

PAPER

Preordering for Chinese-Vietnamese Statistical Machine Translation

Huu-Anh TRAN^{†a)}, Member, Heyan HUANG^{†b)}, Phuoc TRAN^{††c)}, Shumin SHI^{†d)},
and Huu NGUYEN^{†††e)}, Nonmembers

SUMMARY Word order is one of the most significant differences between the Chinese and Vietnamese. In the phrase-based statistical machine translation, the reordering model will learn reordering rules from bilingual corpora. If the bilingual corpora are large and good enough, the reordering rules are exact and coverable. However, Chinese-Vietnamese is a low-resource language pair, the extraction of reordering rules is limited. This leads to the quality of reordering in Chinese-Vietnamese machine translation is not high. In this paper, we have combined Chinese dependency relation and Chinese-Vietnamese word alignment results in order to pre-order Chinese word order to be suitable to Vietnamese one. The experimental results show that our methodology has improved the machine translation performance compared to the translation system using only the reordering models of phrase-based statistical machine translation.

key words: Chinese-Vietnamese machine translation, preordering, word alignment, Chinese grammatical relations, dependency relations

1. Introduction

Word order is one of the hardest problems in machine translation [1]. On the other hand, for different language pairs, the reordering method is also different. However, in general, the word reordering is centered on two branches: (1) for the language pairs having short distance (such as English-French), and (2) for the language pairs having a long distance (such as English-Japanese). In the case of (1), if the bilingual corpora for machine translation are large enough, the distance-based reordering model (DRM) or the lexicalized-based reordering model (LRM) [1] of the phrase-based statistical machine translation (P-SMT) can overcome. In the case of (2), P-SMT cannot process successfully.

Chinese and Vietnamese have the same isolated

language type, have close linguistic relationships [2]. Some linguistic relationships are an uninflected word, the main grammatical forms are word order and function word, spaces do not determine word boundary, using the function words to denote negative or question. In addition to these similarities, Chinese and Vietnamese have a big difference in word order. The difference is expressed in some aspects, including preposition phrase order, noun phrase order, “的” word order, and some specific structures. Figure 1 illustrates the difference between the preposition phrase order between the Chinese and Vietnamese.

On the other hand, because Chinese-Vietnamese is a low-resource language pair, the performance of word reordering of the translation system is not high. Both DRM and LRM models are not enough for the system to reorder correctly. This led to a case that the system mistranslates the word order in some sentences whether these sentences have the same structure as the sentences existing in the training corpus. For example, in Fig. 2, the system mistranslated the word order of the case (a) and it translated the case (b) correctly. These two instances have the same preposition phrase structure, in the form: <preposition> + <noun phrase> + <verb> (<跟> + <女朋友> + <约会> (case (b))), and they are translated into Vietnamese in the form: <verb> + <preposition> + <noun phrase> (<约会> + <跟> + <女朋友> (case (b))). However, due to the training corpus covered the case (b), the word order of this sentence is translated correctly. On the contrary, the system

我 给 他 打 电话
tôi gọi điện_thoại cho anh_ta

Fig. 1 Illustration of the difference of preposition phrase order between Chinese and Vietnamese.

我只 有一个 愿望 就是 跟 她 结婚
Tôi chỉ có một nguyên_vọng là với cô_ấy kết_hôn
(a)
是 跟 女朋友 约会 吗 ?
Có hẹn với bạn_gái phải không ?
(b)

Fig. 2 An example of translation of two sentences having the same structure of P-SMT. (a) shows the incorrect translation case, and (b) illustrates the correct translation case.

Manuscript received June 13, 2018.

Manuscript revised September 19, 2018.

Manuscript publicized November 12, 2018.

[†]The authors are with Department of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China.

^{††}The author is with the NLP-KD Lab, Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Vietnam.

^{†††}The author is with the Faculty Information Technology, Ho Chi Minh City University of Food Industry, Ho Chi Minh City, Vietnam.

a) E-mail: anhuni1006@gmail.com

b) E-mail: hhy63@bit.edu.cn (Corresponding author)

c) E-mail: tranthanhphuoc@tdtu.edu.vn (Corresponding author)

d) E-mail: bjssm@bit.edu.cn

e) E-mail: huunt@cntp.edu.vn

DOI: 10.1587/transinf.2018EDP7211

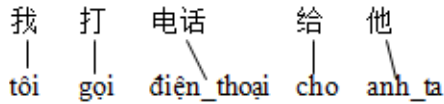


Fig. 3 A Chinese sentence is preordered by our model.

cannot translate the case (a) successfully (the correct order is that “tôi chỉ có một nguyện vọng là kết hôn với cô ấy”). English meaning of the case (a) is “I only have one wish is to marry her” and one of the case (b) is “Do you have an appointment with a girlfriend?”.

