ICIC Express Letters
Part B: Applications
Volume **9**, Number **3**, March **2018**

ICIC International ©2018 ISSN 2185-2766
pp. **265–272**

DEPENDENCY-BASED PRE-ORDERING OF PREPOSITION PHRASES IN CHINESE-VIETNAMESE MACHINE TRANSLATION

Anh Tran Huu[1], Phuoc Tran[2,*], Dien Dinh[3], Vinh Van Vu[4] and Tan Le[5]

[1]Department of Computer Science and Technology
Beijing Institute of Technology
No. 5, South Zhongguancun Street, Haidian Dist., Beijing 100081, P. R. China
anhuni1006@gmail.com

[2]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

[3]Faculty of Information Technology
VNU-HCM University of Science
No. 227, Nguyen Van Cu Street, Ward 4, District 5, Ho Chi Minh City, Vietnam
ddien@fit.hcmus.edu.vn

[4]Faculty of Information Technology
Ho Chi Minh City University of Food Industry
140 Le Trong Tan Street, Tay Thanh Ward, Tan Phu District, Ho Chi Minh City, Vietnam
vinhvv@cntp.edu.vn

[5]Faculty of Information Technology
Université Du Québec À Montréal
201, Avenue du Président-Kennedy, Local PK 4150, H2X 3Y7 Montreal, Quebec, Canada
le.ngoc_tan@courrier.uqam.ca

Received October 2017; accepted December 2017

Abstract. *Word order is one of the biggest differences between Chinese and Vietnamese languages. In particular, the order of the preposition phrase is a grammar type which has a big difference compared to other grammar types. The state-of-the-art phrase-based statistical machine translation cannot overcome the mistakes of the word order of Chinese-Vietnamese translation. Moreover, Chinese-Vietnamese is considered as a low-resource language pair, so the errors of word order in the translation system are more serious than the other rich-resource language pairs. In this paper, we propose an approach by using Chinese dependency relation in order to preorder Chinese word order to be suitable to Vietnamese word order. The experimental results show that our approach has improved the performance of machine translation system compared to the machine translation system using only the reordering model of the phrase-based statistical machine translation.*

Keywords: Chinese-Vietnamese machine translation, Pre-ordering, Word alignment, Chinese grammatical relations, Preposition phrases

1. **Introduction.** Word order is one of the most difficult problems in machine translation [1]. On the other hand, for different language pairs, the word reordering method is also different. Generally, the word reordering focuses on two branches: (1) reordering for language pairs having a short distance (such as English-French), and (2) reordering for the language pairs having a long distance (namely English-Japanese). 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 the state-of-the-art phrase-based statistical machine translation (P-SMT) can be overcome, but it cannot be overcome for the case (2).

Chinese and Vietnamese have the same isolated language type. They have multiple close linguistic relationships, namely: (1) the words are inflected words, (2) the main grammar methods are word order and function words, (3) the spaces do not define the word boundary, and (4) using function word expresses the negation or the question. In addition to these similarities, Chinese and Vietnamese have a big difference about the word order. This difference is indicated in several aspects such as preposition phrase order, noun phrase order, and “的” (DE) word order, in which, the preposition phrase order is a grammatical type which often appears in Chinese text and is often mistranslated when translating it into Vietnamese. Figure 1 illustrates the difference of the word order between Chinese and Vietnamese. (English meaning of the sentence is “we call him”.)

