Taylor & Francis
Taylor & Francis Group

Applied Artificial Intelligence
An International Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uaai20

An Efficient Method for Generating Synthetic Data for Low-Resource Machine Translation

An empirical study of Chinese, Japanese to Vietnamese Neural Machine Translation

Thi-Vinh Ngo, Phuong-Thai Nguyen, Van Vinh Nguyen, Thanh-Le Ha & Le-Minh Nguyen

To cite this article: Thi-Vinh Ngo, Phuong-Thai Nguyen, Van Vinh Nguyen, Thanh-Le Ha & Le-Minh Nguyen (2022) An Efficient Method for Generating Synthetic Data for Low-Resource Machine Translation, Applied Artificial Intelligence, 36:1, 2101755, DOI: 10.1080/08839514.2022.2101755

To link to this article: https://doi.org/10.1080/08839514.2022.2101755

Published online: 02 Aug 2022.

APPLIED ARTIFICIAL INTELLIGENCE
2022, VOL. 36, NO. 1, e2101755 (3006 pages)
https://doi.org/10.1080/08839514.2022.2101755

Taylor & Francis
Taylor & Francis Group

An Efficient Method for Generating Synthetic Data for Low-Resource Machine Translation

An empirical study of Chinese, Japanese to Vietnamese Neural Machine Translation

Thi-Vinh Ngo[a], Phuong-Thai Nguyen[b], Van Vinh Nguyen[c], Thanh-Le Ha[d], and Le-Minh Nguyen[e]

[a]Department of Computer Engineering, Thai Nguyen University of Information and Communication Technology, Thai Nguyen, Vietnam; [b]Institute of Artificial Intelligence, University of Engineering and Technology, Vietnam National University, Hanoi, Vietnam; [c]Department of Computer Science, The Faculty of Information Technology, University of Engineering and Technology, Vietnam National University, Hanoi, Vietnam; [d]Institute of Anthropomatics and Robotics, Faculty of Informatics, Karlsruhe Institute of Technology, Karlsruhe, Germany; [e]School of Information Science, Japan Advanced Institute of Science and Technology, Japan

ABSTRACT

Data sparsity is one of the challenges for low-resource language pairs in Neural Machine Translation (NMT). Previous works have presented different approaches for data augmentation, but they mostly require additional resources and obtain low-quality dummy data in the low-resource issue. This paper proposes a simple and effective novel for generating synthetic bilingual data without using external resources as in previous approaches. Moreover, some works recently have shown that multilingual translation or transfer learning can boost the translation quality in low-resource situations. However, for logographic languages such as Chinese or Japanese, this approach is still limited due to the differences in translation units in the vocabularies. Although Japanese texts contain Kanji characters that are derived from Chinese characters, and they are quite homologous in sharp and meaning, the word orders in the sentences of these languages have a big divergence. Our study will investigate these impacts in machine translation. In addition, a combined pre-trained model is also leveraged to demonstrate the efficacy of translation tasks in the more high-resource scenario. Our experiments present performance improvements up to +6.2 and +7.8 BLEU scores over bilingual baseline systems on two low-resource translation tasks from Chinese to Vietnamese and Japanese to Vietnamese.

ARTICLE HISTORY
Received 30 April 2022
Accepted 1 July 2022

Introduction

Neural Machine Translation (NMT) systems (Bahdanau, Cho, and Bengio 2015; Sutskever, Vinyals, and Le 2014; Vaswani et al. 2017), recently, have shown state of the art in many translation tasks. Language pairs that are high-

CONTACT Thi-Vinh Ngo ntvinh@ictu.edu.vn Thai Nguyen University of Information and Communication Technology, Z115 Street, Quyet Thang, Thai Nguyen, Vietnam

resource have presented impressive results; otherwise, low-resource language pairs have shown poor performance due to the lack of bilingual data. In some cases, datasets for research purposes are also absent. To solve this problem, using monolingual data is considered an effective strategy to enhance translation quality in low-resource situations.

Typically (Edunov et al. (2018; Sennrich, Haddow, and Birch 2016b) propose the Back Translation method that is popularly used in recent NMT systems. This technique requires a translation system that is trained on a seed parallel corpora. Nevertheless, for low-resource language pairs, the seed corpus is normally small; thus, the generated hypotheses are often inaccurate. Although monolingual data are always inexhaustible, we only use an amount equal to the size of the bilingual data to avoid degraded performance in the low-resource situation. To alleviate this problem, in this study, we propose a new technique for generating synthetic data in low-resource issues and compare it to the Back Translation strategy in the experiments. Our proposal could be applied to the NMT systems before they are utilized for inference of monolingual data in the Back Translation method.

On the other way, Ha, Niehues, and Waibel (2016) show the techniques that create synthetic data by making a copy of the target sentence on the source side. In this way, they hope that NMT systems can learn name entities or terms better when both source and target language share the same or similar alphabets. (Sánchez-Cartagena et al. 2021) practice a similar idea by copying the source sentences to the target side. However, in our issue, the source languages use logographic alphabets that have big divergences from the Latin alphabet on the target side; therefore, this approach is restricted. Our proposal does not depend on the language pairs in various alphabets.

In addition, some other works recently suggest simple and fast approaches to make dummy bilingual data by swapping (Artetxe et al. 2018), dropping (Xia et al. 2019), or replacing (Gao et al. 2019; Xie et al. 2017) words in the source or target sentences. These strategies continue to be revised in (Duan et al. 2020; Sánchez-Cartagena et al. 2021). While (Duan et al. 2020) use dependency parsers to determine dropout or replacement words during the noise generation process, (Sánchez-Cartagena et al. 2021) examine more circumstances in terms of reordering words or reversing the sentences.

The aforementioned approaches mainly utilize external resources such as dictionaries, monolingual data, or more complex additional tools such as dependency parse, language models, and pre-trained NMT models. To deal with these restrictions, in the first contribution, we introduce a new simple, and fast idea for generating synthetic data for low-resource NMT systems. Different from previous studies, we do not use any additional resources besides only a small amount of available bilingual corpus for data

augmentation. We suggest employing artificial translation units (ATUs) for the data augmentation while still preserving the word order and context of the sentence. *ATUs are labels that are generated from standard translation units based on a monolingual vocabulary*. Our experiments show substantial improvements in low-resource translation tasks.

In the second contribution, we investigate the robustness when combining Chinese and Japanese texts in the translation task from Chinese, Japanese to Vietnamese. Japanese and Chinese are logographic languages, in which, characters are constructed from ideographs or strokes (Zhang and Komachi 2018). Japanese texts often use three alphabets to transmit information, including Hiragana, Katakana, and Kanji. Hiragana and Kanji letters are mixed to depict the content of a sentence while katakana letters are used to transcript loan words or name entities. According to our knowledge, Kanji characters derive from Chinese characters; thus, their meaning is often similar or relevant to ones in Chinese with the same sharp. Although traditional Chinese and Japanese texts share a set of similar words, the structures of their sentences are in reverse. A verb follows immediately its subject in the Chinese sentence, while this verb is at the end of the Japanese sentence. This opposition is adverse for NMT if texts in these languages are concatenated for simultaneous training. To the best of our knowledge, although Japanese and Chinese are mentioned in many prior studies, none of them has examined the NMT systems that are trained in the combination of Chinese and Japanese texts on the source (Zoph and Knight 2016) or target side in particular as the observations of similar languages in (Saleh et al. 2021; Tan et al. 2019). They are often either source or target language as in (Zhang and Komachi 2018; Zhang and Matsumoto 2017), or they are mixed with other languages in multilingual NMT systems in (Aharoni, Johnson, and Firat 2019), therefore, their reciprocal impact has not been clearly assessed yet. Our work examines combined translation systems by using various recipes for tokenizing Japanese texts. The experiments show that our NMT systems still obtain improvements in performance in the low-resource issue regardless of their opposite structures.

