ICIC Express Letters
Part B: Applications
Volume **9**, Number **10**, October **2018**

 ISSN 2185-2766
pp. **983–990**

PRE-ORDERING OF "DE PHRASE" IN CHINESE-VIETNAMESE MACHINE TRANSLATION

Phuoc Tran[1,*], Le Nguyen[2] and Dien Dinh[3]

[1]NLP-KD Lab
Faculty of Information Technology
Ton Duc Thang University
No. 19, Nguyen Huu Tho Street, Tan Phong Ward, District 7, Ho Chi Minh City, Vietnam
[*]Corresponding author: tranthanhphuoc@tdt.edu.vn

[2]Faculty of Information Technology
Ho Chi Minh City University of Food Industry
No. 140, Le Trong Tan Street, Tay Thanh Ward, Tan Phu District, Ho Chi Minh City, Vietnam
lenv@cntp.edu.vn

[3]Faculty of Information Technology
VNU-HCM University of Science
No. 280, An Duong Vuong Street, Ward 4, District 5, Ho Chi Minh City, Vietnam
ddien@fit.hcmus.edu.vn

Received April 2018; accepted July 2018

Abstract. *DE (的) is a very common structure in Chinese text. These phrases containing this structure are often reordered when translating them into other languages, including Vietnamese. In addition, state-of-the-art phrase-based statistical machine translation cannot overcome word order errors when translating DE phrases from Chinese into Vietnamese. Moreover, Chinese-Vietnamese is a low-resource language pair, and the errors of word order in Chinese-Vietnamese machine translation are more aggravated. In this paper, we propose to use Chinese dependency relation to pre-order Chinese DE phrases in accordance with their Vietnamese word order. The experimental results show that our method has improved performance of translation system compared to the translation systems using only a reordering model of current phrase-based statistical machine translation.*

Keywords: Chinese-Vietnamese SMT, Preordering, Dependency relations, DE phrase

1. **Introduction.** Word order is one of the most difficult problems in machine translation [2]. The word reordering focuses on two branches, including (1) reordering for language pairs having a short distance, and (2) reordering for the language pairs having a long distance. For the case (1), if a bilingual corpus for machine translation is large enough, the distance-based reordering model (DRM) or the lexicalized reordering model (LRM) [1] of state-of-the-art phrase-based statistical machine translation (P-SMT) can overcome; however; they cannot overcome for the case (2).

Grammar structures related to the DE word (called DE phrases) are very common in Chinese text and they are often mistranslated when translating them from Chinese into English [1] as well as into Vietnamese. In the CLC[1] (Computational Linguistic Center) bilingual corpus of 14,507 sentence pairs, there are 3,199 DE words. When translating from Chinese into Vietnamese, the DE phrases are usually reordered. Figure 1 represents a case of word order related to the DE phrase. The English meaning of the Chinese sentence in Figure 1 is "My sister's bicycle is broken".

DOI: 10.24507/icicelb.09.10.983

[1]http://www.clc.hcmus.edu.vn/?page_id=467&lang=en

FIGURE 1. An example of word order of the DE phrases

FIGURE 2. An example of the DE phrase in Figure 1 is pre-ordered.

In this paper, we propose a pre-ordering method to make word order of Chinese DE phrases suitable to their Vietnamese one. Firstly, the system will parse dependence relation (DR) for a Chinese sentence, and then, the DR related to the DE word will be extracted. The system will cluster two phrases in relation to the two words in this DR. Finally, the two clustered phrases will be reordered each other before training and translation processes of P-SMT. This method has significantly improved P-SMT performance. Figure 2 shows a result of the Chinese sentence in Figure 1 which is pre-ordered by our system.

The rest of the paper is structured as follows. Section 2 indicates some related work. Section 3 provides a detailed description of our proposed method. Section 4 shows and discusses the results of our experiments. And, Section 5 summarizes our work and gives our main conclusions.

2. **Related Work.** Reordering model is one of three main models of P-SMT. DRM and LRM are used in P-SMT (such as Moses). However, both models have problems when they are applied to a long distance language pairs. In order to improve the quality of word reordering in P-SMT, the two popular approaches, which are pre-ordering and post-ordering, were applied.

Until the present time, there are three methods using the pre-ordering approach, including pre-ordering based on part of speech information [4,10-12], based on syntax information [3,5,8,13], and based on dependency relation [9,14,18]. Against the pre-ordering approach, this approach focuses on reordering after translation step. This approach was initially done in Japanese-English machine translation [15-17] and gave good results.

