IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 15, 2020, accepted June 24, 2020, date of publication June 30, 2020, date of current version July 9, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3006129

Revisiting Back-Translation for Low-Resource Machine Translation Between Chinese and Vietnamese

HONGZHENG LI[1], JIU SHA[1], AND CAN SHI[2]
[1]School of Computer Science, Beijing Institute of Technology, Beijing 100081, China
[2]School of Foreign Language, Qingdao University, Qingdao 266071, China

Corresponding author: Can Shi (yimufan@126.com)

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61902024, and in part by the China Postdoctoral Science Foundation under Grant 2018M640069.

ABSTRACT Back-translation (BT) has been widely used and become one of standard techniques for data augmentation in Neural Machine Translation (NMT), BT has proven to be beneficial for improving the performance of translation effectively, especially for low-resource scenarios. While most previous works related to BT mainly focus on European languages with high relatedness, few of them study less-related languages in other areas around the world. In this paper, we choose the language pair with less relatedness in Asia: Chinese and Vietnamese, to investigate the impacts of BT on extremely low-resource machine translation between them. We first discuss the similarities and differences between the two languages, then evaluate and compare the effects of different sizes of back-translated data on NMT and Statistical Machine Translation (SMT) models for Chinese-Vietnamese and Vietnamese-Chinese, with both character-based and word-based settings, and conduct further analysis on the translation outputs from several aspects. Some conclusions from previous works are partially confirmed and we also draw some new findings and conclusions, which are beneficial to understand BT further and deeper for translation between less-related low-resource languages.

INDEX TERMS Back-translation, Chinese, low-resource languages, machine translation, Vietnamese.

I. INTRODUCTION

The great success of Neural Machine Translation (NMT) is heavily dependent on large scale parallel data. However, parallel corpora with both good quality and quantity are not always available especially for low-resource language pairs. Under the scenario where bi-texts are limited but much larger amounts of monolingual data are available, there have been extensive works to improve models with monolingual data. In which back-translation (BT) has been wildly used, and is helpful in improving the performance of translation effectively, especially for low-resource languages.

BT is a simple yet effective approach. For a translation goal from source language S to target language T, it first trains another intermediate system to translate extra target monolingual data(T) into the source language(S'), and then S' and T will be combined as new parallel corpus(S',T), known as the *synthetic* corpus, which will be added to the authentic parallel data to train a final system from S to T. BT has become one of standard technologies in the pipeline of NMT.

BT was originally introduced to phrase-based MT and has been recently proposed for NMT in 2016 [1]. Since then, various works have been focusing on improving the performance of machine translation with BT, as a method of data augmentation. However, most of the works pay more attention to languages in European instead of other areas around the world. On the other hand, these European languages are usually language-related, belonging to the same or similar language families, or using the similar writing systems with Latin alphabets. As a result, it's more easier for them to share similar vocabulary or embeddings during the process of NMT. We believe that some conclusions based on these European languages do not necessarily suitable for other low-resource languages, especially for those which are less-related in term of language relatedness.

In this paper, we would like to investigate and evaluate effects of BT for machine translation between the

The associate editor coordinating the review of this manuscript and approving it for publication was Hao Ji.

low-resource language pair in Asia, i.e., Chinese and Vietnamese, which not only belong to different language families, but also use different writing systems, leading to more challenges for low-resource MT.

The relations between China and Vietnam have always been very close in the history, and there have been extensive exchanges in many fields. Vietnam is also one of the countries along the *Belt and Road* Initiative. It is necessary to improve the performance of translation between Chinese and Vietnamese. We conduct SMT and NMT experiments with both character-based and word-based settings by training models with extremely low-resource datasets, providing comparisons for Chinese-Vietnamese and Vietnamese-Chinese translations, we also conduct several further analysis, including N-gram F1 scores, Error rate and linguistic analysis, to draw new conclusions. To the best of our knowledge, this is the first work to comprehensively study the effects of BT for low-resource machine translation between these two languages.

Our main contribution in this paper is threefold:

(1) We present the first comprehensive and systematic comparison of the effects of synthetic data on low-resource MT specially for Chinese-to-Vietnamese and Vietnamese-to-Chinese.

(2) We evaluate various sizes of synthetic data in MT models with both character-based and word-based translation settings.

(3) We try to answer the question that ‘‘With BT data, which settings (character-based and word-based) are more suitable for different translation directions?’’ and provide some recommendation based on the experimental results.

The rest of the paper are organized as follow: section 2 briefly presents some previous related works; section 3 describes the similarities and differences between Chinese and Vietnamese; section 4 conducts the experiments and analysis in detail; Finally, the paper ends with some conclusion and future work.

A. RELATED WORK

In this section, we will briefly introduce some works on BT and Chinese↔Vietnamese machine translation.

B. BACK TRANSLATION

Back-translation was first proposed for NMT in [1], and has shown its great effectiveness in improving the performance of translation. As it is particularly useful when parallel data is scarce, BT has wide application in low-resource scenarios to leverage monolingual data [2].

Recently many works have aroused to understand why BT is beneficial for better NMT performance. For example, Edunov *et al.* [3] investigate several methods to generate synthetic source sentences and their respective effects in NMT. Park *et al.* [4] build the NMT model only using synthetic parallel data from both source side and target side. Reference [5] draw an empirical roadmap to observe how the amounts of BT data impact the performance of the final system, they further investigate more factors of BT data in different SMT and NMT approaches, as well as the amounts of data [6]. Although back translation is proved effective, some works like [7] have demonstrated that back translation increasing the amount of monolingual data improves the translation quality only up to some point, and then it starts to degrade.

