English Syntactic Reordering for English-Thai Phrase-Based Statistical Machine Translation

Nawaphol Labutsri[1], Rapeeporn Chamchong[1],
Richard Booth[1], Annupan Rodtook[2]

[1]Faculty of Informatics, Mahasarakham University
Mahasarakham, 44150, Thailand
Email: nawaphol_labutsri@yahoo.com

[2] Computer Science Department, Faculty of Science, Ramkhamheang University
Bangkok, 10240, Thailand

Abstract

In language pairs which have different word-orders, accuracy of translations in phrase-based statistical machine translation (SMT) systems will decrease. Syntactic reordering approaches can improve phrase-based SMT systems by reordering words in sentences to make word-orders of source language sentences similar to word-orders of target language sentences. This paper proposes reordering rules for an English-Thai phrase-based SMT system. Our reordering approach is the first that is tested in an English-Thai phrase-based SMT system. The reordering rules transform both training and test English sentences in a preprocessing step. After the preprocessing step, word-orders of English sentences are more similar to word-orders of Thai sentences. The reorder approach improves accuracy of English-Thai translation in the Moses phrase-based SMT system. In the system, the BLEU score increases clearly from 40.05% to 57.45%.

Key Words: reordering, English-Thai translation, phrase-based SMT

1. Introduction

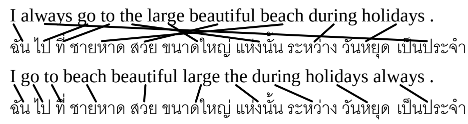
Phrase-based statistical machine translation (SMT) [1] improves upon word-based SMT [2]. Phrase-based SMT can translate from source phrases to target phrases directly. However one limitation of phrase-based SMT systems is that they use little or no syntactic information during translation. This causes the systems to have little ability to handle different word-orders between both source and target language sentences. Reordering approaches can handle different word-orders of language pairs.

```
(ROOT
 (S
   (NP (PRP I))
   (ADVP (RB always))
   (VP (VBP go)
     (PP (TO to)
       (NP (DT the)
         (JJ large)
         (JJ beautiful)
         (NN beach)))
     (PP (IN during)
       (NP (NNS holidays))))
   (. .)))
```

a

```
(ROOT
 (S
   (NP (PRP I))
   (VP (VBP go)
     (PP (TO to)
       (NP (NN beach)
         (JJ beautiful)
         (JJ large)
         (DT the)))
     (PP (IN during)
       (NP (NNS holidays)))
     (ADVP (RB always)))
   (. .)))
```

b



c

Figure 1. (a). Original parse tree, (b). Reordered parse tree for English sentence “I always go to the large beautiful beach during holidays.”, (c) Alignments before and after reordering respectively, between the English and Thai sentences.

Reordering approaches are used in a preprocessing step. First, source language sentences are parsed by a parser. Next the parse trees of source language sentences are transformed by reordering rules for making word-orders of the source sentences more similar to word-orders of the target sentences. In many studies, reordering approaches have improved accuracies of translations from French to English [3], from German to English [4] and from Chinese to English [5] in phrase-based SMT systems.

This paper proposes reordering rules for an English-Thai phrase-based SMT system. Fig. 1 shows the parse trees of an English sentence:

- “I always go to the large beautiful beach during holidays.”

and in Thai the word-order is:

Table 1. Penn English phrase tags

NP	noun phrase
VP	verb phrase
ADJP	adjective phrase
ADVP	adverb phrase
PP	preposition phrase
WHADVP	adverb question phrase
WHNP	noun question phrase
S	sentence
SQ	question sentence

- "I go to beach beautiful large the during holidays always."

To transform the first parse tree into the second we use reordering rules – for example, the adverb phrase "always" is moved to the last position of the verb phrase "go to the large beautiful beach during holidays". Phrase-based SMT systems usually have little ability to handle this type of reordering.

