NLPRL System for Very Low Resource Supervised Machine Translation

Rupjyoti Baruah, Rajesh Kumar Mundotiya, Amit Kumar, Anil Kumar Singh
Department of Computer Science & Engineering
Indian Institute of Technology (BHU)
Varanasi, India
{rupjyotibaruah.rs.cse18, rajeshkm.rs.cse16}@iitbhu.ac.in
{amitkumar.rs.cse17, aksingh.cse}@iitbhu.ac.in

Abstract

This paper describes the results of the system that we used for the WMT20 very low resource (VLR) supervised MT shared task. For our experiments, we use a byte-level version of BPE, which requires a base vocabulary of size 256 only. BPE based models are a kind of sub-word models. Such models try to address the Out of Vocabulary (OOV) word problem by performing word segmentation so that segments correspond to morphological units. They are also reported to work across different languages, especially similar languages due to their sub-word nature. Based on BLEU cased score, our NLPRL systems ranked ninth for HSB to GER and tenth in GER to HSB translation scenario.

1 Introduction

We report the results for our system that was used for our participation in the WMT20 shared task (Barrault et al., 2019) on very low resource Machine Translation (MT). The MT systems were built for the language pair Upper Sorbian (HSB) and German (GER) in both translation directions.

The Sorbian languages are the West Slavic branch of the Indo-European languages, which have further categorized into two closely related languages, Upper Sorbian and Lower Sorbian. The categories of this language are recognized as a different and distinct language in the European Charter for Regional or Minority languages (Dolowy-Rybinska, 2011). Upper Sorbian is a minority language of Germany that is spoken by $10,000$ to $15,000$ speakers (Elle, 2010), although this number is continually declining (Dołowy-Rybińska, 2018). To counter this, attempts are being made to increase the number of Sorbian speakers through bilingual educational scenarios and MT[1].

Low resource MT was being attempted even before Neural Machine Translation (NMT) became the state-of-the-art. Several methods are used to improve the accuracy and quality of the low-resource SMT systems by using comparable corpora (Irvine and Callison-Burch, 2013; Babych et al., 2007), pivot language (English or non-English) technique (Ahmadnia et al., 2017; Paul et al., 2013), and using related resource-rich language (Nakov and Ng, 2012).

We use a byte-level version of Byte Pair Encoding based model with a Transformer for our experiments. The main motivation was to try out this model for the shared task and see how it works under a shared task setting.

2 Background

NMT is an end-to-end learning system (Bahdanau et al., 2015), based on the data-driven approach of machine translation, that requires a massive amount of parallel data for training.

To overcome the lack of such data, several techniques have been tried out which are based on semi-supervised learning (Zheng et al. (2019)), unsupervised learning (Sun et al. (2020)), data augmentation (Siddhant et al. (2020)), transfer learning (Aji (2020)), meta-learning (Li et al. (2020)), pivot-based (Kim et al. (2019)), and multilingual machine translation (Dabre et al. (2020)).

A model-agnostic meta-learning algorithm (Finn et al., 2017) for low-resource NMT exploits the multilingual high-resource language tasks (Gu et al., 2018b). Gu et al. (2018a) achieved significant improvement in performance by utilizing a transfer-learning approach for extremely low resource languages.

Another proposed solution is to use word segmentation units, e.g. characters (Chung et al., 2016), mixed word/characters (Luong and Man-

[1] https://minorityrights.org/minorities/sorbs/

Proceedings of the 5th Conference on Machine Translation (WMT), pages 1075–1078
Online, November 19–20, 2020. ©2020 Association for Computational Linguistics

ning, 2016), or more intelligent sub-words (Sennrich et al., 2016). It is claimed that an NMT model using such an approach is capable of open-vocabulary translation by encoding rare and unknown words as sequences of sub-word units.

The purpose of our experiments was to try out a supervised NMT system for the low resource language like HSB to GER and vice-versa for the WMT20 shared task.

3 System Description

The standard Transformer architecture proposed by Vaswani et al. (2017) is used for this experiment. This architecture is able to handle long-term dependencies among input tokens, output tokens and between input-output by multi-head attention mechanism. Our method based on the model architecture of Wang et al. (2020), which had used the **B**yte-level BPE (BBPE).