The back-translated data may encounter the problem that the quality of synthetic data are not very good and have negative impacts on the performance of translation in the long run. In order to address the problem, Hoang *et al.* [8] propose the iterative back-translation approach, which employs the monolingual data for more than once to improve the performance. Currey and Heafield [9] combine BT with pivot-based MT method to translate monolingual pivot languages into source and target languages to train the NMT models.

C. CHINESE ↔ VIETNAMESE MACHINE TRANSLATION

There are not too many works on Chinese↔Vietnamese machine translation, in which most of them use phrase-based SMT models and few involves NMT models. Zhao *et al.* [10] solve Vietnamese to Chinese MT task by adopting Chinese characters as the pivot. For Chinese-Vietnamese translation, some works focus on unknown words of name entities [11], [12] [13], word segmentation [14] and other challenging problems which can have negative impacts on improving the translation. Tran *et al.* [15] proposes a character-based and word-based approach for Chinese-Vietnamese SMT to address the word segmentation challenge, and some other works use syntactic information to improve the performance of translation [16], for example, Gao *et al.* [17] propose an effective tree-to-tree syntax-aware method for Chinese-Vietnamese MT, Tran *et al.* [18] present word preordering approach to adjust orders in Chinese be suitable for Vietnamese first and then train SMT models with the pre-ordered data.

To the best of our knowledge, there haven’t works on comparison of impacts of synthetic BT data on translations between Chinese and Vietnamese.

II. COMPARISON BETWEEN CHINESE AND VIETNAMESE

As the national language of Vietnam, Vietnamese is written in Latin alphabets with additional diacritics for tones and certain letters since 20th century, and a phonetic syllable corresponds only to a single Vietnamese word.

Vietnamese has close relations with Chinese and has been deeply impacted by Chinese characters and Chinese culture historically. Just like Japanese and Korean, Vietnamese also borrows various of characters and words from China. There are still many Sino-Vietnamese words("Từ Hán Việt") nowadays, accounting for more than 60% in whole Vietnamese vocabulary, and are commonly used in written language.

Next, we will analyze some obvious similarities and differences between these two languages.

A. SIMILARITIES

Syntactically, both of them are analytic languages, and have the basic SVO grammatical structures. As a result, they typically lack morphological changes, and express grammatical structures and meanings mainly by function words and word orders.

Lexically, homonym words are very common in the two languages. That means, a Chinese character or word with the same pronunciation may have different meanings, such phenomenon is also common in Vietnamese, but happens only at monosyllabic level.

For example, the three different Chinese characters (“财, 才, 材”) have different meanings (“money, talent and material” respectively), but with the same pronunciation (cái), and the three characters are all represented as the same syllable (tài) in Vietnamese. Such phenomena maybe more likely to cause ambiguity and mistranslation in the process of translation.

B. DIFFERENCES

First of all, the language families are different. Chinese belong to Sino-Tibetan language family, while Vietnamese belongs to Austroasiatic language family. Unlike many European languages in Indo-European languages family, these two languages are less language-related. Furthermore, they also have different written systems, i.e., characters VS. Latin alphabets. As a result, it's more difficult to mapping them into the same joint embedding representation space, and have far less shared vocabularies or embedding during NMT.

Another significant difference between them is the orders of many words and phrases. We list several common situations as follow where the orders are different.

(1) Noun Phrase (NP). NPs in Chinese can have various grammatical structures, such as: (a) Noun(N)1+Noun(N)2; (b) Adjective(ADJ)+N; (c) NP with the function word “的 (DE)”. “的 (DE)” is an important function word widely used in Chinese, especially in NP. Typical structures with “的 (DE)” include but not limited to "N1+DE+N2", "ADJ+DE+N" or "Pronoun+DE+N".

All the modifiers in these NPs will be reordered after the head noun in Vietnamese.

(2) Position of preposition phrase(PP) in the sentence. PP usually appear after NP and before VP in Chinese sentences, i.e., sentence(S) = NP+PP+VP. But in Vietnamese, just like English, PP usually follows the VP at the end of the sentence: S = NP+VP+PP.

(3) PP structure with the preposition “把 (BA)”. This is another common but unique PP structure in Chinese, which is expressed by the special word “把 (BA)”, usually following by a NP or other elements. The word has no corresponding translation word in many languages including English and Vietnamese. When translating sentences with such structure S = NP1+ 把 (BA)+NP2+VP, they need be reordered as: S = NP1+VP+NP2, and the word “把 (BA)” will be deleted.

The differences have brought more challenges for translation between the two languages.

As mentioned above, although Vietnamese is written in Latin alphabets and syllables are separated by white space, however, unlike many other typical Latin-alphabet based languages (e.g. English), the space cannot be used to determine word boundaries. Thus, word segmentation should be taken into consideration for MT and other Natural Language Processing(NLP) tasks in Vietnamese. Both word-based and character-based Chinese-Vietnamese and Vietnamese-Chinese MT have their own advantages and drawbacks respectively. For example, as previous work discussed in [15], some entity names in Chinese, like person names (PER), must be translated as a whole into Sino-Vietnamese words, which can be achieved in word-based translation, and gives better result, but word-based translation is more likely generates many unknown words. While in character-based translation, some characters in the entity may be mistranslated.

Thus it's really necessary to investigate which translation setting (character-based and word-based) is better and beneficial for MT between the two languages. This is also one of the questions we would like to answer in following section.