In the last contribution, we incorporate the pre-trained BERT model (Devlin et al. 2019), which is trained on the combination of Chinese and Japanese monolingual data to evaluate the efficacy of the translation systems in the more high-resource situation. Our NMT systems achieve interesting improvements in the translation tasks.

In summary, our main contributions are the follows:

- We propose a strategy for data augmentation in NMT and present its effectiveness in two low-resource translation tasks: Chinese to Vietnamese and Japanese to Vietnamese.

• We firstly investigate the translation systems with the combination of Chinese and Japanese texts in the source languages and have achieved substantial improvements in the low-resource situations regardless of the opposition in the structure of their sentences.

• We leverage a BERT pre-trained model in order to enhance the translation performance and estimate the efficacy of the combined translation systems in the more high-resource issue. The translation tasks show the interesting improvements in the translation performance.

In the next section, we discuss previous works related to our approaches for low-resource translation tasks. Our proposal techniques are described in section 3. The training systems, datasets, and preprocessing are presented in section 4. The results are denoted in section 5. Discussion and further analysis are shown in section 6. Finally, we conclude and suggest future works in Section 7.

Related Work

To build under-resourced machine translation systems, previous works have presented several ideas for data augmentation to improve translation quality. Using the trained NMT models to generate synthetic data is a widely used approach such as in (Edunov et al. 2018; Sennrich, Haddow, and Birch 2016b) – called Back Translation. They use a backward model to produce the hypotheses of the source language to raise data for translation systems. As an otherwise idea, Zhang et al. (2018) use a forward model to predict the translations in the target language – called Self-learning. These researches require translation systems that are trained on an initial bilingual corpus; therefore, the generated pseudo data are often low quality in data sparsity issues. Our proposal also generates pseudo data for translation systems, but we do not require pre-trained backward or forward models for inference hypotheses. In addition, these methods often leverage external monolingual data to enrich the encoder or decoder while our technique only uses data from the available bilingual corpus. We hope that our method could help initial systems better in generating dummy data for Back Translation or Self-learning, and in future works, we will consider more such experiments.

On the other hand, Ha, Niehues, and Waibel (2016) show a simple technique, called Mix-source which makes a copy of target sentences on the source side; otherwise, Sánchez-Cartagena et al. (2021) inversely make a copy of source sentences into the target side. These methods benefit NMT systems when common terms or names are shared between source and target languages, and they can be translated more accurately. These studies also allow to leverage the monolingual data from the available bilingual corpus the same as our idea; however, we use ATUs instead of making copies of

source or target data in pseudo data. Moreover, in our situation, the source languages use the logographic alphabet (Chinese and Japanese) while the target language (Vietnamese) uses the Latin alphabet; therefore, these recipes are less efficient.

Other studies propose fast and simple approaches to make dummy bilingual data by swapping (Artetxe et al. 2018), dropping (Xia et al. 2019), or replacing (Gao et al. 2019; Xie et al. 2017) words in the sentences. In detail, for the unsupervised translation task, Artetxe et al. (2018) suggest random swaps of contiguous words in the input sentence to create noisy sentences. Xia et al. (2019) create pseudo data by dropping or swapping words randomly of the input sentences. The other strategies for making dummy data from the original sentences, Xie et al. (2017) utilize probabilities from an n-gram model to determine replacements of words and have shown the benefits in language model and machine translation. Gao et al. (2019) propose to use a language model to select appropriate words for substitution words in the sentences in NMT. In the same line, recently, Duan et al. (2020) and Sánchez-Cartagena et al. (2021) revise swap, dropout, and replacement strategies of words for data augmentation in their works. Duan et al. (2020) employ a dependency parse tree to identify these modifications, while (ánchez-Cartagena et al. (2021) experimentalize more cases such as reversing the sentence order or exploiting an alignment dictionary to produce synthetic data. Our idea is similar to these approaches in terms of modification of the native sentences for generating synthetic data and is closest to the proposal in (Sánchez-Cartagena et al. 2021) when the target sentences are transformed while the source sentences are unaltered. However, the aforementioned works mostly require external resources for generating synthetic data in the low-resource issue such as the unsupervised NMT system, language model, probabilistic model, or dependency parser. These techniques may propagate errors from the auxiliary tasks to the translation task, while our method does not face this problem. In addition, random swap, dropout, or employ alignment dictionaries as in Artetxe et al. (2018; Sánchez-Cartagena et al. 2021) may produce non-fluent sentences, which are not encouraged in NMT, while our approach does not change the word order in dummy sentences. Besides Tu et al. (2017) have proposed a reconstruction based on the hidden states of the decoder to help NMT generate better translations. This idea is then considered by Niu, Xu, and Carpuat (2019) in the bi-directional NMT. This approach is near to our idea when leveraging the information on the target side, but it requires a separate reconstruction with an objective function.

In brief, our data augmentation strategy is different from previous works in the following:

- Our method does not require pre-trained NMT models (backward or forward) for inference of synthetic data to avoid the generation of inaccuracy translations in the low-resource issue.

• We do not employ external resources such as additional monolingual data, manual or alignment dictionaries, unsupervised NMT system, language model, dependency parser, etc., to alleviate the error propagation from the auxiliary tasks to the translation task.

• Previous works mostly modify source sentences for generating synthetic data to avoid the production of non-fluent translations in the target language, while our technique transforms target sentences and does not change the word order of sentences in the pseudo data.

Along with the data augmentation, we investigate the efficacy of low-resource translation tasks in terms of the combination of Chinese and Japanese texts on the source side in the translation task from Chinese, Japanese to Vietnamese (called the combined training system in our work). We expect that the translation tasks can take advantages of the sharing of common translation units between these two languages. Previous works mostly experimentalize one of them either the source language or the target language Zhang et al. (Zhang and Komachi 2018; Zhang and Matsumoto 2017). In the other words, they are investigated in multilingual NMT systems with more mixed other languages (Aharoni, Johnson, and Firat 2019). Therefore, the efficacy of this strategy has not been clearly investigated yet.

Furthermore, we incorporate the BERT language model that is introduced by (Devlin et al. 2019) to our NMT systems. The BERT model has demonstrated its effectiveness for translation tasks in (Clinchant, Jung, and Nikoulina 2019; Zhu et al. 2020). Different from previous works our BERT model is trained on the combination of Japanese and Chinese texts that have inverse structures in terms of grammar. Our aim is to investigate the efficiency of our combined training systems in more high-resource situations.

Improving Low-Resource Neural Machine Translation

In this section, we present techniques to enhance translation quality in the low-resource issue. Alongside our proposed data augmentation, the combined training system from Chinese, Japanese to Vietnamese is shown. Then, a data selection method that is used to filter monolingual data and adaptation of the combined pre-trained BERT model to the NMT system is also described.

Generating Synthetic Data for low-resource Neural Machine Translation

We introduce a simple and fast strategy for creating pseudo data from the available low-resource bilingual corpus.

The proposed method includes the following steps:

(1) Extracting a vocabulary from the target text of the original bilingual corpus. This vocabulary contains standard translation units (can be words, sub-words, or characters) with their frequencies side by side.

(2) Generating artificial translation units (ATUs, or artificial tokens) by tagging a label to each standard translation unit in the vocabulary and obtaining a labeled vocabulary. Each label or ATU is unique and does not overlap with standard translation units. In the experiments, we use special symbols such as id0, id1, etc., to present ATUs.

(3) For an original sentence pair, we generate corresponding dummy sentences from the target sentence based on ATUs. These dummy target sentences are then paired with the source sentence in the native sentence pair to make synthetic data.

Table 1 illustrates dummy sentence pairs that are generated from an original sentence pair, and ATUs correspond to the native tokens in the target sentence as in Figure 1.