The BBPE encoding is deployed on the Byte Pair Encoding (BPE) (Sennrich et al., 2016), which is a subword algorithm to find a way to represent the given entire text dataset with a small number of tokens. BPE tries to find a balance between character- and word-level hybrid representations, enabling the encoding of any rare words in the vocabulary with appropriate subword tokens without introducing any "unknown" tokens. These segmented byte sequences are encoded into variable-length tokens, i.e., n-grams, which leads to the generation of the BPE vocabulary with byte n-grams. Before being fed to the Transformer model, the learned BBPE passes through bidirectional GRU, which enables to retain contextualization between byte representation of BPE.

4 Experimental Setup

We use the Fairseq[2] (Ott et al., 2019) library to train the Transformer with the same learning rate as in the original paper.

4.1 Dataset and Preprocessing

Our models were trained on the data provided by the Workshop on Machine Translation (WMT) 2020. The statistics about the training, validation and test sets are 60000, 2000 and 2000, respectively for both directional pairs (HSB - GER).

We obtained 1727916 and 1710293 tokens of the GER and HSB, respectively, from the train set for preprocessing. The BPE vocabulary, Byte vocabulary and Character vocabulary are 16384, 2048 and 4096, respectively, for generating binary dataset by using fairseq-preprocess. The BBPE used as a subword BPE tokenizer, where preprocessing was performed using lowercasing only. This is beneficial from the low resource point of view, but it loses the case information for German, which could have affected the results.

4.2 Training Details

We trained the Transformer model with Bi-GRU embedding, in which contextualization using the number of encoder and decoder layers are 2 with the dropout value 0.3. We trained our model with a batch size of 100, with the aid of Adam optimizer at 0.0005 learning rate. The learning rate has warmup update by 4000 to label smoothed cross-entropy loss function with label-smoothing value 0.1.

5 Results and Analysis

The BBPE based Transformer model was evaluated on the blind test set at five different metrics provided by the task organizer, namely BLEU (Papineni et al., 2002), BLEU-cased, TER (Snover et al., 2006), BEER2.0 (Stanojević and Sima'an, 2014), and CharacTER (Wang et al., 2016).

The obtained metrics score for each pair to each experiment is specified in Table 1. The prediction of the test set was generated by performing the best validation checkpoint. However, while comparing the BLEU score of the valid set with the test set, we obtained a difference of +3.21 for HSB→GER and +0.15 for GER→HSB pairs.

Before submitting the predictions of the test set, the BLEU scores of best and last checkpoints were almost equal, as shown in Table 2. Moreover, the vocabulary size plays a crucial role in data-driven approaches of MTs as well. Hence, we have increased the vocabulary size from 2048 to 4096 for generating BBPE, which led to a small decrement in the BLEU score. One possible reason for such decrement is the small vocabulary size that generates generalized BBPE for low-resource language.

6 Conclusion and future work

We have report the results for a Transformer-based MT system for the pair of HSB↔GER in very low resource settings. The introduced MT system works on Byte-level Byte Pair Encoding (BBPE), which yields 48.4 and 46.5 on HSB→GER and

[2] https://github.com/pytorch/fairseq

Pair	BLEU	BLEU cased	TER	BEER2.0	CharacTER
HSB-GER	48.4	47.9	0.383	0.706	0.335
GER-HSB	46.5	45.9	0.389	0.696	0.323

Table 1: Obtained scores of different metrics on the test set, provided by the task organizers

Vocab	Pair	Valid		Test
		Checkpoint (last)	Checkpoint (best)	Checkpoint (best)
2048	HSB-GER	45.92	45.19	48.4
	GER-HSB	46.62	46.35	46.5
4096	HSB-GER	45.77	45.09	-
	GER-HSB	46.96	46.24	-

Table 2: Effect on BLEU by increasing vocabulary size

GER→HSB, respectively, as the BLEU score on the test set at the vocabulary size of 2048. When the vocabulary size was increased from 2048 to 4096, lower performance was obtained on the system on either side of the pair on the validation set.

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

Benyamin Ahmadnia, Javier Serrano, and Gholamreza Haffari. 2017. Persian-Spanish Low-Resource Statistical Machine Translation Through English as Pivot Language. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 24–30.