This paper proposes reordering rules to handle the different word-orders between English and Thai sentences by reordering words in both training (in the parallel corpus) and test English sentences. After reordering, word-orders of English sentences are more similar to word-orders of Thai sentences. We do experiments on a hand-made English-Thai parallel corpus in the Moses phrase-based system [6]. The reordering approach improves the BLEU score [7] from 40.05% to 57.45%. We also report on experiments concerning the accuracies of our reordering rules, as well as look more closely at the BLEU scores in different categories of the English sentences (i.e., affirmative sentences and different categories of interrogative sentences). We also give results of human evaluation.

The plan of this paper is as follows. In next section, we review related works which use reordering approaches in a preprocessing step. In Section 3, we introduce the reordering approach. Then, in Section 4, we describe the reordering rules in details. In the last two sections, the experiments are reported then we conclude the experiments and discuss future work respectively.

2. Related works

Many studies [3, 4, 5, 8, 9] have described syntactic reordering approaches in a preprocessing step. The study of Berger et al [8] was limited to a type of reordering French phrases of the form $Noun_1$ *de* $Noun_2$. Niessen and Ney [9] studied the reordering of German question sentences for German-English translation. They also proposed combining rules for verbs and their particles. Xia and McCord [3] described an approach to extract the reordering rules for French from a parallel corpus automatically. However, after organizing and filtering, there were over 56,000 patterns of reordering rules and they only studied French-English pairs which have very similar word-orders. Collins et al [4] described a reordering approach for German which has very different word-orders to English and they extracted the reordering rules manually. Wang et al [5] described a reordering approach for Chinese in a Chinese-English phrase-based system. They also extracted the reordering rules manually for reordering Chinese sentences. The approach which is most similar to the present paper is that in [4] and [5]. We extract reordering rules for English sentences in an English-Thai phrase-based SMT system manually. However, there are differences between our approach and theirs. We also extract the reordering rules for interrogative sentences because they have different word-orders to Thai interrogative sentences.

3. The reordering approach

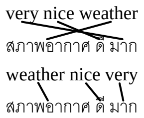
In this section, we give an overview of the steps involved in our reordering approach. First, the parallel corpus for training in the phrase-based SMT system is newly constructed manually. The parallel corpus consists of 4,621 English-Thai sentence pairs. The English sentences in the parallel corpus are obtained from High School English books in Thailand. The Thai sentences in the parallel corpus are translated from the English sentences and segmented manually. Segmentation is the process of dividing written text into meaningful units, such as words. Segmentation is for natural languages which do not have word boundary markers, such as Chinese, Japanese and Thai. The English sentences in the training set (the parallel corpus) are classified into different categories, such as affirmative sentences, interrogative sentences begin with "verb to be", interrogative sentences begin with "What", etc, and are parsed by the Stanford Parser [10]. Then we get the parse trees of the English sentences. Reordering rules are extracted from the classified parse trees of the training set. Note that reordering rules are extracted from the classified parse trees more easily than from the not-classified parse tree. Then reordering rules are stored in a computer program (written in C#) for transforming the English parse trees to make word-orders of the English sentences more similar to word-orders of the Thai sentences. After this, the phrase-based SMT system is trained by the training set, which is reordered. After that, the reordered test set is translated by the reordered phrase-based SMT system.

4. English syntactic reordering rules

We classify reordering rules into two main categories: reordering rules for affirmative sentences and interrogative sentences. In the following, we discuss each of the two main categories in more

```
(NP
 (ADJP (RB very) (JJ nice))
 (NN weather))
```
a

```
(NP
 (NN weather)
 (ADJP (JJ nice) (RB very)))
```
b

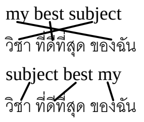


c

Figure 2. (a) Original parse tree for NP "very nice weather", (b) Reordered parse tree, (c) Alignments before and after reordering respectively, between the English and Thai sentences.

```
(NP
 (PRP$ my) (JJS best)
 (NN subject))
```
a

```
(NP
 (NN subject)
 (JJS best) (PRP$ my))
```
b



c

Figure 3. (a) Original parse tree for NP "my best subject", (b) Reordered parse tree, (c) Alignments before and after reordering respectively, between the English and Thai sentences.

```
(S
  (NP (PRP I))
  (VP (VBP am)
    (VP (VBG learning)
      (NP (NNP English))
      (ADVP (RB now))))
  (. .)))
```

Figure 4. Examples an ADVP "now" don't need to be repositioned.

details. A list of important Penn English phrase tags used in this paper is in Table 1.

