Japanese-Russian TMU Neural Machine Translation System using Multilingual Model for WAT 2019

Aizhan Imankulova Masahiro Kaneko Mamoru Komachi
Tokyo Metropolitan University
6-6 Asahigaoka, Hino, Tokyo 191-0065, Japan
{imankulova-aizhan, kaneko-masahiro}@ed.tmu.ac.jp
komachi@tmu.ac.jp

Abstract

We introduce our system that is submitted to the News Commentary task (Japanese↔Russian) of the 6th Workshop on Asian Translation. The goal of this shared task is to study extremely low resource situations for distant language pairs. It is known that using parallel corpora of different language pair as training data is effective for multilingual neural machine translation model in extremely low resource scenarios. Therefore, to improve the translation quality of Japanese↔Russian language pair, our method leverages other in-domain Japanese-English and English-Russian parallel corpora as additional training data for our multilingual NMT model.

1 Introduction

News Commentary shared task of the 6th Workshop on Asian Translation (Nakazawa et al., 2019) addresses Japanese↔Russian (Ja↔Ru) news translation. It is a very challenging task considering: (a) extremely low resource setting, the size of parallel data is only 12k parallel sentences; (b) how distant given language pair is, in terms of different writing system, phonology, morphology, grammar, and syntax; (c) difficulty of translating news from various topics which leads to large presence of unknown tokens in such extremely low-resource scenario.

Usually, neural machine translation (NMT) (Cho et al., 2014; Sutskever et al., 2014; Bahdanau et al., 2015; Vaswani et al., 2017) enables end-to-end training of a translation system requiring a large amount of training parallel data (Koehn and Knowles, 2017). Therefore, there are different techniques of involving other pivot languages to increase the accuracy of low-resource MT such as pivot-based SMT (Utiyama and Isahara, 2007), transfer learning (Zoph et al., 2016; Kocmi and Bojar, 2018), and multilingual modeling (Firat et al., 2016). Recently, a simple multilingual modeling (MultiNMT) was proposed by Johnson et al. (2017) which translates between multiple languages using a single model and an artificial token indicating a target language, taking advantage of multilingual data to improve NMT for all languages involved. Imankulova et al. (2019) showed that incorporating MultiNMT (Johnson et al., 2017) provided better BLEU scores than unidirectional and pivot-based PBSMT approaches and that domain mismatch had a negative effect on low-resource NMT.

Therefore, we use MultiNMT modeling for an extremely low-resource Ja↔Ru translation involving English (En) as the pivoting third language (Utiyama and Isahara, 2007). Considering the importance of domain matching, we focus on only news domain of additional Ja↔En and Ru↔En auxiliary parallel corpora, which we will refer as pivot parallel corpora. And we investigate how translation results are improved by using in-domain pivot parallel corpora (Ja↔En and Ru↔En) in MultiNMT modeling. As a result, in-domain pivot parallel corpora increases the coverage of Ja and Ru vocabulary, and it is clarified that the new tokens introduced from in-domain pivot corpora could be translated successfully.

2 Related Work

The existing state-of-the-art NMT model known as the Transformer (Vaswani et al., 2017) works well on different scenarios (Lakew et al., 2018; Imankulova et al., 2019). MultiNMT using the artificial token approach (Johnson et al., 2017) is known to help the language pairs with relatively lesser data (Lakew et al., 2018; Rikters et al., 2018)

Proceedings of the 6th Workshop on Asian Translation, pages 165–170
Hong Kong, China, November 4, 2019. ©2019 Association for Computational Linguistics

Lang.pair	Source	Partition	#sent.	#tokens	#types
Ja↔Ru	Global Voices	train	12,356	341k / 229k	22k / 42k
	News Commentary	development	486	16k / 11k	2.9k / 4.3k
	News Commentary	test	600	22k / 15k	3.5k / 5.6k
Ja↔En	Global Voices	train	47,082	1.27M / 1.01M	48k / 55k
	Jiji	train	200,000	5.84M / 5.11M	45k / 78k
	News Commentary	development	589	21k / 16k	3.5k / 3.8k
Ru↔En	Global Voices	train	82,072	1.61M / 1.83M	144k / 74k
	News Commentary	train	279,307	7.00M / 7.41M	214k / 89k
	News Commentary	development	313	7.8k / 8.4k	3.2k / 2.3k

Table 1: Statistics on our in-domain parallel data.

and outperform bi-directional and uni-directional translation approaches (Imankulova et al., 2019). Similarly, we exploit MultiNMT approach with Transformer architecture.