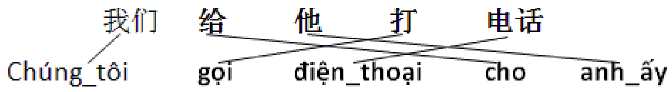


Figure 1. Illustrating a difference of preposition phrase order between Chinese and Vietnamese

On the other hand, because Chinese-Vietnamese is a low-resource language pair, the word reordering performance of the system is not high. Both of DRM and LRM are not enough for the system to reorder accurately local order. This leads to the cases that the system translates wrongly the word order in some sentences whether these sentences have the same structure with sentences existing in the training corpus. For example, in the case of Figure 2, the P-SMT mistranslated the word order in the case (a) and it translated successfully the word order in the case (b). Both cases have the same preposition phrase structure, in the form of <preposition> + <noun> + <verb> (<跟> + <女朋友> + <约会>), and the structure is reordered in the form of <verb> + <preposition> + <noun> (<约会> + <跟> + <女朋友>). However, because the training corpus has covered the case (b), the word order is correctly translated. In contrast, the system mistranslated the word order in the case (a) (the correct order is “tôi chỉ có một nguyện_vọng là kết_hôn với cô_ấy”). The English meaning of the sentence (a) is “I have only one desire that is to marry her” and that of (b) is “Do you have an appointment with your girlfriend?”.

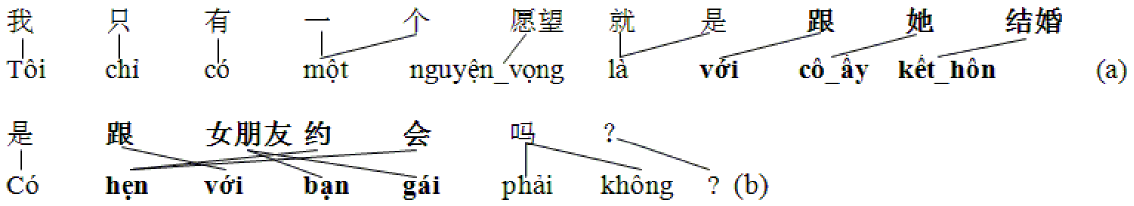


Figure 2. Examples of translating two Chinese sentences having the same structure of P-SMT. The case (a) illustrates an incorrect sample, and the case (b) indicates a correct sample.

In order to improve the quality of word reordering in P-SMT, the two popular approaches are pre-ordering and post-ordering. Until the present time, there are three methods using the pre-ordering approach, including pre-ordering based on part of speech information [6-9], based on syntax information [10-14], and based on dependency relation [15-17,20]. The post-ordering approach was initially done in Japanese-English machine translation [18,19] and gave good results.

The part of speech-based approach has the advantage of reordering between words in a specific phrase and is inefficient in long-distance reordering. The syntax-based and dependency relation-based approaches have overcome the long-distance reordering, but these approaches require translation system to have a large bilingual corpus and a good parser. Chinese-Vietnamese is a low-resource language pair, and the Vietnamese parser is not really good. Therefore, in this paper, we only use the Chinese parser to identify words in a preposition phrase having word reordering. Not all prepositions labeled by the Chinese parser have word reordering when they are translated into Vietnamese, and we use the word alignment result from Chinese-Vietnamese small bilingual corpus to extract reordering prepositions.

Figure 3 shows the pre-ordered result of the Chinese sentence in Figure 1 by our model. In this sentence, the relationship between the preposition "给" and the verb "打" is a preposition phrase DR. The system will cluster the words that have the relationships with these words and preorder them. The phrase related to "给" is "给他", and the phrase related to the verb "打" is "打电话".

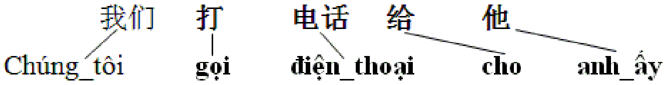


FIGURE 3. A Chinese sentence pre-ordered by our model

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

2. Preposition Phrase Pre-ordering.

2.1. Chinese preposition phrase dependency relation.

The dependency relation (also known as grammatical relation) is similar to Stanford's dependency tree [5]. There are 45 named grammatical relations and one default DR (dep DR). If a DR does not match to any type, it is assigned as *dep*.

