Phrase-Based Chinese-Vietnamese Pseudo-Parallel Sentence Pair Generation

Jiaxin Zhai[1,2], Zhengtao Yu[1,2(✉)], Shengxiang Gao[1,2], Zhenhan Wang[1,2], and Liuqing Pu[1,2]

[1] School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China
ztyu@hotmail.com
[2] Artificial Intelligent Key Laboratory of Yunnan Province, Kunming University of Science and Technology, Kunming 650500, China

Abstract. The lack of Chinese-Vietnamese parallel corpus has resulted in poor translation of Chinese-Vietnamese neural machine translation. In order to solve this problem, we propose a phrase-based Chinese-Vietnamese pseudo-parallel sentence pair generation method. This method expands the corpus of Chinese-Vietnamese neural machine translation and improves the performance of Chinese-Vietnamese neural machine translation. Firstly, based on the small-scale Chinese-Vietnamese parallel corpus, the method selects the phrase module according to the phrase syntactic structure information. Then this method combines word alignment information with replacement rules. Finally, the method achieves the expansion of Chinese-Vietnamese pseudo-parallel corpus. Experiments show that this method can effectively generate Chinese-Vietnamese pseudo-parallel sentence pairs and improve the performance of Chinese-Vietnamese neural machine translation.

Keywords: Phrase structure syntax · Phrase replacement · Pseudo-parallel sentence pair generation · Chinese-Vietnamese · Neural machine translation

1 Introduction

Neural mechanical translation can only achieve better results by training large-scale parallel corpora. Chinese-Vietnamese neural machine translation is a neural machine translation of resource scarcity types. It is difficult to obtain large-scale parallel corpus of Chinese-Vietnamese in a short time. Pseudo-parallel sentence pair generation is one of the important methods to extend pseudo-parallel corpus. Many researches have shown that pseudo-parallel corpora can also effectively improve the performance in neural machine translation of resource scarcity types.

There are three methods to generate pseudo-parallel corpora now. They are the method of back translation [1], the method of retelling [2–5], and the method of data augmentation [6]. These methods use a small amount of parallel corpus to generate pseudo-parallel corpora. But these specific methods are different. The back translation based method uses monolingual corpus resources to generate pseudo-parallel corpora in the iterative process of the neural machine translation model. The retelling based

S. Huang and K. Knight (Eds.): CCMT 2019, CCIS 1104, pp. 58–68, 2019.
https://doi.org/10.1007/978-981-15-1721-1_6

method uses external resources to reproduce the bilingual parallel corpus. The method based on data enhancement uses the information of parallel corpus, and replaces the module under the certain rules to realize the generation of pseudo-parallel corpus.

The method based on data enhancement can achieve better results without introducing additional resources. Therefore, this paper uses the method of data enhancement to realize the generation of Chinese-Vietnamese pseudo-parallel corpora. This method is also called a phrase-based Chinese-Vietnamese pseudo-parallel sentence pair generation method. This method firstly realizes the word alignment and phrase syntactic structure analysis for small-scale Chinese-Vietnamese parallel corpus. Then the method extracts the Chinese-Vietnamese aligned noun phrase (NP) and verb phrase (VP) according to the word alignment information and the phrase syntactic structure information of the parallel sentence pair, and form a collection of Chinese-Vietnamese alignment phrases. Finally, according to the phrase syntactic parsing tree of the Chinese and Vietnamese parallel sentence pairs, we find the NP and VP structures at different depths. At the same time, we use the set of aligned phrases to replace the phrases in the sentence, and use the language model to verify the newly generated sentences, and finally generate the Chinese and Vietnamese pseudo-parallel sentence pairs.

