Can Monolingual Pre-trained Encoder-Decoder Improve NMT for Distant Language Pairs?

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

Recently, several studies have proposed pre-trained encoder-decoder models using monolingual data, such as BART, which can improve the accuracy of seq2seq tasks via fine-tuning with task-specific data. However, the effectiveness of pre-training using monolingual data requires further verification, as previous experiments on machine translation have focused on specific languages with overlapping vocabularies and particular translation directions. Additionally, we hypothesize that the effects of pre-trained models differ depending on the syntactic similarity between languages for pre-training and fine-tuning, as in transfer learning.

To this end, we analyze BART fine-tuned with languages exhibiting different syntactic proximities to the source language in terms of the translation accuracy and network representations. Our experiments show that (1) BART realizes consistent improvements regardless of language pairs and translation directions. Contrary to our hypothesis, there is no significant difference in the translation accuracy based on the syntactic similarity. However, when syntactically similar, BART achieves approximately twice the accuracy of our baseline model in the initial epoch. Furthermore, we demonstrate that (2) syntactic similarity correlates with closeness of the encoder representations; in a syntactically similar language pair, the representations of the encoder do not change after fine-tuning. The code used in our experiments has been published.[1]

1 Introduction

Neural machine translation (NMT) can realize high translation accuracy via training on large-scale bilingual data. However, the lack of bilingual data affects the translation accuracy (Koehn and Knowles, 2017). Previous studies have proposed various methods such as back-translation (Sennrich et al., 2016) and transfer learning (Zoph et al., 2016) to address this problem.

Recently, several studies proposed pre-trained encoder-decoder models using monolingual data; certain models were applied to the NMT task to improve translation accuracy. For example, Lewis et al. (2020) proposed BART, which is pre-trained with monolingual data of the target language (English). It was demonstrated that BART could improve Romanian→English translation. We believe that there is room for further validation. For example, the languages used in their experiments have subword overlap; however, other languages (without subword overlap) were not investigated. Additionally, the translation direction where the source language matches the language of the pre-trained model was not investigated.

In this study, we examine the effects of BART on NMT using different experimental settings from the previous study (Lewis et al., 2020). We use source or target languages that have no subword overlap with the language used pre-train BART, and experiment with both translation directions. In addition, we hypothesize that BART is more effective when the language pair for fine-tuning is syntactically similar with the pre-training language, as in transfer learn-

[1] https://github.com/hwichan0720/Monolingual-Pretrain-for-NMT

Language	Character	Word Order
English	Latin alphabet	SVO
French	Latin alphabet	SVO
Japanese	Hiragana/Kanji etc.	SOV
Korean	Hangul	SOV

Table 1: The languages and their features. In addition, English and French are fusional languages, whereas Japanese and Korean are agglutinative languages.

ing (Zoph et al., 2016; Dabre et al., 2017; Murthy et al., 2019). Therefore, we analyze BART fine-tuned using language pairs with varying syntactic proximities. We observe the following aspects of translation accuracy and network representations.

- BART realizes consistent improvements regardless of the language pairs and translation directions used. Contrary to our hypothesis, there is no significant difference in the translation accuracy based on the level of syntactic similarity. However, when languages are syntactically similar, BART can yield approximately twice the accuracy of our baseline model in the initial epoch.

- The representations of the encoder remain unchanged after fine-tuning when high syntactic similarity prevails between pre-training and fine-tuning languages; however, the representations of the decoder change regardless of syntactic similarity.

2 Related Work

When applying pre-trained encoder models like BERT (Devlin et al., 2019) to the NMT task, sophisticated techniques, such as two-stage optimization (Imamura and Sumita, 2019) or a fusion method as input embedding (Zhu et al., 2020), are required to improve the accuracy of the models. In contrast, pre-trained encoder-decoder models such as MASS (Song et al., 2019) can improve the translation accuracy via fine-tuning with bilingual data. MASS uses monolingual data from both the source and target languages, unlike BART.

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.

Transfer learning is very efficient in improving the quality of low-resource-language translations. Previous studies have demonstrated that transfer learning works most efficiently when the source languages of the parent and child models are linguistically similar (Zoph et al., 2016; Dabre et al., 2017; Nguyen and Chiang, 2017). Murthy et al. (2019) reported that a divergent word order between the parent and child model languages limits the benefits of transfer learning.

