Findings of the LoResMT 2020 Shared Task on Zero-Shot for Low Resource languages

Atul Kr. Ojha[1,5], Valentin Malykh[2], Alina Karakanta[3],Chao-Hong Liu[4]
[1]Data Science Institute, NUIG, Galway, [2]Huawei Noah's Ark lab & Kazan Federal University,
[3]Fondazione Bruno Kessler / University of Trento, [4]Iconic Translation Machines, RWS Group,
[5]Panlingua Language Processing LLP, New Delhi
atulkumar.ojha@insight-centre.org,valentin.malykh@phystech.edu,
akarakanta@fbk.eu,ch.liu@acm.org

Abstract

This paper presents the findings of the LoResMT 2020 Shared Task on zero-shot translation for low resource languages. This task was organised as part of the 3rd Workshop on Technologies for MT of Low Resource Languages (LoResMT) at AACL-IJCNLP 2020. The focus was on the zero-shot approach as a notable development in Neural Machine Translation to build MT systems for language pairs where parallel corpora are small or even non-existent. The shared task experience suggests that back-translation and domain adaptation methods result in better accuracy for small-size datasets. We further noted that, although translation between similar languages is no cakewalk, linguistically distinct languages require more data to give better results.

1 Introduction

Research and development in Statistical and Neural Machine Translation has rapidly emerged over in the last one decade especially after the availability of several open source machine translation (MT) toolkits like: Moses (Koehn et al., 2007), OpenNMT (Klein et al., 2017), Nematus (Sennrich et al., 2017), Marian (Junczys-Dowmunt et al., 2018), etc. For the past few years, researchers, developers, users and commercial organizations are widely using Neural Machine Translation (NMT) (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Luong et al., 2015; Bahdanau et al., 2016; Vaswani et al., 2017) to enhance the performance of their MT systems. NMT has become the preferred paradigm due to its ability to produce better results. Despite the multiple advantages of using NMT and Statistical Machine Translation (SMT) methods, several challenges are also encountered, the main challenge being the lack of quality data for training the systems. Both SMT and NMT require large-sized parallel corpora. However, out of 7,117 languages,[1] most languages are low resourced or even endangered. This particular challenge has long posed a strong impediment for developing NMT systems for low resource languages (Koehn and Knowles, 2017).

Overcoming this obstacle is an extensive and tedious process. The preparation of a parallel corpus in any language can be a very costly procedure that demands the expertise of language professionals at several levels. It is therefore necessary to exploit the existing resources to build MT systems in low resource languages. Methodologies such as data augmentation, exploitation of monolingual data, cross-lingual transfer etc. are preferred approaches under the aforesaid circumstances.

In the preceding year, a shared task (Karakanta et al., 2019) was organised where a monolingual and parallel corpus for the low-resource languages Bhojpuri, Magahi, Sindhi, and Latvian was provided to create NMT/SMT systems[2]. This year's workshop in an extension to the same objective as last year, but this time the team has focused on the zero-shot approach (Firat et al., 2016) for building quality MT systems. In addition to pivot MT, the zero-shot approach is one notable development in NMT to build MT systems for language pairs where parallel corpora are small or even non-existent. However, the performance of zero-shot NMT is low compared to pivot MT in general. In this paper, we discuss the results of the LoResMT 2020 shared task, organised as part of the 3rd Workshop on Technologies for MT of Low Resource Languages (LoResMT)[3] at AACL-IJCNLP 2020[4]. In this task, we solicited participants to submit

[1] https://www.ethnologue.com/guides/how-many-languages
[2] https://sites.google.com/view/loresmt/loresmt-2019
[3] https://sites.google.com/view/loresmt/
[4] http://aacl2020.org/

Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages, pages 33–37

novel zero-shot NMT systems for the following language pairs:

- Hindi↔Bhojpuri
- Hindi↔Magahi
- Russian↔Hindi

The remaining paper is organized as follows. Section 2 presents the setup and schedule of the LoResMT 2020 shared task and Section 3 presents the dataset used in the competition. Section 4 describes the approaches used by participants of the competition and Section 5 presents and analyzes the results they obtained. Finally, 6 concludes this paper and presents avenues for future work.

