Unsupervised Approach for Zero-Shot Experiments: Bhojpuri–Hindi and Magahi–Hindi@LoResMT 2020

Amit Kumar and Rajesh Kumar Mundotiya and Anil Kumar Singh
IIT (BHU), Varanasi
{amitkumar.rs.cse17, rajeshkm.rs.cse16, aksingh.cse}@iitbhu.ac.in

Abstract

This paper reports a Machine Translation (MT) system submitted by the NLPRL team for the Bhojpuri–Hindi and Magahi–Hindi language pairs at LoResMT 2020 shared task. We used an unsupervised domain adaptation approach that gives promising results for zero or extremely low resource languages. Task organizers provide the development and the test sets for evaluation and the monolingual data for training. Our approach is a hybrid approach of domain adaptation and back-translation. Metrics used to evaluate the trained model are BLEU, RIBES, Precision, Recall and F-measure. Our approach gives relatively promising results, with a wide range, of 19.5, 13.71, 2.54, and 3.16 BLEU points for Bhojpuri to Hindi, Magahi to Hindi, Hindi to Bhojpuri and Hindi to Magahi language pairs, respectively.

1 Introduction

For the past few years, Neural Machine Translation (NMT) (Sutskever et al., 2014) has been the favoured approach for both high resource and low resource languages translation. A large number of variations of the Transformer (Vaswani et al., 2017) model, such as BERT (Devlin et al., 2018), BART (Lewis et al., 2019), MBART (Liu et al., 2020) have been developed, giving state-of-the-art result for translation. However, successful NMT is only possible if a large parallel corpus is available to train the model. However, not so much parallel corpus is available for extremely low resource languages and there are still many languages in which no parallel corpora are available.

Creating data, even parallel corpus, for extremely low resource languages is a time and labour consuming process. Instead of creating the data from scratch for such low resource languages, there is a possibility to use the available resources and tools of high resource languages for creating the resources and tools for low resource languages.

In this paper, we use the available Nepali–Hindi low resource data for domain adaptation (Chu and Wang, 2018), followed by a back-translation (Hoang et al., 2018) approach for creating the synthetic dataset and NMT tools for Bhojpuri–Hindi and Magahi–Hindi zero-shot languages.

We evaluate our method using a Transformer-based NMT system (Vaswani et al., 2017; Kumar and Singh, 2019) for domain adaptation and back-translation of monolingual data. Our approach gives new state-of-the-art results of 19.5, 13.71, 2.54, and 3.16 BLEU points for Bhojpuri → Hindi, Magahi → Hindi, Hindi → Bhojpuri and Hindi → Magahi language pairs, respectively.

2 Background

Ojha (2019) describe the resources created for Bhojpuri languages and it also covers some results on Bhojpuri–English SMT systems. In 2019, LoResMT (Karakanta et al., 2019) organized the shared task for Bhojpuri–English and Magahi–English language pairs. However, previous works contains some parallel sentences. The work in this paper is primarily based on unsupervised domain adaptation (Miller, 2019), followed by back-translation of monolingual data (Guzmán et al., 2019). Unsupervised domain adaptation is the task of training a model on labelled data from a source domain to achieve better performance on data from a target domain, with access to only unlabeled data in the target domain. Suppose $D_S = \{X, Y\}$ is domain of source task, where X is a source instance, and Y is a labelled instance, and $D_T = \{Z\}$ is the target domain, where Z is an unlabeled instance. Then, the goal of unsupervised domain adaptation is to improve the accuracy of the target domain with the help of the source domain, where source

Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages, pages 43–46
Decmber 04, 2020. ©2020 Association for Computational Linguistics

and target domains are two different tasks that are related to each other in some relevant way.

We apply domain adaptation to similar languages. Hindi, Nepali, Bhojpuri and Magahi are similar languages and they are also orthographically similar [1] (Mundotiya et al., 2020). All these languages use the Devanagari script and they share a lot of cognates and loan words among each other. This is the main motivation behind using unsupervised domain adaptation on similar languages in our paper.

