Hindi-Marathi Cross Lingual Model

Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji, Partha Pakray, Sivaji Bandyopadhyay
Department of Computer Science and Engineering
National Institute of Technology Silchar
Assam, India
{sahinur_rs, abdullah_ug, partha}@cse.nits.ac.in, sivaji.cse.ju@gmail.com

Abstract

Machine translation (MT) is a vital tool for aiding communication between linguistically separate groups of people. The neural machine translation (NMT) based approaches have gained widespread acceptance because of its outstanding performance. We have participated in WMT20 shared task of similar language translation on Hindi-Marathi pair. The main challenge of this task is by utilization of monolingual data and similarity features of similar language pair to overcome the limitation of available parallel data. In this work, we have implemented NMT based model that simultaneously learns bilingual embedding from both the source and target language pairs. Our model has achieved Hindi to Marathi bilingual evaluation understudy (BLEU) score of 11.59, rank-based intuitive bilingual evaluation score (RIBES) score of 57.76 and translation edit rate (TER) score of 79.07 and Marathi to Hindi BLEU score of 15.44, RIBES score of 61.13 and TER score of 75.96.

1 Introduction

MT is a well-known task of natural language processing (NLP) wherein automatic translation is performed between different languages. Broadly, MT is categorized into rule-based and corpus-based, where rule-based is based on a pre-defined rules on the concerned languages and corpus-based finds a generalized approach after being trained on a large corpus. MT switches from rule-based approach to the corpus-based which blots out the need for linguistic expertise. In the corpus-based approach, example-based machine translation (EBMT), statistical machine translation (SMT) and NMT techniques are available. The disadvantage of EBMT is that even though the corpus is large, all examples are not covered. To mitigate the issues of the contemporary approach SMT is introduced Brown et al. (1990); Koehn (2010). The SMT based system makes an assumption based on probability scores of the translated text. And hence, the ranking is done. SMT also faces many issues like system complexity, long term dependency problem, context-analyzing inability, word-alignment and the rare word problem. The inefficiency of SMT leads to the development of the NMT Devlin et al. (2014). But like SMT, the NMT based model also suffers the requirement of sufficient training parallel corpus, which is a challenge in the case of low resource languages. For this reason, there is a demand for direct translation among similar language pairs by utilizing similarity features and monolingual data, so that less availability of the parallel data does not pose a challenge. However, the NMT technique achieves state-of-the-art approach in MT because of its transformer model Vaswani et al. (2017). For low resource language pair translation, NMT models have been improved with monolingual corpus Sennrich et al. (2016b); Burlot and Yvon (2018); Wu et al. (2019). In this work, we have adopted cross-lingual language model (XLM) Conneau and Lample (2019) to implement an NMT model for Hindi-Marathi similar language translation task because XLM shows significant improvements for low-resource languages by utilizing the monolingual corpora.

2 Related Work

Hindi-Marathi translation lacks background work. However, similar work is found on Hindi-Nepali pair at WMT19 shared task of similar language translation Laskar et al. (2019). The literature survey mainly focuses on NMT for low resource language pairs since NMT outperforms conventional SMT on low resource pairs like English to Mizo, English to Hindi, English to Punjabi, and English to Tamil Pathak et al. (2018); Pathak and Pakray (2018); Laskar et al. (2019). It is noticed that train-

Proceedings of the 5th Conference on Machine Translation (WMT), pages 396–401
Online, November 19–20, 2020.