The DR prep is a grammatical relation in these 45 relations, and it shows the relationship between preposition and main verb phrase which follows this preposition. For example, for the Chinese sentence "他在大学工作." (Vietnamese meaning: "Anh ta làm việc tại trường đại học", English meaning: "He works at the university"), the Stanford Parser tags *prep* DR (prep(工作-4, 在-2)) for the words "工作" (Vietnamese meaning: "làm việc", English meaning: "work") and "在" (Vietnamese meaning: "tại", English meaning: "at").

2.2. Extracting high reordering probability prepositions.

A prep DR is a reordering candidate if the reordering probability between the preposition and the verb in this DR is larger than a given threshold. Our system uses the word alignment result of the two words to determine whether they have the word reordering or not. To reduce the wrong alignments as well as to increase the quality of candidate reordering prepositions identification, we only use 1-1 or 1-n alignment that the words inside n must be continuous to count the number of prepositions and the number of reordering prepositions.

A *prep* DR is in the form $prep(w_1\text{-}fid, w_2\text{-}sid)$, in which, w_2 is a preposition and w_1 is a verb having dependency relation with the preposition w_2. The method to find the total of prepositions and the total of reordering prepositions is as follows:

- Conditions:
  (1) *fid* > *sid*.
  (2) *fid* and *sid* belong to 1-1 alignment or 1-n alignment which the words inside n are continuous together.
  (3) $Pos(fid)$ and $Pos(sid)$ are positions in the Vietnamese sentence aligned to *fid* and *sid*, correspondingly. If *fid* or *sid* is aligned with many Vietnamese continuous words, $Pos(fid)$ is the largest position of the Vietnamese words aligned with *fid* and $Pos(sid)$ is the smallest position of the Vietnamese words aligned with *sid*.

– Formulas:
  + The total of prepositions is total of prep DR prep containing this preposition and satisfying the conditions (1) and (2).
  + The total of reordering prepositions is total of prep DR prep containing the three conditions and satisfying constraints $Pos(fid) < Pos(sid)$.

Table 1 illustrates the number of prepositions in the corpus of 14,000 sentence pairs of CLC[1] which has reordering times greater than and equal to 30%.

TABLE 1. Candidate reordering prepositions

ID	Preposition	Total	Having reordering	Percentage
1	在	751	323	43.01%
2	给	273	125	45.79%
3	跟	138	96	69.57%
4	比	111	84	75.68%
5	和	57	25	43.86%
6	向	43	25	58.14%

2.3. The method of the preposition phrase pre-ordering.

– Assumptions:
  + Prep DR form: prep ***prep*** $(<\boldsymbol{W_i}>$-$<\boldsymbol{i}>, <\boldsymbol{W_j}>$-$<\boldsymbol{j}>)$
  + $< j >$ must be less than $< i >$.
  + The verb $< \boldsymbol{W_i} >$ and the preposition $< \boldsymbol{W_j} >$ must be included in a complete phrase. In this paper, we use the punctuation to determine a complete phrase.
  + *PREP_SET* is a set of Chinese prepositions which have reordering when they are translated into Vietnamese. In the experiment, the *PREP_SET* includes the prepositions in Table 1.
  + The phrase containing prep DR is like $X(W_j)Y(W_i)$, where $X(W_j)$ and $Y(W_i)$ are the phrases containing preposition W_j, the verb W_i and the words having dependency relation with them.

– The reordering method:

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

Figure 4 illustrates a prep DR based word reordering. For Chinese sentence "我给东京的朋友打电话" (I call my friend in Tokyo), with $X(W_j)$ being "给东京的朋友" ("cho người bạn ở Đông_Kinh", "my friend in Tokyo") and $Y(W_i)$ being "打电话" ("gọi điện_thoại", "call"), the two phrases will be reordered each other when they are translated into Vietnamese.