III. EXPERIMENTS AND ANALYSIS

In this section, we will conduct character-based and word-based translation experiments with synthetic data using SMT and NMT models for Chinese(zh)-Vietnamese(vi) and Vietnamese-Chinese. The performance of translation are compared in terms with two automatic evaluation metrics: BLEU [19] and METEOR [20].

A. DATA SETTING

As Chinese and Vietnamese belong to low-resource language pair, there are not many public corpora and datasets available. Datasets in previous works are also unavailable. As a result, we decide to use the news domain parallel data developed by our team as the datasets for our experiments. This also means, our results in this work are unable to compare with those in previous works.

The datasets(marked as *auth_D*) are all collected from multilingual news websites and processed by following steps:

(1) Crawling bilingual news websites to obtain Chinese and corresponding Vietnamese articles according to publishing date and news titles.

(2) Separating the parallel documents into sentences by punctuations (such as full stops, question marks and exclamatory marks) respectively to form preliminary parallel sentence alignments.

(3) Asking the native Vietnamese speakers to manually check the parallel sentence alignments to guarantee correct alignments and high translation quality of the parallel sentence pairs.

The processed datasets contains 56,610 sentence pairs from politics, economics and cultural and social news. After shuffling, 50,000 sentences pairs are randomly selected as training set. Then the remaining 6610 sentences are evenly divided into valid set (3305) and test set (3305) respectively. Table 1 shows the statistics of the datasets.

Besides the authentic dataset, we also use monolingual Chinese and Vietnamese data to generate back-translated synthetic data. We first train character-based and

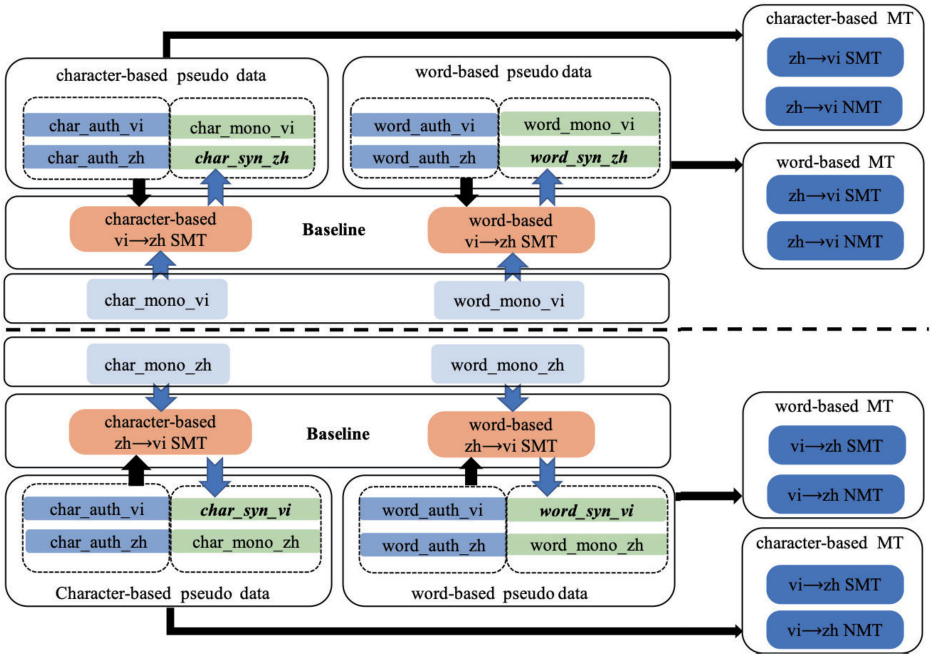

FIGURE 1. Experiment architecture in this paper. The arrows in blue colors represent the processes of generating back-translated data with monolingual Chinese and Vietnamese data respectively. The black vertical arrows indicate data used for training the Baseline SMT models.

TABLE 1. Statistic of datasets in the experiments.

Data	Sentence pairs	Tokens(zh)	Tokens(vi)
Training	50,000	13,991	10,586
Dev	3305	2044	2101
Test	3305	1857	2012

word-based Chinese↔Vietnamese SMT models respectively with *auth_D*, using the same configuration as described in next subsection, then translate the monolingual Chinese data[1](*mono_zh*) provided by WMT2019 to generate *synthetic* back-translated Vietnamese (*syn_vi*) data, building pseudo parallel sentence pair{*mono_zh*,*syn_vi*}; and translate the monolingual Vietnamese data (*mono_vi*) to generate *synthetic* Chinese (*syn_zh*), building pseudo parallel sentence pair{*mono_vi*,*syn_zh*}. It's worth noting that, BT data are usually generated by NMT models in most previous works, considering the extremely small size of dataset here, we decide to produce BT data with SMT models instead of NMT models to guarantee higher quality of BT data.

For the word-based experiment groups, all the data are preprocessed with word segmentation and Byte-Pair Encoding (BPE) [21] using 10,000 merge operations. HanLP toolkit[2] is used for Chinese word segmentation, and VnCoreNLP toolkit[3] is used for segmentation of Vietnamese sentences. After word segmentation, a Vietnamese word composed of more than one syllable is represented in the form of syllables connected by the "_".

B. MODELS IN THE EXPERIMENTS

1) MODELS WITH AUTHENTIC DATA

We first build character-based and word-based baseline models with authentic *auth_D* dataset.

The SMT models used to generate the BT data mentioned above can serve as SMT baseline models. We use GIZA++ for word alignment, build the 5-gram language models with data in *auth_D* using the KenLM toolkit [22], train the translation models by using the Moses toolkit [23] with default settings, and tuning the models with MERT [24].