2 Related Work

In recent years, domestic and foreign scholars have studied the methods of corpus generation for small-scale parallel corpora, and have achieved a series of results. On the premise of not introducing additional resources, He et al. [5] proposed a paraphrase method based on dependency analysis and sentence generation. The method obtains a dependency tree by performing dependency analysis on sentences, and then generates multiple natural language sentences from the dependency tree. The sentence generated by this method has no lexical change compared to the original sentence. However, this method has changed the word order and improved the quality of machine translation without introducing additional resources. Fadae et al. [6] proposed the method of TDA (Translation Data Augmentation) to generate pseudo-parallel sentence pairs. The method first replaces the common words in the parallel sentence pairs with the rare words, and obtains the pseudo-parallel sentence pairs. To ensure that the pseudo-parallel sentence pairs are grammatically and semantically correct, the method uses a language model to filter pairs the pseudo-parallel sentences pairs. The pseudo-parallel sentence pair through the screening mechanism is the training corpus that can be used as a neural machine translation. Cai et al. [7] use data enhancement technology to expand the training data of resource-starved languages. The method first blocks the sentence and then finds the two most similar modules in the sentence. Finally, by forming a new sentence by adjusting their position, we have realized the extension of the pseudo-parallel sentence pair.

These results have effectively expanded the scale of translation corpus and improved the performance of machine translation. He et al. [5] adjust the order of statistical machine translation by changing word order. Fadae et al. [6] and Cai et al. [7] did not consider sentence structure complexity when they use module substitution to generate pseudo-parallel corpora. This method leads to grammatical semantic errors in

the sentence. We believe that the granularity of words is too small, and there is a one-to-many problem in the process of word alignment. Therefore, there will be grammatical and semantic errors in the sentence during the replacement process. There is also the problem that the replaced alignment words do not match in the sentence. The smallest translation unit is composed of multiple words. It is difficult to have a one-to-many problem. However, if we perform module replacement without the instruction of syntactic information, it is prone to grammatical errors.

In order to solve these problems, this paper proposes a phrase-based method for generating Chinese and Vietnamese pseudo-parallel sentence pairs. In this method, we use the phrase syntax structure information to guide the phrase replacement process. This approach not only avoids one-to-many problems, but also avoids syntactic errors in the replacement process.

3 Phrase-Based Extension Model of Chinese-Vietnamese Pseudo-Parallel Sentences

This section focuses on the Chinese and Vietnamese phrase extraction and alignment, as well as the phrase replacement rules. Figure 1 is the overall frame diagram of this document.

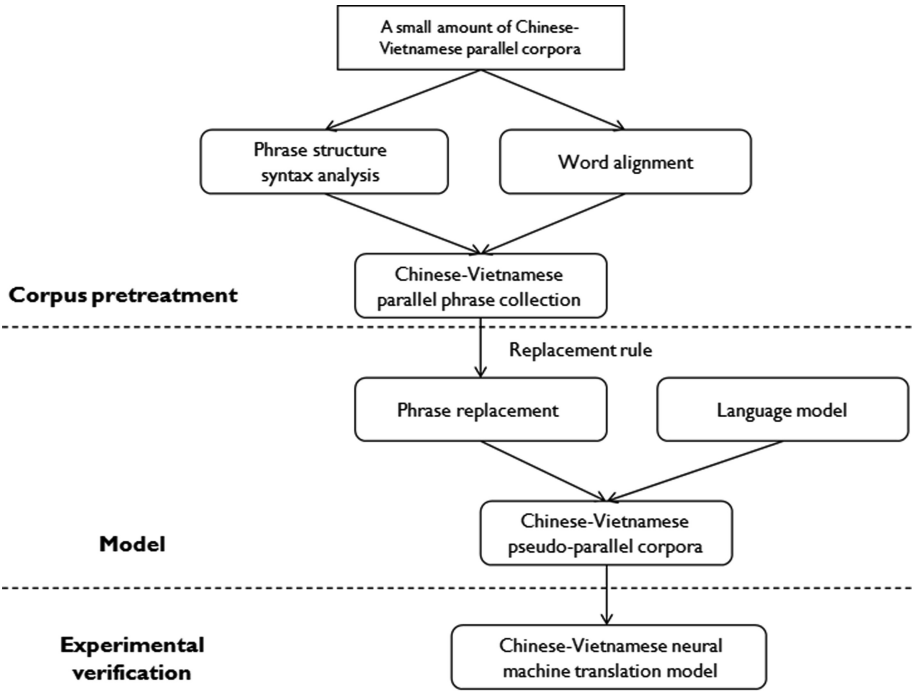

Fig. 1. Phrase-based Chinese and Vietnamese pseudo-parallel sentence pair generation model

3.1 Chinese-Vietnamese Sentence Structure

The main syntactic components of Chinese and Vietnamese are arranged in the same order, and the order of the modifiers is inconsistent in most cases. Modern linguistics has found that all languages in the world seem to have the same structure [8].