The part of speech-based method has the advantage of reordering between words in a specific phrase and is inefficient in long-distance reordering. The syntax-based method has overcome the long-distance reordering, but the method requires translation system to have a large bilingual corpus. Because Chinese-Vietnamese is a low-resource language pair, the two approaches are inefficient. In this paper, we have applied the dependency method to pre-ordering Chinese DE phrases.

Typical work for reordering the DE word was conducted by Chang et al. [1]. There are many labels related to the DE word in Chinese Treebank Tag set, and the authors only focused to handle the DE word having noun phrase forms (NPs). The authors divided this label into five classes, and for each class they proposed a specific method to reorder. In our work, we use Chinese dependency relation to extract DE phrases, and DE phrases are not only NPs but also verb phrase (rcmod DR, presented in Section 3.2.b).

Dependency relation (DR) based approach for pre-ordering was used in [14]. In this work, the authors pre-ordered Chinese sentence based on 45 DR taggers to be suitable to English word order. This work defined two word pre-ordering rules to extract reordered DRs. This is a work that is considered to be close to our research. We also use the DR to pre-order Chinese word. However, our approach is different, as indicated below.

– Pre-ordering approach depends on the target language. Our target language is Vietnamese; therefore, the DR candidates as well as pre-ordering rules of our work are not similar to this work.
– There are not any Chinese DRs whose two phrases are completely reordered when translating them into Vietnamese. Our work will add some linguistic constraints to DE phrase to increase the accuracy of pre-ordering.

3. **Pre-Ordering DE Phrase.** Figure 3 shows a process building the DE phrases reordering rules.

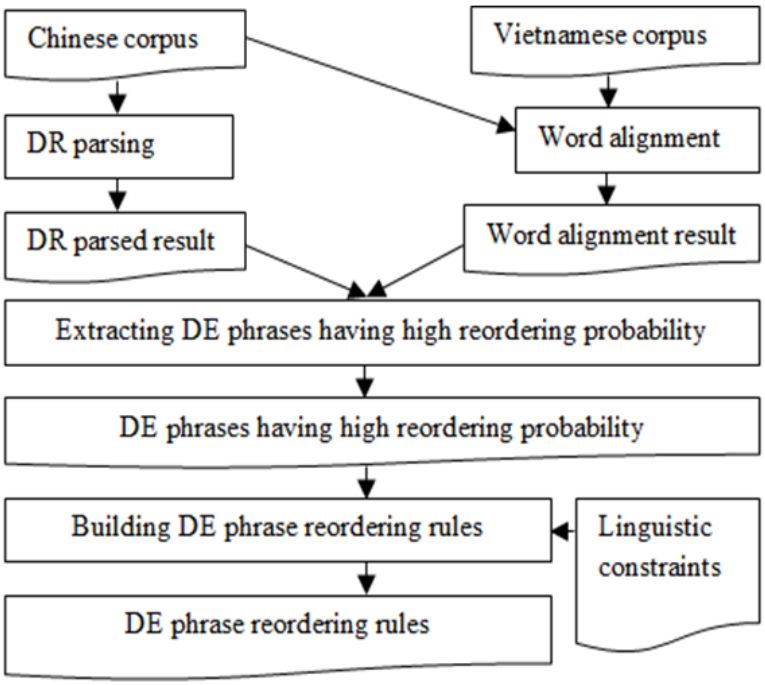

FIGURE 3. Building the DE phrases reordering rules framework

3.1. **Extracting the DE phrases having high reordering probability.** A candidate DE phrase is a phrase which reordering probability of the two words in this phrase is greater than a given threshold. Table 1 lists DE DRs as well as their reordering probability when translating them into Vietnamese.

TABLE 1. Illustrating reordering probability of DE DRs

ID	DR name	Number of reordered DR	Number of DR	Percentage
1	assm	1,004	1,341	0.75
2	assmod	990	1,325	0.75
3	rcmod	219	434	0.50
4	cpm	202	480	0.42

The assm and assmod DRs usually exist together, assm DR represents a relationship between the preceding noun and the DE word, and assmod DR expresses a relationship between two nouns in this DR. Similar to the assmod and assm DRs, rcmod and cpm DRs also often exist together and they are often reordered in Chinese-Vietnamese translation. cpm DR usually expresses a relationship between the preceding verb or adjective and the DE word, while rcmod DR represents a relationship between the verb or adjective preceding the DE word and the noun following the DE word.

