following subsections. The dataset description is presented in Section 3.1. For our system, we have used the OpenNMT-py toolkit (Klein et al., 2017) for the data preprocessing, training and testing.

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 284–287
November 10–11, 2021. ©2021 Association for Computational Linguistics

Corpus	Type	Sentences	Tokens		Source
			Tamil	Telugu	
Parallel	Train	40147	588919	625308	WMT21 Organizer
	Validation	1261	25443	25844	
	Test	1735	33911	35895	
Monolingual	Tamil	31542481	488507451		IndicNLP
	Telugu	47877462	574131374		

Table 1: Dataset Statistics

3.1 Dataset

The parallel corpus for Tamil-Telugu pair is provided by the WMT21 organizer[1]. It consists of 40147, 1261, 1735 sentence pairs for train, validation and test set. Apart from this, we also collected Monolingual data from the IndicNLP[2] corpus. It consists of 31542481 Tamil sentences and 47877462 Telugu sentences. This monolingual corpus is specifically used for deriving pretrained embeddings to use in the model. The dataset statistics are described in the table 1.

3.2 Data Preprocessing

The OpenNMT-py toolkit is used to preprocess the parallel data and then generates a vocabulary of size: 50002 for the source-target sentences by tokenizing and indexing in a dictionary. It was done in both ways independently, considering Tamil as source and Telugu as target and then with Telugu as source and Tamil as the target to train models for both ways for translation in either direction. We have used GloVe (Pennington et al., 2014) to pretrain on the monolingual corpora to obtain word vectors. These word vectors are specifically used in the form of word embeddings in the transformer model during the training process.

3.3 System Training

After the data preprocessing, the pre-trained embeddings and parallel dataset are used for training our model for both Tamil-to-Telugu and Telugu-to-Tamil. We have adopted a transformer model to implement both of the trained models separately. The transformer model consists of a self-attention mechanism, encoder, and decoder layers. The self-attention comes into play, where the relevancy of one word to other words of the sentence is represented as an attention vector that contains the context between words in that sentence. Multiple such attention vectors are calculated, and the weighted average is taken so that the interactions with other words are captured properly rather than their value. More specifically, the embeddings are converted into three spaces: query, key, and value. The dot product of its query vector and all the key vectors are calculated for every embedding. Since the hidden state of the previous embedding is not needed in calculating the current embedding's hidden state, the self-attention can be done in parallel for all embeddings. Thus, it can be run in parallel for all embeddings simultaneously. This speeds up the training and translation process a lot. Now, the target sentences are passed to the decoder layers similarly to the encoders and then passed to the self-attention block. The difference is that in attention layers, the next word of the target sentence is masked so that the word will be predicted using previous results for learning. It is called a masked multi-head attention block. The attention vectors thus produced and the outputs from the encoder layers are then passed to another attention block called encoder-decoder attention block. The attention vectors for every word in the sentences are the output. Then we pass it through a feed-forward network for making output acceptable for further layers.

Our transformer model consists of six layers for both encoders and decoders and eight attention heads. We used adam optimizer with a learning rate 0.001 and a drop-out of 0.1 for normalization. The rest of the parameters were selected as the default configuration of the toolkit. This configuration is used for both models, the Tamil-to-Telugu and vice versa.

3.4 System Testing

The obtained trained models are used in system testing, where the test data is used to obtain the predicted translation for both Tamil-to-Telugu and vice-versa independently.

[1] http://statmt.org/wmt21/similar.html

[2] https://indicnlp.ai4bharat.org/corpora/

Translation	System Type	BLEU	RIBES	TER
Tamil to Telugu	Primary	4.05	24.80	97.24
Telugu to Tamil	Primary	4.05	24.80	97.24

Table 2: Our System results for Tamil-Telugu pair at WMT21

4 Result

Our system's outputs were submitted to the organizer for evaluation. Consequently, the results of the shared task on ”Similar Language Translation” were announced separately for Tamil-to-Telugu[3] and Telugu-to-Tamil[4]. The ranking of the systems is mainly based on BLEU score, while the RIBES and TER scores are also given. Our team name is CNLP-NITS. For the Tamil-to-Telugu translation system, we achieved 4th rank with a BLEU score of 4.05 and 6th rank with a BLEU score of 4.05 for the Telugu-to-Tamil translation. The results of our system are reported in the Table 2. The system performance is identical for both translation directions. We need to perform a human evaluation in future work to identify the test set and predicted output are identical or not.

5 Conclusion and Future Work

This work reports our system description along with results, which we have participated in the WMT19 shared task of similar language pair: Tamil-Telugu. Both direction of translations, transformer model based NMT is used and utilized monolingual data through pre-trained word embeddings. We will investigate multilingual NMT approach in future to improve such low-resource translation quality.

Acknowledgement

Authors would like to thank WMT21 Shared task organizers for organizing this competition and also, thank Center for Natural Language Processing (CNLP) and Department of Computer Science and Engineering at National Institute of Technology, Silchar for providing the requisite support and infrastructure to execute this work.

[3] https://mzampieri.com/workshops/wmt/2021/TA_TE.pdf

[4] https://mzampieri.com/workshops/wmt/2021/TE_TA.pdf

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*.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Shubhanker Banerjee, Richard Saldanha, John P. McCrae, Anand Kumar M, Parameswari Krishnamurthy, and Melvin Johnson. 2021. Findings of the shared task on machine translation in Dravidian languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 119–125, Kyiv. Association for Computational Linguistics.

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. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. 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.

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.

S. R. Laskar, A. Dutta, P. Pakray, and S. Bandyopadhyay. 2019. 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. 2020a. 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, Abdullah Faiz Ur Rahman Khilji, Partha Pakray, and Sivaji Bandyopadhyay. 2020b. Hindi-Marathi cross lingual model. In *Proceedings of the Fifth Conference on Machine Translation*, pages 396–401, Online. Association for Computational Linguistics.

Sahinur Rahman Laskar, Partha Pakray, and Sivaji Bandyopadhyay. 2019. 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.

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.

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.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543, Doha, Qatar. ACL.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

Prajit Ramachandran, Peter Liu, and Quoc Le. 2017. Unsupervised pretraining for sequence to sequence learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 383–391, Copenhagen, Denmark. 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 *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA. MIT Press.

Dušan Variš and Ondřej Bojar. 2019. Unsupervised pretraining for neural machine translation using elastic weight consolidation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 130–135, Florence, Italy. Association for Computational Linguistics.

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.

Rongxiang Weng, Heng Yu, Shujian Huang, Weihua Luo, and Jiajun Chen. 2019. Improving neural machine translation with pre-trained representation. *CoRR*, abs/1908.07688.

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.