Our work is heavily based on Imankulova et al. (2019). They proposed a multi-stage fine-tuning approach that combines multilingual modeling and domain adaptation. They utilize out-of-domain pivot parallel corpora to perform domain adaptation on in-domain pivot parallel corpora and then perform multilingual transfer for a language pair of interest. However, instead of utilizing out-of-domain pivot parallel corpora, we investigate the impact of other in-domain pivot parallel corpora.

Pseudo-parallel data can be used to augment existing parallel corpora for training, and previous work has reported that such data generated by so-called back-translation can substantially improve the quality of NMT (Sennrich et al., 2016). However, this approach requires base MT systems that can generate somewhat accurate translations (Imankulova et al., 2017). Therefore, instead of creating noisy pseudo-parallel corpora, we take advantage of other in-domain pivot parallel corpora.

3 Experimental Settings

3.1 Data

To train MultiNMT systems we used the news domain data provided by WAT2019[1]. More specifically, we used Global Voices[2] as a training data for Ja↔Ru, Ja↔En and Ru↔En, and manually aligned, cleaned and filtered News Commentary data was used as development and test sets.[3] Additionally, we utilized Jiji[4] and News Commentary[5] data for Ja↔En and Ru↔En, respectively. Table 1 summarizes the size of train/development/test splits used in our experiments.

We tokenized English and Russian sentences using *tokenizer.perl* of `Moses` (Koehn et al., 2007).[6] To tokenize Japanese sentences, we used `MeCab`[7] with the IPA dictionary. After tokenization, we eliminated duplicated sentence pairs and sentences with more than 100 tokens for all the languages.

3.2 Systems

This section describes our system TMU and our baseline which based on the same MultiNMT architecture (Johnson et al., 2017) but trained on different training corpora (Table 1). Here, MultiNMT translates from multiple source languages into different target languages within a single model. To realize such translation, an artificial token is introduced at the beginning of the input sentence to indicate the target language the model should translate to. Since we have 3 language pairs, we concatenate all pairs in both directions with oversampling to match the biggest parallel data. We add a target language token to the source side of each pair and treat it like a single language-pair case.

We experiment with the following systems:

- **TMU**: Our system is trained on a balanced concatenation of Global Voices, Jiji and News Commentary corpora on 6 translation directions.

- **Only GV**: This is our baseline system which is trained on only Global Voices data on

[1] http://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2019/index.html

[2] https://globalvoices.org/

[3] https://github.com/aizhanti/JaRuNC

[4] http://lotus.kuee.kyoto-u.ac.jp/WAT/jiji-corpus/

[5] http://lotus.kuee.kyoto-u.ac.jp/WAT/News-Commentary/news-commentary-v14.en-ru.filtered.tar.gz

[6] https://github.com/moses-smt/mosesdecoder

[7] http://taku910.github.io/mecab, version 0.996.

6 translation directions, the same as in Imankulova et al. (2019).

Only GV is used as a comparative model to investigate the effect of additional pivot corpora.

3.3 Implementation

We used the open-source `tensor2tensor` implementation of the Transformer model.[8]

Table 2 contains some specific hyper-parameters. The hyper-parameters not mentioned in this table used the default values in `tensor2tensor`. We over-sampled Ja→Ru and Ja→En training data so that their sizes match the largest Ru→En data for each model. However, the development set was created by concatenating those for the individual translation directions without any over-sampling. We also used `tensor2tensor`'s internal sub-word segmentation mechanism. The size of the shared sub-word vocabularies was set to 32k. By default, `tensor2tensor` truncates sentences longer than 256 sub-words to prevent out-of-memory errors during training. We incorporated early-stopping by stopping training if BLEU score for the development set was not improved for 10,000 updates (10 check-points).

At inference time, we averaged the last 10 check-points and decoded the test sets with beam size and a length penalty which were tuned by a linear search on the BLEU score for the development set. Length penalty for Ja→Ru was 1.0 and for Ru→Ja 1.1. Beam size was set to 12 and 3 for Ja→Ru and Ru→Ja, respectively. Although we train our models on 6 translation directions, we only report the BLEU scores on Ja→Ru and Ru→Ja test sets.

4 Results

Table 3 demonstrates the BLEU scores of our baseline Only GV model and proposed TMU model on News Commentary Ja→Ru[9] and Ru→Ja[10] test data for News Commentary shared task. Our TMU system trained on additional pivot parallel corpora exceeded the baseline Only GV model trained without additional pivot parallel corpora by approximately 3 BLEU points on both Ja→Ru and Ru→Ja.