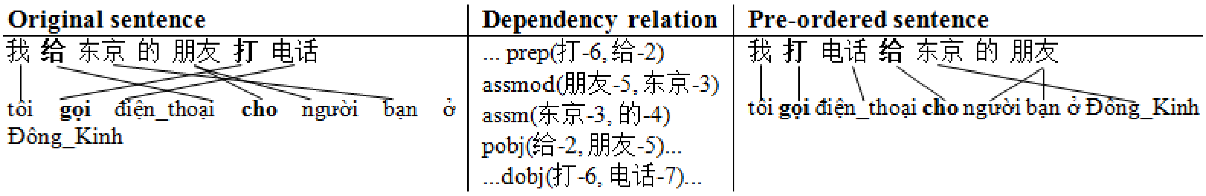


FIGURE 4. Illustrating a prep DR based word reordering

3. Experiments.

3.1. Toolkits in experiment.

We used the Stanford to segment words in Chinese corpus. As for Vietnamese, we used the CLC_VN_WS toolkit to segment words. In addition,

[1] Computational Linguistics Center (http://www.clc.hcmus.edu.vn/?page_id=467&lang=en).

we also use GIZA++ toolkit to align words, Chinese characters and Vietnamese spelling words. SRILM[2] toolkit is used to train the language model and the-state-of-the-art Moses[3] toolkit is used for phrase-based SMT.

3.2. **Experimental corpora.** Our experiment bilingual corpus consists of 35,623 Chinese-Vietnamese sentence pairs, which were extracted from Chinese conversational textbooks, online Chinese-Vietnamese forums and Chinese-Vietnamese bilingual websites. Documents in the corpus are mostly communicative text, so the length of the sentences is relatively short. We used 90% of the corpus for training, 5% for testing, and the remaining 5% for developing. We used these corpora to perform two experiments including WS translation (WS-Trans) and preposition pre-ordering translation (Prep-Trans).

– WS-Trans: The words in Chinese corpora and Vietnamese corpora are segmented, and these corpora are trained and tested by Moses toolkit.
– Prep-Trans: We make prep DR-based pre-ordering for the three corpora of P-SMT which are training corpus, developing corpus and testing corpus.

3.3. **Experimental result.** Table 2, Table 3 and Table 4 show the results of the case: in each 20 sentences, there are the first 18 sentences for training, the 19th sentence for developing and the last one for testing.

Table 2. Number of reordering types

	Number of all reordering types	**Monotone (M)**		**Swap-Discontinuous (S-D)**	
		Total	**%**	**Total**	**%**
WS-Trans	269,753	126,074	46.74%	143,679	53.26%
Prep-Trans	271,725	133,194	49.02%	138,531	50.98%

Table 3. Number of UKW of testing corpora

	Number of UKW	**Number of words**	**%**
WS-Trans	1,206	13,186	9.15%
Prep-Trans	1,198		9.08%

Table 4. The BLEU scores of the two translation systems

	BLEU
WS-Trans	35.16
Prep-Trans	35.80

3.4. **Analysis.** Based on the results in Table 2, Table 3 and Table 4, we found that the Prep-Trans has improved the quality of machine translation compared to the WS translation system. The improvement is indicated in three aspects: (1) a number of M reordering type of the Prep-Trans is more than that of the WS-Trans and a number of S-D types are reduced, (2) the BLEU score of the Prep-Trans is higher than that of the WS-Trans, and (3) the Prep-Trans has a less UKW than the WS-Trans. After pre-ordering words for Chinese corpus, the word order between Chinese sentence and Vietnamese sentence of the Prep-Trans is more similar than that of WS-Trans. This leads to that a number of swap-discontinuous reordering type of the Prep-Trans decreases and its number of monotone reordering type increases.

[2] Download at: http://www.speech.sri.com/projects/srilm/download.html
[3] Download at: http://www.statmt.org/moses/?n=Moses.Releases

Figure 5 shows an example about a change of the number of word order types of the Prep-Trans and WS translation systems in the training corpus. In the case of WS-Trans, this system has three M (up 50%) and three S-D (up 50%). Also in this sentence (English meaning is “I will give you some”), however, after pre-ordering words, we have 6 M (up 100%) and no S-D (up 0%).