4.1 Reordering rules for affirmative sentences

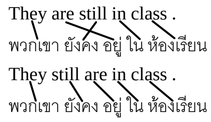
There are three sub categories: reordering in ADJPs (adjective phrases), reordering in NPs (noun phrases) and reordering ADVPs (adverb phrases).

1) ADJPs: words in ADJPs are reordered. In general, words in ADJPs are repositioned conversely. Fig. 2 shows an example of an ADJP in a NP. The front word "very" is repositioned after the back word "nice". After reordering, "very nice" is reordered to "nice very".

2) NPs: words and ADJPs in NPs are reordered. In general, words and ADJPs in NPs are repositioned conversely as same as reordering in ADJPs. In fig. 2 an ADJP, "nice very", is repositioned after the back word "weather". After reordering in the NP "very nice weather" is reordered to "weather nice very". Fig. 3 shows a NP "my best subject" that has many words. NP "my best subject" is reordered to "subject best my". Notice that there are many special cases which don't reorder words in NPs for example: "a day", "an hour" and "a month". We also make the

```
(ROOT
  (S
    (NP (PRP They))
    (VP (VBP are)
     (ADVP (RB still))
     (PP (IN in)
       (NP (NN class))))
    (. .)))
```
a

```
(ROOT
  (S
    (NP (PRP They))
    (ADVP (RB still))
    (VP (VBP are)
      (PP (IN in)
        (NP (NN class))))
    (. .)))
```
b

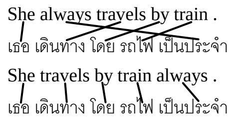


c

Figure 5. Example an ADVP "still" needs to be repositioned after the subject. (a) Original parse tree, (b) Reordered parse tree, (c) Alignments, before and after reordering respectively.

```
(ROOT
  (S
    (NP (PRP She))
    (ADVP (RB always))
    (VP (VBZ travels)
      (PP (IN by)
        (NP (NN train))))
    (. .)))
```
a

```
(ROOT
  (S
    (NP (PRP She))
    (VP (VBZ travels)
      (PP (IN by)
        (NP (NN train)))
      (ADVP (RB always)))
    (. .)))
```
b



c

Figure 6. Example an ADVP "always" needs to be repositioned to the end of the VP. (a) Original parse tree, (b) Reordered parse tree (c) Alignments, before and after reordering respectively.

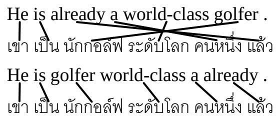
rules to handle these special cases by observing the parse trees of English sentences in the English-Thai parallel corpus.

3) ADVPs: ADVPs can be reordered in many ways because ADVPs may be positioned in English sentences in three different ways: at the beginning position of sentences, in front of verbs and behind verbs. We classify reordering rules for ADVPs in English sentences into the following four cases.

- If ADVPs are at the beginning of the English sentences, they do not need to be repositioned.
- If ADVPs are at end position of VP, they do not need to be repositioned. Fig. 4 shows an example of this case.
- Else some ADVPs are repositioned after the subject of sentences. Example: "never", "still" and "ever". Fig. 5 shows an example of this case. The ADVP "still" is moved after the subject "They".
- Else some ADVPs are repositioned to the end position of VP. Fig. 6 shows an ADVP "always" is moved to the end of the VP "travels by train". Fig. 7 shows an ADVP

```
(ROOT
  (S
    (NP (PRP He))
    (VP (VBZ is)
      (ADVP (RB already))
      (NP (DT a)
        (JJ world-class)
        (NN golfer)))
    (. .)))
```
a

```
(ROOT
  (S
    (NP (PRP He))
    (VP (VBZ is)
      (NP (NN golfer)
        (JJ world-class)
        (DT a)))
      (ADVP (RB already))
    (. .)))
```
b