In this paper, we propose a pre-ordering model to make word order of Chinese sentence in accordance with Vietnamese word order. Based on Chinese-Vietnamese word alignment result and Chinese dependency relation (DR), the system extracts DRs having word reordering when translating them into Vietnamese. Then, the system will build reordering rules for these DR types. Finally, these rules will be applied to pre-order the Chinese corpora before training and translation processes of P-SMT. Figure 3 shows the pre-ordered result of the Chinese sentence in Fig. 1 by our model.

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

2. Related Works

Reordering model is one of the three primary models of P-SMT. The DRM and LRM have been used in P-SMT (such as Moses). However, both models are problematic when they are applied in language pairs with substantial differences in word order. In this section, we will present some other methods to reorder in machine translation, mainly in two directions, namely pre-ordering and post-ordering.

2.1 Pre-Ordering

This is the most common approach to reorder word in machine translation. The goal of this approach is to pre-process the order of the source sentence to be suitable to the target sentence before training and translating of P-SMT. Until the present time, three methods are using the pre-ordering approach, including pre-ordering based on the part of speech (POS) information, based on syntax information, and based on dependency relation.

POS is used quite commonly to pre-order a source language. In [14], Goh et al. proposed separating a long sentence into multiple phrases by using linguistic information, and translating these phrases. Also, the authors also limited the reordering scope so that it could not cross the boundary of the phrase. The authors used POS and punctuations as markers to determine a phrase location.

Also using POS information to pre-order, Li et al. [4] identified two phrases, such as verb phrase and preposition

phrase, and pre-ordered them in the Chinese language to be close to the order in Korean side. In [12], the authors focused on chunk-based verb reordering in Arabic-English machine translation. In [15], Stymne et al. developed a rule set of reordering from word alignment result of a bilingual corpus in which the source language was labeled POS tags. German-English and Haitian Creole-English were tested. For the low-resource Haitian Creole, the authors combined POS information and other labels from the parser to extract the pre-ordering rules.

In [3], the authors combined morphology and syntax information to pre-order for English-Vietnamese machine translation as well as limiting the sparse data. The authors used manual reordering rules for English morphological analysis to limit the sparse data. For syntactic transformation, the authors proposed a Bayesian-based transformation model to adjust the source language order according to the structure of the target language.

In the same approach to reorder based syntax, Badr et al. [6] developed reordering rules for translating from English to Arabic. These rules are built for four tags, including NP (noun phrase), PP (preposition phrase), THE (the article), and VP (verb phrase). Although Habash [5] warned that syntax-based reordering would not improve the quality of machine translation if the quality of the parser is not good enough. However, because this work was built for an English-Arabic language pair with English as a reordered language, the quality of the English parser is reliable and the pre-ordering quality is good.

Machine translation quality will be significantly improved if the text in the source language and target language monotonizes [7]–[9]. To do this, in [10], the authors pre-ordered the source language based on the tree to tree conversion and applied to English-Dutch language pair. The two main tasks performed in this work are (1) creating a bilingual corpus “source-reordered source” ($s \rightarrow s'$) and (2) “reordered-target” ($s' \rightarrow t$). The $s \rightarrow s'$ corpus was used to train the reordering model for the source sentence in line with the target one. The $s' \rightarrow t$ bilingual corpus was devoted to P-SMT training.

In [16], the authors combined the advantages of the lexicon-based reordering model and the syntax-based reordering models. The advantage of the lexicon-based reordering model is local reordering (short-distance), and the syntax-based reordering models are more efficient for long distance reordering.

The work [18] also pre-ordered based on the syntactical tree and the word alignment result of the bilingual corpus. The approach of this work is to examine and discover the reordering rules on multiple nodes of the syntax tree.

The dependency relation-based pre-ordering approach has been successfully applied to some languages. In [13], Jiang et al. used a DR to pre-order the phrases related to the Chinese function words. This paper extracted reordering patterns in relation to the three function words “的” (DE), “把” (BA), and “被” (BEI). The reordering patterns are also extracted based on the syntactic parsing and word alignment

results. Genzal [11] also used a dependency parser to analyze the source sentence. The corpus with the reordered source sentences was used for training and testing in P-SMT. This work was applied to many language pairs and gave encouraging results. As for English-Vietnamese SMT, Viet et al. [26] applied this approach to pre-order English word order to be suitable to Vietnamese one.

The same approach to pre-ordering Chinese based on DR, Cai et al. [17] pre-ordered Chinese sentence based on 45 DR tags to match the English word order. The work defined two types of reordering rules, such as “ $x : y$ ” (1) and “ $x - y$ ” (2), where x and y are two DRs. X and Y are two dependent structures corresponding to x and y . The dependency structure consists of the dependency words and the subtree of this dependency. The $X \setminus Y$ is the structure including X except for Y . For the structure (1), Y is sub-structure of X , the reordering rule is to convert $X \setminus Y$ to in front of Y . For the structure 2, X and Y are sibling, the reordering rule is move X behind Y . The rule extraction method is as follows:

- Search DRs that match the two forms.
- Remove the structures that occur less than 5,000 times and irrespective DRs such as root, conj, cc, and so on.
- For the remaining structures: select some sample dependency analysis trees, reconstruct the analysis trees in one of the two forms and evaluate reordered phrases. If this new order is close to the English order, this structure will be candidate reordering rule.
- Do SMT experiment and rely on BLEU score to eliminate ineffective rules.