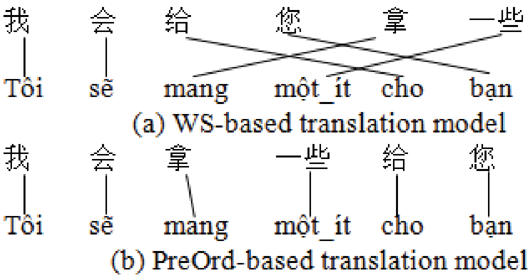


FIGURE 5. A change of reordering types between the Prep-Trans and WS-Trans

The quality of machine translation will be significantly improved if the texts in the source language and the target language are monotonous together [2-4]. The Prep-Trans has more Monotone ratio than the WS-Trans (Table 2), so the translation quality of the Prep-Trans is better than the WS one (reflected by BLEU score in Table 4). Figure 6 shows an improvement in the testing corpus of the Prep-Trans compared to the WS-Trans. Vietnamese and English meanings of this Chinese sentence are “có vấn_đề gì thì liên_hệ với chủ_nhà” and “Having any problem, you should contact the landlord”.

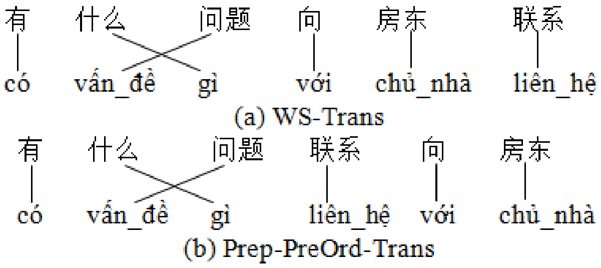


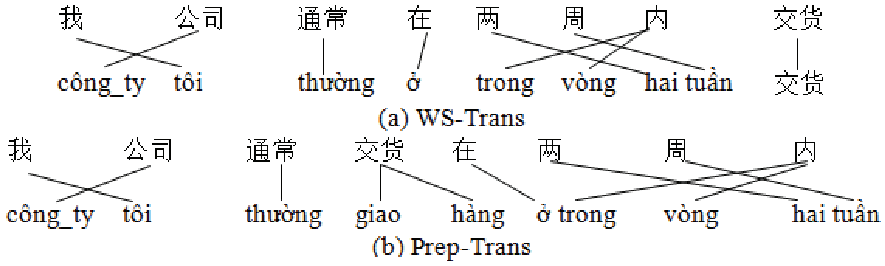
FIGURE 6. Indicating an improvement of translation quality of the Prep-Trans compared to the WS-Trans

In the case (a) of Figure 6, the WS-Trans has reordered the two words “什么” (what) and “问题” (problem) correctly because the words already exist in its phrase table. However, for the phrase “向房东联系” (contact the landlord), the WS-Trans reordered wrongly because the reordering rule is not learned. Therefore, instead of the correct translation being “liên_hệ với chủ_nhà” (contact the landlord), the WS-Trans translated the phrase to be “với chủ_nhà liên_hệ” (the landlord contact).

In the Prep-Trans, the Chinese sentence will be pre-ordered at the words level based on DR before translating it into Vietnamese. This sentence is parsed, including an important DR prep(联系-6, 向-4). Based on the word reordering rules of prep DR, the Prep-Trans will reorder the two phrases “向房东” and “联系” each other. After this step, the phrase “向房东联系” is converted into “联系向房东”. The new word order fits Vietnamese one, and its translation result is more accurate than that of the WS-Trans.

Another improvement of the Prep-Trans compared to the WS-Trans is that the Prep-Trans obtains less UKW than the WS-Trans. Because the corpus of the Prep-Trans contains more monotone order types than the corpus of the WS-Trans, the extracting phrase process of the Prep-Trans gains some useful phrases that the WS-Trans does not obtain. Figure 7 shows the less UKW result of the Prep-Trans compared to the WS-Trans. English meaning of this sentence is “My company usually delivers within two weeks”.