There are several methods to calculate the neural networks' similarity based on canonical correlation analysis (Raghu et al., 2017; Morcos et al., 2018). However, Kornblith et al. (2019) demonstrated these methods can not measure meaningful similarities if the data points are fewer than the dimension of the representations. Therefore, they proposed a novel method called centered kernel alignment (CKA), which does not suffer from this limitation. Recently, many studies have conducted analyses of neural networks using the CKA (Wu et al., 2020; Vulić et al., 2020; Conneau et al., 2020; Muller et al., 2021). In this study, we measure the similarity between before and after fine-tuning BART using the CKA.

3 Experimental Settings

We use the English BART base (EnBART)[2] and Japanese BART base v1.1 (JaBART)[3] as the non-

[2] https://github.com/pytorch/fairseq/tree/master/examples/bart

[3] https://github.com/utanaka2000/fairseq/tree/japanese_bart_pretrained_model

	Ko→Ja		Ja→Ko		En→Ja		Ja→En	
	dev	test	dev	test	dev	test	dev	test
Baseline	67.40±.08	71.51±.16	67.81±.02	71.10±.14	38.70±.08	42.53±.15	37.63±.11	40.87±.23
JaBART	**68.75±.10**	**72.76±.14**	**68.56±.07**	**72.11±.06**	**39.14±.07**	**43.72±.05**	**38.39±.06**	**41.94±.08**
Δ	+1.35	+1.25	+0.75	+1.01	+0.44	+1.19	+0.76	+1.07
	Fr→En		En→Fr		Ja→En		En→Ja	
	dev	test	dev	test	dev	test	dev	test
Baseline	35.42±.30	35.26±.11	34.63±.11	34.81±.15	38.13±.24	40.49±.03	38.31±.26	42.65±.18
EnBART	**36.65±.29**	**36.27±.33**	**35.91±.13**	**36.12±.13**	**38.88±.11**	**42.86±.09**	**39.68±.15**	**44.41±.13**
Δ	+1.22	+1.01	+1.28	+1.31	+0.75	+2.37	+1.37	+1.76

Table 3: BLEU scores of each language pair of baseline and fine-tuned BART. These BLEU scores are the averages of the three models. We indicate the best scores in bold. The Δ scores indicate the BLEU-score gains of the fine-tuned BART over that of the baseline model.

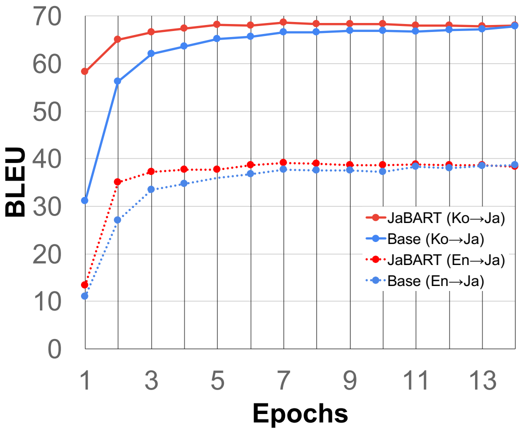


Figure 1: BLEU scores of the Ko→Ja and En→Ja models for each epoch.

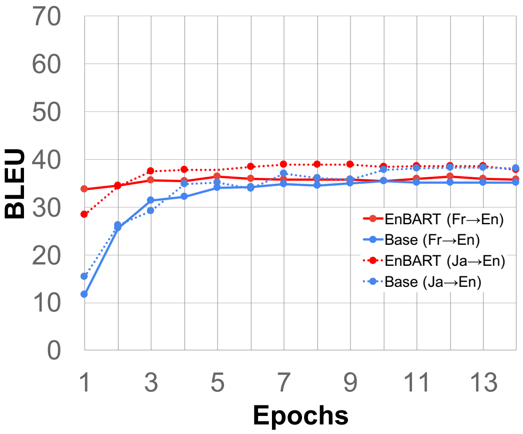


Figure 2: BLEU scores of the Fr→En and Ja→En models for each epoch.