The synthetic corpus is then concatenated with the native corpus for training NMT systems. Our overall method is described in Figure 2. If all standard translation units in the original vocabulary are replaced by their ATUs, the vocabulary size of the data augmented NMT system will be doubled, this raises the size as well as the time training of NMT models. To reduce this adverse impact, we use an appropriate threshold of frequency to determine a standard translation unit that could be replaced or stayed. In this way, our technique enables us to produce a set of diverse dummy sentences from the original sentence by using different thresholds as in Table 1. This alleviates the parameters of NMT models substantially while upgrading

Table 1. An example of synthetic data is generated from the original sentence pair on our **TED Talk** datasets using various thresholds. If the threshold (ths) is *ths* = *x*), standard translation units (tokens) whose frequency is greater than *x* in the target vocabulary will be replaced by their ATUs respectively. The source sentence is preserved while the target sentence is transformed in the dummy data pairs.

Original pair	src: 这 是 我 在 印度 的 拉达克 地区 观看 到 的 喜马拉雅山脉 。 (This is an observatory in the Himalayas, in Ladakh in India .) tgt: Vietnamese đây là đài quan sát ở Himalayas, và ở Ladakh, Ấn Độ .
Dummy pair (*ths* = 0)	src: 这 是 我 在 印度 的 拉达克 地区 观看 到 的 喜马拉雅山脉 。 tgt: id41 id4 id1456 id83 id427 id28 id6004 id0 id2 id28 id12596 id0 id781 id833 id1
Dummy pair (*ths* = 7)	src: 这 是 我 在 印度 的 拉达克 地区 观看 到 的 喜马拉雅山脉 。 tgt: id41 id4 id1456 id83 id427 id28 id6004 id0 id2 id28 **Ladakh** id0 id781 id833 id1

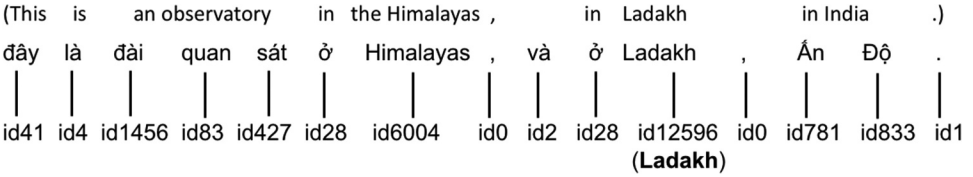

Figure 1. ATUs are corresponding to the standard translation units in the target sentence.

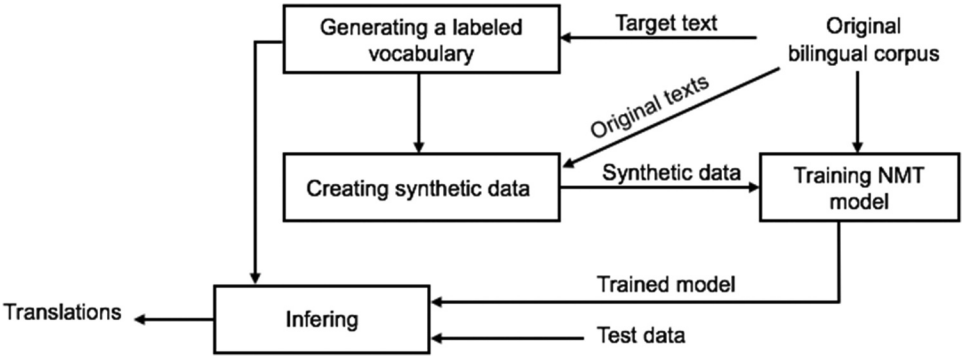

Figure 2. Our overall method for generating synthetic data and integrating it into the NMT system.

translation quality. In reality, the data augmentation for low-resource language pairs in the mentioned studies may be also increased the parameters of NMT systems due to additional resources or auxiliary tasks, so this is a normal phenomenon in our proposal. Additionally, due to the biased tendency of NMT in translating high-frequency translation units, the usage of the frequency threshold allows us to mask them to facilitate the lower-frequency others to be considered.

Moreover, our recipe does not depend on types of translation units (words, sub-words, or characters); therefore, Sennrich's BPE (byte-pair-encoding) (Sennrich, Haddow, and Birch 2016a) technique can be earlier applied to the texts to minimize the parameters. Our proposal only applies to the training corpus and does not affect the development or evaluation data. In the inference, ATUs can be predicted instead of standard translation units, the labeled vocabulary in step 2 is then used to discover the corresponding standard translation units.

Our experiments present interesting improvements when using a double synthetic data of native parallel corpus in two low-resource translation tasks from Chinese to Vietnamese and Japanese to Vietnamese on the TED Talks datasets.[1] In particular, the Japanese to Vietnamese translation task gains divergent improvements when Japanese texts are segmented by three different strategies. We aim to investigate the translation quality in terms of variable sentence lengths due to segmentation and sharing translation units with Chinese in combined systems in section 3.2. In addition, our method shows competitive results compared to the NMT systems that utilize the Back Translation strategy for data augmentation.

Combined Training of Chinese and Japanese in Neural Machine Translation

As mentioned in the previous sections, Chinese and Japanese texts share information through Kanji characters in Japanese. Texts written in these languages do not use spaces to delimit the boundaries of words. Therefore,

they need to be segmented in the preprocessing of natural language processing tasks in general as well as in machine translation. As shown in (Phuoc Tran, NGUYEN, and Long 2016), Chinese is known as a non-morphology language or monosyllabic language. Thus, we anticipate that its translation quality in NMT may be less affected by different segmentation tools. In contrast, Japanese is a morphologically rich language with different variations of a word. Moreover, auxiliary words are also supplemented to indicate diverse contexts of the sentence. This makes Japanese sentences tend to be long and may be affected by divergent segmentation.

Our study will focus on several segmentation for Japanese texts using kytea,[2] mecab,[3] and spacy.[4] We then estimate the effectiveness of these segmented texts when they are mixed with Chinese texts for training NMT systems. From the point of grammar view, the same as English or Vietnamese, Chinese has the sentence structure in terms of **SVO** (subject-verb-object); otherwise, this is **SOV** (subject-object-verb) in Japanese sentence. This inversion often reduces the efficiency of NMT translation systems when these languages are concatenated for training. In the low-resource issue, however, we still achieve substantial improvements in the translation performance of two translation tasks from Chinese to Vietnamese and Japanese to Vietnamese. To authenticate this, two low-resource datasets including TED Talks and ALT (Riza et al. 2016) are used for our evaluation.

Incorporating BERT into Low-Resource Neural Machine Translation

Our research explores BERT (Devlin et al. 2019) for two aims: (1) leveraging the benefit of the pre-trained model to improve the translation performance in the data-sparse situation as shown in prior works; (2) evaluating the further impact of the opposition in word orders to NMT systems when combining Chinese and Japanese texts for training in the high-resource situation.

Due to the availability of monolingual resources, BERT models are usually trained on large-scale data. However, the monolingual data often contain sentences in the general domain (out-domain), and therefore, to obtain the best efficacy in the TED Talks domain (in-domain), it is filtered before training the pre-trained BERT model for incorporation into the translation systems.

Filtering Monolingual Data

We use Chinese and Japanese monolingual data from CCAligned[5] (El-Kishky et al. 2020) for training the BERT model. It is available for both Japanese and Chinese. Then, the TF-IDF (Term Frequency-Inverse Document Frequency) measure (Salton and Yang 1973) is used to select sentences from the out-domain that are relevant to the in-domain (Dou, Anastasopoulos, and Neubig 2020; Eck, Vogel, and Waibel 2005; Silva et al. 2018). This method is simple,

fast, and effective Dou, Anastasopoulos, and Neubig (2020), in which, TF (Term Frequency) denotes the ratio between the times of a term in a document and the total of terms in this one, while IDF (Inverse Document Frequency) is the ratio between the total number of documents and the number of documents containing the term.

In our case, we call G and I are out-domain and in-domain corpus, respectively. TF_w of particular word w in the sentence $s \in G$ is the ratio between the number of occurrences of w in I and total words in s while IDF indicates the rate of all sentences in I and the sentences that contain the given word.