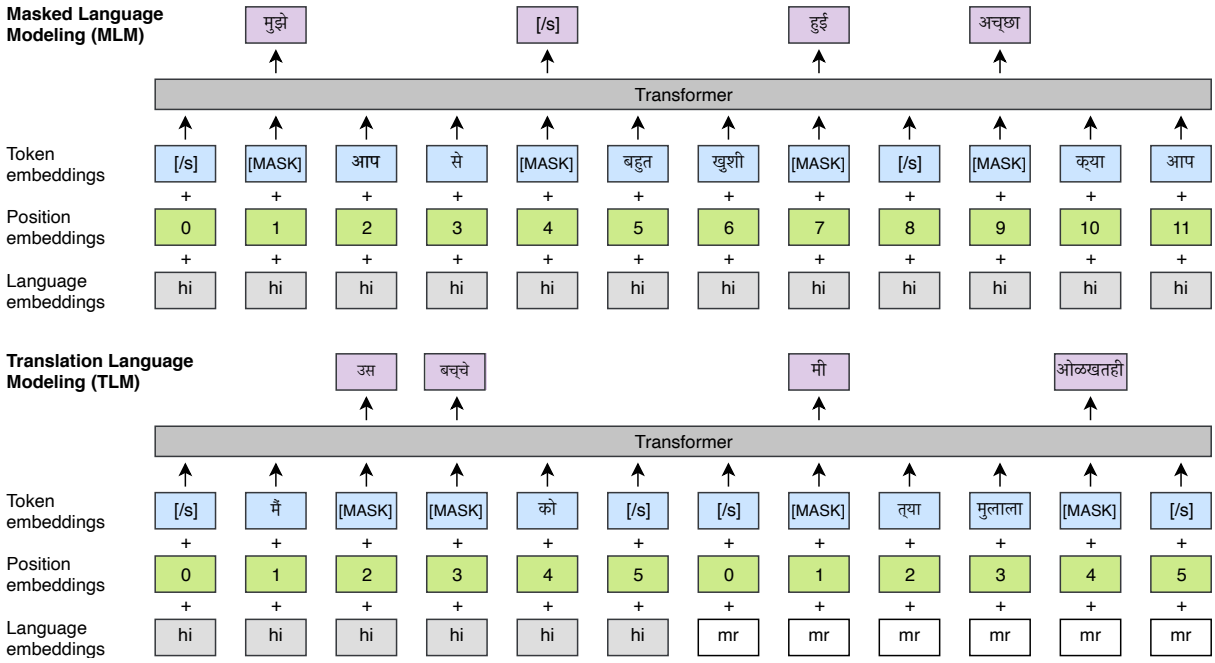

Figure 1: MLM pre-training inspired from Devlin et al. (2018) and TLM fine-tuning objective which extends the MLM task to parallel sentences as used by Conneau and Lample (2019). Diagram adapted from (Conneau and Lample, 2019) after suitable changes.

ing performance improves while parallel training data increases. For low resource languages, it is difficult to collect parallel data unlike monolingual data which is easily found through online sources. Hence, monolingual based NMT systems are introduced to enhance the translation quality of low resource language pair translation Sennrich et al. (2016b); Burlot and Yvon (2018); Wu et al. (2019). To get the advantage of monolingual data, unsupervised pre-train methods are introduced Ramachandran et al. (2017); Variš and Bojar (2019). Conneau and Lample (2019) proposed XLM based on bidirectional encoder representations from transformers (BERT) where the contextual language model is built with words based on preceding and succeeding context. No work has been done on Hindi-Marathi low resource language pair with such advanced NMT based approach, from the best of our knowledge. Our work investigates XLM model on Hindi-Marathi low resource language pair translation.

3 Dataset

3.1 Description

The organizers of WMT20 provided parallel and monolingual corpus for both Hindi and Marathi. The training dataset available for the WMT20, Hindi-Marathi task was obtained from three main sources viz. Indic WordNet, News, and PM India. Having 11,188, 12,349, and 25,897 parallel sentences (total 49434 sentences) respectively. The validation and test set contain 1941 and 1411 sentences. The Hindi monolingual dataset contains about 96 million sentences at about 32GB whereas the Marathi dataset is much smaller at only 4.72 million sentences totalling to around 2GB of corpus.

3.2 Preprocessing

We have removed unwanted symbols like URLs, email IDs and English text from the monolingual corpora of both the languages if any were to be present. In addition to this, since Hindi and Marathi languages share many common Devnagiri characters and hence to leverage this idea we have pre-processed the dataset obtained from Section 3.1 by a common vocabulary prepared via byte pair encoding (BPE) Sennrich et al. (2016a) on the same data provided by the organizer. Such an approach greatly helps in aligning the embedding space as shown in Lample et al. (2017). BPE learning is performed as used by Conneau and Lample (2019). The BPE is thus learnt after joining random sentences from the monolingual corpora. Following Conneau and Lample (2019) the text is sampled using a multinomial distribution. The distribu-

tion is as shown in Equation 1. The probabilities of the distribution are $p_{i=1...N}$. The BPE codes are generated and applied using the C++ implementation[1] of Sennrich et al. (2015).

$$p_i = \frac{q_i^\alpha}{\sum_{j=1}^{N} q_j^\alpha} \quad (1)$$

and p_i is as defined in Equation 2.

$$q_i = \frac{n_i}{\sum_{k=1}^{N} n_k} \quad (2)$$

α is taken as 0.5.