As a result, the system extracted eight pre-ordering rules belonging to three DRs. Those are *plmod* (localizer modifier of a preposition), *rcmod* (relative clause modifier), and *prep* (preposition modifier). This is the most related work to our work. We also use DR to pre-order the Chinese sentence. However, our study differs from this work in several respects:

- The pre-ordering approach depends on target language, corresponding to each target language, the reordering rule of the source language will be different. Because the target language of our work is Vietnamese, our candidate DRs, as well as pre-ordering rules will not be the same for this work.
- There is no Chinese DR whose two phrases are entirely reordered when translating into Vietnamese. There are two reasons for this case, including (1) the dependency parser analyzed incorrectly, and (2) the two phrases in the DR itself are not reordered when translating them into Vietnamese. Therefore, our work will add some constraints to the DR to increase the accuracy of pre-ordering.

Recently, the problem of increasing the computation performance of pre-ordering is also of interest. Typical works for this approach are [19] and [20]. In [19], Nakagawa et al. used the Bracketing transduction grammar to pre-order. Meanwhile, in [20], Gispert et al. used neural

networks to speed up the execution and accuracy of pre-order in SMT. Both of these works were used in many language pairs and gave satisfactory results.

2.2 Post-Ordering

Researchers are also investing in this approach in recent times. Contrary to the pretreatment approach, which is reordered before translation, this approach focuses on processing word order after translation.

Typical work for this approach is [21] which was done by Sudoh et al. Instead of pre-ordering the source sentence before translation, here, the authors proposed to process word order after translation. Japanese-English machine translation is done in this work. This approach is implemented in two steps, including (1) translating the Japanese sentence into HFE (Head-Final English) using the Japanese-HFE model, and (2) translating the HFE into English using the HFE-English translation model. HFE corpora are generated based on English pre-ordering model in the bilingual corpus [22].

The same post-ordering approach, Goto et al. [23] used a parsing model to translate from HFE to English. First, the HFE sentences will be analyzed into binary trees, and then the nodes with the suffix “_SW” will be permuted to create the final English sentence.

3. Building Pre-Ordering Rules

3.1 Chinese Dependency Relation

Chinese dependency relation (DR), also called grammatical relation is similar to Stanford’s dependency tree [24]. There are a total of 45 named grammatical relations and one default DR. If a DR does not match any pattern, it is assigned as *dep* DR. A DR will represent a dependency relation between two words in a sentence. For example, DR *nn* (大学-8, 北京-7), this DR depicts the dependence between the two nouns 大学 (“đại học”: university) and 北京 (“Bắc Kinh”: Beijing) in the Chinese sentence. Numbers 8 and 7 following these two nouns are their positions in the sentence.

3.2 Candidate DR Extraction

We extract the candidate DRs as follows: Using word alignment result and DR tagging result to extract the candidate DRs. A DR is a pre-ordering candidate if the probability of the two words in this DR is higher than a given threshold. In practice, we chose the threshold to be 25%.

First, our system calculates the total of DRs and the total of DRs having reordering. The words in these DRs must have 1-1 or continuous 1-n alignment. Our system used Stanford Parser tool to label DR. A DR has the following form: <DR name> (<word 1> - <fid>, <word 2> - <sid>). Where, fid, sid, in turn, is the position of words <word 1> and <word 2> in <DR name>. <DR name> is the name of a dependency relation, such as *nn*, *prep*, etc. DR *nn* (身体-5,

Table 1 Candidate DRs

ID	DR name	NR-DR	N-DR	Percentage
1	nn	1,276	2,319	0.55
2	assm	1,004	1,341	0.75
3	assmod	990	1,325	0.75
4	det	990	1,452	0.68
5	lobj	293	469	0.62
6	lccomp	67	121	0.55
7	plmod	92	264	0.35
8	ordmod	63	76	0.83
9	prep	634	1,965	0.32
10	amod	309	494	0.63
11	rcmod	219	434	0.50
12	cpm	202	480	0.42
13	clf	1,070	1,798	0.60

妈妈-4) is an example, in which, 身体 (“sức khỏe”: health) is a <word 1>, 5 is a fid; 妈妈 (“mẹ”: mother) is a <word 2>, and 4 is a sid.