Specifically, for each sentence $s \in G$, we calculate the similarity score of each word w in s to I as in the formula 1:

$$TF - IDF_w = \frac{F_w^I}{L_s} . \frac{N_I}{N_w} \tag{1}$$

where F_w^I is the number of occurrences w in I, L_s is the total words in of s, and N_I is the total sentences of I, N_w is the number of sentences that contain w in I.

The similarity weight of each $s \in G$ to I as:

$$score_{s \in G}^I = \sum_{i=1}^{L_s} TF - IDF_{w_i} \tag{2}$$

Thus, if a sentence contains many words that appear in I, its score will be higher, and vice versa. These scores are then used to rank sentences in corpus G. The sentences that have the highest scores nearest to the in-domain corpus will be selected.

To remove less potential data for NMT from CCAligned corpus, we filter out sentences with lengths below 8 or above 100 prior to the TF-IDF filtering. The top 2.5MB of sentences with the highest scores from filtered datasets of each language (Japanese or Chinese) are chosen. They are subsequently concatenated together and mixed with the bilingual in-domain corpus to train our BERT model.

Incorporating BERT into Neural Machine Translation

We train a BERT model relying on the filtered monolingual data in the section 3.3.1. The output vectors of the trained BERT model are then used as the inputs of NMT systems instead of initializing random embeddings corresponding to the sentences. Using a pre-trained model, we hope that our translation systems will be obtained further improvements in the translation performance.

Moreover, the BERT model is trained on mixed data of Chinese and Japanese texts, thus we aim to investigate the translation quality in the high-resource scenarios when these languages have the structural contradiction in terms of grammar. In practice, we observe that the Japanese to Vietnamese

translation task increase translation quality while this decreases in the Chinese to Vietnamese translation task. The settings of our BERT models are presented in section 4.1.

Experiments

Toolkits and Experimental Settings

We conduct NMT experiments using NMTGMinor[6] which implements the transformer architecture relying on (Vaswani et al. 2017). Our baseline systems exploit six layers for both encoder and decoder, the size of hidden units as well as embeddings are 512 dimensions for each layer. Adam optimizer (Kingma and Ba 2014) is utilized for parameter optimization with the initial learning rate at 1.0 and the *warmup* is set to 480 for the ALT corpus, and 4800 for the TED Talks corpus. The size of mini-batches in sentences is 64 pairs for bilingual systems and the systems trained on the ALT corpus, while this is 128 for the remaining systems. The vocabulary size is 40K of the most frequency tokens for both source and target sides in the bilingual and Back Translation systems, while this is 60K for the remaining systems. The beam search algorithm is employed to predict translations in the inference process with a beam size of 10 and an alpha factor of 0.2. Other settings follow the defaults of NMTGMinor.

For the pre-trained BERT model, we use the BERT framework[7] introduced by Devlin in Devlin et al. (2018), it is based on the tensorflow library. Our BERT system includes six layers and its vocabulary is extracted from Chinese and Japanese texts of bilingual datasets with all available tokens. The hidden size is 512, and other settings are default following the framework. The NMT baseline architecture is then modified to integrate the trained BERT model into its encoder. We remove the embedding layer from the original NMT system, alternately, the output vectors from the BERT model will be inputs of the encoder in the modified translation system.

We set the length of sentences in the training process to 150 tokens for both NMT and BERT systems. To evaluate the accuracy of translation systems, the BLEU measure (Papineni et al. 2002) which is implemented in SacreBLEU[8] (Post 2018) is utilized to estimate the difference between 1-best hypothesis and its reference translation.

Datasets and Prepocessing

We extract bilingual datasets from TED Talks[9] for two language pairs: Chinese-Vietnamese, and Japanese-Vietnamese. For the Chinese-Vietnamese pair, we separate it into the train, validation, and test sets. The validation and test sets have not been published before. For the Japanese-Vietnamese pair, we

Table 2. The number of the sentence pairs is in our bilingual datasets.

Datasets	Domain	Training	dev	test
Chinese-Vietnamese	TED Talks	244076	1316	1296
	ALT	18088	1000	1018
Japnese-Vietnamese	TED Talks	244417	568	1220
	ALT	18088	1000	1018

utilize developer and test sets in (Ngo et al. 2018), thus, they are cleaned from collected data to obtain the training set. In addition, to verify the effectiveness of the NMT system from Chinese, Japanese to Vietnamese, ALT datasets[10] are also utilized. The statistics of the datasets are listed in Table 2.

Monolingual datasets of Chinese and Japanese for training our BERT model from CCAlinger include 15MB sentences. They are then filtered using the method in section 3.3.1 . After that, we get the top 2.5MB sentences in terms of the highest scores in each language to mix them before concatenating them with ones from the TED Talks domain for training the BERT model. Table 3 shows specific datasets to train the BERT model in sentences. For Back Translation NMT systems, 10MB of monolingual data for Vietnamese from VLSP 2020 (Ha, Tran, and Nguyen 2020) is also filtered utilizing TF-IDF in section 3.3.1 to select sentences near the TED Talks domain.

In the pre-processing, Chinese texts are segmented using jieba,[11] Japanese texts are segmented exploiting different tools for comparison purpose, including kytea, spacy, and mecab. Vietnamese texts are tokenized and true-cased using moses scripts.[12]

To reduce the vocabulary size, the texts are applied to Sennrich's BPE (Sennrich, Haddow, and Birch 2016a) with 30,000 merge operations that are learned from their corresponding ones of the bilingual corpus, exceptionally Vietnamese texts in ALT corpus, we do not utilize this.

The synthetic datasets are generated from the native parallel corpus using the recipe in the section 3.1 with random frequency thresholds of 0 and 7.

Training

Our experiments are conducted on an NVIDIA GeForce GTX 1080 with 12GB VRAM. All NMT systems to Vietnamese are trained after 40 epochs for comparison purposes. The backward models for Back Translation are trained

Table 3. The number of monolingual sentences is used for training our BERT model.

Monolingual Datasets	Domain	Sentences
Chinese	TED Talk	244076
	CCAlinger	2500000
Japnese	TED Talk	244417
	CCAlinger	2500000
The total	–	5488493

after 70 epochs to obtain the convergence of the perplexity measure in training datasets. Our BERT model is trained for 150.000 steps. After that, the last model is converted to pytorch format before being adapted into our NMT systems. These NMT systems are then trained after 70 epochs for convergence.

The development sets are often used for early stopping in the training process. However, ATUs are only applied for training sets in our data augmentation systems. Therefore, the best model in terms of its accuracy in training data from the NMT systems is always used to infer the test sets in our experiments.

Results

Our experimental results are presented in the Table from 4 to 11 in BLEU scores.

Baseline Systems

The baseline systems are only trained on bilingual datasets of each language pair after 40 epochs. For the Chinese (Cn) to Vietnamese (Vi) translation task, it obtains 17.4 BLEU scores. For Japanese (Ja) to Vietnamese translation task, we show the various results when Japanese texts are segmented by kytea, or spacy, or mecab. Our aim is to estimate the impact of these segmentation strategies on translation performance with the variation of vocabularies and sentence length. In Table 4, we see that the systems that utilize spacy (15.9 BLEU scores) or mecab (15.4 BLEU scores) are better than the one using kytea

Table 4. Our data augmentation systems overcome the baseline systems on **TED Talks** datasets with the frequency threshold of 7 for replacement ATUs by standard translation units in the synthetic data. Our method is also compared to the Back Translation technique.