We use back-translation to increase the parallel corpus that helps in improving the BLEU points (Papineni et al., 2002) for the translation task.

3 Methodology

We used an unsupervised hybrid approach to train the model for zero parallel training data.

3.1 Domain adaptation using similar languages

We used Nepali–Hindi language pairs to train the model in all translation directions. The motive behind using Nepali–Hindi language pairs is the relatedness between this language pair with Bhojpuri–Hindi and Magahi–Hindi language pairs. Here, we use language pairs as one domain and evaluate it on another domain. So, we use directly trained model on Bhojpuri–Hindi and Magahi–Hindi language pairs to evaluate the model.

3.2 Two-iteration back-translation using monolingual data

For Bhojpuri to Hindi, we used Hindi→Nepali trained model to translate the monolingual Hindi data to Nepali. Then we append this predicted parallel data with Nepali→Hindi and trained the Bhojpuri→Hindi model and evaluate the model with Bhojpuri→Hindi development and test sets provided by Organizer.

For Hindi to Bhojpuri, we used Nepali→Hindi trained model to translate the monolingual Bhojpuri data to Hindi. Then we append this predicted parallel data with Hindi→Nepali and trained the Hindi→Bhojpuri model and evaluated the model with Hindi→Bhojpuri development and test set provided by the shared task organizers.

Similarly, we repeat the above steps for Magahi–Hindi language pairs.

[1] https://nepalgo.de/post/121994474120/guffgaff-nepali-vs-hindi

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

4 Corpus Description

We used the Nepali–Hindi language pairs collected from WMT 2019[2] similar language translation task, as well as from Opus[3] and TDIL[4]. There are a total of 136991 Nepali–Hindi language parallel sentences with 3000 development sets. We also used the 91131 Bhojpuri, 148606 Magahi, and 473605 Hindi monolingual sentences provided by LoResMT 2020 shared task (Ojha et al., 2020) organizers for Zero-shot purpose. Finally, we used 500 size development set and 500 size test set provided by LoResMT 2020 shared task organizers to evaluate the data.

5 Experimental Settings

We used fairseq (Ott et al., 2019) sequence modelling toolkit for training the transformer-based NMT model. The hyper-parameter settings used in the paper are described in table 1.

6 Results

The shared task organizers used BLEU (Papineni et al., 2002), RIBES (Isozaki et al., 2010), Precision, Recall and F-measure metrics to evaluate the performance of the system which is shown in figure 1 and table 2. We use unsupervised domain adaptation approach followed by Back-translation for Bhojpuri → Hindi and Magahi → Hindi language pairs, whereas for Hindi → Bhojpuri and Hindi