The method for finding the total of DRs and the total of DRs having reordering is as follows:

- Conditions:
 - + (1) fid > sid.
 - + (2) fid and sid belong to 1-1 alignments or continuous 1-n alignments.
 - + (3) Pos (fid) and Pos (sid) are positions in the Vietnamese sentence aligned to fid and sid, respectively. If fid or sid belong to the continuous 1-n alignments, Pos (fid) is the smallest position of the Vietnamese words aligned with fid and Pos (sid) is the largest position of the Vietnamese words aligned with sid.
- Formulas:
 - + The total of the DR is the total of the DR satisfying the condition (1) and (2).
 - + The total of reordering DR is total of the DR containing the three conditions and satisfying constraints Pos (fid) < Pos (sid).

Table 1 shows a statistic of the number of DR having reordering from 30% or more in our 14k corpus. In which, NR-DR is a number of Reordered DR, N-DR is a number of DR.

In Table 1, there are 13 DRs having reordering probability from 25% or more. Also from Table 1, we found that there is no DR that the reordering rate of the two words in this DR reaches a maximum of 100%. The reason is that the word segmentation and DR tagging are wrong. Moreover, some DRs have both reordering and non-reordering. The following sections will show the pre-ordering rules as well as some of their exceptions.

3.3 Building Pre-Ordering Rules

A DR has the form: <DR name> (<Wi>-<i>, <Wj>-<j>). The general steps of DR-based pre-ordering are as follows:

- Step 1: Defining the phrase X (Wi) containing the word

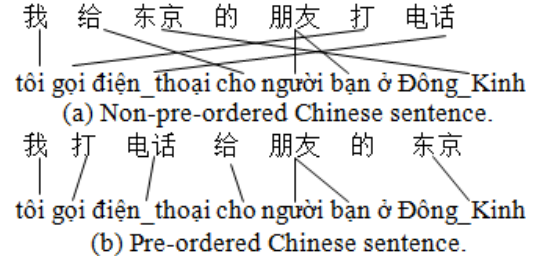


Fig. 4 An example of Chinese sentences before and after being pre-ordered.

Wi and the phrase Y (Wj) containing the word Wj. The phrase X (Wi) consists of the word Wi and the words having a dependency relation with it. Based on the 45 DRs, we easily extract the phrases containing the word Wi. The same way is for the phrase Y (Wj).

- Step 2: Reordering X (Wi) and Y (Wj).

For example, the Chinese sentence “我给东京的朋友打电话” (tôi gọi điện_thoại cho người bạn ở Đông_Kinh: I called my friend in Tokyo) has a different order compared to the Vietnamese one (Fig. 4.a). This sentence is parsed into some DRs, including prep (打-6, 给-2), assmod (朋友-5, 东京-3), assm (东京-3, 的-4), pobj (给-2, 朋友-5), dobj (打-6, 电话-7). There are two DRs having reordering in these DRs, which are assmod and prep DRs. This Chinese sentence is pre-ordered as follows:

- Prep-based pre-ordering: This DR consists of two words, including Wi 打 (“gọi”: call) and Wj 给 (“cho”: for). The phrases X (打) = “打电话” (“gọi điện_thoại”: call) and Y (给) = “给东京的朋友” (“cho người_bạn ở Đông_Kinh”: friend in Tokyo).
- Reordering the two phrases X (打) and Y (给): After this reordering, the Chinese sentence is “我打电话给东京的朋友”.
- assmod-based pre-ordering: This DR consists of two words 东京 (“Đông_Kinh”: Tokyo) and 朋友 (“người_bạn”: friend). The phrase X (东京) = “东京” and Y (朋友) = “朋友”.
- Reordering the two phrases X (东京) and Y (朋友). The Chinese sentence has the order “我打电话给朋友的东京”, this order is also its order when translating it into Vietnamese (Fig. 4.b).

4. Experiments

4.1 Toolkits in the Experiments

We used the Stanford Segmenter[†] to segment word in Chinese corpus. As for Vietnamese, we used the CLC_VN_WS toolkit^{††} to segment word. In addition, we also use GIZA++ toolkit to align word. SRILM toolkit^{†††} is

[†]<https://nlp.stanford.edu/software/segmenter.html>

^{††}http://www.clc.hcmus.edu.vn/?page_id=471&lang=en

^{†††}Download at: <http://www.speech.sri.com/projects/srilm/download.html>

Table 2 Illustration of number of words as well as number of words per a sentence.

	Cn Training	Cn Developing	Cn Testing
NoW	236,598	13,071	13,186
NoS	32,061	1,781	1,781
NoW/NoS	7.4	7.3	7.4
	Vn Training	Vn Developing	Vn Testing
NoW	296,256	16,328	16,509
NoS	32,061	1,781	1,781
NoW/NoS	9.2	9.2	9.3

used to train the language model, and state-of-the-art Moses toolkit[†] is used for P-SMT.

4.2 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 sentences for training, 5% of the sentences for testing, and the remaining 5% of the sentences for developing. We used these corpora to perform two experiments including WS translation (WS-Trans) and Pre-ordering translation (PreOrd-Trans).