Translation tasks	Systems	Dev	Test
Cn→ Vi	**Bilingual baseline**	17.1	17.4
	Our data augmentation (*ths = 7*)	**17.2** (+0.1)	**17.9** (+0.5)
	Back Translation	**18.0** (+0.9)	**18.5** (+1.1)
	Our data augmentation + Back Translation	**18.3** (+1.2)	**18.6** (+1.2)
Ja→ Vi	**Bilingual baseline (ja-kytea)**	14.1	15.1
	Our data augmentation (ja-kytea, *ths = 7*)	**16.8** (+2.7)	**18.0** (+2.9)
	Bilingual baseline (ja-spacy)	14.9	15.9
	Our data augmentation (ja-spacy, *ths = 7*)	**17.6** (+2.7)	**18.4** (+2.5)
	Back Translation (ja-spacy)	13.4 (−1.5)	14.3 (−1.6)
	Bilingual baseline (ja-mecab)	14.0	15.4
	Our data augmentation (ja-mecab, ths = 7)	**18.1** (+4.1)	**19.4** (+4.0)
	Back Translation (ja-mecab)	13.7 (−0.3)	14.6 (−0.8)
	Our data augmentation (*ths = 7*) + Back Translation (ja-mecab)	**15.4** (+1.4)	**16.6** (+1.2)

Table 5. Our data augmentation systems overcome the baseline systems on **ALT** datasets with the frequency threshold of 7 for replacement ATUs by standard translation units in the synthetic data.

Translation tasks	Systems	Dev	Test
Cn→ Vi	**Bilingual baseline**	9.9	9.5
	Our data augmentation (*ths = 7*)	**11.8** (+1.9)	**11.6** (+2.1)
Ja→ Vi	**Bilingual baseline (ja-kytea)**	8.5	8.5
	Our data augmentation (ja-kytea, *ths = 7*)	**9.7** (+1.2)	**9.7** (+1.2)
	Bilingual baseline (ja-spacy)	8.2	8.0
	Our data augmentation (ja-spacy, *ths = 7*)	**9.8** (+1.6)	**9.4** (+1.4)
	Bilingual baseline (ja-mecab)	8.0	7.8
	Our data augmentation (ja-mecab, *ths = 7*)	**9.5** (+1.5)	**9.4** (+1.6)

(15.1 BLEU scores) on TED Talks corpus. For ALT datasets in Table 5, when spacy is used to segment Japanese texts, and the baseline systems for two translation tasks are 9.5 and 8.0 BLEU scores, for mecab and kytea as 7.8 and 8.5 BLEU points.

Our Data Augmentation Systems

The data augmentation systems are trained on the concatenation of native parallel corpus and their respective synthetic datasets, which are created using our method in section 3.1. We set the frequency threshold for the replacement of ATUs as 7, and our statistics in Table 6 shows that most of the sentences on the target side contain ATUs. Thus, the size of the synthetic datasets is quite the same as the size of the native bilingual datasets. In our observation, all translation tasks have obtained improvements over the baseline systems in Table 4 on the TED Talks domain. The translation system from Chinese to Vietnamese gains of +0.5 BLEU score while the ones from Japanese to Vietnamese get different improvements and achieve the best result when using mecab for Japanese text segmentation (+4.0 BLEU scores). We will analyze the further cause of these divergences in the discussion section.

In Table 5, both translation tasks have achieved improvements in ALT datasets with +2.1 BLEU scores in the Chinese to Vietnamese system, while the translation systems from Japanese to Vietnamese do not have difference improvements when using three segmentation tools.

Table 6. The number of sentences in the target side (Vietnamese) contains ATUs when using the *threshold = 7* in TED Talk and ALT datasets.

Bilingual datasets	All sentences	Number of sentences contain ATUs
TED Talk Chinese → → Vietnamese	244076	244075
TED Talk Japanese → Vietnamese	244417	244417
ALT Chinese or Japanese → Vietnamese	18088	18088

For comparison purposes, we train backward models from Vietnamese to Chinese and Vietnamese to Japanese after 70 epochs to obtain convergent models. These models are then used to infer Vietnamese monolingual sentences into corresponding Chinese or Japanese sentences to produce synthetic data. For fair evaluation, the size of the pseudo dataset is also equal to the size of the bilingual dataset as well as the size of the pseudo dataset in our method. The synthetic datasets are concatenated to original bilingual datasets of each language pair to train translation systems from Chinese or Japanese to Vietnamese. We show that the translation system from Chinese to Vietnamese with Back Translation gains bigger BLEU scores (+1.1) than our proposal. For the translation systems from Japanese to Vietnamese, we experiment with the Back Translation technique in the situations that Japanese texts are segmented using spacy and mecab, and find that the translation performance is degraded substantially (−1.6 and −0.8 BLEU scores). We speculate that the backward translation system (from Vietnamese to Japanese) may not be powerful enough to produce good-quality translations. The data augmentation in our proposal does not face this problem.

Furthermore, we examine the combination of our method with Back Translation in translation systems. In this situation, our strategy is first applied to the backward models (from Vietnamese to Chinese, or Vietnamese to Japanese) for inference of synthetic data and then applied to the forward models (from Chinese, or Japanese to Vietnamese) again. In Table 4, we observe a further improvement of +1.2 BLEU scores in the Chinese to Vietnamese. Although the combination of our method and Back Translation has gained performance improvement, this is not significant compared to the systems that only utilize either our strategy or Back Translation while it uses double training data. In the Japanese to Vietnamese, due to the performance degradation of the Back Translation technique, we only consider the translation system that employs mecab for segmentation of Japanese texts. An improvement of (+1.2) BLEU scores has been obtained over the baseline system; however, this is still below the translation system that only employs our proposal. Thus, NMT systems do not have benefits when applying the Back Translation strategy in our low-resource scenario, while the proposed approach can deal with this problem. In addition, our proposal could be used in combination with Back Translation for data augmentation in the sparse data situation to obtain better improvements.

For more detail about backward models, their BLEU scores of them are shown in Table 7, in this table, vanilla backward models are applied to our proposed method. The vanilla backward system from Vietnamese to Japanese has achieved a big improvement of +5.38 BLEU points, while this is equal to the baseline system in the Vietnamese to Chinese.

Table 7. The BLEU scores in backward models from baseline systems and vanilla systems in **TED Talk** Datasets. We use spacy for segmentation Japanese texts.

		Vi → Cn		Vi → Ja	
No.	Systems	dev	test	dev	test
1	Baseline Backward	10.5	11.30	12.00	16.02
2	Vanilla Backward	**11.1** (+0.6)	11.30	**22.10** (+10.10)	**21.40** (+5.38)

Chinese and Japanese to Vietnamese Translation Systems

We concatenate the original bilingual corpus of two language pairs for training together (combined training). The results are shown in the Tables 8 and 9. In this case, for the Japanese to Vietnamese translation task, the bilingual system that gains the best BLEU score is chosen as the baseline system (using spacy). To demonstrate the effectiveness of this strategy, ALT datasets are also utilized. In this case, Japanese texts are only segmented by spacy. Improvements are gained in the performance of all translation systems. Specifically, on the TED Talks dataset, we obtained a better gain of **+1.3** BLEU points in the Chinese to Vietnamese task when kytea is used for segmentation of Japanese texts and **+0.9** BLEU points in the Japanese to Vietnamese task when mecab is used for the same purpose. On ALT datasets, these BLEU points are **+0.8** and **+1.4** for respective translation tasks. Thus, both Japanese to Vietnamese and Chinese to Vietnamese translation tasks are beneficial when using combined training in the NMT systems from Chinese, Japanese to Vietnamese regardless of the opposition in structures of Chinese and Japanese sentences.

Our data augmentation method continues to be applied to the combined training systems to demonstrate its effectiveness in the translation tasks with the same frequency threshold of 7 for the replacement of ATUs. On the TED Talks datasets, we achieve further improvements in the Japanese to Vietnamese translation task for all various segmentation (the best one is **+2.7**) while these are not substantial in the Chinese to Vietnamese translation task compared to combined training systems, though they overcome the baseline systems. On the ALT domain, gains are **+1.8** and **+1.9** for these translation tasks, respectively.

Table 8. The practical results of combined translation systems from Chinese, Japanese to Vietnamese on the **TED Talks** datasets.