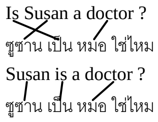


c

Figure 7. Example an ADVP "already" needs to be repositioned to the end of the VP. (a) Original parse tree, (b) Reordered parse trees, (c) Alignments, before and after reordering respectively.

```
(ROOT
  (SQ (VBZ Is)
    (NP (NNP Susan))
    (NP (DT a) (NN doctor))
    (. ?)))
```
a

```
(ROOT
  (SQ
    (NP (NNP Susan)) (VBZ is)
    (NP (NN doctor) (DT a))
    (. ?)))
```
b



c

Figure 8. Example parse trees of an interrogative sentence beginning with "verb to be" ("Is"). (a) Original parse tree, (b) Reordered parse tree (c) Alignments, before and after reordering respectively.

"already" is moved to the end of the VP.

We perform the reordering rules in the third and the fourth cases by listing adverbs in ADVPs and classifying which adverbs belong to the third or fourth case.

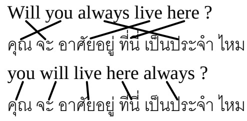
4.2 Reordering rules for interrogative sentences

Here there are many sub categories classified by the beginning words of interrogative English sentences. The beginning words need to be reordered to new positions. In the following, we discuss each category in more details.

1) Begin with "verb to be", "verb to do", "verb to have", and "auxiliary verb": Example "is", "am", "are", "was", "were", "do", "does", "did", "can", "may", "will" and "would". The "verb to be", "verb to do", "verb to have", and "auxiliary verb" is reordered after the subject of interrogative sentences. This makes the structure of the interrogative sentence become like the affirmative sentences. After that, we use reordering rules for affirmative sentences given in the previous subsection. Fig. 8 shows parse trees of an interrogative sentence which begins with "Is". "Is" is repositioned after the subject "Susan" then reordered by using the reordering rules for affirmative sentences. Fig. 9 shows parse trees of an interrogative sentence that begins with "Will".

```
(ROOT
  (SQ (MD Will)
    (NP (PRP you))
    (ADVP (RB always))
    (VP (VB live)
      (ADVP (RB here)))
    (. ?)))
```
a

```
(ROOT
  (SQ
    (NP (PRP you)) (MD will)
    (VP (VB live)
      (ADVP (RB here))
      (ADVP (RB always)))
    (. ?)))
```
b

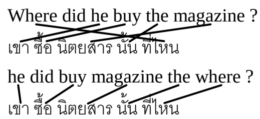


c

Figure 9. Example parse trees of an interrogative sentence beginning with "auxiliary verb" ("Will"). (a) Original parse tree, (b) Reordered parse tree, (c) Alignments, before and after reordering respectively.

```
(ROOT
  (SBARQ
    (WHADVP (WRB Where))
    (SQ (VBD did)
      (NP (PRP he))
      (VP (VB buy)
        (NP (DT the)
          (NN magazine))))
    (. ?)))
```
a

```
(ROOT
  (SBARQ
    (SQ
      (NP (PRP he)) (VBD did)
      (VP (VB buy)
        (NP (NN magazine) (DT the))
        (WHADVP (WRB Where)))
    (. ?)))
```
b



c

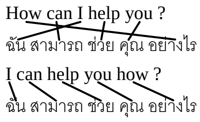
Figure 10. Example parse trees of an interrogative sentence begins with "Where". (a) Original parse tree, (b) Reordered parse tree, (c) Alignments, before and after reordering respectively.

2) Begin with "WHADVP": WHADVPs are phrases for interrogative sentences. They are the same as ADVPs. Examples of WHADVPs are "Where", "When", "How" and "Why". In this rule, first "verb to be", "verb to do", "verb to have", and "auxiliary verb" in SQ are repositioned the same as with rule 1. Then WHADVPs are repositioned to the end of VP of sentences. Finally, reordering rules for affirmative sentences are used for reordering words in ADJPs and NPs and reordering ADVPs. Fig. 10 and Fig. 11 show parse trees of interrogative sentences which begin with "Where" and "How" and reordered parse trees respectively. Notice that the interrogative sentences which begin with "Why" do not need to reorder "Why" because "Why" is at the beginning of the interrogative sentences in Thai.