3.2. The method of DE phrases pre-ordering.

a. assmod DR based pre-ordering

– Assumptions:
  + assmod DR form: *assmod* ($< W_i >$-$< i >$, $< W_j >$-$< j >$)
  + $< j >$ must be less than $< i >$.
  + The phrase containing assmod DR is like $X(W_i)$ 的 $Y(W_j)$, where $X(W_j)$ and $Y(W_i)$ are the phrases containing W_i and W_j as well as other words having dependency relation with them. The words in the $X(W_j)$ and $Y(W_i)$ phrases will also be reordered.

– The reordering method:

$$X(W_j) \text{ 的 } Y(W_i) \rightarrow Y(W_i) \text{ 的 } X(W_j) \tag{1}$$

Figure 4 illustrates a pre-ordering result of assmod(自行车-4, 妹妹-2) in Chinese sentence "我 妹妹 的 自行车 坏 了" (Xe đạp của em gái tôi hư rồi: My sister's bicycle is broken) to be suitable to Vietnamese word order. $X(W_i)$ in this case is "我 妹妹" (em gái tôi: my sister), and $Y(W_j)$ is "自行车" (xe đạp: bicycle).

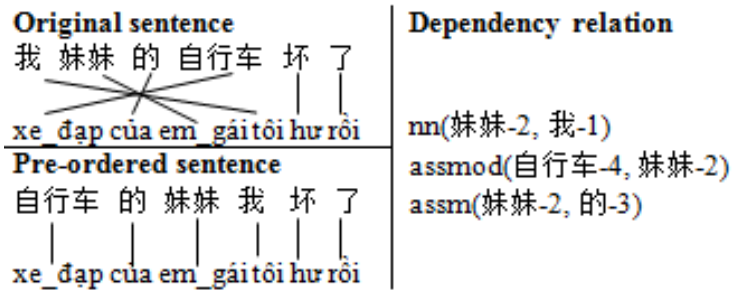

FIGURE 4. An example of assmod-based pre-ordering

b. rcmod DR based pre-ordering

– Assumptions:
  + rcmod DR form: *rcmod* ($< W_i >$-$< i >$, $< W_j >$-$< j >$)
  + $< j >$ must be less than $< i >$.
  + The phrase containing rcmod DR is like $W_j[X]$ 的 $Y(W_i)$, where $[X]$ is a phrase following the word W_j (optional) and preceding the DE word, and $Y(W_i)$ is a phrase containing W_i as well as other words having dependency relation with it.

– The reordering method:

$$W_j[X] \text{ 的 } Y(W_i) \rightarrow \begin{array}{ll} Y(W_i) \text{ 的 } W_j[X] & \text{(a)} \\ W_j Y(W_i) \text{ 的 } [X] & \text{(b)} \end{array} \tag{2}$$

Chinese sentences in this case will be pre-ordered into two separate sentences and they are translated by P-SMT. Based on 2-gram language model, the system will choose the best sentence. Figure 5 presents two cases of pre-ordering based on rcmod DR. In the case 1, a Chinese sentence "这 是 学习 汉语 的 好 机会" (Đây là cơ hội tốt để học tiếng Hoa: This is a good opportunity to learn Chinese), W_j is the verb "学习" (học: learn), $[X]$ is the noun "汉语" (tiếng Hoa: Chinese), and $Y(W_i)$ is the noun phrase "好 机会" (cơ hội tốt: good opportunity). Pre-ordering method in this case is as Formula (2)(a), and the noun phrase $Y(W_i)$ (好 机会) is reordered with the verb W_j (学习). In the case 2, part of a Chinese sentence "听说 上海 的 小吃..." (Nghe nói thức_ăn nhẹ của Thượng_Hải ...: I heard that Shanghai snacks ...), W_j is the verb "听说" (nghe nói: heard), $[X]$ is the noun "上海" (Thượng Hải: Shanghai), and $Y(W_i)$ is the noun "小吃" (thức_ăn nhẹ: snack). The method of pre-ordering for this case is based on Formula (2)(b), and the phrase $Y(W_i)$ (小吃) is reordered with the verb W_j (听说).

