TMU NMT System with Japanese BART for the Patent task of WAT 2021

Hwichan Kim and **Mamoru Komachi**
Tokyo Metropolitan University
6-6 Asahigaoka, Hino, Tokyo 191-0065, Japan
kim-hwichan@ed.tmu.ac.jp, komachi@tmu.ac.jp

Abstract

In this paper, we introduce our TMU Neural Machine Translation (NMT) system submitted for the Patent task (Korean⇆Japanese and English⇆Japanese) of 8th Workshop on Asian Translation (Nakazawa et al., 2021). Recently, several studies proposed pre-trained encoder-decoder models using monolingual data. One of the pre-trained models, BART (Lewis et al., 2020), was shown to improve translation accuracy via fine-tuning with bilingual data. However, they experimented only Romanian→English translation using English BART. In this paper, we examine the effectiveness of Japanese BART using Japan Patent Office Corpus 2.0. Our experiments indicate that Japanese BART can also improve translation accuracy in both Korean⇆Japanese and English⇆Japanese translations.

1 Introduction

Neural Machine Translation (NMT) has achieved high translation accuracy in large-scale data conditions. However, translation accuracy of NMT drops in the lack of bilingual data (Koehn and Knowles, 2017). There are several approaches such as back-translation (Sennrich et al., 2016) and transfer learning (Zoph et al., 2016) to address this problem. Furthermore, in addition to these methods, there are some approaches to use pre-trained models using only monolingual data.

BERT (Devlin et al., 2019), which is the most typical pre-trained model, can boost the accuracy of many downstream tasks compared to models without BERT via fine-tuning with the task-specific training data. However, applying BERT to NMT in fine-tuning form like the other tasks requires two-stage optimization and does not provide significant improvement (Imamura and Sumita, 2019). Recently, several studies proposed pre-trained encoder-decoder models using a monolingual data. Lewis et al. (2020) proposed BART, which is one of the pre-trained encoder-decoder models. They demonstrated that BART works well for not only comprehension tasks such as GLEU (Wang et al., 2018) and SQuAD (Rajpurkar et al., 2016) but also text generation tasks such as text summarization and translation. However, they reported only the effect of English BART, so they did not investigate BART trained by monolingual data of another language. Furthermore, in the translation task, they experimented with only Romanian→English translation, which have subword overlap. Therefore, the effect in translations between language pairs without subword overlapping is not clear. Furthermore, they did not experiment in translation direction where the source language matches the language of the pre-trained model.

Additionally, we consider that fine-tuning pre-training models such as BART in translation task is similar to transfer learning (Zoph et al., 2016). Transfer learning in NMT is a method that trains the network of the parent language pair (the parent model) as the initial network and then fine-tunes it for the child language pair (the child model). In the terminology of transfer learning, the pre-trained BART and fine-tuned model are the parent model and child model, respectively. Previous studies have shown that transfer learning works most efficiently when the source languages of the parent and child models are syntactically similar (Dabre et al., 2017; Nguyen and Chiang, 2017). Therefore, we hypothesize that BART is more effective when the language pair for fine-tuning is syntactically similar to the pre-training language.

In this study, we examine the effects of Japanese BART on the translation task. We use Korean/Japanese and English/Japanese bilingual data of Japan Patent Office Patent Corpus 2.0 (JPO corpus) for fine-tuning. We also experiment in both translation directions of Ko⇆Ja and En⇆Ja.

Proceedings of the 8th Workshop on Asian Translation, pages 133–137
Bangkok, Thailand (online), August 5-6, 2021. ©2021 Association for Computational Linguistics

Language pair	Partition	Sent.	Tokens
Korean / Japanese	train	1,000,000	31,569,641 / 37,282,300
	dev	2,000	104,493 / 124,871
	test	5,230	271,744 / 320,584
English / Japanese	train	1,000,000	21,071,895 / 25,695,404
	dev	2,000	524,88 / 64,838
	test	5,668	169,023 / 198,039

Table 1: Data statistics.

2 Related Work

There are some approaches pre-trained encoder models like BERT (Devlin et al., 2019) to the NMT task. Imamura and Sumita (2019) used BERT as an encoder and demonstrated the effectiveness of two-stage optimization, which first trains parameters without BERT encoder, and then fine-tunes all parameters. Zhu et al. (2020) used BERT representations as input embedding and showed more effectiveness than using BERT as the encoder.