		Cn → Vi		Ja → Vi	
No.	Systems	dev	test	dev	test
1	Bilingual baseline	17.1	17.4	14.9	15.9
2	Combined training (ja-kytea)	18.2 (+1.1)	**18.7** (+1.3)	15.2 (+0.3)	16.4 (+0.5)
3	Combined training (ja-spacy)	**18.4** (+1.3)	18.2 (+0.8)	14.7 (−0.2)	16.3 (+0.4)
4	Combined training (ja-mecab)	18.0 (+0.9)	18.0 (+0.6)	**15.8** (+0.9)	**16.8** (+0.9)
5	Combined training (ja-kytea) + Our data augmentation (*ths* = *7*)	17.9 (+0.8)	18.4 (+1.0)	16.8 (+1.9)	17.9 (+2.0)
6	Combined training (ja-spacy) + Our data augmentation (*ths* = *7*)	**18.3** (+1.2)	**18.6** (+1.2)	**17.1** (+2.2)	**18.6** (+2.7)
7	Combined training (ja-mecab) + Our data augmentation (*ths* = *7*)	17.7 (+0.6)	18.4 (+1.0)	16.8 (+1.9)	18.5 (+2.6)

Table 9. The practical results in combined training systems from Chinese, Japanese to Vietnamese on the **ALT** datasets.

No.	Systems	Cn → Vi		Ja → Vi	
		dev	test	dev	test
1	Bilingual baseline	9.9	9.5	8.2	8.0
2	Combined training (ja-spacy)	**10.3** (+0.4)	**10.3** (+0.8)	**9.2** (+1.0)	**9.4** (+1.4)
3	Combined training (ja-spacy) + Our data augmentation	**11.4** (+1.5)	**11.3** (+1.8)	**10.4** (+2.2)	**9.9** (+1.9)

Incorporating BERT Model to NMT Systems

We incorporate a BERT model that is trained on a larger monolingual dataset of Chinese and Japanese into the combined translation systems with the proposal data augmentation. The results are described in the Tables 10 and 11. As aforementioned purpose in the previous sections, our translation systems perform better when the pre-trained BERT model is integrated. The great improvements are shown on both TED Talks and ALT datasets in the Japanese to Vietnamese translation task with **+5.4** and **+5.8** of BLEU scores. The Chinese to Vietnamese translation task only presents the efficacy on the ALT corpus with **+5.3** of BLEU scores while this score on the TED Talks domain is −0.4.

After 70 epochs, the systems gain convergence on the perplexity measure of the development and training sets, we again observe further improvements of **+6.2** and **+6.4** BLEU scores for two translation tasks on the ALT corpus. On the TED Talks domain, the same as BERT incorporated translation systems, an improvement of **+7.8** points of BLEU scores is found in the Japanese to Vietnamese translation task while the performance in the Chinese to Vietnamese translation task tends to degradation.

Discussion

For further analysis, Table 4 shows various effectiveness for different segmentation strategies of Japanese texts. We find that Japanese sentences segmented by kytea are lengthier than the ones segmented by spacy or mecab. The statistics are detailed in Table 12.

Japanese sentences segmented by mecab have the shortest lengths. Prior works have proved that NMT systems learn well for shorter sentences than for longer sentences. Therefore, this leads to better translations. This also reduces the number of sentences that are discarded from the training datasets and the number of dummy sentence pairs is also increased when employing our data augmentation. In ALT datasets, the number of sentences are quite the same for all segmentation tools, and the distances among the average lengths are shorter, so the BLEU scores do not have big differences. The examples of translations from Japanese to Vietnamese NMT systems on the TED Talks domain are presented in Table 13.

Table 10. The practical results of translation systems when incorporating the BERT model on the **TED Talks** datasets. The system in (4) is then trained continuously to epoch 70 to achieve the state of the art.

		Cn → Vi		Ja → Vi	
No.	Systems	dev	test	dev	test
1	Bilingual baseline	17.1	17.4	14.9	15.9
2	Pre-trained BERT (ja-spacy) + Combined training	16.9 (–0.2)	17.2 (–0.2)	**17.2** (+2.3)	**17.6** (+1.7)
3	Pre-trained BERT (ja-kytea) + Combined training	**17.2** (+0.1)	**17.6** (+0.2)	**17.4** (+2.5)	**17.1** (+1.2)
4	Pre-trained BERT (ja-spacy) + Combined training + Our data augmentation (*ths* = *7*)	16.5 (–0.6)	17.0 (–0.4)	**20.9** (+6.0)	**21.3** (+5.4)
5	Pre-trained BERT (ja-spacy)+ Combined training+ Our data augmentation (*ths* = *7*)+ Continue to epoch 70	15.8(-1.3)	16.1(-1.3)	**23.3**(+8.4)	**23.7**(+7.8)

Table 11. The practical results of translation systems when incorporating the BERT model on the **ALT** datasets. The system in (3) is also trained continuously to epoch 70.

		Cn → Vi		Ja → Vi	
No.	Systems	dev	test	dev	test
1	Bilingual baseline	9.9	9.5	8.2	8.0
2	Pre-trained BERT (ja-spacy) + Combined training	**13.9** (+4.0)	**13.8** (+4.3)	**13.0** (+4.8)	**12.8** (+4.8)
3	Pre-trained BERT (ja-spacy) + Combined training + Our data augmentation (*ths* = *7*)	**14.9** (+5.0)	**14.8** (+5.3)	**14.5** (+6.3)	**13.8** (+5.8)
4	Pre-trained BERT (ja-spacy) + Combined training + Our data augmentation (*ths* = *7*) + Continue to epoch 70	**15.6**(+5.7)	**15.7** (+6.2)	**15.1** (+6.9)	**14.4** (+6.4)

In Table 13, the translation (****) which uses mecab for segmentation texts in our data augmented NMT system is more accurate than the other ones. It helps us understand the correct meaning of the source sentence. The translations in (*) and (**) are the same and inaccurate. It denotes the wrong meaning of the source sentence. Although the translation in (***) is better than in (*) and (**), it is still incorrect.

For our proposal data augmentation, the target vocabulary size may be reached the maximum value (be double its original size) when all tokens in target texts are replaced by their ATUs. In this situation, we can produce many pseudo sentences from an original target sentence to reinforce training data for translation systems with various thresholds. Due to time limitations, we only investigate NMT systems in the issue that a synthetic sentence is generated from an original one. Tables 14 and 15 show the BLEU scores and the corresponding vocabulary sizes in two datasets with frequency thresholds of replacement are 0, and 7 in combined training systems.

In Tables 14 and 15, the translation systems that use pseudo data have larger sizes of the target vocabularies compared to the systems in (1). The systems in (2) (threshold = 0) have double the size of the target vocabularies in (1), while the systems in (3) (threshold = 7) only possess the vocabulary sizes, which are much smaller than the ones in (2). Nevertheless, the translation performances of the systems in (3) are still equal to or better than the ones in (2). Therefore, we can conclude that it is not necessary to use all artificial tokens, we still achieve substantial improvements in translation performance. This reduces the size of the target vocabulary and saves memory.

For combined training systems in Table 8, we observe that the longer Japanese sentences are segmented (the statistics in Table 14), the better translations in the translation task from Chinese to Vietnamese are predicted. Thus, the combined translation systems that use mecab for segmenting Japanese texts obtain lower performance in this task (18.0 BLEU points), while these are 18.2 for spacy or 18.7 for kytea. Otherwise, using kytea for

Table 12. The number of sentence pairs in the bilingual training dataset of Japanese – Vietnamese when Japanese texts are segmented by kytea, or spacy, or mecab with the limitation in lengths of 150 source tokens in the training process.

Datasets	Total sentence pairs	by kytea (sent pairs per average length)	by spacy (sent pairs per average length)	by mecab (sent pairs per average length
TED Talks	244417	243079/29.91 (30.15 (after bpe)	243330/26.302 26.58 (after bpe)	243363/25.35 25.74 (after bpe)
ALT	18088	18076/38.12 (38.98 (after bpe)	18077/35.24 36.20 (after bpe)	18077/33.85 35.04 (after bpe)

Table 13. An example of translations in NMT systems from Japanese to Vietnamese employing our data augmentation (aug) method in the Table 1 from bilingual systems.

