Zero-Shot Neural Machine Translation: Russian-Hindi @LoResMT 2020

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_ug, abdullah_ug, partha}@cse.nits.ac.in,
sivaji.cse.ju@gmail.com

Abstract

Neural machine translation (NMT) is a widely accepted approach in the machine translation (MT) community, translating from one natural language to another natural language. Although, NMT shows remarkable performance in both high and low resource languages, it needs sufficient training corpus. The availability of a parallel corpus in low resource language pairs is one of the challenging tasks in MT. To mitigate this issue, NMT attempts to utilize a monolingual corpus to get better at translation for low resource language pairs. Workshop on Technologies for MT of Low Resource Languages (LoResMT 2020) organized shared tasks of low resource language pair translation using zero-shot NMT. Here, the parallel corpus is not used and only monolingual corpora is allowed. We have participated in the same shared task with our team name CNLP-NITS for the Russian-Hindi language pair. We have used masked sequence to sequence pre-training for language generation (MASS) with only monolingual corpus following the unsupervised NMT architecture. The evaluated results are declared at the LoResMT 2020 shared task, which reports that our system achieves the bilingual evaluation understudy (BLEU) score of 0.59, precision score of 3.43, recall score of 5.48, F-measure score of 4.22, and rank-based intuitive bilingual evaluation score (RIBES) of 0.180147 in Russian to Hindi translation. And for Hindi to Russian translation, we have achieved BLEU, precision, recall, F-measure, and RIBES score of 1.11, 4.72, 4.41, 4.56, and 0.026842 respectively.

1 Introduction

The end-to-end recurrent neural network (RNN) based NMT (Cho et al., 2014b,a) approach attracts attention in MT because it deals with many challenges like variable-length phrases using sequence to sequence learning concept, long-term dependency problem adopting long short term memory (LSTM) (Sutskever et al., 2014), attention mechanism (Bahdanau et al., 2015; Luong et al., 2015) which pays attention globally and locally to all source words. The RNN based NMT approach is not able to process all the input words parallelly, to solve parallelization transformer-based NMT (Vaswani et al., 2017) is proposed by using a self-attention mechanism. Despite modifying NMT architecture, it needs reasonable parallel training data which is a challenge for low resource language pair translation. Generally, language pairs can be considered as low-resource when training data is less than a million (Kocmi, 2020). For low resource language pair translation, pivot-based NMT (Kim et al., 2019) is an effective approach where an intermediate language is considered as a pivot language (source to pivot and pivot to target). (Johnson et al., 2017) introduced a zero-shot approach to language pair translation without considering the parallel data using multilingual-based NMT. In this paper, we have participated in the LoResMT 2020 shared task of zero-shot NMT approach on Russian-Hindi pair using the only monolingual corpus and the same has been implemented using MASS-based unsupervised NMT (Song et al., 2019). The reason behind choosing MASS-based unsupervised NMT is that it achieves state-of-the-art performance on the unsupervised English-French pair translation.

2 Related Work

There is a lack of background work on Russian-Hindi translation. However, the literature survey finds work on unsupervised NMT using MASS (Song et al., 2019) which outperform previous unsupervised approaches (Lample and Conneau, 2019; Lample et al., 2018). (Song et al., 2019) without us-

Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages, pages 38–42
Decmber 04, 2020. ©2020 Association for Computational Linguistics

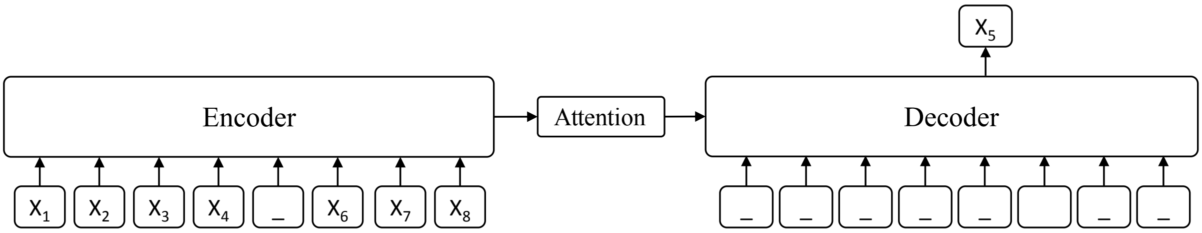


Figure 1: The encoder-decoder framework of the MASS model used (as adopted from (Song et al., 2019))

Type	Language	Sentences	Tokens
Train	Hindi	473,605	7,092,870
	Russian	154,589	1,007,029
Valid	Hindi	500	2,538
	Russian	500	2,000
Test	Hindi	500	3,150
	Russian	500	9,057

Table 1: Data Statistics provided by the LoResMT 2020 shared task organizer (Ojha et al., 2020)

ing bilingual corpus and only utilizing monolingual data achieves BLEU score 37.50, 34.90 on English to French and French to English translation and for English to German and German to English, it attains BLEU score 28.30, 35.20 respectively. Transformer architecture (Vaswani et al., 2017) based MASS implemented in two steps: pre-training on the monolingual data and then fine-tuning with the self-generated back translation data which acts as a pseudo bilingual corpus during the training process.