3) Begin with "WHNP": WHNPs are phrases for interrogative sentences. They are the same as NPs. Then WHNPs are repositioned after the verb or repositioned after the preposition (IN) which does not have nouns (NP). Examples of WHNPs are "What", "Who", "What animal" and "What food". In this rule first "verb to be", "verb to do", "verb to have", and "auxiliary verb" in SQ are repositioned the same as

```
(ROOT
  (SBARQ
    (WHADVP (WRB How))
    (SQ (MD can)
      (NP (PRP I))
      (VP (VB help)
        (NP (PRP you))))
    (. ?)))
              a
```

```
(ROOT
  (SBARQ
    (SQ
      (NP (PRP I)) (MD can)
      (VP (VB help)
        (NP (PRP you))
        (WHADVP (WRB How))))
    (. ?)))
              b
```

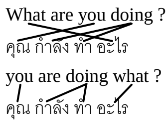


c

Figure 11. Example parse trees of an interrogative sentence begins with "How". (a) Original parse tree, (a) Reordered parse trees, (c) Alignments, before and after reordering respectively.

```
(ROOT
  (SBARQ
    (WHNP (WP What))
    (SQ (VBP are)
      (NP (PRP you))
      (VP (VBG doing)))
    (. ?)))
              a
```

```
(ROOT
  (SBARQ
    (SQ
      (NP (PRP you)) (VBP are)
      (VP (VBG doing)
        (WHNP (WP what))))
    (. ?)))
              b
```



c

Figure 12. Example parse trees of an interrogative sentence begins with "What" and reordering "What" after the verb of the sentence. (a) Original parse tree, (b) Reordered parse tree, (c) Alignments, before and after reordering respectively.

with rule 1. Then WHNPs are repositioned according to the following conditions:

- Reorder WHNPs after the verb of the sentences. An example is shown in fig. 12.
- If there is an incomplete PP (PP that doesn't have NP after IN), reorder WHNPs after IN in the PP. An example is shown in fig. 13.

After reordering WHNPs, then reordering rules for affirmative sentences are used for reordering words in ADJPs and NPs and reordering ADVPs.

5. Experiments

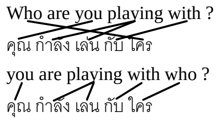
5.1 Accuracy of the reordering rules

The purpose of this experiment was to see how reordering rules given in previous section are accurate when they were applied to the test set. The test set was obtained randomly from High School English books. In the test set, Lengths of the sentences are between 3 and 12 words. Average length is 5.96 words. The test set was split into five subsets:

- 50 affirmative sentences which have only short-distance reordering, such as reordering in NPs and ADJPs.

```
(ROOT
  (SBARQ
    (WHNP (WP Who))
    (SQ (VBP are)
      (NP (PRP you))
      (VP (VBG playing)
        (PP (IN with))))
    (. ?)))
              a
```

```
(ROOT
  (SBARQ
    (SQ
      (NP (PRP you)) (VBP are)
      (VP (VBG playing)
        (PP (IN with)
          (WHNP (WP Who))))
    (. ?)))
              b
```



c

Figure 13. Example parse trees of an interrogative sentence begins with "Who" and reordering "Who" after the IN "with" in PP. (a) Original parse tree, (b) Reordered parse tree, (c) Alignments, before and after reordering respectively.

Table 2. Accuracy of the reordering rules

English test subsets	Accuracy
Affirmative Sentences which have only short-distance reordering	94.00%
Affirmative Sentences which have long-distance reordering	92.00%
Interrogative sentences which begin with "verb to be", "verb to do", "verb to have" and "auxiliary verb"	70.00%
Interrogative sentences which begin with WHADVP	82.00%
Interrogative sentences which begin with WHNP	88.00%

- 50 affirmative sentences which have long-distance reordering, such as reordering ADVPs.
- 50 interrogative sentences which begin with "verb to be", "verb to do", "verb to have" and "auxiliary verb".
- 50 interrogative sentences which begin with WHADVPs.
- 50 interrogative sentences which begin with WHNPs.