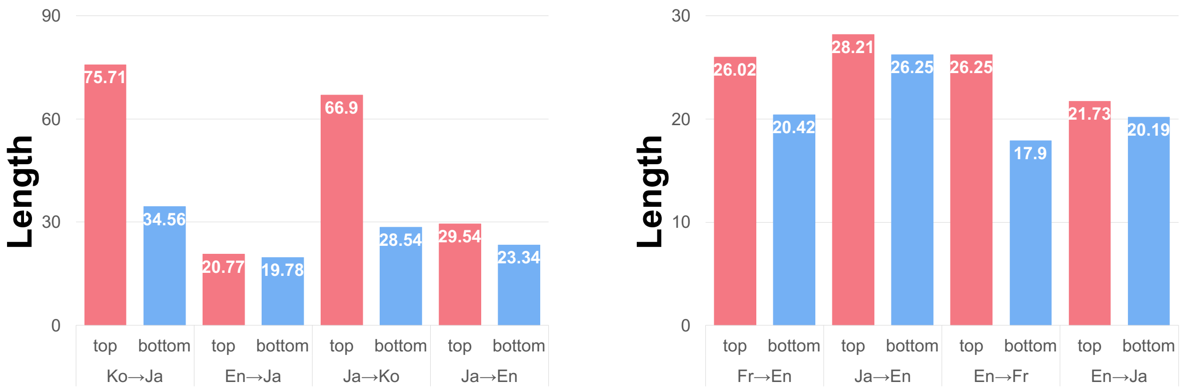
English BART, which was trained using Japanese Wikipedia sentences (18M sentences). In this study, we use Japanese, Korean, and English, as these languages do not exhibit subword overlap; Korean (Ko) and French (Fr) are used as the syntactically similar languages to Japanese (Ja) (Shibatani, 1990; Jeong et al., 2007) and English (En), respectively. Table 1 presents the languages used in our experiments and their linguistic features.

We fine-tune the BART for each language[4] with Ko⇆Ja, En⇆Ja, and Fr⇆En as the language pairs and train baseline models consisting of the same architecture as that of BART. We use the same hyperparameters presented in Table 2 for both fine-tuning BART and training the baseline model. We fine-tune and train the models using the fairseq implementation (Ott et al., 2019).

For the Ko⇆Ja and En⇆Ja language pairs, we use Japan Patent Office (JPO) Corpus 2.0[5]. For the Fr⇆En language pair, we use 1M parallel sentences, which are sampled randomly from the Europarl Parallel Corpus (Koehn, 2005) such that they match the size of the training data obtained from the JPO corpus. For Japanese pre-processing, we use the

[4] In this study, we do not use an additional encoder, unlike Lewis et al. (2020). Instead, we add randomly initialized embeddings for each unknown subword in JaBART to both the encoder and decoder. We share the embeddings for characters that match across languages, such as numbers.

[5] `http://lotus.kuee.kyoto-u.ac.jp/WAT/patent`



(a) JaBART

(b) EnBART

Figure 3: The average lengths of the source sentences for which the BLEU scores of each BART model are higher (top) and lower (bottom) than the baseline in the initial epoch.

JaBART tokenizer. For Korean, English[6] and French, we tokenize sentences using MeCab-ko[7] and Moses scripts[8]. Next, we apply SentencePiece (Kudo and Richardson, 2018) with a vocabulary size of 32k.

4 Discussions

4.1 Translation Accuracy

Effect of language pair and translation directions. Table 3 presents the BLEU scores of the evaluation data and the gains of the fine-tuned BART over the baseline model. The fine-tuned BART models achieve consistent improvements for all language pairs and directions. In particular, the BLEU scores of fine-tuned JaBART improve by approximately 0.56–1.22 in En⇆Ja translations, and 0.74–1.35 in Ko⇆Ja translations, compared to the scores of the baseline model.

Traning process. We also investigate the training process for each epoch. Figures 1 and 2 present the BLEU scores on dev data of the En→Ja, Ko→Ja and Ja→En, Fr→En models for each epoch. In the En→Ja and Ja→En models (dotted line), no significant differences in the improvement of the BLEU scores per epoch are observed between the baseline and fine-tuned BART. However, in the Ko→Ja and Fr→En models (solid lines), there are significant differences in the BLEU scores of the initial epochs between the baseline and fine-tuned BART. Additionally, in the Fr→En model, the EnBART already achieves approximately equal BLEU scores with the final epoch in the initial epoch. Notably, the same training-process trend is observed in the opposite direction.

	Ko→Ja		Ja→Ko	
	JaBART	Baseline	JaBART	Baseline
Long	58.18	19.04	59.22	21.61
Short	57.44	43.50	56.97	46.51
Δ	+0.74	-24.46	+2.25	-24.90
	Fr→En		En→Fr	
	EnBART	Baseline	EnBART	Baseline
Long	35.32	11.19	29.03	7.80
Short	40.58	17.71	31.81	15.30
Δ	-5.26	-6.52	-2.77	-7.49

Table 4: BLEU scores of the subsets of long and short sentences. The Δ scores indicate the differences between the long and short subsets.