Parameter	Value
Word Embedding size	512
Multi-Head number	8
Layer size	6
Hidden size	512
Optimizer	Adam
Adam β_1	0.9
Adam β_2	0.997
Warmup steps	16,000
Learning rate	0.2
Dropout	0.2
Weight decay	0.0
Label smoothing	0.1
Batch size	6144

Table 2: Hyper parameter values of transformer models.

Models	Ja→Ru	Ru→Ja
Only GV	3.66	8.79
TMU	**6.59**	**11.00**

Table 3: Evaluation results: BLEU scores. **Bold** indicates the best BLEU score for each translation direction.

5 Discussion

We investigate the effect of adding Jiji and News Commentary corpora as pivot parallel corpora to original Global Voices training data. In extremely low-resource machine translation in the news domain, unknown tokens become a serious issue due to vocabulary coverage. Adding the pivot parallel corpora to training data can be expected to increase vocabulary coverage.

Therefore, we investigate how much vocabulary coverage was improved by using pivot parallel corpora. For that purpose, we investigate the following vocabulary sets $\mathcal{A}$ and $\mathcal{B}$:

$$\mathcal{A} = \mathcal{T} \cap \mathcal{G} \quad (1)$$

$$\mathcal{B} = \mathcal{T} \cap (\mathcal{G} \cup \mathcal{P}) \quad (2)$$

$\mathcal{T}$ is a set of unknown tokens from test data not included in the direct Ja↔Ru 12k training data, $\mathcal{G}$ is pivot Gloval Voices vocabulary set and $\mathcal{P}$ is Jiji and News Commentary training vocabulary set. $\mathcal{A}$ is the test data unknown tokens set covered by pivot Global Voices training data. $\mathcal{B}$ is the test data unknown tokens set covered by concatenated vocabulary of Jiji and News Commentary pivot paral-

[8] https://github.com/tensorflow/tensor2tensor, version 1.6.6.

[9] http://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/list.php?t=66o=4

[10] http://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/list.php?t=67o=1

	Ja→Ru				Ru→Ja			
	$\mathcal{A}$ (Only GV)		$\mathcal{B}$ (TMU)		$\mathcal{A}$ (Only GV)		$\mathcal{B}$ (TMU)	
	#tokens	#types	#tokens	#types	#tokens	#types	#tokens	#types
Coverage in data	1,467	1,220	2,072	1,751	481	362	596	450
Correctly translated	85	65	191	147	26	21	31	24

Table 4: The coverage of tokens from additional pivot parallel data and the number of correctly translated tokens and types of distinct words by each system calculated for test set.

(a)	**Source**	Должны ли акционеры быть королями ?
	Target	[株主] が、王様 に なる べき か? (Should [shareholders] be kings ?)
	Only GV	この акционер が 社会 の 中心 と なっ て いる の だろ う か？ (Is this акционер the center of society?)
	TMU	[株主] は 王 を 持つ べき な の か？ (Should [shareholders] have a king?)
(b)	**Source**	Преемственность всегда оставалась сугубо семейным делом , и все споры оставались за закрытыми дверями .
	Target	これ まで、継承 者 は、厳格 に 首長 家 から 選ば れる もの と さ れ、いかなる 論争 も [表立っ] て さ れる こと は なかっ た。 (The succession was always strictly a family affair , and no disputes have [emerged].)
	Only GV	家族 経営 の ドライクリーニング 店 で、常習 的 な 商事 に は 至っ て い ない。 (It is a family-run dry cleaning shop, and it has not become a regular business.)
	TMU	この よう な 虐待 は 日々 くり 返さ れ て い た。 (Such abuse was repeated every day.)

Table 5: Examples of translating [unknown tokens] included in pivot parallel data $\mathcal{C}$ from Russian into Japanese.

lel corpora added to $\mathcal{A}$. By comparing the number of tokens and types of distinct words of $\mathcal{A}$ and $\mathcal{B}$, you can see how much the coverage has increased. In addition, we investigate how correctly the tokens added by Jiji corpus and News Commentary are translated. If a token from vocabulary set of $\mathcal{A}$ or $\mathcal{B}$ appeared in both the gold sentence and the translated sentence of the system, it was counted as being correctly translated.