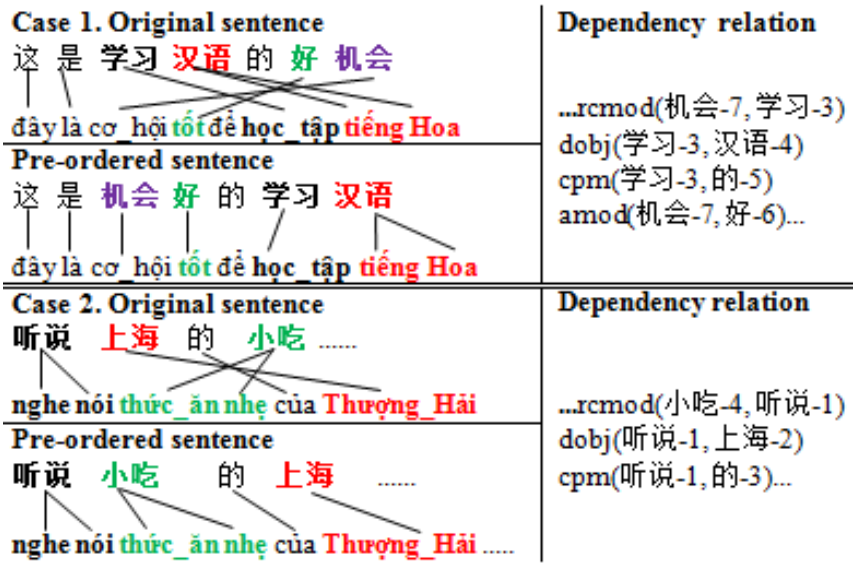

FIGURE 5. Illustrating pre-ordering based on rcmod DR

4. Experiments.

4.1. **Experimental corpora.** Our experiment bilingual corpus consists of 35,623 Chinese-Vietnamese sentence pairs of CLC. We used 90% of the sentences for training, 5% of the sentences for testing, and the remaining 5% of the sentences for tuning the machine translation parameters. We used these corpora to perform two experiments including WS translation and DE based pre-ordering translation.

- WS translation (WS-Trans): The words in Chinese corpora and Vietnamese corpora are word-segmented.
- DE-based pre-ordering translation (DE-Trans): We pre-order Chinese corpora (training, developing, and testing) based on the DRs in relation to word DE.

4.2. **Experimental results.** We evaluated impact of DE-based pre-ordering on translation result based on the following criteria:

- Precision of DE-based pre-ordering.
- Number of unknown words (UKW).
- BLEU and TER scores.

Table 2, Table 3, and Table 4 show the experimental results.

TABLE 2. Precision of DE-based pre-ordering

	Number of DE phrases	Correct translation cases		Incorrect translation cases	
		Total	Percentage	Total	Percentage
WS Trans	393	216	54.96%	177	45.04%
DE Trans		327	83.21%	66	16.79%

TABLE 3. Number of UKW of testing corpora

	Number of UKW	Number of words	Percentage
WS Trans	1,206	13,186	9.15%
DE Trans	1,198		9.08%

TABLE 4. BLEU and TER scores of the two translation systems

	BLEU	TER
WS-Trans	34.90	48.16
DE-Trans	35.40	47.13

4.3. **Analysis.** Based on the experimental results in Table 2, Table 3, and Table 4, we found that the DE-Trans improved machine translation compared to the WS-Trans. The improvements are expressed in three aspects, (1) the DE-Trans has higher precision of DE-based pre-ordering, (2) the DE-Trans has higher BLEU score and lower TER score, and (3) the DE-Trans has less UKW. After pre-ordering for Chinese corpora, the word order of DE phrases of the DE-Trans is more similar to the word order of Vietnamese phrases than the one of the WS-Trans. Therefore, the DE-Trans translates correctly the sentences containing the DE phrases although these phrases do not exist in the training corpus.

Figure 6 shows a case of correct translation of a sentence containing a DE phrase which does not exist in the training corpus. In the WS-Trans, because this DE phrase ("你 的 汽车": xe hơi của bạn: your car) does not exist in the training corpus, this translation system cannot correctly reorder when translating this phrase into Vietnamese. A correct translation is "xe hơi của bạn đã chạy bao_nhiêu dặm?" (How many miles did your car run?), while the WS-Trans translates into "bạn chạy xe bus bao_nhiêu dặm?" (similar to "you drive a bus how many miles?"). On the contrary, in the DE-Trans, the DE phrase is pre-ordered in accordance with Vietnamese word order. Consequently, the DE-Trans translates this phrase correctly because each word in this DE phrase appears in the corpus and the word order of this phrase is similar to Vietnamese one.