Sentence length. Next, we examine the types of sentences translated in a better manner by the BART model than the baseline model in the initial epoch. We measure the sentence-level BLEU scores on dev data in the first epoch and report the differences in scores between the fine-tuned BART and baseline models. Subsequently, we sort these values in descending order and calculate the average sentence

[6] When we fine-tune the EnBART and train the baseline models for comparison with the fine-tuned EnBART models, we use the EnBART tokenizer.

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

[8] `https://github.com/moses-smt/mosesdecoder/tree/RELEASE-4.0`

		Length	BLEU
Reference	… に 設け られた ベース １０ …下部 リンク ２ ｄ が 設け られる 。	67	-
Source	… 에 설치 된 베이스 (10) … 하부 링크 (2 d) 가 설치 된다 .	86	-
English	… the base 10 shown on … the lower link 2d are also provided on …	52	-
Baseline	… に 設置 さ れた ベースベース １０ ａ … を 示す が 、右 側 (図 ２ ａ) 。	65	13.44
JaBART	… に 設け られた ベース １０ … 下部 リンク ２ ｄ が 設け られる 。	52	**90.73**
Reference	燃料 タンク の 説明 図 である 。	7	-
Source	도 15 는 연료 탱크 의 설명도 이 다 .	10	-
English	Figure 15 shows an illustration of the fuel tank .	10	-
Baseline	燃料 タンク の 説明 図 である 。	7	**100.00**
JaBART	図 １５ は 燃料 タンク の 説明 図 である 。	10	63.89

Table 5: Examples of Ko→Ja translations in the first epoch and the corresponding English translations.

		Length	BLEU
Reference	図 ２ に 方向 決定 部 １３ の 機能 ブロック 図 を 示す 。	14	-
Source	FIG. 2 shows a function block diagram of the direction determination portion 13 .	14	-
Baseline	図 ２ は 、ブロック １３ の ブロック １３ の 機能 機能 を 示す 図 である 。	17	13.67
JaBART	図 ２ は 、方向 決定 部 １３ の 機能 ブロック 図 である 。	14	**57.57**
Reference	次に 、クリップ １０ の 構成 に ついて 説明 する 。	11	-
Source	Next, the configuration of the clip 10 will be described .	11	-
Baseline	次に 、クリップ １０ の 構成 に ついて 説明 する 。	11	**100.00**
JaBART	これ に より 、ヘッド １０ の 構成 が 説明 さ れる 。	10	14.21

Table 6: Examples of En→Ja translations in the first epoch.

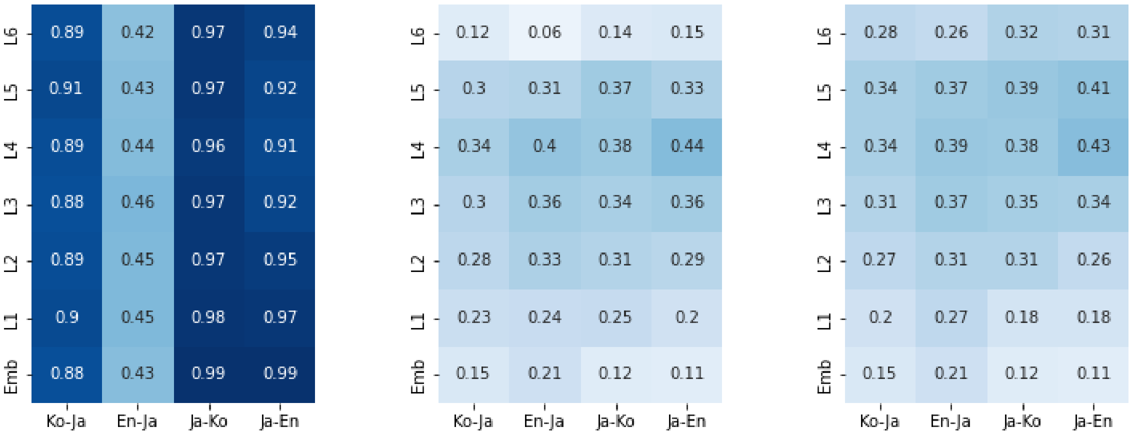
length of the top and bottom 300 source sentences. Figure 3 presents the average top and bottom sentence lengths for each model. This figure reveals significant differences between the length of top and bottom sentence lengths in the translations of syntactic similar language pairs (Ko→Ja, Ja→Ko and Fr→En, En→Fr) with each BART's language compared to other languages. These results indicate that the fine-tuned BART models are good at translating long sentences even in the first epoch. Tables 5 and 6 present the top (upper) and bottom (lower) translation examples of the Ko→Ja and En→Ja language pairs in the JaBART setting. Notably, there is a translation error in the reference within the lower example in Table 5. The words "도 15 (Figure 15)" in the source sentence should be translated to "図 15" in the reference sentence; notably, these words are not translated. Therefore, the baseline BLEU score is higher than that of JaBART; however, JaBART translates "도 15" more adequately.