Original pair	source	今年はチャールズ・ダーウィンの生誕200年です
	ref	Năm nay chúng ta kỷ niệm 200 năm ngày sinh của Charles Darwin. (This year we celebrate the 200th birthday of Charles Darwin.
Bilingual (ja-mecab, *ths = 7*)	source	今年 は チャ@@ ー@@ ルズ・ダ@@ ー@@ ウィン の 生誕 200 年 です
	trans(*)	Năm nay là ngày sinh của Charles Darwin. (This year is the date of birth of Charles Darwin.)
Our data aug (ja-kytea, *ths = 7*)	source	今年 は チャ@@ ー@@ ルズ・ダ@@ ー@@ ウィン の 生誕 200 年 で す
	trans(**)	Năm nay là ngày sinh của Charles Darwin. (This year is the date of birth of Charles Darwin.)
Our data aug (ja-spacy, *ths = 7*)	source	今年 は チャ@@ ー@@ ルズ・ダ@@ ー@@ ウィン の 生誕 200 年 です
	trans(***)	Năm nay là ngày sinh 200 của Charles Darwin. (This year is the 200th date of birth of Charles Darwin)
Our data aug (ja-mecab, *ths = 7*)	source	今年 は チャ@@ ー@@ ルズ・ダ@@ ー@@ ウィン の 生誕 200 年 です
	trans(****)	Năm nay là 200 năm ngày sinh của Charles Darwin. (This year is 200 years of Charles Darwin's birthday)

Table 14. Results in the BLEU score and the target vocabulary sizes of combined translation systems with frequency thresholds of replacement are *ths* = 0 and *ths* = 7 on the **ALT** datasets. Japanese texts are segmented by spacy.

No.	Systems	Cn → Vi		Ja → Vi		Target vocabulary size (token)
		dev	test	dev	test	
1	Combined training (ja-spacy)	10.3	10.3	9.2	9.4	17706
2	Combined training (ja-spacy) + Our data aug *(ths = 0)*	11.1	11.1	10.1	**10.0**	35408
3	Combined training (ja-spacy) + Our data aug *(ths = 7)*	**11.4**	**11.3**	**10.4**	9.9	21160

Table 15. Results in the BLEU score and the target vocabulary sizes of combined translation systems which are augmented dummy data with frequency thresholds of replacement are *ths* = 0 and *ths* = 7 on **TED Talks** datasets. Japanese texts are segmented by spacy.

No.	Systems	Cn → Vi		Ja → Vi		Target vocabulary size (token)
		dev	test	dev	test	
1	Combined training (ja-spacy)	18.4	18.2	14.7	16.3	26571
2	Combined training (ja-spacy) + Our data augmentation *(ths = 0)*	17.8	18.5	**17.1**	18.2	53138
3	Combined training (ja-spacy) + Our data augmentation *(ths = 7)*	18.3	**18.6**	**17.1**	**18.6**	39298

the segmentation, the translation task from Japanese to Vietnamese gains the least improvements +0.5 and +2.0 in lines (2) and (5). Line (3) and line (6) in this table show good improvements for both translation tasks. Therefore, we recommend that spacy should be utilized for the segmentation of Japanese texts in the combined translation systems. The experiments in Table 9 also confirm the translation efficiency of this approach on the ALT datasets.

Table 8 also shows that the data augmented translation systems only achieve equivalent improvements to the combined translation systems in the translation task from Chinese to Vietnamese, while the translation task from Japanese to Vietnamese gets bigger improvements. For this reason, we use a BERT model that is trained from the amount of larger monolingual data of Chinese and Japanese texts to improve the translation quality for these tasks. Tables 10 and 11 again present larger improvements in the translation task from Japanese to Vietnamese on both TED Talks and ALT datasets. On the other hand, the performance in the translation task from Chinese to Vietnamese tends to degradation in the TED Talks domain. We think that the contrariety in the structure of sentences in Japanese and Chinese may be the cause of this problem. When the translation systems are continued to train after 70 epochs, the degradation of performance in this translation task continues to rise in the TED Talks domain (−1.3). We can see that when continuing training, the combined training systems bring more benefits to the translation performance in the translation task from Japanese to Vietnamese than in the translation task from Chinese to Vietnamese.

Conclusion

We have proposed an effective method for generating pseudo-bilingual datasets to advance the performance of translation systems in the low-resource issue. Our proposed method is fast, robust, and does not require supplementing external resources such as dictionaries, pre-trained models, rules, language models, etc. as in previous works. This strategy may increase the number of parameters of the translation model due to the augmentation of the target vocabulary size. To reduce the adverse impact of this problem, an appropriate frequency threshold for the replacement of artificial ATUs is used to degrade the size of the vocabulary while still guaranteeing translation performance. Our method may be used in combination with the Back Translation technique to improve the performance of the translation system. Moreover, the proposed method does not depend on languages, it can be applied to divergent language pairs.

Furthermore, we investigate the combined translation systems of logographic languages from Chinese, Japanese to Vietnamese to leverage the share of common translation units regardless of opposition in grammatical structure. Our experiments demonstrate that these systems achieve substantial

improvements in two datasets in the low-resource situation by using various segmentation strategies for Japanese texts. spacy segmentation tool may be a good choice to balance the translation quality in both translation tasks from Chinese to Vietnamese and Japanese to Vietnamese.

For further implementations, a BERT model is leveraged to upgrade the translation quality of NMT systems in data sparsity. From experimental results, we conclude that in the high-resource scenario, the translation task from Japanese to Vietnamese gains great benefits on translation performance in the combined training system; otherwise, this tends to decrease in the translation task from Chinese to Vietnamese.

In the future, we would like to experiment with more NMT systems with different frequency thresholds for synthetic data generation to further consider the effectiveness of the proposed method.

Notes

1. We collected bilingual datasets Chinese-Vietnamese, Japanese-Vietnamese from the TED Talks domain and release at https://github.com/ngovinhtn/Low-resource-Machine-Translation.git
2. www.phontron.com/kytea/
3. https://taku910.github.io/mecab/
4. https://spacy.io/
5. https://opus.nlpl.eu/CCAligned.php
6. https://github.com/quanpn90/NMTGMinor
7. https://github.com/google-research/bert
8. https://github.com/mjpost/sacrebleu
9. https://www.ted.com/talks
10. https://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/
11. https://github.com/fxsjy/jieba
12. https://github.com/moses-smt/mosesdecoder/tree/master/scripts

Acknowledgments

We would like to thank the anonymous reviewers for carefully reading our paper and giving detailed comments

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

This work has been supported by the Ministry of Science and Technology of Vietnam under Program KC 4.0, No. KC-4.0.12/19-25.

ORCID

Thi-Vinh Ngo http://orcid.org/0000-0001-8764-6688

References

Aharoni, R., M. Johnson, and O. Firat. 2019. Massively multilingual neural machine translation. 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), 3874-–3884, Minneapolis, Minnesota: Association for Computational Linguistics. June. doi: 10.18653/v1/N19-1388.

Artetxe, M., G. Labaka, E. Agirre, and K. Cho. 2018. Unsupervised neural machine translation. 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April - May 30 - 3, 2018, Conference Track Proceedings. https://openreview.net/pdf?id=Sy2ogebAW

Bahdanau, D., K. Cho, and Y. Bengio. 2015. Neural machine translation by jointly learning to align and translate. Proceedings of International Conference on Learning Representations, ICLR 2015, May 7 - 9, 2015, San Diego, CA, United States.