Alham Fikri Aji. 2020. In Neural Machine Translation, What Does Transfer Learning Transfer? In *Proceedings of the 2020 Annual Conference of the Association for Computational Linguistics*. Association for Computational Linguistics.

Bogdan Babych, Anthony Hartley, and Serge Sharoff. 2007. Translating from under-resourced languages: Comparing direct transfer against pivot translation. *Proceedings of the MT Summit XI*, pages 412–418.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR 2015*.

Loïc Barrault, Ondřej Bojar, Marta R Costa-Jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, et al. 2019. Findings of the 2019 Conference on Machine Translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61.

Junyoung Chung, Kyunghyun Cho, and Yoshua Bengio. 2016. A Character-level Decoder without Explicit Segmentation for Neural Machine Translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1693–1703.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A Survey of Multilingual Neural Machine Translation. *ACM Comput. Surv.*, 53(5).

Nicole Dolowy-Rybinska. 2011. A model minority. *Insight Academia*.

Nicole Dołowy-Rybińska. 2018. Learning Upper Sorbian. The problems with minority language education for non-native pupils in the Upper Sorbian grammar school in Bautzen/Budyšin. *International Journal of Bilingual Education and Bilingualism*, pages 1–15.

Ludwig Elle. 2010. Sorben–demographische und statistische Aspekte. *Vogt, Matthias Theodor, Neyer, Jürgen, Bingen, Dieter et Jan Sokol (éds.), Minderheiten als Mehrwert, Peter Lang GmbH, Frankfurt am Main*, pages 309–318.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor OK Li. 2018a. Universal Neural Machine Translation for Extremely Low Resource Languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 344–354.

Jiatao Gu, Yong Wang, Yun Chen, Victor OK Li, and Kyunghyun Cho. 2018b. Meta-Learning for Low-Resource Neural Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3622–3631.

Ann Irvine and Chris Callison-Burch. 2013. Combining Bilingual and Comparable Corpora for Low Resource Machine Translation. In *Proceedings of the eighth workshop on statistical machine translation*, pages 262–270.

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 865–875.

Rumeng Li, Xun Wang, and Hong Yu. 2020. MetaMT, a Meta Learning Method Leveraging Multiple Domain Data for Low Resource Machine Translation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8245–8252. AAAI Press.

Minh-Thang Luong and Christopher D Manning. 2016. Achieving Open Vocabulary Neural Machine Translation with Hybrid Word-Character Models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1054–1063.

Preslav Nakov and Hwee Tou Ng. 2012. Improving Statistical Machine Translation for a Resource-Poor Language Using Related Resource-Rich Languages. *Journal of Artificial Intelligence Research*, 44:179–222.

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. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.

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 of the Association for Computational Linguistics*, pages 311–318.

Michael Paul, Andrew Finch, and Eiichrio Sumita. 2013. How to Choose the Best Pivot Language for Automatic Translation of Low-Resource Languages. *ACM Transactions on Asian Language Information Processing (TALIP)*, 12(4):1–17.

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.

Aditya Siddhant, Ankur Bapna, Yuan Cao, Orhan Firat, Mia Chen, Sneha Kudugunta, Naveen Arivazhagan, and Yonghui Wu. 2020. Leveraging Monolingual Data with Self-Supervision for Multilingual Neural Machine Translation. In *Proceedings of the 2020 Annual Conference of the Association for Computational Linguistics*, pages 2827–2835. 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 *Proceedings of association for machine translation in the Americas*, volume 200. Cambridge, MA.

Miloš Stanojević and Khalil Sima'an. 2014. BEER: BEtter Evaluation as Ranking. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 414–419.

Haipeng Sun, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2020. Knowledge Distillation for Multilingual Unsupervised Neural Machine Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3525–3535. Association for Computational Linguistics.

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.

Changhan Wang, Kyunghyun Cho, and Jiatao Gu. 2020. Neural Machine Translation with Byte-Level Subwords. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9154–9160. AAAI Press.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. CharacTer: Translation Edit Rate on Character Level. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 505–510.

Zaixiang Zheng, Hao Zhou, Shujian Huang, Lei Li, Xin-Yu Dai, and Jiajun Chen. 2019. Mirror-Generative Neural Machine Translation. In *International Conference on Learning Representations*.