- WS-Trans: The words in Chinese corpora and Vietnamese corpora are segmented, and these corpora are aligned by the GIZA++ toolkit.
- PreOrd-Trans: We pre-order three corpora of the P-SMT, which are training corpus, developing corpus, and testing corpus.

Table 2 shows the number of words (NoW), number of sentences (NoS) and number of words per sentence (NoW/NoS) in the experimental corpora of the two translation systems. These corpora are divided as follows: In every 20 sentences, there are the first 18 sentences for training, the 19th sentence for developing and the last one for testing.

4.3 Evaluation

We evaluated the impact of the PreOrd-Trans on translation result based on the following criteria:

- A number of reordering type, including non-reordering (monotone: M) and reordering (swap and discontinuous: S-D).
- Number of unknown word (UKW).
- BLEU scores.

4.3.1 A number of Reordering Types

For example, on line 4 of Fig. 5, we have five Monotone types (including 1-0 and 1-1, 1-1 and 2-2, 5-5 and 6-6, 6-6

1	0-0	1-1	3-2	2-3	4-4	5-5	6-6	7-6	8-7					
2	0-0	1-0	2-1	3-2	6-3	5-4	4-5	7-6						
3	0-0	0-1	1-2	2-3	3-4	4-5	4-6	7-7	5-8	6-9	6-10	8-11		
4	0-0	1-0	1-1	2-2	4-3	3-4	5-5	6-6	7-7	8-8				
5	0-0	1-1	2-2	3-3	4-4	5-5	6-6	7-7						
6	1-0	0-1	2-2	3-3	4-4	5-5								

Fig. 5 Example of alignment file.**Table 3** Number of reordering types.

	WS-Trans	PreOrd-Trans
Number of all reordering types	269,753	268,467
Number of M	126,074	150,415
Percentage of M	46.74%	56.03%
Number of S-D	143,679	118,052
Percentage of S-D	53.26%	43.97%

Table 4 Number of UKW of testing corpora.

	N-UKW	NW	Percentage
WS-Trans	1,206	13,186	9.15%
PreOrd-Trans	1,153	13,186	8.74%

Table 5 BLEU scores.

Translation systems	WS-Trans	PreOrd-Trans
BLEU score	35.83	36.04

and 7-7, 7-7 and 8-8), three swap – discontinuous types (2-2 and 4-3, 4-3 and 3-4, 3-4 and 5-5).

4.3.2 Number of UKW and BLEU Score

Number of UKW is the number of Chinese tokens in the Vietnamese translation. After improving word alignment, some word alignment will be removed and some are added. The generation of more or less UKW does not indicate the performance of SMT (illustrated through BLEU scores), but it represents the vocabulary coverage of the SMT's phrase table. The BLEU score is calculated based on the work of Papineni K. A. [25].

4.4 Experimental Results

Tables 3–5 show the experimental results. In which, N-UKW is a number of UKW, NW is a number of words.

4.5 Analysis

Based on the experimental results in Table 3, Table 4 and Table 5, we found that the PreOrd-Trans has improved translation quality compared to the WS-Trans. The improvement is expressed in three aspects, such as (1) the number of Monotone reordering of the PreOrd-Trans is greater the WS-Trans' one, the number of Swap-Discontinuous reordering is less than the WS-Trans' one, (2) the PreOrd-Trans has higher BLEU score, and (3) the PreOrd-Trans has fewer UKW. After pre-ordering for Chinese corpus, the word order of the Chinese sentence of the PreOrd-Trans is more similar to the word order of the Vietnamese sentence than

[†]Download at: <http://www.statmt.org/moses/?n=Moses.Releases>

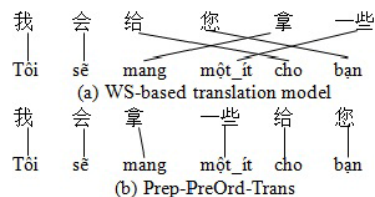


Fig. 6 Illustrating of changing in the number of reordering types between the WS-Trans and PreOrd-Trans.

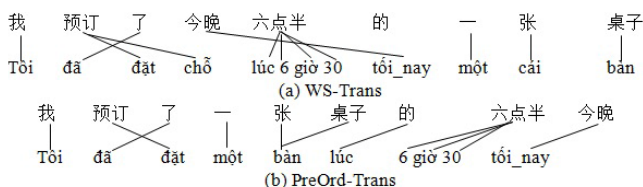


Fig. 7 An example of translation quality improvement of the PreOrd-Trans compared to the WS-Trans.

the WS-Trans. This led that the number of Swap - Discontinuous reordering decreased and of course the number of Monotone increased.

Figure 6 shows a case of changing in the number of reordering types of the WS-Trans and PreOrd-Trans in the training corpus. In the WS-Trans, the system has three M types (50%) and three S-D types (50%). Also in this sentence, after pre-ordering, the PreOrd-Trans has six M types (100%) and no S-D type (0%). English meaning of the Chinese sentence in Fig. 6 is “I will give you some”.