2 Task Setup and Schedule

Based on a detailed call for participation, researchers were asked to register themselves. The choice of language pair was left to the participants. These registered participants were sent the links to the training (train) dataset including monolingual and development (dev) data, along with a description of the format and statistics of the dataset. They were allowed to use only additional monolingual data to train the system, with the condition that the additional monolingual dataset should be publicly available. Moreover, participants were allowed to use pretrained word embeddings, and publicly-existing linguistic models. The participants were given 24 days to experiment and develop a system. After this period, the test set was released and the participants had 5 days to test and upload their system using the following abbreviations:

- *"-a"* - Only provided development and monolingual corpora.
- *"-b"*- Any provided corpora, plus publicly available different/similar language's monolingual corpora and/or pretrained/linguistics model (e.g. systems used pretrained word2vec, UDPipe, etc. model).
- *"-c"* - Any provided corpora, plus any publicly external monolingual corpora.

The complete timeline of the shared task is given in Table 1.
Each team was allowed to submit any number of systems for evaluation and their best 3 systems were included in the final ranking presented in this report. Each submitted system was evaluated on standard automatic MT evaluation metrics; BLEU (Papineni et al., 2002), Precision, Recall, F-measure and RIBES (Isozaki et al., 2010).

Date	Event
August 15, 2020	Announcement & registration
September 15, 2020	Train (monolingual & dev set release)
October 09, 2020	Test set release
October 13, 2020	MT system submission due
October 17, 2020	Results announcement
October 24, 2020	System description paper

Table 1: Schedule of LoResMT Shared Task

3 Datasets

The dataset of this shared task comprises three domains: news, subtitling and/or literature. Details of the collected sources are described below:

- **Monolingual dataset**: Bhojpuri data was extracted from Wikipedia and online newspapers (Ojha, 2019). Magahi data was collected from blogs (Kumar et al., 2018). Russian data was extracted from the Opensubtitles (OPUS)[5] website. Hindi data was compiled from Wikipedia, pmindia (Haddow and Kirefu, 2020) and OPUS.
- **Dev and test dataset**: Each language pair's dev and test dataset was built on monolingual data which were manually translated and validated by professional translators, native speakers of the target languages.

The participants of the shared task were provided with more than one million words of monolingual data for each language pair, while 500 manually translated and validated parallel sentences were provided for dev and test set. The complete shared task datasets are available at GitHub[6]. The detailed statistics of the dataset in each language is provided in Table 2.

Language	Sentences	Words	Characters
Bhojpuri	91131	1562465	20002174
Hindi	473605	7092870	86982827
Magahi	148606	2178424	25692432
Russian	154589	1007029	8261212

Table 2: Statistics of the monolingual data of Bhojpuri, Hindi, Magahi and Russian

[5] http://opus.nlpl.eu/tools.php
[6] https://github.com/panlingua/loresmt-2020

Team	Hindi↔Bhojpuri	Hindi↔Magahi	Hindi↔Russian	System Description
CNLP-NITS	-	-	✓	(Laskar et al., 2020)
IIT(BHU)-NLPRL	✓	✓	✓	-
NLPRL	✓	✓	-	(Kumar et al., 2020)
SU-NLP	✓	✓	✓	-
vandan mujadia	✓	✓	✓	-
Total	**4**	**4**	**4**	**2**

Table 3: Details of the participated teams in the LoResMT 2020 Shared Task

System	Task description	BLEU	PRECISION	RECALL	F-MEASURE	RIBES
Bhojpuri-Hindi	Bho2Hi-Transformer-b	19.5	24.44	25.32	24.87	0.79
Magahi-Hindi	Mag2Hi-Transformer-b	13.71	18.51	18.95	18.73	0.71
Russian-Hindi	Ru2Hi-MASS-a	0.51	3.19	4.83	3.84	0.12
	Ru2Hi-MASS-c	0.59	3.43	5.48	4.22	0.18
Hindi-Bhojpuri	Hi2Bho-Transformer-b	2.54	6.02	6.16	6.09	0.03
Hindi-Magahi	Hi2Mag-Transformer-b	3.16	6.84	7.03	6.93	0.04
Hindi-Russian	Hi2Ru-MASS-a	0.59	4.48	4.23	4.35	0.025
	Hi2Ru-MASS-c	1.11	4.72	4.41	4.56	0.02

Table 4: Result of submitted systems at Bhojpuri, Hindi, Magahi and Russian

4 Participants and Methodology

A total of 5 participants registered for the shared task, with most of the teams registering to participate for Hindi↔Bhojpuri and Hindi↔Magahi language pair except 1 team (see table 3). Out of these, finally a total of 6 systems were submitted by CNLP-NITS and NLPRL teams. All the teams who submitted their system were invited to submit the system description paper, describing the experiments conducted by them. Table 3 lists the participating teams and the language they took part in.

Next, we give a short description of the approach taken by each team for building their system(s). More details about the approaches can be found in the papers submitted by the respective teams.