For NMT, we train the baseline LSTM models using Pytorch version of OpenNMT [25] with default parameters, i.e., 2-layer LSTM with 500 hidden units, and 500-dimensional word embeddings. As the datasets are extremely small to train better models with more complicated architectures, we decide not to use the popular Transformer architecture [26], which will be one part of our future work.

2) MODELS WITH BT DATA

As shown in Fig. 1, we combine the pseudo parallel sentence pairs with *auth_D* to build new synthetic data sets(*syn_D*) to train new character-based and word-based SMT and NMT models. In order to investigate the impacts of different sizes of BT data on the performance of models, we train the models with increasing sizes of synthetic data pairs range from 60k(50k authentic data and 10k pseudo data) to 100k(50k authentic data and 50k pseudo data), adding 10k data to the

[1] http://data.statmt.org/news-crawl/zh/
[2] https://github.com/hankcs/HanLP
[3] https://github.com/vncorenlp/VnCoreNLP

models each time and train five models in total under each setting.

C. RESULTS AND ANALYSIS

1) CHINESE→VIETNAMESE MT RESULTS

The following two tables show the results of SMT and NMT for Chinese-Vietnamese, where the bold items indicate the best scores.

TABLE 2. Character-based Chinese→Vietnamese MT results.

	SMT		NMT	
	BLEU	METEOR	BLEU	METEOR
Baseline	14.29	30.29	11.68	25.32
$+syn\_D$				
60k	16.32	33.65	14.28	28.54
70k	16.51	**33.99**	14.79	29.34
80k	**16.86**	33.86	15.35	30.14
90k	16.42	33.01	**15.98**	30.87
100k	16.50	33.53	15.89	**30.97**

TABLE 3. Word-based Chinese→Vietnamese MT results.

	SMT		NMT	
	BLEU	METEOR	BLEU	METEOR
Baseline	11.71	26.70	9.65	22.77
$+syn\_D$				
60k	12.07	27.29	10.58	24.42
70k	12.08	27.40	11.12	25.24
80k	12.12	27.48	11.48	25.71
90k	12.38	27.62	11.68	26.18
100k	**12.38**	**27.67**	**11.85**	**26.28**

As for the character-based setting in Table 2, increasing sizes of synthetic data can improve the performance, and all the scores are higher than baseline. But the upward trends between SMT and NMT are different. Taking BLEU scores for example, the best score of SMT appears in the mid-size data(80k), while scores of NMT continue to increase until the best score is reached on the largest 100k data. As for the word-based setting in Table 3, with the increase in data sizes, the scores also gradually increase, reaching the peak on the largest data finally. The results in Table 2 and Table 3 indicate that BT is beneficial to both character-based and word-based Chinese-Vietnamese MT. However, the data sizes are still too small to train better models to enhance the performance greatly, especially for NMT.

When comparing the results in the same MT setting, it can be seen that performance of SMT always outperform those of NMT in both character-based and word-based settings. When comparing the results of SMT and NMT in the two tables, it is clearly shown that character-based results in Table 2 achieve better than word-based settings in Table 3 by 2-4 BLEU points.

2) VIETNAMESE→CHINESE RESULTS

Next, we'll take an inside look at the Vietnamese→Chinese MT performance, as shown in Table 4 and Table 5, where the bold items indicate the best scores.

TABLE 4. Character-based Vietnamese→Chinese MT results.

	SMT		NMT	
	BLEU	METEOR	BLEU	METEOR
Baseline	16.50	36.28	13.39	28.57
$+syn\_D$				
60k	**16.10**	**35.34**	13.12	28.45
70k	15.75	35.02	13.98	29.67
80k	15.46	33.87	14.19	30.17
90k	15.88	34.78	14.30	30.14
100k	15.89	35.09	**14.94**	**31.02**

TABLE 5. Word-based Vietnamese→Chinese MT results.

	SMT		NMT	
	BLEU	METEOR	BLEU	METEOR
Baseline	8.58	23.47	7.44	20.28
$+syn\_D$				
60k	6.24	20.72	6.39	19.50
70k	6.31	20.44	6.39	19.44
80k	**6.36**	**20.75**	6.52	19.76
90k	6.32	20.46	6.55	20.06
100k	6.22	20.22	**6.61**	**20.34**

From Table 4 and Table 5, scores of both the character-based and word-based NMT improve with the increasing sizes of synthetic data, and models trained with largest data all achieve the best scores of BLEU and METEOR, although the BLEU scores of word-based NMT are below the baseline.

However, scores of SMT in the two tables are totally different with NMT. For SMT in Table 4, the best score is surprisingly in the minimum data group and scores decrease to the lowest in the mid-size 80k group and then increase from 80k to 100k, but still lower than that of 60k; for SMT in Table 5, BLEU scores first increase to the peak in 80k group then gradually decrease. Note that, all the scores of BLEU and METEOR of SMT results are below the baseline. In other words, adding back-translated Vietnamese generally has negative impacts on the SMT performance.

Comparing Table 4 and Table 5, performance of character-based settings significantly outperform word-based settings regardless of SMT or NMT. When it comes to the results within Table 4, all the results of SMT are better than those of NMT; in Table 5, however, the situation is opposite that performance of word-based NMT are better than those of SMT.

Many previous works have argued that performance of SMT generally tend to outperform NMT under low-resource scenarios. In our experiments for the two translation directions, most results here do support that conclusion, except that the ones in Table 5. But as Sennrich and Zhang [27] also discussed, low-resource NMT is very sensitive to hyperparameters, if tuning well, training competitive NMT systems is also possible to surpass SMT systems. Our experimental results also prove this conclusion.