3 Dataset Description

The LoResMT 2020 shared task organizer (Ojha et al., 2020) provided the Russian-Hindi monolingual dataset of train, valid, and test sets, which is summarized in Table 1. Additionally, we have used external monolingual data set of Hindi (9 GB) from IITB[1] (Kunchukuttan et al., 2018; Bojar et al., 2014) and Russian (9GB) from WMT16[2].

4 System Description

We have adopted MASS based unsupervised NMT (Song et al., 2019) to build our system on a single GPU. Our system consists of two major steps namely the pre-training and then the fine-tuning step which are discussed in the sub-sections 4.1 and 4.2. For BPE (Sennrich et al., 2016) and vocabulary creation, we have used the cross-language model (XLM) (Lample and Conneau, 2019) codebase as given in their repository[3]. Moses is used for tokenization (Koehn and Hoang, 2010). The MASS (Song et al., 2019) based model leverages encode-decoder framework to develop complete sentences from given fractured pieces of sentences as shown in Figure 1. The model details are further described in Section 4.1 and 4.2, where we have shown the pre-training and fine tuning step respectively.

4.1 Pre-training

For the pre-training step, following (Song et al., 2019) we have undertaken the log likelihood objective function (LF) as shown in Equation 1. Here, s belongs to the source sentence corpus S. And in a particular sentence s, the region from u to v is masked, such that the sentence length remains constant.

$$\begin{aligned} LF(\theta;\mathcal{S}) &= \frac{1}{|\mathcal{S}|}\Sigma_{s\in\mathcal{S}} \log P(s^{u:v}|s^{\setminus u:v};\theta) \\ &= \frac{1}{|\mathcal{S}|}\Sigma_{s\in\mathcal{S}} \log \prod_{t=u}^{v} P(s_t^{u:v}|s_{<t}^{u:v}, s^{\setminus u:v};\theta). \end{aligned} \tag{1}$$

Here, the seq2seq model learns the parameter θ to compute the conditional probability. t denotes the word position.

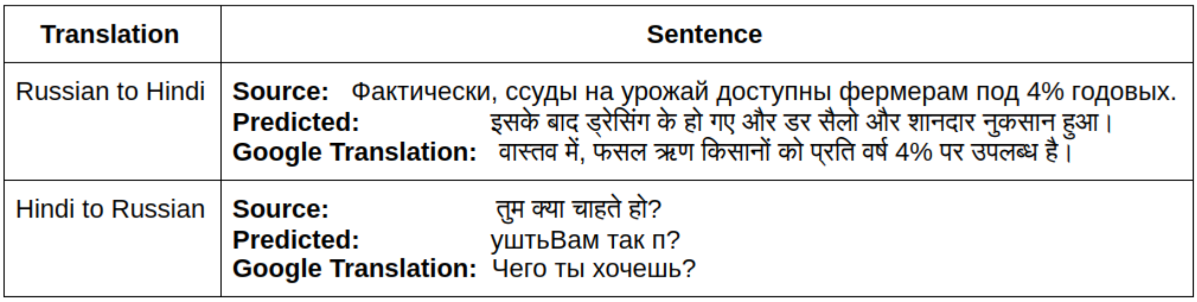
4.2 Fine Tuning

Since, the parallel data is not made available by the LoResMT 2020 organizers for this specific task, we have undertaken the unsupervised approach as followed by (Song et al., 2019). Only the monolingual data is used here. Here simply back-translation is employed to generate pseudo bilin-

[1] http://www.cfilt.iitb.ac.in/iitb_parallel/

[2] http://www.statmt.org/wmt16/translation-task.html

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



Translation	Sentence
Russian to Hindi	**Source:** Фактически, ссуды на урожай доступны фермерам под 4% годовых. **Predicted:** इसके बाद ड्रेसिंग के हो गए और डर सैलो और शानदार नुकसान हुआ। **Google Translation:** वास्तव में, फसल ऋण किसानों को प्रति वर्ष 4% पर उपलब्ध है।
Hindi to Russian	**Source:** तुम क्या चाहते हो? **Predicted:** уштьВам так п? **Google Translation:** Чего ты хочешь?