Because there is a possibility that fine-tuned BART model can not translate well the short sentences only the results of Figure 3, we compare the BLEU scores between the subsets of short and long sentences. We sort dev data in ascending source-sentence-length order and extract the longest and shortest subsets of the 300 sentences. We measure their BLEU scores of the fine-tuned BART and baseline. Table 4 lists the BLEU scores of each model. This table reveals the existence of minimal BLEU-score differences between subsets in the JaBART and EnBART models. Therefore, it is not that the fine-tuned BART cannot translate short sentences; it simply the baseline model cannot translate long sentences adequately.

Summary. BART achieves consistent improvements regardless of language pairs and translation directions[9]. However, syntactic similarity does not affect the enhancement of final BLEU scores.

The JaBART and EnBART models work better

[9] Aji et al. (2020) demonstrated that the model pre-trained for a copy sequence task using monolingual data of the target language can improve translation accuracy of several language pairs. Our experiments indicate the same trend, and also demonstrate the effectiveness in both translation directions.

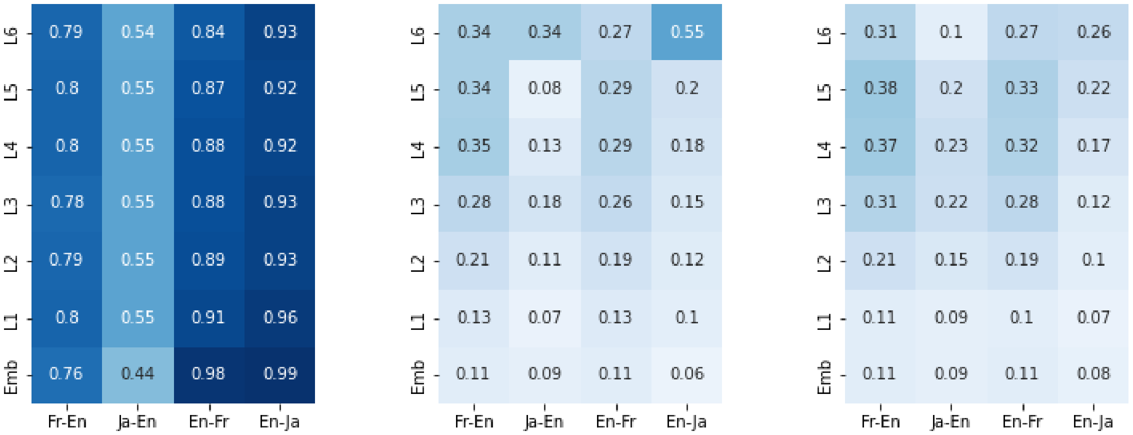


(a) Encoder

(b) Decoder

(c) Up to the self-attention layer

Figure 4: Encoder and decoder's CKA similarity between JaBART and each fine-tuned model.



(a) Encoder

(b) Decoder

(c) Up to the self-attention layer

Figure 5: Encoder and decoder's CKA similarity between EnBART and each fine-tuned model.

with syntactically similar languages (Korean and French) rather than syntactically dissimilar languages (English and Japanese) as the initial network parameters. We observe that the fine-tuned BART is better at translating long sentences compared to the baseline model for syntactically similar pairs(Ko→Ja, Ja→Ko and Fr→En, En→Fr) in the initial epoch.

4.2 Network Representation

As described in the previous section, we observed that the fine-tuned BART achieved high translation accuracy even in the first epoch with syntactically similar language pairs. Additionally, the BLEU-score gains of the final model over the initial model are approximately 10 points and 3 points, in JaBART and EnBART settings, respectively. From these results, we hypothesize that the layer representations do not change significantly before and after fine-tuning on fine-tuning with a syntactically similar language. To confirm our hypothesis, we calculated the similarities between the networks before and after fine-tuning using centered kernel alignment (CKA) (Kornblith et al., 2019). The linear CKA similarity measure is defined as follows:

$$\mathrm{CKA}(X,Y)=\frac{\|Y^TX\|_{\mathrm{F}}^2}{(\|X^TX\|_{\mathrm{F}}\|Y^TY\|_{\mathrm{F}})}$$

where X and Y correspond respectively to the matrices of the d-dimensional mean pooled subword representations at the layer of the n parallel source and target sentences. We randomly sampled 100 sentences from the dev data to calculate the CKA similarity.