Recently, several studies proposed pre-trained encoder-decoder models such as MASS (Song et al., 2019) and BART (Lewis et al., 2020), and these models can improve the translation accuracy via fine-tuning with bilingual data. MASS (Song et al., 2019) uses monolingual data from both the source and target languages for pre-training when applying to the NMT. On the contrary, BART (Lewis et al., 2020) uses only monolingual data of target language, unlike MASS. Liu et al. (2020) trained multilingual BART (mBART) using monolingual data of 25 languages. They indicated that mBART initialization leads significant gains in low resource settings. However, Wang and Htun (2020) showed that mBART cannot obtain improvements in the Patent task.

3 Experimental Settings

3.1 Implementation

Hyperparameter	Value
Embedding dimension	768
Attention heads	12
Layers	6
Feed forward dimension	3072
Optimizer	Adam
Adam betas	0.9, 0.98
Learning rate	0.0005
Dropout	0.1
Label smoothing	0.1
Max tokens	4,098

Table 2: Hyperparameters.

In this study, we use Japanese BART[1] base v1.1 (JaBART) trained using Japanese Wikipedia sentences (18M sentences). For fine-tuning, we do not use an additional encoder like in Lewis et al. (2020)'s method. Instead, we add randomly initialized embeddings for each unknown subword in JaBART to both encoder and decoder. We share the embeddings of characters that match across languages, such as numbers and units. We also train baseline models consisting of the same architecture as that of JaBART. We use the same hyperparameters indicated in Table 2 for both fine-tuning JaBART and training the baseline model. We fine-tune and train the models using the fairseq implementation[2].

3.2 Data

To train and fin-tune the models, we use Ko–Ja and En–Ja datasets of JPO corpus. Korean and English have almost no subword overlaps with Japanese, because these languages use Hangul, Latin alphabets, and Hiragana/Katakana/Kanji characters, respectively. For Japanese pre-processing, we use JaBART tokenizer. For Korean and English, we tokenize sentences using MeCab-ko[3] and Moses scripts[4], respectively. Then, we apply the SentencePiece (Kudo and Richardson, 2018) with a 32k vocabulary size. Table 1 presents the training, de-

[1] https://github.com/utanaka2000/fairseq/blob/japanese_bart_pretrained_model

[2] https://github.com/utanaka2000/fairseq

[3] https://bitbucket.org/eunjeon/mecab-ko

[4] https://github.com/moses-smt/mosesdecodertree/RELEASE-4.0

		Ko→Ja		Ja→Ko	
		dev	test	dev	test
Single	Baseline	67.400±.080 / -	71.510±.166 / 0.947±.001	67.816±.028 / -	71.103±.144 / 0.942±.001
	JaBART	**68.750±.104** / -	**72.760±.140 / 0.949±.000**	**68.563±.065** / -	**72.116±.060 / 0.946±.001**
	Δ	+1.350 / -	+1.250 / +0.002	+0.746 / -	+1.013 / +0.003
Ensemble	Baseline	68.770 / -	73.240 / 0.946	68.590 / -	72.070 / 0.942
	JaBART	**69.570** / -	**73.670 / 0.949**	**69.440** / -	**72.700 / 0.946**
	Δ	+0.800 / -	+0.430 / +0.001	+0.850 / -	+0.630 / +0.002
		En→Ja		Ja→En	
		dev	test	dev	test
Single	Baseline	38.706±.083 / -	42.533±.151 / 0.843±.0.02	37.636±.112 / -	40.873±.231 / 0.843±.001
	JaBART	**39.146±.077** / -	**43.720±.053 / 0.849±.001**	**38.393±.060** / -	**41.943±.084 / 0.851±.001**
	Δ	+0.440 / -	+1.187 /+0.005	+0.757 / -	+1.070 / +0.008
Ensemble	Baseline	**40.360** / -	45.000 / 0.853	39.260 / -	43.140 / 0.853
	JaBART	40.270 / -	**45.240 / 0.855**	**39.660** / -	**43.780 / 0.857**
	Δ	-0.090 / -	+0.240 / +0.002	+0.400 / -	+0.640 /+0.004

Table 3: BLEU / RIBES scores of each single and ensemble of three models. The scores of single are the average of the three models. We indicate the best scores in bold. The scores of Δ indicate the gains of the fine-tuned JaBART's BLEU score over the baseline model.

velopment, and test[5] data statics.