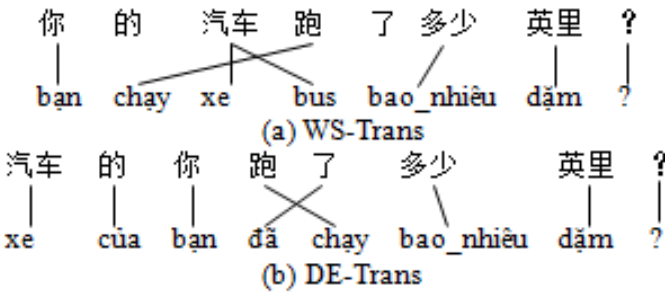

FIGURE 6. Illustrating a case of correct reordering of the DE-Trans compared to the WS-Trans

Quality of machine translation will be improved significantly if the texts in source language and target language are monotonous [6,7]. Because the word order of Chinese and Vietnamese of the DE-Trans is closer than the WS-Trans' one, the number of Monotone reordering types of DE-Trans is greater than the WS-Trans' one. This causes the DE-Trans to get some useful phrases that the WS-Trans does not get. As a result, number of UKW of the DE-Trans is less than the WS-Trans' one. Figure 7 shows a case that the DE-Trans does not generate UKW but the WS-Trans does not overcome. The English meaning of the Chinese sentence in Figure 7 is "I want to take part in your helicopter".

In the case (a) of Figure 7, the dependency parser analyzes it into some important DRs, such as nsubj(想-2, 我-1), rcmod(直升飞机-6, 参加-3), and cpm(参加-3, 的-5). Then, the system will pre-order based on rcmod DR and give a result as the case (b) of Figure 7. After this sentence is pre-ordered, the translation system not only translates its word order correctly but also translates successfully the word "直升飞机" (trực_thăng: helicopter). The correct order must be "trực_thăng của các_bạn" ("your helicopter"). Particularly, for the UKW "直升飞机", a reason for the UKW generation of the WS-Trans is that the

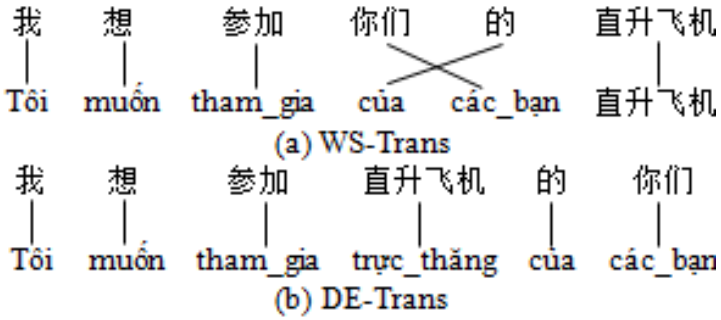

FIGURE 7. Illustrating the less UKW generation of the DE-Trans compared to the WS-Trans

word does not exist in the WS-Trans' phrase table. We found that there are two sentences containing this word in the training corpus, but due to the sparse data and a different word order between Chinese and Vietnamese, the phrase table does not show the word "直升飞机". However, after the training corpus is pre-ordered, the phrase table of the DE-Trans contains this word and it is obvious that the DE-Trans translates this word successfully and does not generate UKW.

In addition to these improvements, our pre-ordering system will not work well for Chinese sentences having complex DE phrases and the parser analyzes incorrectly these phrases. Our system pre-orders DE phrases based on the result of the dependency parsing. If the parser parses wrongly, the system will pre-order incorrectly, and the translation result of the pre-ordered sentence can be worse than the original sentence's one.

5. **Conclusions and Perspectives.** In this paper, we have improved the quality of Chinese-Vietnamese machine translation based on pre-ordering the DE phrases to match Vietnamese word order before training and translation processes of P-SMT. The DE phrases appear commonly in Chinese text and take part in many grammatical roles in Chinese sentence. In some cases, these phrases are reordered and in some cases not. Our system has used Chinese dependency relation to classify the DE phrases having reordering in Chinese-Vietnamese translation. Then, based on some linguistic constraints, we have constructed two word order rules for the two DRs in relation to the DE word, including assmod and rcmod DRs. The pre-ordering has improved performance of P-SMT which is indecated in aspects, for example, DE-Trans' BLEU score is higher than WS-Trans' one, DE-Trans' TER score is lower than WS-Trans' one, and DE-Trans generates less UKW than WS-Trans.