(1) A sentence (ROOT) consists of at least one simple clause (IP);

$$ROOT \rightarrow IP^{*} \tag{1}$$

(2) A simple clause (IP) consists of a noun phrase (NP) and a verb phrase (VP);

$$\text{IP} \rightarrow \text{NP VP} \tag{2}$$

(3) A noun phrase (NP) is composed of the qualifier (det), the adjective (A), and the noun (N);

$$\text{NP} \rightarrow \text{det A}^{*}\text{N} \tag{3}$$

(4) A verb phrase (VP) consists of a noun phrase (NP) and a verb (V).

$$\text{VP} \rightarrow \text{NP VP} \tag{4}$$

There are other phrase structures in Chinese and Vietnamese sentences, such as prepositional phrases (PP). This article mainly uses noun phrases (NP) and verb phrases (VP) as phrases. In particular, the noun phrase (NP) here has only one word.

3.2 Chinese-Vietnamese Phrase Alignment

There are not natural spacers between words in Chinese sentences. Although there are spaces in Vietnamese sentences. However, spaces are used as spacers for syllables. A syllable is probably not a separate word. We use Stanford University's Stanford NLP [9] toolkit for word segmentation and syntactic structure analysis of Chinese and Vietnamese corpora. At the same time, we use GIZA++ [10] to perform the Chinese-Vietnamese word alignment processing, and obtain the Chinese-Vietnamese word alignment information.

After the syntactic parsing of the Chinese-Vietnamese parallel sentence pairs, we can obtain the phrase syntactic structure tree of the parallel sentence pairs. Figure 2 (a) is a syntactic parse of the Chinese phrase structure syntax tree. Figure 2(b) is the corresponding Vietnamese phrase structure syntax tree. The phrase structure syntax tree of the Chinese and Vietnamese parallel sentence pairs is similar. Both the NP phrase and the VP phrase in the sentence are at the same depth in the tree, and the components that make up the phrases are similar.

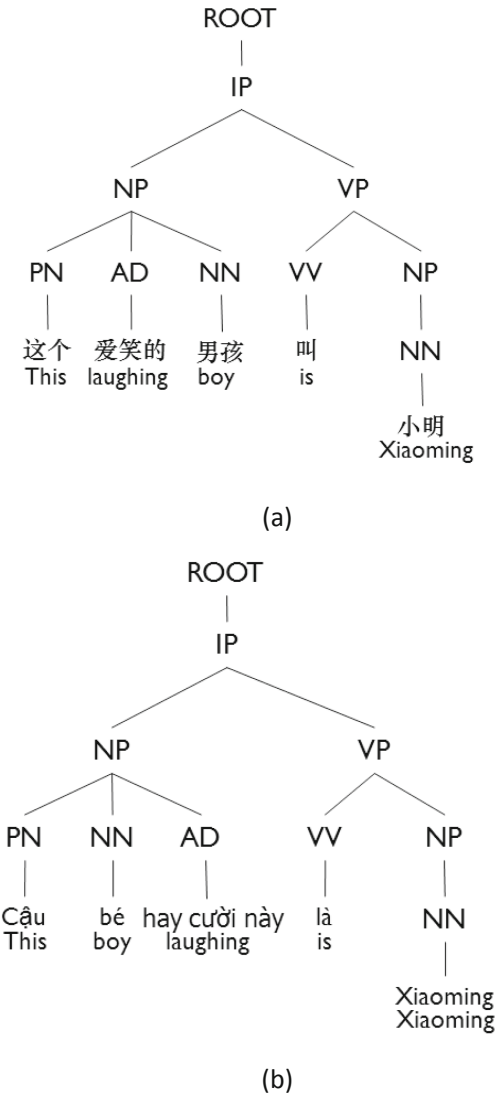

Fig. 2. Chinese and Vietnamese syntax tree.