The quality of machine translation will be significantly improved if the text in the source language and target language monotonizes [7]–[9]. Because PreOrd-Trans has a higher Monotone ratio than WS-Trans (Table 3), the quality of the translation of the PreOrd-Trans is better (shown in BLEU score in Table 5). Figure 7 shows an improvement case of the PreOrd-Trans compared to the WS-Trans in the testing corpus. English meaning of the Chinese sentence in Fig. 7 is “I booked a table at half past six tonight”.

In the case (a) of Fig. 7, the WS-Trans correctly reordered the two phrases “预订了”/“đã đặt chỗ” (booked) and “今晚六点半”/“lúc 6 giờ 30 tối nay” (at half past six tonight) because the two phrases existed in the phrase table of the WS-Trans. However, for the phrase “今晚六点半的一张桌子” (“một bàn lúc 6 giờ 30 tối nay”), the WS-Trans could not properly reorder because this phrase has not been learned (does not exist in the phrase table). The WS-Trans only correctly reorder for the phrase preceding “的” word (今晚六点半 (“6 giờ 30 tối nay”: at half past six tonight)), but it does not correctly reorder the two phrases which are preceding and following “的” word (今晚六点半: “6 giờ 30 tối nay”: “at half past six tonight” and 一张桌子: “một cái bàn”: “a table”). As a result, instead of correct translation is “đặt một bàn lúc 6 giờ 30 tối nay” (booked a table at half past six tonight), the WS-Trans incorrectly translated it into “lúc 6 giờ 30 tối nay một cái bàn” (at half past six tonight a table).

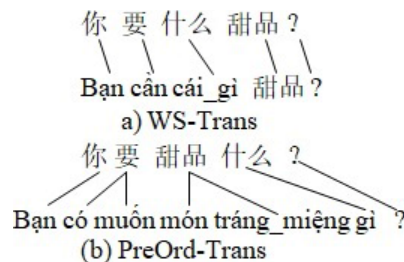


Fig. 8 An example of the PreOrd-Trans has less UKW than the WS-Trans.

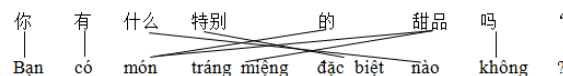


Fig. 9 Word alignment result of the Chinese sentence containing the word “甜品” in the training corpus of the WS-Trans.

In the PreOrd-Trans, this Chinese sentence will be pre-ordered based on DRs before translation. This sentence is parsed into some important DRs, such as nn (六点半-4, 今晚-5), assmod (桌子-6, 六点半-8), nummod (张-8, 一-7), clf (桌子-9, 张-8). Based on the assmod DR, the PreOrd-Trans will pre-order the two phrases preceding and following the “的” word. The phrase “今晚六点半的一张桌子” is converted to “一张桌子的今晚六点半”. Next, the PreOrd-Trans continues to pre-order for the sub-phrase “今晚六点半” based on nn DR (converted to “六点半今晚”). After this step, the phrase “今晚六点半的一张桌子” is pre-ordered to “一张桌子的六点半今晚”. This new word order only has a Monotone type, in line with the Vietnamese order, and the translation result is more accurate than the WS-Trans’ one.

Another improvement of PreOrd-Trans compared to the WS-Trans is that the PreOrd-Trans is less UKW. Because the PreOrd-Trans has more Monotone types than the WS-Trans, the phrase table of the PreOrd-Trans has some useful phrases that the WS-Trans does not obtain. Figure 8 shows the less UKW translation result of the PreOrd-Trans compared to the WS-Trans’ one. In addition, this example also shows that the PreOrd-Trans translated more correctly than the WS-Trans did. English meaning of the Chinese sentence in Fig. 8 is “What dessert do you want?”.

The phrase table in the WS-Trans does not contain the word “甜品”/“món tráng miệng” (dessert) although the training corpus has one sentence containing this word. We found that ten phrases are containing this word, but no phrase contain only “甜品”. Therefore, the WS-Trans did not translate this word. In addition, the WS-Trans translated wrong order of the phrase “什么甜品”. This phrase must be pre-ordered to “甜品什么” before translation. The Chinese sentence containing the word “甜品”, the word alignment result, and the phrase table in relation to the word “甜品” of the WS-Trans are illustrated in Fig. 9 and Fig. 10.

In this case, the fourteen phrases are containing the word “甜品” in the phrase table of the PreOrd-Trans and “甜品” is one of the fourteen phrases. Therefore, the PreOrd-Trans has correctly translated the word “甜品” without

的甜品 || món tráng_miệng
 特别的甜品 || món tráng_miệng đặc_biệt
 什么特别的甜品 || món tráng_miệng đặc_biệt nào
 有什么特别的甜品吗 || món tráng_miệng đặc_biệt nào không
 有什么特别的甜品吗 ? || món tráng_miệng đặc_biệt nào không ?
 有什么特别的甜品 || có món tráng_miệng đặc_biệt nào
 有什么特别的甜品吗 || có món tráng_miệng đặc_biệt nào không
 有什么特别的甜品吗 ? || có món tráng_miệng đặc_biệt nào không ?
 你有什么特别的甜品 || Bạn có món tráng_miệng đặc_biệt nào
 你有什么特别的甜品吗 || Bạn có món tráng_miệng đặc_biệt nào không

Fig. 10 Phrase extraction result containing the word “甜品” in the training corpus of the WS-Trans.