Table 4 shows token and type coverage and correctly translated tokens and types of distinct words on test data for $\mathcal{A}$ and $\mathcal{B}$, respectively. It can be seen that both Ru and Ja have improved $\mathcal{B}$ coverage compared to $\mathcal{A}$. In particular, the coverage of Ru is greatly improved. And by adding Jiji corpus and News Commentary to the training data, you can see that the number of correctly translated tokens has increased. This shows that vocabulary coverage has increased and translation accuracy has improved. On the other hand, the number of correctly translated tokens is few compared to increased coverage from additional parallel data. This is considered to be due to difficulty of directly learning Ja↔Ru translation from added indirect Ja↔En and Ru↔En pivot corpora.

Furthermore, in order to deepen the knowledge about the tokens covered using pivot corpora, we analyze the cases where the newly added tokens by Jiji and News Commentary corpora are translated correctly and incorrectly. By adding Jiji and News Commentary corpora, we define the vocabulary set newly covered by the test data vocabulary as $\mathcal{C}$ as follows:

$$\mathcal{C} = (\mathcal{T} \cap \mathcal{P}) - \mathcal{G} \tag{3}$$

Table 5 shows translation examples of only GV and TMU systems. The [unknown tokens] in each sentence belong to $\mathcal{C}$. The first sentence is an example (a) where TMU was able to correctly translate “株主” compared to Only GV. On the other hand, the second example shows that neither TMU nor Only GV could correctly translate an unknown token “表立っ” included in pivot parallel corpora. It is considered that it cannot be translated because the whole sentence was translated incorrectly.

6 Conclusion

In this paper, we introduced our system submitted to the News Commentary task (Ja↔Ru) of the 6th Workshop on Asian Translation. The difficult part of this shared task is unknown tokens due to difficult news domain covering various topics and extremely low-resource available parallel data. To address this issue, we investigated the coverage of translatable tokens by training MultiNMT using an in-domain pivot parallel corpora. As a result, we found out that our system can translate more tokens by taking advantage of additional pivot parallel corpora. In the future, we will explore whether translation results improve by using other Ja↔Ru (e.g. Tatoeba) and Ru↔En (e.g. UN) corpora.

In the news domain, there is also a problem of completely new tokens, which is a type of unknown tokens, that cannot be dealt by simply increasing training data coverage since new information is out every day. Therefore, we plan to tackle the problem of new tokens that cannot be introduced by using additional corpora.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of the 3rd International Conference on Learning Representations*.

Kyunghyun Cho, Bart van Merriënboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-Way, Multilingual Neural Machine Translation with a Shared Attention Mechanism. In *The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875.

Aizhan Imankulova, Raj Dabre, Atsushi Fujita, and Kenji Imamura. 2019. Exploiting Out-of-Domain Parallel Data through Multilingual Transfer Learning for Low-Resource Neural Machine Translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 128–139.

Aizhan Imankulova, Takayuki Sato, and Mamoru Komachi. 2017. Improving Low-resource Neural Machine Translation with Filtered Pseudo-parallel Corpus. In *Proceedings of the 4th Workshop on Asian Translation*, pages 70–78.

Melvin Johnson, Mike Schuster, Quoc 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.

Tom Kocmi and Ondřej Bojar. 2018. Trivial Transfer Learning for Low-Resource Neural Machine Translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 244–252.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej 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.

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.

Surafel M Lakew, Mauro Cettolo, and Marcello Federico. 2018. A Comparison of Transformer and Recurrent Neural Networks on Multilingual Neural Machine Translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 641–652.

Toshiaki Nakazawa, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Nobushige Doi, Yusuke Oda, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, and Sadao Kurohashi. 2019. Overview of the 6th workshop on Asian translation. In *Proceedings of the 6th Workshop on Asian Translation*, Hong Kong. Association for Computational Linguistics.

Matīss Rikters, Mārcis Pinnis, and Rihards Krišlauks. 2018. Training and Adapting Multilingual NMT for Less-resourced and Morphologically Rich Languages. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, pages 3766–3773.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Controlling Politeness in Neural Machine Translation via Side Constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40.

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*, pages 3104–3112.

Masao Utiyama and Hitoshi Isahara. 2007. A Comparison of Pivot Methods for Phrase-Based Statistical Machine Translation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational*

Linguistics; Proceedings of the Main Conference, pages 484–491.

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 *Proceedings of 30th Advances in Neural Information Processing Systems*, pages 5998–6008.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer Learning for Low-Resource Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575.