Since the Chinese-Vietnamese parallel sentence pairs have similar syntax structures, we find all NP nodes and VP nodes in the tree. And we use each NP node and VP node as the root node to form multiple subtrees, each subtree is the phrase in this article. Then we use the word alignment information, the depth information of the node, and the node information of each subtree to perform the phrase alignment. Table 1 is the phrase after the alignment of the Chinese and Vietnamese parallel sentences.

For the phrase consisting of at least two words, we add it to the collection of Chinese and Vietnamese aligned phrases. For an NP phrase containing only one word, if the word is a rare word (the frequency of occurrence in the corpus is less than C), then we add this NP phrase block to the set of Chinese-Vietnamese aligned phrases.

Table 1. Alignment phrases in Chinese-Vietnamese parallel sentence pairs.

phrase	Chinese	Vietnamese
1	(NP(PN 这个(This))(AD 爱笑的(laughing))(NN 男孩(boy)))	(NP(PN Cậu(This))(NN bé(boy))(AD hay cười này(laughing)))
2	(VP(VV 叫(is))(NP(NN 小明(Xiaoming))))	(VP(VV là(is))(NP(NN Xiaoming(Xiaoming))))
3	(NP(NN 小明(Xiaoming)))	NP(NN Xiaoming(Xiaoming)))

3.3 Phrase Replacement

Phrase structure syntax analysis can transform sentences into tree structures. The structure of this tree puts the words in the right place, and the structure of the tree is modular [8]. In the phrase syntax tree, the noun phrase NP and the verb phrase VP are like the components of a certain shape. According to the rules of the phrase structure syntax tree, we can insert or replace one component (phrase) arbitrarily with another component (phrase).

Therefore, the rules for the replacement of phrases in this article are mainly:

(1) The same phrase can be replaced. That is, the NP phrase in the sentence can only be replaced with the NP phrase, and the VP phrase can only be replaced with the VP phrase.
(2) Each sentence replaces only one phrase at a time, and the new sentence pair no longer replaces the phrase.

This substitution rule can not only change the word frequency information of the corpus, but also change the structural information of the sentence. When we replace the other phrase blocks with a noun phrase consisting of only one rare word, we can increase the frequency of occurrence of rare words and enhance the generalization ability of rare words. When the phrases of different sizes are replaced, the structural information of the sentence is also changed.

Figure 3(a) is a phrase replacement for changing the word frequency information, and Fig. 3(c) is a phrase replacement for changing the syntax structure.

In Fig. 3(a), we also replace the NP phrase in the Chinese-Vietnamese parallel sentence with an NP phrase in the Chinese-Vietnamese aligned phrase set, which changes the corpus frequency information. In Fig. 3(b), we replace the NP phrase consisting of one word in a sentence with a more complex NP phrase, which changes the structural information of the sentence.