The Vietnamese→Chinese results in Table 4 and Table 5 have aroused two questions for us:

TABLE 6. Comparison of word-based SMT, word-based NMT and Google translations of an sentence example in Vietnamese.

Source	Theo Chủ_nhiệm Ủy_ban Đối_ngoại Quốc_hội **Trần_Văn_Hằng**, đây là chuyến thăm Trung_Quốc đầu_tiên trên cương_vị Chủ_tịch Quốc_hội khóa 13 của Chủ_tịch Quốc_hội Nguyễn Sinh Hùng.
English	The head of the Foreign Affairs Committee of the Vietnamese National Assembly, Trần_Văn_Hằng, said that this is the first visit to China by the chairman of the Vietnamese parliament Nguyễn Sinh Hùng since he became the chairman of the 13th National Congress of Vietnam.
Reference	越南 国会 对外 委员会 主任 陈文恒 表示， 这是 越南 国会 主席 阮生雄 出任 越南 第 十三 届 国会 主席 以来 首次 访问 中国。
SMT	越南 国会 对外 委员会 主任 Trần_Văn_Hằng， 这是 首次 访问 中国 十三 届 国会 主席 阮生雄。
NMT	越南 国会 主任 陈文恒 表示， 这 是 越南 国会 主席 阮生雄 以 国会 主席 身份 首次 访问 中国。
Google	据胡志明市人民委员会外交委员会主席说，这是对中国的访问。

(1) Why the performances of character-based and word-based SMT decrease with the increasing size of data, while those of NMT improve?

(2) Why word-based NMT results are better than word-based SMT results in Table 5?

D. FURTHER ANALYSIS

1) ANALYSIS ON VIETNAMESE→CHINESE OUTPUTS

In order to find some answers to the two questions, we analyse the translation outputs.

We first analyze the character-based and word-based Vietnamese→Chinese SMT outputs, and find that: (a) There exits many Vietnamese source words in the Chinese outputs. That means, these Vietnamese words are not translated at all. What's more, with the increasing size of training data, the numbers of Vietnamese words in the translation hypothesis tend to increase accordingly. (b) Word orders in the outputs, especially in the long sentences, are not adjusted well to be consistent with the proper syntactic structure of Chinese. (c) Finally, word segmentation preprocessing for Chinese and Vietnamese also result in inevitable mistakes to a great extent during the translation. We believe that these three aspects can explain why the SMT performance decrease as the training data increase.

We then compare the word-based SMT outputs with word-based NMT outputs, and find that the NMT ones are more advantageous in the following two aspects: (a) the numbers of Vietnamese words that are not translated are much fewer than SMT ones, including the entity names; (b) word orders and overall syntactic structures in NMT outputs are more accurate and more readable.

Table 6 shows an sentence example in Vietnamese and its corresponding translations of word-based SMT and NMT, which can clearly explain why the performance of NMT are better.

In the example, the bold person name in the source sentence is not translated by SMT, but is correctly translated in NMT; on the other hand, the meaning of the second sub-sentence in SMT output is "visits China the chairman of the parliament", indicating inaccurate word orders and wrong syntactic structures, while the NMT translation "the chairman of the parliament visits China" is very reasonable. We also put the source example sentence into the Google Translate, it shows the length of the second sub-sentence is much shorter, because many (important) words are not translated at all.

2) CHARACTER N-GRAM F1 SCORES

In this part, we will use the CHRF metrics [28] to estimate of the quality of character-based Vietnamese-Chinese translations of SMT and NMT by calculating the N-gram F1 scores.

According to Table 4, we choose the highest and lowest BLEU scores in SMT and NMT respectively. That is, in SMT, outputs of models trained with 60k data and 80k data are the highest and lowest; in NMT, the data are in 100k and 60k groups. The F1 scores of N-gram ranges from 1-gram to the default 6-gram in the CHRF toolkit are presented in Figure 2.

From the Figure, in SMT, except the results of 2-gram, in any other rest N-grams($N = 1,3,4,5,6$), it shows the same trend that F1 scores in the outputs of model trained with 60k data(which BLEU score is the highest) are always higher than those of 80k data(lowest BLEU score). In other words, when the size of training data increase, the F1 scores and BLEU scores in the outputs decrease.

When it comes to the NMT group, however, F1 scores in the outputs with highest BLEU score are always higher than the outputs with lowest BLEU score. That means, the larger the training data, the better the N-gram F1 scores and BLEU scores. In a summary, the F1 scores of character N-gram can also explain why the BLEU scores in character-based Vietnamese-Chinese SMT decrease as the training data increases in Table 4.

3) AUTOMATIC ERROR ANALYSIS

In order to better understand the impacts of BT on the performance of translation, we carried out more detailed analysis of all translation outputs. We use the Hjerson toolkit [29] to analyze four error categories: word order, omission, addition and mistranslation. The results are presented in Table 7 and Table 8, the lower the rate indicates the better results.

As shown in Table 7, increasing synthetic data can reduce the orders, omissions and mistranslation error rates for SMT, and it also improves the addition error in NMT. Due to the limitation of the size of the data set, compared with SMT, the effects of synthetic data on the error rates in NMT is not as obvious as those of SMT.