After reordering the test set by the reordering rules, the accuracy values of the reordering rules were collected for each subset on the test set. The accuracy values were given in percentage form. Human evaluation was used for evaluating how accurately the reordering rules are applied to the test set. The annotator was a student from English Department of Mahasarakham University. The annotator evaluated accuracies of the reordering rules by comparing word-orders of reordered English sentences with corresponding Thai sentences and counted the number of the reordered English sentences which have word-orders similar to word-orders of the Thai sentences. Table 2 shows the accuracies of the reordering rules for each subset of English sentences on the test set.

The experiment showed that the most common causes of errors of the reordering rules are incorrect parsing and special-case reordering. Incorrect parsing of the parser is the main reason which causes errors of reordering. However, the accuracies of the reordering rules for English sentences are still high

Table 3. Performance of the reordered system

English test subsets	BLEU (%)	Gain[a] (%)
Affirmative Sentences which have only short-distance reordering	59.09	+15.73
Affirmative Sentences which have long-distance reordering	57.31	+25.23
Interrogative sentences which begin with “verb to be”, “verb to do”, “verb to have” and “auxiliary verb”	42.07	+5.42
Interrogative sentences which begin with WHADVPs	55.37	+23.6
Interrogative sentences which begin with WHNPs	70.50	+13.55

a. Gain is the change of BLEU scores when compare the reordered system with the baseline system.

because the English sentences in the test set are not complex and the English parser is fairly accurate.

5.2 The BLEU scores of the baseline and reordered systems

The purpose of this experiment was to measure performance of the baseline and the reordered systems using the BLEU score, the popular metric for machine translation [7]. In the preprocessing step, English sentences in both the training (parallel corpus) and test sets were parsed into English parse trees first. We used the same test set as in the experiment 1 above. The Stanford Parser [10] was used for parsing the English sentences. Then reordering rules described in Section 4 were applied to the parse trees. In the baseline system, the original training set used to train in the phrase-based SMT system and the original test set was translated. In the reordered system, the reordered training set used to train in the phrased-based SMT and the reordered test set was translated.

In the evaluation step of this experiment, there were two reference sets. Because English sentences can be translated into correspond-meaning Thai sentences in many different ways, the reference sets should be based on the outputs from both the baseline and reordering systems. It is not fair if translation of “The dog eats the meat” which is “สุนัข ทาน เนื้อ นั้น” is classified as incorrect because the reference, “หมา กิน เนื้อ ชิ้นนั้น”, is different from the translation. Hence the original outputs were adjusted better into the reference sets by the students of the English Department in Mahasarakham University. The first reference set was based on the output from the baseline systems and the second reference set was based on the output from the reordered systems.

The BLEU score was used for measuring accuracy of the translation in the baseline and reordered systems. For each of the two outputs, two BLEU scores were measured by the first and the second reference sets. Then, for each of the outputs, the two BLEU scores were averaged. The reordering approach improved the average BLEU scores from 40.05% to 57.45%.

5.3 Effectiveness of reordering for each of the English test subsets

For this experiment, the purpose was to see how the reordering rules impact each test subset separately. This is because each subset has different structure. The average BLEU scores of the translations were collected for each of the test subsets. Table 3 shows the performance of the reordered system. The performance of the system improves clearly because the reordered system can handle the difference in word-orders between English and Thai sentences. Performance of the reordered system for the test subset of interrogative sentences, which begin with “verb to be”, “verb to do”, “verb to have” and “auxiliary verb”, is the lowest because the system cannot handle the continuous tense. An example of an error of the continuous tense translation is that “is moving” translates to the incorrect translation, “คือ เคลื่อนย้าย”, instead of the correct one, “กำลัง เคลื่อนย้าย”. And for this subset it improves little when compared with the other test subsets because the word-orders between English and Thai sentences in this subset are not so different. In addition, the baseline and reordered systems have little ability to handle spurious words when translating to Thai sentences. A spurious word is a Thai word which does not match with any English word. Examples of spurious words are “หรือไม่”, “ไหม” or “ใช่ไหม”.