你 有 甜品 的 特别 什么 吗 ?
 Bạn có món tráng_miệng đặc_biệt nào không ?

Fig. 11 Word alignment result of the Chinese sentence containing the word “甜品” in the training corpus of the PreOrd-Trans.

甜品 || món tráng_miệng
 有甜品 || có món tráng_miệng
 你有甜品 || Bạn có món tráng_miệng
 你有甜品的特别 || Bạn có món tráng_miệng đặc_biệt
 你有甜品的特别什么 || Bạn có món tráng_miệng đặc_biệt nào
 你有甜品的特别什么吗 || Bạn có món tráng_miệng đặc_biệt nào không
 有甜品的特别 || có món tráng_miệng đặc_biệt
 有甜品的特别什么 || có món tráng_miệng đặc_biệt nào
 有甜品的特别什么吗 || có món tráng_miệng đặc_biệt nào không
 有甜品的特别什么吗 ? || có món tráng_miệng đặc_biệt nào không ?
 甜品的特别 || món tráng_miệng đặc_biệt
 甜品的特别什么 || món tráng_miệng đặc_biệt nào
 甜品的特别什么吗 || món tráng_miệng đặc_biệt nào không
 甜品的特别什么吗 ? || món tráng_miệng đặc_biệt nào không ?

Fig. 12 Phrase extraction result containing the word “甜品” in the training corpus of the PreOrd-Trans.

generating UKW. Also, before translation, this sentence is pre-ordered based on det DR (det(甜品-3, 什么-4)), the phrase “什么甜品” is converted to “甜品什么”, the translation result of this phrase is correct to the order of the Vietnamese side. The Chinese sentence containing the word “甜品”, the word alignment result, and the phrase table in relation to the word “甜品” of the PreOrd-Trans are illustrated in Fig. 11 and Fig. 12.

In addition to these improvements, our pre-ordering system will not work well for Chinese sentences having complex structures and the parser analyzes them incorrectly. Our system pre-orders Chinese sentence 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 translation quality of Chinese-Vietnamese P-SMT based on pre-ordering Chinese to be suitable to Vietnamese order before training and translation. The pre-ordering is made based on two factors, such as word alignment result and Chinese dependence relation. The result of pre-ordering has improved P-SMT perfor-

mance, expressed in three aspects: (1) the PreOrd-Trans has more Monotone type and less Swap-Discontinuous types than the WS-Trans, (2) the PreOrd-Trans' BLEU score is higher than the WS-Trans' one, and (3) the PreOrd-Trans has less UKW than the WS-Trans.

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

In the future, we intend to parse the dependency in the Vietnamese side, then combine with the dependent result in the Chinese side to further improve the final pre-ordering in Chinese-Vietnamese machine translation.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grand No. 61732005, 61671064) and the National Key Research and Development Program of China (Grand No. 2017YFB1002103).

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, IWSLT 2005, pp.68–75, 2005.
- [2] P. Tran, D. Dinh, T. Lê, and L.H.B. Nguyen, “Linguistic-Relationships-Based Approach for Improving Word Alignment,” ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) (DOI: 10.1145/3133323), vol.17, no.1, pp.1–16, 2017.
- [3] T.P. Nguyen and A. Shimazu, “Improving Phrased-based SMT with morpho-syntactic analysis and transformation,” Proc. 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. Fourth Workshop on Statistical Machine Translation, pp.190–196, 2009.
- [5] N. Habash, “Syntactic Preprocessing for Statistical Machine Translation,” Proc. Machine Translation Summit (MT-Summit), 2007.
- [6] I. Badr, R. Zbib, and J. Glass, “Syntactic Phrase Reordering for English-to-Arabic Statistical Machine Translation,” Proc. 12th Conference of the European Chapter of the ACL, pp.86–93, 2009.
- [7] M.R. Costa-juss'a and J.A.R. Fonollosa, “Statistical machine reordering,” Proc. HLT/EMNLP'06, pp.70–76, 2006.
- [8] M. Khalilov and K. Sima'an, “Source reordering using maxent classifiers and supertags,” Proc. EAMT'10, pp.292–299, 2010.
- [9] C. Wang, M. Collins, and P. Koehn, “Chinese syntactic reordering for statistical machine translation,” Proc. EMNLP-CoNLL'07, pp.737–745, 2007.
- [10] M. Khalilov and K. Sima'an, “A Discriminative Syntactic Model for Source Permutation Via Tree Transduction,” Proc. SSST-4, Fourth Workshop on Syntax and Structure in Statistical Translation, pp.92–100, COLING 2010.
- [11] D. Genzel, “Automatically Learning Source-side Reordering Rules for Large Scale Machine Translation,” Proc. 23rd International Conference on Computational Linguistics (Coling 2010), pp.376–384, 2010.
- [12] A. Bisazza and M. Federico, “Chunk-based Verb Reordering in