It can be seen from Table 8 that, increasing synthetic data is particularly beneficial for reducing mistranslation in SMT

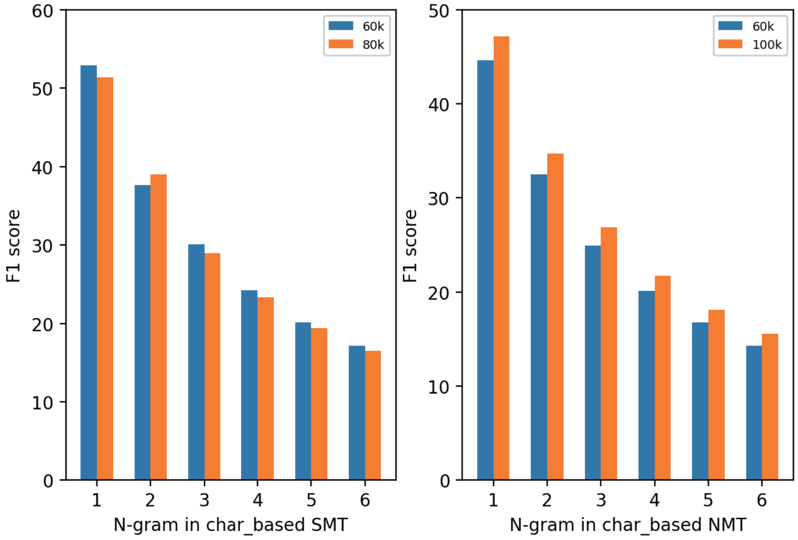

FIGURE 2. Character N-gram F1 scores in character-based Vietnamese-Chinese SMT (left) and NMT (right).

TABLE 7. Error rates(%) in Chinese-Vietnamese MT.

	Chinese→Vietnamese			
	order	omission	addition	mistranslation
character-based SMT				
60k	11.46	11.44	4.87	17.43
70k	11.21	4.77	5.27	16.60
80k	11.3	4.9	5.19	16.79
90k	11.1	5.03	5.06	16.57
100k	10.98	4.88	5.26	16.70
character-based NMT				
60k	9.41	5.53	4.64	17.63
70k	9.42	5.61	4.87	17.54
80k	9.38	6.11	4.38	17.85
90k	9.25	5.71	4.72	17.44
100k	9.37	6.29	4.32	17.75
word-based SMT				
60k	5.52	5.02	4.94	13.09
70k	5.56	5.05	4.96	13.17
80k	5.69	5.41	4.78	13.68
90k	5.54	5.09	4.95	13.15
100k	5.59	5.06	4.97	13.28
word-based NMT				
60k	5.32	5.03	4.71	13.34
70k	5.35	5.18	4.65	13.59
80k	5.33	5.16	4.66	13.58
90k	5.35	5.22	4.54	13.69
100k	5.43	5.16	4.65	13.66

TABLE 8. Error rates(%) in Vietnamese-Chinese MT.

	Vietnamese→Chinese			
	order	omission	addition	mistranslation
character-based SMT				
60k	12.41	5.67	3.92	16.57
70k	12.69	5.56	3.90	16.75
80k	12.19	5.82	3.76	16.84
90k	12.21	5.62	3.86	16.62
100k	12.25	5.54	3.82	16.46
character-based NMT				
60k	10.07	6.33	3.82	16.90
70k	10.06	6.55	1.86	15.83
80k	10.07	6.02	4.00	16.68
90k	10.11	6.27	3.96	16.73
100k	10.11	5.86	4.09	16.57
word-based SMT				
60k	5.07	5.20	5.64	13.86
70k	4.95	5.14	5.48	13.59
80k	5.04	5.13	5.65	13.72
90k	4.78	4.81	5.31	12.98
100k	4.89	4.94	5.50	13.52
word-based NMT				
60k	4.92	5.44	5.51	14.49
70k	4.82	5.56	5.43	14.72
80k	4.86	5.55	5.53	14.51
90k	4.84	5.48	5.59	14.43
100k	4.82	5.42	5.66	14.38

and NMT. It also reduces word ordering error rates for NMT. Most of the error category rates in NMT are much lower than those in SMT, this can also explain why word-based NMT outperform word-based SMT in Table 5.

4) LINGUISTIC ANALYSIS

As mentioned in Section 3, Sino-Vietnamese words account for large proportion in Vietnamese, playing an important role in written language, and an significant difference between Chinese and Vietnamese is the word orders in many phrases.

TABLE 9. Accuracy of Sino-Vietnamese words and NP in Chinese-Vietnamese translation outputs.

	SMT		NMT	
Data	S-V words(%)	NP(%)	S-V words(%)	NP(%)
60k	74.46	71.62	73.56	71.24
80k	74.84	72.21	73.95	71.66
100k	75.33	72.78	74.54	72.13

As word order is one of the factors that affect the performance of translation, here we choose two representative structures from Chinese-Vietnamese translation outputs:

TABLE 10. Examples of Chinese-Vietnamese translation.

Example 1	
Source	越南倾听[$_{NP}$企业、人民的心声]。
Reference	Việt Nam lắng nghe tiếng nói của doanh nghiệp, người dân.
English	Vietnam listens to [$_{NP}$ the voices of enterprises and people.]
System hypothesis	Việt Nam lắng nghe tiếng của người dân.
Example 2	
Source	纺织品服装要求[$_{NP}$良好职业技能和工厂技术的产品] 。
Reference	hàng hóa dệt may đòi hỏi tay nghề cũng nhưcông nghệ của các nhà máy phải tốt.
English	Textiles and garments require [$_{NP}$ products with good professional skills and factory technology.]
System hypothesis	dệt may tập trung yêu cầu tốt về kỹ năng, công nghệ để các sản phẩm nhà máy sản xuất.