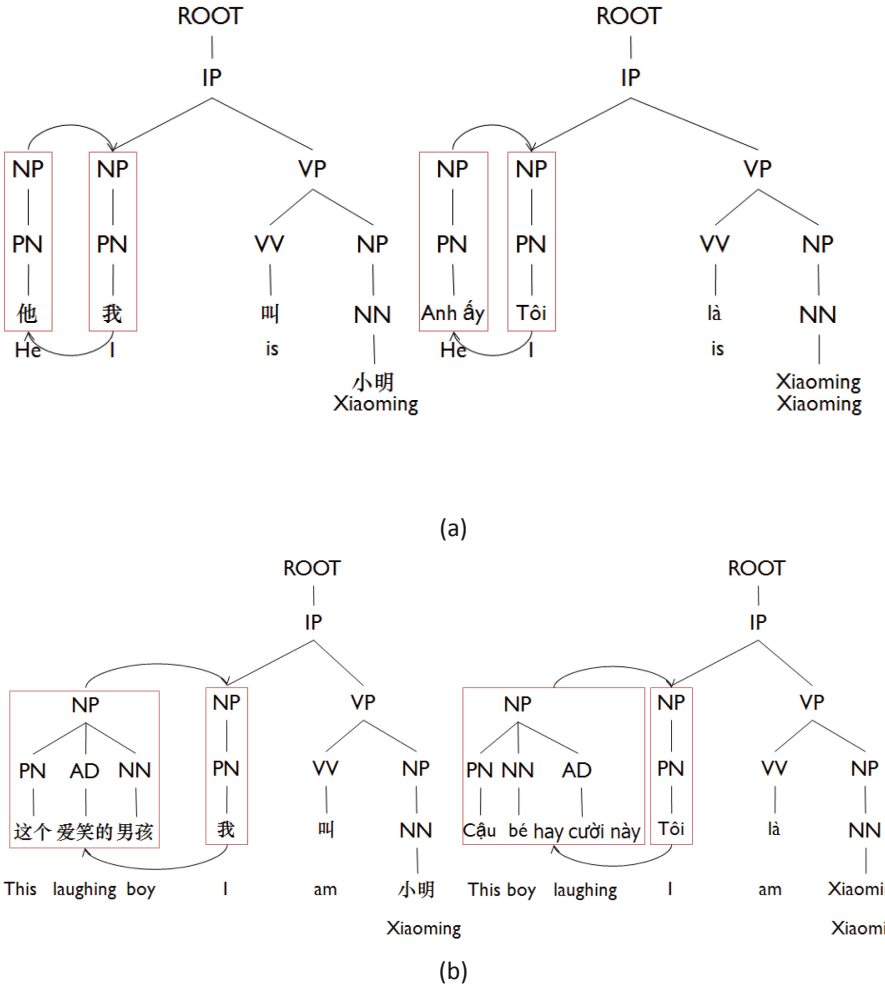

Fig. 3. The rules of phrase replacement

3.4 RNN-Based Language Model Verification Mechanism

We believe that the Chinese and Vietnamese sentence pairs generated under the guidance of the phrase structure syntax information have fewer grammatical errors, but many sentence pairs have semantic errors. In order to judge whether the Chinese and Vietnamese sentences obtained by phrase substitution conform to the grammatical and semantic features, we use the verification mechanism of the RNN-based language model. The verification mechanism can predict the probability of occurrence of the next word according to the context of the text, and can further calculate the probability of occurrence of the entire sentence. In theory, when the grammatical semantics of the corpus of the training language model is correct, sentences with wrong grammatical semantics will get lower scores.

In order to ensure that the generated Chinese-Vietnamese sentence pairs are correct in syntax and semantics, we constructed Chinese and Vietnamese language models for verification. The specific process is shown in Fig. 4.

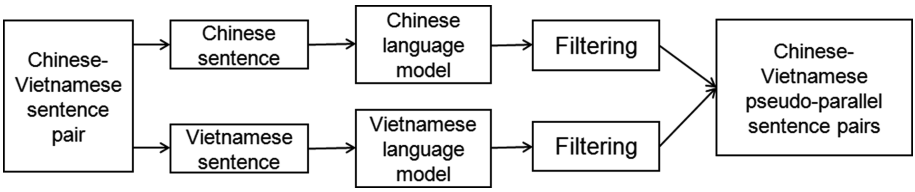

Fig. 4. Language model verification mechanism.

Figure 4 is the flow of the RNN-based language model verification mechanism. The Chinese language model and the Vietnamese language model were trained using Chinese Wikipedia corpus and Vietnamese Wikipedia corpus. For the generated Chinese-Vietnamese sentence pairs, we use the Chinese language model and the Vietnamese language model to score Chinese sentences and Vietnamese sentences. This score is the probability that a sentence will appear. If the score of the sentence is higher, it means that the higher the probability of the sentence appearing, the higher the probability that the sentence is correct in grammatical semantics. When the score of a sentence is less than the threshold we set, we think that the sentence is the wrong sentence, and filter out the sentence of the Chinese-Vietnamese sentences. Only when the Chinese sentences and Vietnamese sentences in the Chinese and Vietnamese sentences are screened by the language model, we use the Chinese and Vietnamese sentence pairs as the Chinese and Vietnamese pseudo-parallel corpus for training the Chinese and Vietnamese neural machine translation models.