One drawback of the system is that pre-ordering is completely based on the dependency parsing results of the parser. The parser sometimes gives wrong results in some complex DE phrase structures. This leads to some incorrect translation results. Improving the performance of the dependency parser will be our near future work.

Acknowledgment. We sincerely thank the Computational Linguistics Center (CLC) has contributed to us the Chinese-Vietnamese bilingual corpus.

REFERENCES

[1] P.-C. Chang, D. Jurafsky and C. D. Manning, Disambiguating "DE" for Chinese-English machine translation, *Proc. of the 4th Workshop on Statistical Machine Translation*, pp.215-223, 2009.

[2] P. Koehn, A. Axelrod, A. Birch, C. Callison-Burch, M. Osborne and D. Talbot, Edinburgh system description for the 2005 IWSLT speech translation evaluation, *International Workshop on Spoken Language Translation*, pp.68-75, 2005.

[3] T. P. Nguyen and A. Shimazu, Improving phrased-based SMT with morpho-syntactic analysis and transformation, *Proc. of the 7th Conference of the Association for Machine Translation in the Americas*, pp.138-147, 2006.

[4] J.-J. Li, J. Kim, D.-I. Kim and J.-H. Lee, Chinese syntactic reordering for adequate generation of Korean verbal phrases in Chinese-to-Korean SMT, *Proc. of the 4th Workshop on Statistical Machine Translation*, pp.190-196, 2009.

[5] I. Badr, R. Zbib and J. Glass, Syntactic phrase reordering for English-to-Arabic statistical machine translation, *Proc. of the 12th Conference of the European Chapter of the ACL*, pp.86-93, 2009.

[6] M. Khalilov and K. Sima'an, Source reordering using maxent classifiers and super tags, *Proc. of EAMT'10*, pp.292-299, 2010.

[7] C. Wang, M. Collins and P. Koehn, Chinese syntactic reordering for statistical machine translation, *Proc. of EMNLP-CoNLL'07*, pp.737-745, 2007.

[8] M. Khalilov and K. Sima'an, A discriminative syntactic model for source permutation via tree transduction, *Proc. of the 4th Workshop on Syntax and Structure in Statistical Translation*, pp.92-100, 2010.

[9] D. Genzel, Automatically learning source-side reordering rules for large scale machine translation, *Proc. of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pp.376-384, 2010.

[10] A. Bisazza and M. Federico, Chunk-based verb reordering in VSO sentences for Arabic-English statistical machine translation, *Proc. of the Joint 5th Workshop on Statistical Machine Translation and Metrics (MATR)*, pp.235-243, 2010.

[11] J. Jiang, J. Du and A. Way, Source-side syntactic reordering patterns with functional words for improved phrase-based SMT, *Proc. of the 4th Workshop on Syntax and Structure in Statistical Translation*, pp.19-27, 2010.

[12] C.-L. Goh, T. Onishi and E. Sumita, Rule-based reordering constraints for phrase-based SMT, *Proc. of the 15th Conference of the European Association for Machine Translation*, pp.113-120, 2011.

[13] U. Lerner and S. Petrov, Source-side classifier pre-ordering for machine translation, *Proc. of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp.513-523, 2013.

[14] J. Cai, M. Utiyama, E. Sumita and Y. Zhang, Dependency-based pre-ordering for Chinese-English machine translation, *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics*, pp.155-160, 2014.

[15] K. Sudoh, X. Wu, K. Duh, H. Tsukada and M. Nagata, Post-ordering in statistical machine translation, *Proc. of the 13th Machine Translation Summit*, pp.316-323, 2010.

[16] H. Isozaki, K. Sudoh, H. Tsukada and K. Duh, A simple reordering rule for SOV languages, *Proc. of WMT-MetricsMATR*, pp.244-251, 2010.

[17] I. Goto, M. Utiyama and E. Sumita, Post-ordering by parsing for Japanese-English statistical machine translation, *Proc. of the 50th Annual Meeting of the Association for Computational Linguistics*, pp.311-316, 2010.

[18] T. H. Viet, N. V. Vinh, V. T. Huyen and N. L. Minh, Dependency-based pre-ordering for English-Vietnamese statistical machine translation, *VNU Journal of Science: Comp. Science & Com. Eng.*, vol.31, no.3, pp.1-13, 2017.