Sino-Vietnamese(S-V) words and Noun Phrases(NP), to conduct linguistic analysis to understand the effects of BT on them.

By using the stratified sampling method, we extract same 600 samples from three translation outputs generated from models trained with 60k, 80k and 100k data respectively to calculate the translation accuracy of Sino-Vietnamese words and NPs.

The results presented in Table 9 show that, with the increasing of data size, the accuracy of the two structures also improve accordingly, which indicates that adding back-translated data is beneficial for enhancing the translation of some specific structures, leading to overall improvement on the whole translation hypothesis finally. The accuracy of Sino-Vietnamese words are all higher than those of NP, we believe the reasons are that, Sino-Vietnamese words typically consist of two to three syllables and have relatively fixed expressions, for example, "Trung Quôc"(China); while the length of syllables in noun phrases is much longer than that of Sino-Vietnamese words, and NPs have complex syntactic structures. As a result, the accuracy of NPs must be lower.

When taking a deeper look inside the NPs, we find that translation accuracy of NPs with different structures are actually different:

(a) For NPs composed of two nouns(N):NP = N1+N2, the accuracy is the highest. Word orders in Most of such NPs have changed correctly during translation, that is, the first noun is adjusted after the second noun.

(b) For the typical NPs with the function word "的 (DE)" and two Nouns: NP = N1+DE+N2, the accuracy is also good. "的 (DE)" will be translated as "của", and word orders are reordered.

(c) For the NPs with several modifiers, or with complicated and nested syntactic structures, the accuracy is much lower, and word orders are not adjusted well. There are mainly two kinds of errors in the translation: first, one or more modifiers are not translated; second, the meanings in the translation are not the same with those of source language sentences.

Table 10 shows two examples which contain NPs with more than one modifier. In the first example, the word "doanh nghiệp"("enterprises") is missed and not translated. The second example sentence contains NP composed of multiple noun phrases as nested modifiers, and the head noun is "products". The literal meaning of the translation is "Textiles require skills, factory technology and products." Which is different with the meaning of source sentence, what's more, word orders and syntactic structures are not very good and acceptable.

E. EXPERIMENTAL CONCLUSION

Based on the preliminary experimental results and analysis above, we can draw some conclusions as follow:

(1) For Chinese→Vietnamese translation, adding synthetic Chinese data has ***positive*** impacts on the performance of both SMT and NMT under character-based and word-based settings.

(2) For Vietnamese→Chinese translation, adding synthetic Vietnamese data generally has ***negative*** impacts on the performance of both character-based and word-based SMT, but has ***positive*** impacts on the performance of NMT.

(3) For bidirectional Chinese↔Vietnamese translation, performance of SMT outperform those of NMT in most cases.

(4) Character-based setting is more recommended and more suitable for both Chinese-to-Vietnamese and Vietnamese-to-Chinese translations.

IV. CONCLUSION AND FUTURE WORK

In this paper, we present systematic empirical investigation on the effects of synthetic back-translated data for low-resource machine translation between the language pair with less relatedness: Chinese and Vietnamese. We first discuss some similarities and differences between the two languages, then evaluate the performance of character-based and word-based SMT and NMT models trained with increasing size of synthetic back-translated data and conduct further analysis on the translation outputs.

Some findings from previous works are partially confirmed, we also draw the new findings and conclusions, which can provide good clues and directions for future work on back-translation and low-resource machine translation. In the future, we would like to expand the dataset size to conduct experiments with more advanced model architectures and deeper analysis, and we will incorporate back-translation with other approaches such as transfer learning to improve translation performance of low-resource translation.

ACKNOWLEDGMENT

The authors would like to thank Dr. Huu-Anh Tran for providing the Chinese and Vietnamese parallel data for this work, and they thank Dr. Alberto Poncelas from the School of Computing, Dublin City University (DCU) for the kind discussion and suggestions.

REFERENCES

[1] R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, Berlin, Germany, 2016, p. 86–89.

[2] I. Gibadullin, A. Valeev, A. Khusainova, and A. Khan, "A survey of methods to leverage monolingual data in low-resource neural machine translation," 2019, *arXiv:1910.00373*. [Online]. Available: http://arxiv.org/abs/1910.00373

[3] S. Edunov, M. Ott, M. Auli, and D. Grangier, "Understanding back-translation at scale," in *Proc. EMNLP*, Brussels, Belgium, 2018, p. 489–500.

[4] J. Park, J. Song, and S. Yoon, "Building a neural machine translation system using only synthetic parallel data," 2017, *arXiv:1704.00253*. [Online]. Available: https://arxiv.org/pdf/1704.00253.pdf

[5] A. Poncelas, D. Shterionov, A. Way, G. Maillette de Buy Wenniger, and P. Passban, "Investigating backtranslation in neural machine translation," 2018, *arXiv:1804.06189*. [Online]. Available: http://arxiv.org/abs/1804.06189

[6] A. Poncelas, M. Popovic, D. Shterionov, G. Maillette de Buy Wenniger, and A. Way, "Combining SMT and NMT back-translated data for efficient NMT," 2019, *arXiv:1909.03750*. [Online]. Available: http://arxiv.org/abs/1909.03750

[7] F. Stahlberg, J. Cross, and V. Stoyanov, "Simple fusion: Return of the language model," 2018, *arXiv:1809.00125*. [Online]. Available: http://arxiv.org/abs/1809.00125

[8] V. C. D. Hoang, P. Koehn, G. Haffari, and T. Cohn, "Iterative back-translation for neural machine translation," in *Proc. 2nd Workshop Neural Mach. Transl. Gener.*, Melbourne, VIC, Australia, 2018, , p. 18.