- **CNLP-NITS** (Laskar et al., 2020) uses unsupervised masked sequence-to-sequence pre-training for language generation (MASS) (Song et al., 2019) for Hindi-Russian and Russian-Hindi language pair. They used additional Hindi and Russian monolingual data on the same method. This system is submitted as a constrained system.

- **NLPRL** (Kumar et al., 2020) uses unsupervised domain adaptation and back-translation for Hindi-Bhojpuri, Bhojpuri-Hindi, Hindi-Magahi and Magahi-Hindi using similar Hindi-Nepali data.

5 Results

As previously mentioned, the participants were allowed to use monolingual datasets, other than that provided. However, due to the lack of a similar substitute monolingual dataset for Bhojpuri and Magahi, participants used only one of the provided data by the shared task organisers. The NLPRL team used orthographically similar Hindi-Nepali data to build their system. On the other hand, the CNLP-NIST team only used additional Hindi and Russian monolingual data for the constrained system submission. As mentioned earlier, for the evaluation of the system, 500 sentences were given to the participants in each language pair for each direction.

The results of the participating teams on Hindi-Bhojpuri, Hindi-Magahi and Hindi-Russian language pairs is presented in Table 4.

6 Conclusion

In this paper, we have reported the findings of the LoResMT 2020 Shared Task on zero-shot translation for low resource languages, organized as part of the 3rd LoResMT workshop at AACL-IJCNLP 2020. All the systems submitted used the unsupervised method. We conclude that the use of domain adaptation and back translation methods provides better results for MT system training where the datasets are small-sized. Another concluding point is that the Masked sequence-to-sequence pre-

training method provides comparatively low performance on all measures: BLEU, Precision, Recall, F-measure and RIBES. Bhojpuri to Hindi has provided better accuracy scores than vice versa for both the teams who selected Bhojpuri and Hindi as their language pairs. The systems trained for Hindi and Russian did not provide desired results in any language direction, despite them having larger datasets than the other two languages in the shared task. This understanding should be accompanied with the knowledge that Russian and Hindi are completely dissimilar languages belonging to separate language families. We also believe that a human evaluation could provide better insights than automatic evaluation metrics. In the next version of the Shared Task, we are planning to introduce human evaluation of the systems, in order to extend and improve the findings of our Shared Task on low resource languages.

7 Acknowledgements

This publication has emanated from research in part supported by the EU H2020 programme under grant agreements 731015 (ELEXIS-European Lexical Infrastructure). We are also grateful to Panlingua Language Processing LLP to provide Hindi, Bhojpuri, Magahi monolingual and parallel corpora.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate.

Orhan Firat, Baskaran Sankaran, Yaser Al-onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277, Austin, Texas. Association for Computational Linguistics.

Barry Haddow and Faheem Kirefu. 2020. Pmindia – a collection of parallel corpora of languages of india.

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.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent convolutional neural networks for discourse compositionality. *arXiv preprint arXiv:1306.3584*.

Alina Karakanta, Atul Kr Ojha, Chao-Hong Liu, Jonathan Washington, Nathaniel Oco, Surafel Melaku Lakew, Valentin Malykh, and Xiaobing Zhao. 2019. Proceedings of the 2nd workshop on technologies for mt of low resource languages. In *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*.

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.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. 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.

Amit Kumar, Rajesh Kumar Mundotiya, and Anil Kumar Singh. 2020. Unsupervised approach for zero-shot experiments: Bhojpuri–Hindi and Magahi–Hindi@loresmt 2020. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 43–46, Suzhou, China. Association for Computational Linguistics.

Ritesh Kumar, Bornini Lahiri, Deepak Alok, Atul Kr. Ojha, Mayank Jain, Abdul Basit, and Yogesh Dawar. 2018. Automatic identification of closely-related Indian languages: Resources and experiments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji, Partha Pakray, and Sivaji Bandyopadhyay. 2020. Zero-shot neural machine translation: Russian-Hindi @loresmt 2020. In *Proceedings of the 3rd Workshop on Technologies for MT of Low*

Resource Languages, pages 38–42, Suzhou, China. Association for Computational Linguistics.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Atul Kr. Ojha. 2019. *English-Bhojpuri SMT System: Insights from the Karaka Model*. Ph.D. thesis, Ph D thesis, Jawaharlal Nehru University, New Delhi, India.

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.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, et al. 2017. Nematus: a toolkit for neural machine translation. *arXiv preprint arXiv:1703.04357*.

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*.

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, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.