4 System Description

Our approach consists of the two principal approaches viz. the pre-training step and the fine-tuning step which are discussed in the following sub-sections 4.1 and 4.2. .

4.1 Pretraining our Model

For the pre-training step we have followed the steps of (Conneau and Lample, 2019) and utlized the masked language modeling (MLM) objective of (Devlin et al., 2018). Thus, following the work of Devlin et al. (2018) we have sampled 15% of BPE tokens randomly from the textual data and masked then by a `[MASK]` token roughly 80%. Also from the remaining 20%, the 10% component is randomly replaced and the rest part remains unchanged. The difference our approach has from the work of Devlin et al. (2018) is that, we have used lengths truncated to a fixed number (256 in our case), whereas the former uses pairs of sentences. To create a balance between the rare and commonly occuring BPE tokens like punctuation marks, the frequent outputs were subsampled using a multinomial distribution, where the weights are proportional to the inverse square root of the frequencies (an approach similar to Mikolov et al. (2013)). The pretraining objective is illustrated in Figure 1.

4.2 Fine Tuning

The model pre-training step follows an unsupervised approach and requires only the monolingual data. Since, the principal task for our work was to build a MT system, we need to leverage parallel data. Following, (Conneau and Lample, 2019) we used the translation language modeling (TLM) for fine-tuning the model obtained from Section 4.1. Here, instead of the truncated monolingual corpora we utlize the concatenation of parallel data as shown in Figure 1. Since the parallel sentences are concatenated for the concerned TLM task, we can mask and predict simultaneously from both Hindi and Marathi sentences. Enabling better placement of Hindi and Marathi word representations. Specifically as shown by Conneau and Lample (2019), this enables the model to leverage the context even if single handedly the source or target sentence is insufficient to decipher the sentence.

5 Experimental Setup

We have trained the transformer based cross language model (XLM) (Conneau and Lample, 2019) also known as MLM + TLM task. We have used 6 layers with 8 attention heads. An embedding layer is also used with size 256. Given the comparatively smaller Marathi dataset as discussed in Section 3.1, and limited availability of computational resources[2] we trained the smaller model instead of the usual 12 layers and 16 attention heads as proposed by Conneau and Lample (2019). Batch size of 32 was used. Following settings of Conneau and Lample (2019), attention dropout was set to 0.1, gelu activation was used. Also, adam was used as an optimizer with an initial learning rate of 0.0001. Rest of the parameters are same as used by Conneau and Lample (2019) in their experiments and as given in their GitHub repository[3].

6 Result and Analysis

The WMT20 organizer declared result for the shared task of similar language translation on Hindi to Marathi[4] and Marathi to Hindi[5] and the results of our system's is reported in Table 3. Our team's name is NITS-CNLP. The participated systems are evaluated by BLEU Papineni et al. (2002), RIBES Isozaki et al. (2010) and TER Snover et al. (2006) and the tracks are ranked by BLEU score. A total of 21 teams participated in Hindi to Marathi translation track and 23 teams for Marathi to Hindi translation track including both primary and contrastive system types. Our system's rank is 10 with BLEU score 11.59 for Hindi to Marathi translation