[9] A. Currey and K. Heafield, "Zero-resource neural machine translation with monolingual pivot data," in *Proc. 3rd Workshop Neural Gener. Transl.*, Hong Kong, 2019, pp. 99–107.

[10] H. Zhao, T. J. Yin, J. Y. Zhang, "Vietnamese to chinese machine translation via chinese character as pivot," in *Proc. PACLIC*, 2013, pp. 250–259.

[11] P. T. Tran and D. Dinh, "Retranslating number expression unknown word in chinese-vietnamese statistical machine translation," *J. Comput. Sci. Cybern.*, vol. 30, no. 2, p. 127, Jul. 2014.

[12] P. Tran, D. Dinh, T. Le, and T. Nguyen, "Handling organization name unknown word in chinese-vietnamese machine translation," in *Proc. RIVF Int. Conf. Comput. Commun. Technol. Res., Innov., Vis. for Future (RIVF)*, Nov. 2013, p. 242.

[13] P. T. Tran, D. Dinh, and L. Tran, "Resolving named entity unknown word in Chinese-Vietnamese machine translation," in *Knowledge and Systems Engineering*. Cham, Switzerland: Springer, 2014, pp. 273–284.

[14] P. Tran, D. Dinh, and L. H. B. Nguyen, "Word re-segmentation in chinese-vietnamese machine translation," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 16, no. 2, pp. 1–22, Nov. 2016.

[15] P. Tran, D. Dinh, and H. T. Nguyen, "A character level based and word level based approach for chinese-vietnamese machine translation," *Comput. Intell. Neurosci.*, vol. 2016, pp. 1–11, 2016.

[16] J. He, Z. T. Yu, C. T. Lv, H. Lai, S. X. Gao, and Y. Zhang, "Language post positioned characteristic based Chinese-Vietnamese statistical machine translation method," in *Proc. IALP*, 2017, p. 180–184.

[17] S. Gao, J. Huang, M. Xue, Z. Yu, Z. Wang, and Y. Zhang, "Syntax-based chinese-vietnamese Tree-to-Tree statistical machine translation with bilingual features," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 18, no. 4, pp. 1–20, Aug. 2019.

[18] H. Tran, H. Y. Huang, P. Tran, S. M. Shi, and H. Nguyen, "Preordering for Chinese-Vietnamese Statistical Machine Translation," *IEICE Trans. Inf. Syst.*, vol. 102, no. 2, p. 375–382, 2019.

[19] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, Philadelphia, PA, USA, 2001, p. 311.

[20] A. Lavie and A. Agarwal, "Meteor: An automatic metric for MT evaluation with high levels of correlation with human judgments," in *Proc. 2nd Workshop Stat. Mach. Transl.*, 2007, p. 65.

[21] R. Sennrich, B. Haddow and A. Birch, "Neural machine translation of rare words with subword units," in *Proc. ACL*, Berlin, Germany, 2016, p. 1715–1725.

[22] K. Heafield, "KenLM: Faster and smaller language model queries," in *Proc. WMT*, Edinburgh, Scotland, 2011, p. 187–197.

[23] P. Koehn, "Moses: Open source toolkit for statistical machine translation," in *Proc. 45th Annu. Meeting Assoc. Comput. Linguistics Companion*, Prague, Czech, 2007, pp. 177–180.

[24] F. J. Och, "Minimum error rate training in statistical machine translation," in *Proc. 41st Annu. Meeting Assoc. Comput. Linguistics*, Sapporo, Japan, 2003, p. 160.

[25] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. Rush, "OpenNMT: Open-source toolkit for neural machine translation," in *Proc. ACL*, Vancouver, BC, Canada, 2017, p. 67.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[27] R. Sennrich and B. Zhang, "Revisiting low-resource neural machine translation: A case study," in *Proc. ACL*, Florence, Italy, 2019, pp. 211–221.

[28] M. Popovic, "CHRF: Character n-gram F-score for automatic MT evaluation," in *Proc. 10th Workshop Stat. Mach. Transl.*, Lisboa, Portugal, 2015, pp. 392–395.

[29] M. Popovic, "Hjerson: An open source tool for automatic error classification of machine translation output," *Prague Bull. Math. Linguistics*, vol. 96, pp. 59–68, Oct. 2011.

HONGZHENG LI was born in Shandong, China, in 1990. He received the Ph.D. degree in linguistics and applied linguistics from Beijing Normal University, Beijing, China, in 2018.

Since 2018, he has been a Postdoctoral Researcher and an Assistant Professor with the School of Computer Science, Beijing Institute of Technology. His research interests include machine translation and natural language processing.

Dr. Li received the sponsorship of the China Postdoctoral Science Foundation, in 2018, and the sponsorship of the National Natural Science Foundation of China (NSFC), in 2019.

JIU SHA was born in Gansu, China, in 1994. He received the B.S. degree in computer science from the Minzu University of China, Beijing, China, in 2018. He is currently pursuing the master's degree with the School of Computer Science, Beijing Institute of Technology.

His research interests include machine translation and natural language processing.

CAN SHI was born in Jining, China, in 1988. She received the Ph.D. degree in literary theory from Beijing Normal University, Beijing, China, in 2017.

She is an Assistant Professor and a Supervisor of master's degree students with the School of Foreign Language, Qingdao University. Her research interest includes translatology, especially machine translation and the English translation of Chinese classical poetry.