3.3 Results

Table 3 shows that the BLEU and RIBES scores of each single and ensemble model.

In the single model, the fine-tuned JaBART achieves the highest scores for dev and test data in both language pairs and translation directions of Ko⇆Ja and En⇆Ja. Specifically, the BLEU scores of the dev and test data reveal improvements of 0.440-1.350 and 1.013-1.250 from the baseline models, respectively. The RIBES scores also reveal improvements of 0.001-0.007, but there is no significant difference between the fine-tuned BART and baseline models.

In the ensemble model[6], the fine-tuned JaBART improves the BLEU and RIBES scores approximately 0.440-0.850 and 0.001-0.008, respectively, in the dev and test of Ko⇆Ja and Ja→En translations. However, in En→Ja translation, the BLEU score of the fine-tuned JaBART decreases 0.09 in the dev and improves 0.240 in the test data. Thus, in the ensemble scenario, the fine-tuned JaBART model can improve translation accuracy except for En→Ja translation.

4 Discussions

We hypothesize that JaBART is more effective when the language pair for fine-tuning is syntactically similar to the pre-training language, as in transfer learning. In our experimental settings, Korean and English are syntactically similar and different languages with Japanese, respectively [7]. Therefore, we expect that JaBART is more effective in the Ko⇆Ja translations than in the En⇆Ja translations. However, Table 3 shows no significant differences in Δ scores between the Ko⇆Ja and En⇆Ja translations. These results indicate that syntactic similarity does not affect the enhancement in the final BLEU scores.

5 Conclusions

In this paper, we described our NMT system submitted to the Patent task (Ko⇆Ja and En⇆Ja) of the 8th Workshop on Asian Translation. We compared the baseline and fine-tuned JaBART models, and demonstrated that the fine-tuned JaBART achieves consistent improvements of BLEU scores in language pairs with no subword overlapping, and irrespective of translation directions.

Contrary to our hypothesis, our experiments indicated no significant difference in the translation accuracy depending on the syntactic similarity. However, we consider that there are some differences in

[5] In this study, we use test-n data, a union of test-n1, test-n2, and test-n3 data, for evaluation.

[6] We submitted the En⇆Ja ensemble models as the target for human evaluation.

[7] Japanese and Korean are SOV and agglutinative languages, whereas English is SVO and fusional language (Masayoshi, 1990; Jeong et al., 2007).

another aspect such as training process per epoch and network representations. Therefore, we attempt to analyze BART fine-tuned using language pairs with varying syntactic proximities in detail in the future.

Acknowledgments

This work has been partly supported by the programs of the Grant-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (JSPS KAKENHI) Grant Number 19K12099.

References

Raj Dabre, Tetsuji Nakagawa, and Hideto Kazawa. 2017. An empirical study of language relatedness for transfer learning in neural machine translation. In *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*, pages 282–286.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Kenji Imamura and Eiichiro Sumita. 2019. Recycling a pre-trained BERT encoder for neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 23–31.

Hyeonjeong Jeong, Motoaki Sugiura, Yuko Sassa, Tomoki Haji, Nobuo Usui, Masato Taira, Kaoru Horie, Shigeru Sato, and Ryuta Kawashita. 2007. Effect of syntactic similarity on cortical activation during second language processing: a comparison of English and Japanese among native Korean trilinguals. *Human Brain Mapping*, 28(3):195–204.

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.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

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. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Shibatani Masayoshi. 1990. *The Languages of Japan.* Cambridge University Press.

Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, and Sadao Oda, Yusuke Kurohashi. 2021. Overview of the 8th workshop on Asian translation. In *Proceedings of the 8th Workshop on Asian Translation*.

Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: Masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning*, pages 5926–5936.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355. Association for Computational Linguistics.

Dongzhe Wang and Ohnmar Htun. 2020. Goku's participation in WAT 2020. In *Proceedings of the 7th Workshop on Asian Translation*, pages 135–141.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tieyan Liu. 2020. Incorporating BERT into neural machine translation.

In *International Conference on Learning Representations*.

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.