[1] https://github.com/glample/fastBPE

[2] The model was trained on a Quadro P200 GPU having 5GB of GPU RAM

[3] https://github.com/facebookresearch/XLM

[4] http://mzampieri.com/workshops/wmt/HI-MR.pdf

[5] http://mzampieri.com/workshops/wmt/MR-HI.pdf

Type	Source: Hindi Target: Marathi	
Short	Source Test Sentence	अवसरों की समानता है।
	Predicted Test Sentence	संधीची समानता आहे.
	Google Translation	संधींची समानता आहे.
	Bing Translation	संधींची समानता आहे.
Medium	Source Test Sentence	यह मेरे लिए एक बहुत ही सुखद अनुभूति रही है।
	Predicted Test Sentence	ही माझ्यासाठी अतिशय सुखद अनुभूती आहे.
	Google Translation	ही माझ्यासाठी खूप आनंददायी भावना आहे.
	Bing Translation	माझ्यासाठी ही खूप सुखद भावना आहे.
Long	Source Test Sentence	बल्कि यह एक सकारात्मक शांति है जहां हम सब करुणा और ज्ञान के आधार पर संवाद, सद्भाव और न्याय को बढ़ावा देने के लिए काम करते हैं।
	Predicted Test Sentence	ही एक सकारात्मक शांतता आहे जिथे आपण करुणा आणि ज्ञानाच्या आधारे संवाद सद्भावनेला प्रोत्साहन देण्यासाठी काम करतो.
	Google Translation	उलट ही एक सकारात्मक शांती आहे जिथे आपण सर्व करुणा आणि ज्ञानावर आधारित संवाद, सुसंवाद आणि न्यायाला चालना देण्यासाठी कार्य करीत आहोत.
	Bing Translation	उलट ही एक सकारात्मक शांती आहे जिथे आपण सर्वजण करुणा आणि ज्ञानावर आधारित संवाद, सामंजस्य आणि न्याय ाला प्रोत्साहन देण्याचे काम करतो.

Table 1: Best Performance examples for Hindi to Marathi translation.

Type	Source: Hindi Target: Marathi	
Long	Source Test Sentence	साथियो, जीएसटी की व्यवस्था को और सशक्त, और सरल करने के प्रयास लगातार चल रहे हैं।
	Predicted Test Sentence	मित्रांनो वस्तू आणि सेवा कर व्यवस्था अधिक सशक्त आणि सुलभ करण्याचे
	Google Translation	मित्रांनो, जीएसटी कारभारास आणखी बळकटी आणि सुलभ करण्यासाठी प्रयत्न सुरू आहेत.
	Bing Translation	मित्रांनो, जीएसटी प्रणाली अधिक सक्षम करण्यासाठी, सोपे करण्यासाठी प्रयत्न सुरू आहेत.

Table 2: Worst Performance examples for Hindi to Marathi translation.

Translation	System Type	BLEU	RIBES	TER
Hindi to Marathi	Primary	11.59	57.76	79.07
Marathi to Hindi	Primary	15.44	61.13	75.96

Table 3: Our system's results.

track and for Marathi to Hindi translation track, the rank is 15 with BLEU score 15.44 in primary configuration.

Analysis We have attained a lower BLEU score for Hindi to Marathi translation as compared to Marathi to Hindi translation as shown in Table 3. This is because we have used more Hindi monolingual corpus than Marathi monolingual corpus. As a result of this our NMT system encoded more frequency of Hindi words as compared to Marathi words and thus, decoder could be able to generate better target Hindi words than Marathi target words. To examine the best performance, we have considered sample source test sentences and corresponding predicted, Google[6], Bing[7] translated sentences for Hindi to Marathi translation in three different types of sentences such as short, medium and long sentences as shown in Table 1. Table 2 shows the worst performance of our NMT system in case of long type sentences. In Table 2, Google translation is better than our predicted test sentence and Bing translation.

7 Conclusion and Future Work

Our NMT system adopts cross lingual model for a similar language translation task of Hindi-Marathi pair in both forward and backward directions. The evaluated result and in-depth analysis of the predicted sentences shows that our NMT system performs well for the short and medium types of sentences and shows poor performance in long sentences. However, our NMT system needs more Marathi monolingual corpus and in the future works, multilingual NMT system will be developed to overcome the limitation of corpus for such low resource language pair translation.

Acknowledgement

Authors would like to thank WMT20 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.

[6]https://translate.google.co.in/

[7]https://www.bing.com/translator

References

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Franck Burlot and François Yvon. 2018. Using monolingual data in neural machine translation: a systematic study. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 144–155, Brussels, Belgium. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7059–7069.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. 2014. Fast and robust neural network joint models for statistical machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1370–1380, Baltimore, Maryland. 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. 2010. *Statistical Machine Translation*, 1st edition. Cambridge University Press, USA.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

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, 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.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

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.

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.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh neural machine translation systems for wmt 16. *arXiv preprint arXiv:1606.02891*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. 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.

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.

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.