For the selection of the language model threshold, we use the language model to score the monolingual corpus of the corresponding language, and separately calculate the lowest score in the monolingual corpus, which is used as the threshold of the corresponding language model.

4 Experiment

4.1 Data Settings

This paper is based on the small-scale parallel sentence pairs of Chinese and Vietnamese to generate Chinese-Vietnamese pseudo-parallel sentence pairs. Therefore, we will use the 120,000 Chinese-Vietnamese parallel sentence pairs crawled from the Internet as the sentence pairs to be expanded.

Before performing pseudo-parallel sentence pairs based on phrase substitution, we also need to perform a series of pre-processing work on Chinese-Vietnamese parallel corpus, including word segmentation, word alignment, phrase extraction, and phrase alignment.

4.2 Experimental Results

In this paper, we mix Chinese-Vietnamese parallel corpus with Chinese-Vietnamese pseudo-parallel corpus in different proportions. Through this approach, we verify the influence of the generated Chinese and Vietnamese pseudo-parallel corpus on the translation of the Chinese and Vietnamese neural machines. The benchmark experiments in this paper are RNNSearch [11], GNMT [12], and Transformer [13]. The benchmark experiment was trained by the Chinese and Vietnamese parallel corpora, and the corpus size was 125 k parallel sentence pairs. According to the ratio of 2:1, 1:1, 1:2, 1:5 (parallel corpus: pseudo-parallel corpus), we mix parallel corpus and pseudo-parallel corpus. Then we trained the RNNSearch, GMT, and Transformer models with the mixed corpus. Table 2 shows the experimental results of the baseline model and the addition of pseudo-parallel corpus. The evaluation index is the value of BLEU.

Table 2. Experimental results after adding pseudo-parallel corpus.

Mixed ratio (parallel: pseudo-parallel)	RNNSearch	GNMT	Transformer
–	13.43	14.21	18.63
2:1	13.86	14.49	18.86
1:1	14.08	14.83	19.15
1:2	14.70	15.20	19.63
1:5	14.55	15.57	20.06

After joining the Chinese-Vietnamese pseudo-parallel corpora, the performance of the Chinese and Vietnamese neuromachine translations has generally improved. For the RNNSearch model, when the mixing ratio of the Chinese-Vietnamese parallel corpus and the Chinese-Vietnamese pseudo-parallel corpus is 1:2, the value of the BLEU is the highest. For the GNMT model, when the mixing ratio of the Chinese-Vietnamese parallel corpus and the Chinese-Vietnamese pseudo-parallel corpus is 1:5, the value of the BLEU is the highest. For the Transformer model, when the mixing ratio of the Chinese-Vietnamese parallel corpus and the Chinese-Vietnamese pseudo-parallel corpus is 1:5, the value of the BLEU is the highest. In general, the more pseudo-parallel corpora generated by the method of this paper, the better the performance of Chinese-Vietnamese neural machine translation. Table 3 is a partial pseudo-parallel sentence pair generated by the phrase block replacement to generate pseudo-parallel sentence pairs.

From the experimental results, the generated Chinese-Vietnamese pseudo-parallel sentence pairs have higher quality. In the pseudo-parallel sentence pairs between Chinese-Vietnamese, there may be cases where Chinese and Vietnamese cannot be completely translated because of the word alignment. However, since this situation is rare, we believe that the generated Chinese-Vietnamese pseudo-parallel sentence pairs have higher quality.

Table 3. Pseudo-parallel sentence pair generation results based on phrase substitution.