[2] http://www.statmt.org/wmt19/

[3] http://opus.nlpl.eu/

[4] https://www.tdil-dc.in/index.php?lang=en

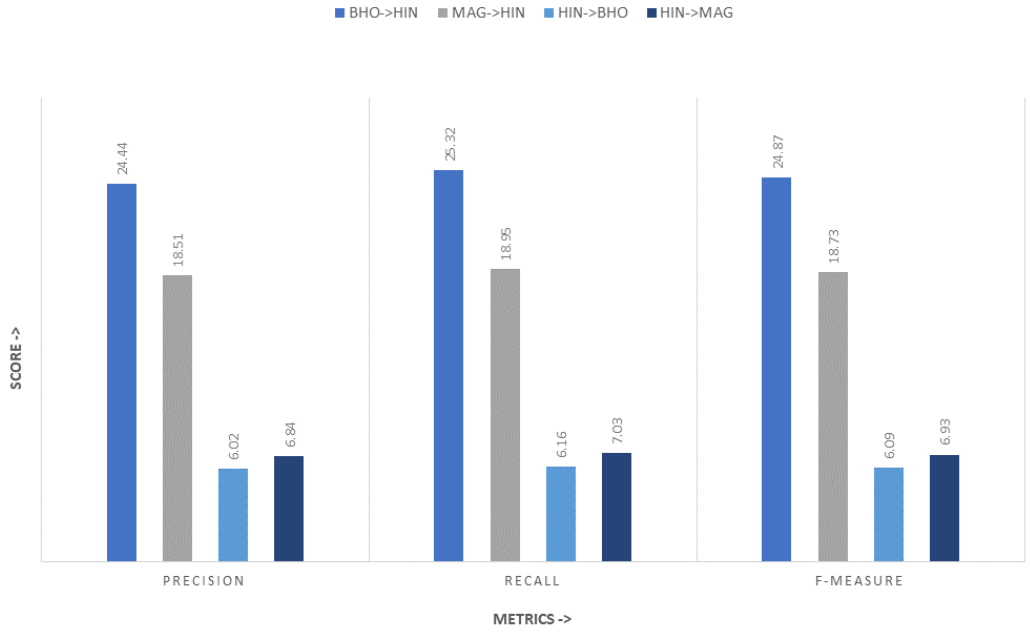

Figure 1: Bar diagram represents Precision, Recall, and F-measure

System	BLEU	RIBES
Bhojpuri → Hindi	19.5	0.794593
Magahi → Hindi	13.71	0.714497
Hindi → Bhojpuri	2.54	0.031503
Hindi → Magahi	3.16	0.040177

Table 2: Scores of our system evaluated by task organizers

System	BLEU	RIBES
Bhojpuri → Hindi	17.8	0.779195
Magahi → Hindi	12.58	0.691139
Hindi → Bhojpuri	2.61	0.030563
Hindi → Magahi	2.89	0.035748

Table 3: Scores of baseline (Vanilla Transformer) system

→ Magahi language pairs, we used only unsupervised domain adaptation without back-translation because for these pairs translations degraded the performance with back-translations.

7 Conclusion

Our results show that by using domain adaptation for similar languages, an unsupervised approach gives varying degrees of improvement for translation of zero-shot languages. We note that similar language data and models can play a significant role in future for creating the resources and tools for extremely low or zero resource languages. We also used back-translation of monolingual data that also gives an improvement in scores. This work is evaluated on the dataset provided by LoResMT 2020 shared task organizers. Evaluation is performed using the metrics BLEU, RIBES, Precision, Recall and F-measure. Results are reproducible using the publicly available fairseq toolkit.

Acknowledgments

The support and the resources provided by PARAM Shivay Facility under the National Supercomputing Mission, Government of India at the Indian Institute of Technology, Varanasi are gratefully acknowledged.

References

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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*.

F Guzmán, PJ Chen, M Ott, J Pino, G Lample, P Koehn, V Chaudhary, and M Ranzato. 2019. Two new evaluation datasets for low-resource machine translation: Nepali-English and Sinhala-English. arxiv 2019. *arXiv preprint arXiv:1902.01382*.

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.

Alina Karakanta, Atul Kr. Ojha, Chao-Hong Liu, Jonathan Washington, Nathaniel Oco, Surafel Melaku Lakew, Valentin Malykh, and Xiaobing Zhao, editors. 2019. *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*. European Association for Machine Translation, Dublin, Ireland.

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.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation.

Timothy Miller. 2019. Simplified neural unsupervised domain adaptation. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2019, page 414. NIH Public Access.

Rajesh Kumar Mundotiya, Manish Kumar Singh, Rahul Kapur, Swasti Mishra, and Anil Kumar Singh. 2020. Basic linguistic resources and baselines for bhojpuri, magahi and maithili for natural language processing. *arXiv preprint arXiv:2004.13945*.

Atul Kr. Ojha. 2019. English-bhojpuri smt system: Insights from the karaka model.

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.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*.

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.

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.