The phrase table of the WS-Trans does not contain the word “交货” although the training corpus includes the sentences containing this word. In the training corpus of the WS-Trans, we discover there are five sentences containing this word, but due to the sparse data and a different word order between Chinese and Vietnamese, the phrase table does not show the word. Therefore, the WS-Trans did not translate this WS.



(a) WS-Trans

(b) Prep-Trans

Figure 7. Illustrating the less UKW generation of the Prep-Trans compared to the WS-Trans

Also in this sentence, the Prep-Trans has ten phrases containing the word “交货”, especially, its phrase table includes the two phrases “交货” (giao hàng) and “通常交货” (thường giao hàng). So this translation system translated the word “交货” and does not generate UKW. Also, before translating, this sentence is pre-ordered based on the DR prep(交货-8, 在-4), the phrase “在两周内交货” is reordered into “交货在两周内”, and the translation result of this sentence is suitable to Vietnamese order.

4. **Conclusions and Perspectives.** In this paper, we have improved the quality of Chinese-Vietnamese machine translation based on pre-ordering Chinese word to be suitable to Vietnamese word order before the training and translation stages. The preordering is done based on two factors: Chinese-Vietnamese word alignment result and Chinese dependency relation. The preordering has improved the P-SMT performance. This improvement is demonstrated by three aspects: (1) the monotone order type of the Prep-Trans is more than that of the WS-Trans and the swap-discontinuous order type of the Prep-Trans is less than the WS-Trans, (2) the BLEU score of the Prep-Trans is higher than that of the WS-Trans and (3) the Prep-Trans generates less UKW than WS-Trans.

One point of concern is that performance of P-SMT will be affected, even may be reduced, if the quality of the parse toolkit is not good. Our system uses a popular Chinese parser at the present time, namely, Stanford parser. However, based on experiments, we found that the parser is still wrong in some cases in which the two phrases in prep DR do not have the reordering when they are translated into Vietnamese. Therefore, to increase the quality of pre-ordering, our system has adopted some linguistic constraints on the parsing results of Stanford parser. The aim consists in limiting parsing errors and increasing the performance of pre-ordering.

In the future, we plan to parse dependency relation in Vietnamese side, and then we combine it with the result of the Chinese side to improve the quality of Chinese preordering in Chinese-Vietnamese machine translation.

REFERENCES

[1] 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.

[2] M. R. Costa-jussa and J. A. R. Fonollosa, Statistical machine reordering, *Proc. of the 2006 Conference on Empirical Methods in Natural Language Processing*, pp.70-76, 2006.

[3] M. Khalilov and K. Sima’an, Source reordering using MaxEnt classifiers and supertags, *Proc. of the 14th Annual Conference of the European Association for Machine Translation*, pp.292-299, 2010.

[4] C. Wang, M. Collins and P. Koehn, Chinese syntactic reordering for statistical machine translation, *Proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp.737-745, 2007.

[5] P.-C. Chang, H. Tseng, D. Jurafsky and C. D. Manning, Discriminative reordering with Chinese grammatical relations features, *Proc. of SSST-3, the 3rd Workshop on Syntax and Structure in Statistical Translation*, pp.51-59, 2009.

[6] 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.

[7] 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 MetricsMATR*, pp.235-243, 2010.

[8] 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.

[9] S. Stymne, Clustered word classes for preordering in statistical machine translation, *Proc. of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp.28-34, 2012.

[10] 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.

[11] 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.

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

[13] G. Wu, Y. Zhang and A. Waibel, Rule-based preordering on multiple syntactic levels in statistical machine translation, *Proc. of the 11th International Workshop on Spoken Language Translation*, 2014.

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

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

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

[17] 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.

[18] 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.

[19] 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.

[20] 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.