Chinese	Vietnamese
这是西班牙的EUPHORE粉尘和烟雾研究实验室。(This is the EUPHORE dust and smoke research laboratory in Spain.)	ĐâylàPhòngnghiêncứukhóibụi EUPHORE ở Tây Ban Nha. (This is the EUPHORE dust and smoke research laboratory in Spain.)
这是西班牙的科学仪器。(This is the Spanish scientific instruments.)	Đâylàmộtcôngcụ khoa họctừTây Ban Nha. (This is the Spanish scientific instruments.)
这是西班牙的最大的科学会议。(This is the largest scientific conference in Spain.)	Đâylàhộinghị khoa họclớnnhất ở Tây Ban Nha. (This is the largest scientific conference in Spain.)
这是西班牙的头条新闻。(This is the headline news of Spain.)	Đây là tin tứctiêuđềcủaTây Ban Nha. (This is the headline news of Spain.)
这是一场全球性的品牌推广活动。(This is a global branding event.)	Đâylàmộtsựkiệnthươnghiệutoàncầu. (This is a global branding event.)
这是我的第二本书。(This is my second book.)	Đâylàcuốnsáchthứhaicủatôi. (This is my second book.)
这是一场狩猎游戏吗? (Is this a hunting game?)	Đâycóphảimộttròsănìmkhông? (Is this a hunting game?)

5 Summary

This paper proposes a pseudo-parallel corpus generation method based on small-scale Chinese-Vietnamese parallel corpora. We transform the generation of pseudo-parallel corpus into the replacement and recombination of elements between sentences. We combine the noun phrase (NP) and the verb phrase (VP) in the phrase structure syntax tree into a phrase block. Then we reorganize the sentences based on the principle that phrases of the same nature can be replaced. Finally, we use the language model to grammatically and semantically constrain the newly generated sentences, and achieve the purpose of generating pseudo-parallel sentence pairs with correct grammatical semantics. The experimental results show that the proposed method can generate Chinese-Vietnamese pseudo-parallel corpus with high quality, which effectively improves the performance of Chinese-Vietnamese neural machine translation.

Acknowledgements. The work was supported by National key research and development plan project (Grant Nos. 2018YFC0830105, 2018YFC0830100), National Natural Science Foundation of China (Grant Nos. 61732005, 61672271, 61761026, and 61762056), Yunnan high-tech industry development project (Grant No. 201606), and Natural Science Foundation of Yunnan Province (Grant No. 2018FB104).

References

1. Sennrich, R., Haddow, B., Birch, A.: Improving Neural Machine Translation Models with Monolingual Data. arXiv preprint arXiv:1511.06709 (2015)
2. He, W., Zhao, S.Q., Wang, H.F., et al.: Enriching SMT training data via paraphrasing. In: Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP), 8–13 November 2011, Chiang Mai, Thailand, pp. 803–810 (2011)
3. Bond, F., Nichols, E., Appling, D.S., et al.: Improving statistical machine translation by paraphrasing the training data. In: Proceedings of the International Workshop on Spoken Language Translation (IWSLT), 20–21 October 2008, Honolulu, Hawaii, USA, pp. 150–157 (2008)
4. Nakov, P.: Improved statistical machine translation using monolingual paraphrases. In: Proceedings of the 18th European Conference on Artificial Intelligence (ECAI), 21–25 July 2008, Patras, Greece, pp. 338–342 (2008)
5. He, W., Liu, T.: Parse-realize based paraphrasing and SMT corpus enriching. J. Harbin Inst. Technol. **45**(5), 45–50 (2013)
6. Fadaee, M., Bisazza, A., Monz, C.: Data Augmentation for Low-Resource Neural Machine Translation. arXiv preprint arXiv:1705.00440 (2017)
7. Cai, Z.L., Yang, M.M., Xiong, D.Y.: Data augmentation for neural machine translation. J. Chin. Inf. Process. **32**(7) (2018)
8. Pinker, S.: The Language Instinct: How the Mind Creates Language, pp. 101–105. Penguin, UK (2003)
9. Manning, C.D., Mihai, S., John, B., et al.: The Stanford CoreNLP natural language processing toolkit. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL), 22–27 June 2014, Baltimore, MD, USA, pp. 55–60 (2014)
10. Och, F.J.: Giza++: training of statistical translation models (2001). http://www.informatik.rwth-aachen.de/Colleagues/och/ software/GIZA++.html
11. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate (2014). arXiv:1409.0473
12. Wu, Y., Schuster, M., Chen, Z., et al.: Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. arXiv preprint arXiv:1609.08144 (2016)
13. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017 (NISP), 4–9 December 2017, Long Beach, CA, USA, pp. 6000–6010 (2017)