Clinchant, S., K. W. Jung, and V. Nikoulina. 2019. On the use of BERT for neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, 108–117, Hong Kong: Association for Computational Linguistics. November. doi: 10.18653/v1/D19-5611.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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), 4171–4186, Minneapolis, Minnesota: Association for Computational Linguistics. June. doi: 10.18653/v1/N19-1423.

Dou, Z.-Y., A. Anastasopoulos, and G. Neubig. 2020. Dynamic data selection and weighting for iterative back-translation. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 5894–5904, Online: Association for Computational Linguistics. November. doi: 10.18653/v1/2020.emnlp-main.475.

Duan, S., H. Zhao, D. Zhang, and R. Wang. 2020. Syntax-aware data augmentation for neural machine translation. CoRR, abs/2004.14200. https://arxiv.org/abs/2004.14200

Eck, M., S. Vogel, and A. Waibel. 2005. Low cost portability for statistical machine translation based on n-gram frequency and TF-IDF. Proceedings of the Second International Workshop on Spoken Language Translation, Pittsburgh, Pennsylvania, USA, October 24-25. https://aclanthology.org/2005.iwslt-1.7

Edunov, S., M. Ott, M. Auli, and D. Grangier. 2018. Understanding back-translation at scale. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 489–500, Brussels, Belgium: Association for Computational Linguistics, October-November. doi: 10.18653/v1/D18-1045.

El-Kishky, A., V. Chaudhary, F. Guzmán, and P. Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 5960–5969, Online: Association for Computational Linguistics, November. doi: 10.18653/v1/2020.emnlp-main.480.

Gao, F., J. Zhu, L. Wu, Y. Xia, T. Qin, X. Cheng, W. Zhou, and T.-Y. Liu. 2019. Soft contextual data augmentation for neural machine translation. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 5539–5544, Florence, Italy: Association for Computational Linguistics, July. doi: 10.18653/v1/P19-1555.

Ha, T., J. Niehues, and A. H. Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. CoRR, abs/1611.04798. http://arxiv.org/abs/1611.04798

Ha, T.-L., V.-K. Tran, and K.-A. Nguyen. 2020. Goals, challenges and findings of the vlsp 2020 English-vietnamese news translation shared task. *VLSP* 2020, Hanoi, Vietnam, 99–105, https://aclanthology.org/2020.vlsp-1.18.pdf

Kingma, D., and J. Ba. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Ngo, T.-V., T.-L. Ha, P.-T. Nguyen, and L.-M. Nguyen. 2018. Combining advanced methods in japanese-vietnamese neural machine translation. 2018 10th International Conference on Knowledge and Systems Engineering (KSE), Nov 1-3, 2018, Ho Chi Minh City, Vietnam, 318–322.

Niu, X., W. Xu, and M. Carpuat. 2019. Bi-directional differentiable input reconstruction for low-resource neural machine translation. 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), 442–448, Minneapolis, Minnesota: Association for Computational Linguistics, June. doi: 10.18653/v1/N19-1043.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 311–318, Philadelphia, Pennsylvania, USA: Association for Computational Linguistics. July. doi: 10.3115/1073083.1073135.

Phuoc Tran, D. D., L. H. B. NGUYEN, and H. B. Long. 2016. Word re-segmentation in Chinese-vietnamese machine translation. *ACM Transactions on Asian and Low-Resource Language Information Processing* 16 (2):1–22. doi: 10.1145/2988237.

Post, M. 2018. A call for clarity in reporting BLEU scores. Proceedings of the Third Conference on Machine Translation: Research Papers, 186–191, Brussels, Belgium: Association for Computational Linguistics. October. doi: 10.18653/v1/W18-6319.

Riza, H., M. P. Gunarso, T. Uliniansyah, A. A. Ti, S. M. Aljunied, L. C. Mai, V. T. Thang, N. P. Thai, V. Chea, and R. Sun, et al. 2016. Introduction of the asian language treebank. 2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA) Oct 26-28, Bali, Indonesia, 1–6, doi: 10.1109/ICSDA.2016.7918974.

Saleh, F., W. Buntine, G. Haffari, and L. Du. 2021. Multilingual neural machine translation: Can linguistic hierarchies help? Findings of the Association for Computational Linguistics: EMNLP 2021, 1313–1330, Punta Cana, Dominican Republic: Association for Computational Linguistics. November. doi: 10.18653/v1/2021.findings-emnlp.114.

Salton, G., and C. S. Yang. 1973. On the specification of term values in automatic indexing. *Journal of Documentation* 290 (4):0 351-–372.

Sánchez-Cartagena, V. M., M. Esplà-Gomis, J. A. Pérez-Ortiz, and F. Sánchez-Martnez. 2021. Rethinking data augmentation for low-resource neural machine translation: A multi-task learning approach. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 8502–8516, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. November. doi: 10.18653/v1/2021.emnlp-main.669.

Sennrich, R., B. Haddow, and A. Birch. 2016a. Neural machine translation of rare words with subword units. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1715–1725, Berlin, Germany: Association for Computational Linguistics. August. doi: 10.18653/v1/P16-1162 .

Sennrich, R., B. Haddow, and A. Birch. 2016b. Improving neural machine translation models with monolingual data. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 86–96, Berlin, Germany: Association for Computational Linguistics. August. doi: 10.18653/v1/P16-1009.

Silva, C. C., C.-H. Liu, A. Poncelas, and A. Way. 2018. Extracting in-domain training corpora for neural machine translation using data selection methods. Proceedings of the Third Conference on Machine Translation: Research Papers, 224–231, Brussels, Belgium: Association for Computational Linguistics.October. doi: 10.18653/v1/W18-6323.

Sutskever, I., O. Vinyals, and Q. V. Le. 2014. Sequence to sequence learning with neural networks. CoRR, abs/1409.3215. http://arxiv.org/abs/1409.3215

Tan, X., Y. Leng, J. Chen, Y. Ren, T. Qin, and T. Liu. 2019. A study of multilingual neural machine translation. *CoRR*, abs/1912.11625. http://arxiv.org/abs/1912.11625

Tu, Z., Y. Liu, L. Shang, X. Liu, and H. Li. 2017. Neural machine translation with reconstruction. Proceedings of the AAAI Conference on Artificial Intelligence, 310 (1), February. doi: 10.1609/aaai.v31i1.10950.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762. http://arxiv.org/abs/1706.03762

Xia, M., X. Kong, A. Anastasopoulos, and G. Neubig. 2019. Generalized data augmentation for low-resource translation. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 5786–5796, Florence, Italy: Association for Computational Linguistics. July. doi: 10.18653/v1/P19-1579.

Xie, Z., S. I. Wang, J. Li, D. Lévy, A. Nie, D. Jurafsky, and A. Y. Ng. 2017. Data noising as smoothing in neural network language models. 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. http://arxiv.org/abs/1703.02573

Zhang, J., and T. Matsumoto. 2017. Improving character-level japanese-Chinese neural machine translation with radicals as an additional input feature. 2017 International Conference on Asian Language Processing (IALP) December, 5-7, Singapore, 172–175.

Zhang, L., and M. Komachi. 2018. Neural machine translation of logographic language using sub-character level information. Proceedings of the Third Conference on Machine Translation: Research Papers, 17–25, Brussels, Belgium: Association for Computational Linguistics. October. doi: 10.18653/v1/W18-6303.

Zhang, Z., S. Liu, M. Li, M. Zhou, and E. Chen. 2018. Joint training for neural machine translation models with monolingual data. In The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18) Association for the Advancement of Artificial Copyright Intelligence (AAAI 2018), February Louisiana, USA: https://www.aaai.org/. 555–562.

Zhu, J., Y. Xia, L. Wu, D. He, T. Qin, W. Zhou, H. Li, and T. Liu. 2020. Incorporating BERT into neural machine translation. *CoRR*, abs/2002.06823. https://arxiv.org/abs/2002.06823

Zoph, B., and K. Knight. 2016. Multi-source neural translation. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 30–34, San Diego, California: Association for Computational Linguistics. June. doi: 10.18653/v1/N16-1004.