- VSO Sentences for Arabic-English Statistical Machine Translation,” Proc. Joint 5th Workshop on Statistical Machine Translation and MetricsMATR, pp.235–243, 2010.
- [13] J. Jiang, J. Du, and A. Way, “Source-side Syntactic Reordering Patterns with Functional Words for Improved Phrase-based SMT,” Proc. SSST-4, Fourth Workshop on Syntax and Structure in Statistical Translation, pp.19–27, 2010.
- [14] C.-L. Goh, T. Onishi, and E. Sumita, “Rule-based Reordering Constraints for Phrase-based SMT,” Proc. 15th Conference of the European Association for Machine Translation, pp.113–120, 2011.
- [15] S. Stymne, “Clustered Word Classes for Preordering in Statistical Machine Translation,” Proc. 13th Conference of the European Chapter of the Association for Computational Linguistics, pp.28–34, 2012.
- [16] U. Lerner and S. Petrov, “Source-Side Classifier Preordering for Machine Translation,” Proc. 2013 Conference on Empirical Methods in Natural Language Processing, pp.513–523, 2013.
- [17] J. Cai, M. Utiyama, E. Sumita, and Y. Zhang, “Dependency-based Pre-ordering for Chinese-English Machine Translation,” Proc. 52nd Annual Meeting of the Association for Computational Linguistics, pp.155–160, 2014.
- [18] G. Wu, Y. Zhang, and A. Waibel, “Rule-Based Preordering on Multiple Syntactic Levels in Statistical Machine Translation,” Proc. 11th International Workshop on Spoken Language Translation (IWSLT 2014), 2014.
- [19] T. Nakagawa, “Efficient Top-Down BTG Parsing for Machine Translation Preordering,” Proc. 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pp.208–218, 2015.
- [20] A. de Gispert, G. Iglesias, and B. Byrne, “Fast and Accurate Pre-ordering for SMT using Neural Networks,” Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pp.1012–1017, 2015.
- [21] K. Sudoh, X. Wu, K. Duh, H. Tsukada, and M. Nagata, “Post-ordering in Statistical Machine Translation,” Proc. 13th Machine Translation Summit, pp.316–323, 2010.
- [22] H. Isozaki, K. Sudoh, H. Tsukada, and K. Duh, “A simple reordering rule for SOV languages,” Proc. WMT-MetricsMATR, pp.244–251, 2010.
- [23] I. Goto, M. Utiyama, and E. Sumita, “Post-ordering by Parsing for Japanese-English Statistical Machine Translation,” Proc. 50th Annual Meeting of the Association for Computational Linguistics, pp.311–316, 2010.
- [24] P.-C. Chang, H. Tseng, D. Jurafsky, and C.D. Manning, “Discriminative Reordering with Chinese Grammatical Relations Features,” Proc. SSST-3, Third Workshop on Syntax and Structure in Statistical Translation, pp.51–59, 2009.
- [25] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: a method for automatic evaluation of machine translation,” Proc. 40th annual meeting of the association for computational linguistics, pp.311–318, 2002.
- [26] 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.



Huu-Anh Tran received the B.S. degree in Computer Science from Hue University of Education in 2007 and received M.S. degree from Hanoi National University of Education, Viet Nam, in 2011. He has been teaching and researching in TBU (ThaiBinh University, Viet Nam) since 2008. He is currently a Ph.D. student in the School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China.



Heyan Huang received her Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, in 1989. She is currently a professor and the dean of the School of Computer Science and Technology, Beijing Institute of Technology, Beijing. She is also the director of Beijing Engineering Research Center of High Volume Language Information Processing and Cloud Computing Applications.



Phuoc Tran received the M.S. and Ph.D. degrees in Computer Science from VNU-HCMC University of Science in 2012 and 2018, respectively. During 2006–2014, he stayed in HUFU (Ho Chi Minh city University of Food Industry). He is now a lecturer and researcher in NLP-KD Lab, Faculty of Information Technology, Ton Duc Thang University, Viet Nam.



Shumin Shi received her Ph.D. degree in computer application from Nanjing University of Science and Technology, Nanjing, in 2008. She is currently a lecturer in the School of Computer Science and Technology, Beijing Institute of Technology.



Huu Nguyen received the B.S. degree in Computer Science from Ho Chi Minh City University of Technology in 2012 and received M.S. degree from Le Quy Don Technical University in 2017, respectively. He has been teaching and researching in HUFU (Ho Chi Minh city University of Food Industry, Viet Nam) since 2009.