Figure 2: Example Sentences of Translation.

Experiment	Task	BLEU	Precision	Recall	F-measure	RIBES
Russian to Hindi	Ru2Hi-MASS-a	0.51	3.19	4.83	3.84	0.129554
Russian to Hindi	Ru2Hi-MASS-c	0.59	3.43	5.48	4.22	0.180147
Hindi to Russian	Hi2Ru-MASS-a	0.59	4.48	4.23	4.35	0.025767
Hindi to Russian	Hi2Ru-MASS-c	1.11	4.72	4.41	4.56	0.026842

Table 2: Results of our systems

gual corpus for the training step. Auto encoder with denoising is not used. Initial learning rate is 0.0004 together with Adam optimizer.

5 Experimental Setup

During pre-processing of the data, following (Song et al., 2019) and using the code provided by (Lample and Conneau, 2019), we used fastBPE[4] to learn byte pair encoding (BPE) vocabulary with 50,000 codes. Also, for leveraging the model features, we have followed the settings of (Song et al., 2019). In the pre-training step, we have followed the default settings of Transformer model-based Mass (Song et al., 2019), where 6 layers with 8 attention heads are used. Due to limited computational resources, we have used 256 embedding layers with batch size 32, tokens per batch 500 and dropout 0.1. The obtained pre-trained model from 4.1 are fine-tuned with pseudo bilingual corpus through self-generated back-translation data following default settings of (Song et al., 2019).

6 Result and Analysis

The LoResMT 2020 shared task organizer declared the evaluation result[5] of zero-shot NMT on the language pairs namely, Hindi-Bhojpuri, Hindi-Magahi, and Russian-Hindi, and participated by two teams only. For the Russian-Hindi language pair, only our team participated and our team name is CNLP-NITS. The results are evaluated using automatic evaluation metrics, BLEU (Papineni et al., 2002), precision, recall, F-measure and RIBES (Isozaki et al., 2010). We have submitted two systems result, one only using provided monolingual data (extension -a) and another with external monolingual data addition of provided monolingual data (extension -c) and the same have been reported in Table 2. From Table 2, it is observed that our scores are very low. However, it is to be noted that with increasing monolingual data, the performance of our systems improves. Moreover, from the predicted translation as shown in Figure 2, it is quite clear that the translation accuracy is very poor in terms of adequacy but better in the fluency factor of translation. To achieve better translation accuracy, we need to improve both adequacy as well as fluency of predicted translations. In this work, we have used the default tokenizer i.e. Moses. In future, we will use IndicNLP tokenizer (Kunchukuttan, 2020). This tokenizer is specifically designed for Indic languages, in order to improve the overall performance of predictive models in Hindi languages.

7 Conclusion and Future Work

This paper presents a zero-shot NMT task on the Russian ⇔ Hindi translation, this system was used to participate in the LoResMT 2020 shared task. We have used unsupervised NMT approach of MASS (Song et al., 2019) to build a single model that can translate in both the directions i.e. Russian to Hindi and vice-versa. The obtained scores and

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

[5] https://bit.ly/3l2Hc1h

closely observed predicted output remarks that our future works require significant improvement to achieve better translation accuracies in both directions.

Acknowledgement

We would like to thank Center for Natural Language Processing (CNLP) and Department of Computer Science and Engineering at National Institute of Technology, Silchar, India for providing the requisite support and infrastructure to execute this work. Also, thank to LoResMT 2020 shared task organizers.

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*.

Ondřej Bojar, Vojtěch Diatka, Pavel Rychlý, 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, Reykjavik, Iceland. European Language Resources Association (ELRA).

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014a. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014b. 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.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Yunsu Kim, Petre Petrov, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2019. Pivot-based transfer learning for neural machine translation between non-English languages. 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 866–876, Hong Kong, China. Association for Computational Linguistics.

Tom Kocmi. 2020. Exploring benefits of transfer learning in neural machine translation.

Philipp Koehn and Hieu Hoang. 2010. Moses. *Statistical Machine Translation System, User Manual and Code Guide*, page 245.

Anoop Kunchukuttan. 2020. The IndicNLP Library. `https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf`.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. The IIT Bombay English-Hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium. 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.

Atul Kr. Ojha, Valentin Malykh, Alina Karakanta, and Chao-Hong Liu. 2020. Findings of the loresmt 2020 shared task on zero-shot for low-resource languages. In *"Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages"*. 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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. In *International Conference on Machine Learning*, pages 5926–5936.

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, pages 3104–3112, Cambridge, MA, USA. MIT Press.

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 *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.