Encoder representation. The heat maps in Figures 4a and 5a reveal the CKA similarities of each layer on the encoder side before and after fine-tuning JaBART and EnBART.

When fine-tuning the Ja→Ko, Ja→En models in Figure 4a and the En→Fr, En→Ja models in Figure 5a, the CKA similarities are very high, indicating that the layer representations are approximately the same. Contrary to our hypothesis, the representations of the encoder side do not change not only in the Ja→Ko and En→Fr models, but also in the Ja→En and En→Ja models, in the JaBART and EnBART settings, respectively.

When fine-tuning the En→Ja model in Figure 4a and the Ja→En model in Figure 5a, the CKA similarities are lower than other translation directions.This indicates that the network representations change after fine-tuning for the language pair with the most dissimilar source and target languages. Furthermore, the CKA similarities with the fine-tuned Ko→Ja and Fr→En models are very high, in the JaBART and EnBART settings, respectively. Specifically, the similarity scores exceed 0.88 and 0.76 in each setting. This can be attributed to the fact that Korean and French are syntactically similar to Japanese and English, respectively. This result is consistent with our hypothesis.

Decoder representation. Figures 4b and 5b depict the CKA similarity of each layer of the decoder side of JaBART and EnBART. Figure 4b reveals that the CKA similarities with the fine-tuned Ja⇆En and Ja⇆Ko models are lower than those on the encoder side. Furthermore, Figure 5b reveals that the similarities with the fine-tuned En⇆Fr models are lower than the corresponding similarities on the encoder side.

We consider that the change in representations of the decoder side, especially when the target language is the same as the language of each BART, is caused by the information input received in different languages by the cross-attention layer from the encoder side. Therefore, we additionally calculate the similarity using the representations up to the self-attention layer, and illustrate the resulting heat maps in Figures 4c and 5c. In the final layer of Figure 4c, the similarity scores for the representation up to the self-attention layer are higher than the scores indicated in Figure 4b. However, there are almost no differences in the other layers and the layers of Figure 5c.

Summary. The representations of the encoder side do not change when the source language is the same as or syntactically similar to the target language; however, the representations of the decoder side change regardless of the target language.

5 Conclusions

In this study, we analyzed the effect of BART on NMT in detail. Our experiments showed that BART realizes consistent improvements regardless of language pairs with no subword overlapping, and irrespective of translation directions. Furthermore, we observed that BART was adequate as an initial network representation and the representations of the encoder do not change after fine-tuning when the languages were syntactically similar. Additionally, our experimental results revealed that the representations of the decoder change after fine-tuning, regardless of the target language.

We believe that the difference between the pre-training and fine-tuning tasks causes the change in decoder representations. Guu et al. (2020) reported that intermediate pre-training, similar to a downstream task before fine-tuning, can improve final performance. Therefore, in the future, we will attempt to perform an intermediate pre-training similar to NMT and investigate what types of pre-training methods using monolingual data are suited for machine translation tasks.

Acknowledgments

This work was partly supported by JSPS KAKENHI Grant Number 19K12099.

References

Alham Fikri Aji, Nikolay Bogoychev, Kenneth Heafield, and Rico Sennrich. 2020. In neural machine translation, what does transfer learning transfer? In *Proceed-*

ings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7701–7710.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034.

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.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR, 13–18 July.

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.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86. AAMT, AAMT.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3519–3529.

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.

Ari Morcos, Maithra Raghu, and Samy Bengio. 2018. Insights on representational similarity in neural networks with canonical correlation. In *Advances in Neural Information Processing Systems 31*, pages 5732–5741.

Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. 2021. First align, then predict: Understanding the cross-lingual ability of multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2214–2231.

Rudra Murthy, Anoop Kunchukuttan, and Pushpak Bhattacharyya. 2019. Addressing word-order divergence in multilingual neural machine translation for extremely low resource languages. 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 3868–3873.

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.

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.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. 2017. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems 30*, pages 6076–6085.

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.

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

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.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240.

John Wu, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2020. Similarity analysis of contextual word representation models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4638–4655.

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.