5.4 Human evaluation

Human evaluation was also used for comparing translation qualities of both the baseline and reordered systems. For this experiment we followed the same general procedure as in [4]. Four test subsets were used for evaluating. The affirmative sentences were selected randomly from the two test subsets of affirmative sentences in the experiment 1 for the first test subset in this experiment and the remaining three test subsets in this experiment were same as the remaining three test subsets in the experiment 1. The two annotators were students from the English Department of Mahasarakham University. The annotators compare qualities of translations of the test subsets from the baseline and reordered systems. The annotators did not know which translations were from the baseline system or from the reordered system to prevent the annotators' bias. The annotators evaluated the pair of translations from the baseline and reordered systems by stating one of two choices:

Table 4. Agreements between the annotators for qualities of the translations

English test subsets	Annotator 1	Annotator 2		
Affirmative Sentences		***R***[a]	***E***[b]	***B***[c]
	R	34	6	0
	E	0	8	0
	B	0	2	0
Interrogative sentences which begin with “verb to be”, “verb to do”, “verb to have” and “auxiliary verb”		***R***	***E***	***B***
	R	5	12	0
	E	0	20	0
	B	2	1	10
Interrogative sentences which begin with WHADVPs		***R***	***E***	***B***
	R	21	7	0
	E	3	15	0
	B	0	1	3
Interrogative sentences which begin with WHNPs		***R***	***E***	***B***
	R	15	7	0
	E	1	24	1
	B	1	1	0

a. R is counts of translations from the reordered system which the annotator considers better.

b. E is counts of translations from both systems which the annotator considers equal of qualities.

c. B is counts of translations from the baseline system which the annotator consider better.

- Quality of the translation is better than the other.
- Qualities of both the translations are equal.

Table 4 shows agreements between the annotators for qualities of the translations. For example 34 translations of the affirmative sentences from the reordered system are considered better by both the annotators. The translation qualities of the interrogative sentences, which begin with “verb to be”, “verb to do”, “verb to have” and “auxiliary verb”, are nearly equal in both the systems. The results of this experiment are in same direction with the experiment 3. Overall, the reordered system is considered better than the baseline system.

6. Conclusions and future works

This paper proposes reordering rules for English-Thai phrase-based SMT system. Reordering rules can improve English-Thai translation qualities of both the affirmative and interrogative sentences. The reordering approach improved the BLEU score for the MOSES system from 40.05% to 57.45% on average although there are some errors of reordering because of incorrect parsing and special-case reordering. However, the reordered system has little ability to handle the continuous tense and the spurious words. In the future we may investigate more special-case reordering rules and restructure approaches which have reordering, insertion and deletion instead only reordering for an English-Thai phrase-based SMT system.

7. Acknowledgement

We would like to thank Mahasarakham University (MSU), in Thailand, who support fund for us to make the English-Thai parallel corpus used in our experiments. We also thank the students from English Department of MSU who make the parallel corpus and evaluate qualities of the translations in the experiments.

8. References

[1] P. Koehn, F. J. Och, and D. Marcu, “Statistical phrase based translation”, In Proceedings of HLT-NAACL, 2003.

[2] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer, “The Mathermatics of Statistical Machine Translation: Parameter Estimation”, Association for Computational Linguistics, 1993.

[3] F. Xia, and M. McCord, “Improve a statistical MT system with automatically learned rewrite patterns”, In Proceedings of COLING, 2004.

[4] M. Collins, P. Koehn, and I. Kucerova, “Clause restructuring for statistical machine translation”, In Proceedings of EMNLP, 2005.

[5] C. Wang, M. Collins, and P. Koehn, “Chinese syntactic reordering for statistical machine translation”, In Proceedings of EMNLP-CoNLL, 2007, pp. 737–745.

[6] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constrantin, and E. Herbst, “Moses: Open source toolkit for statistical machine translation”, In Proceedings of ACL, Demonstration Session, 2007.

[7] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, “BLEU: a Method for Automatic Evaluation of Machine Translation”, Association for Computational Linguistics, 2002, pp. 311-318.

[8] A. L. Berger, S. A. D. Pietra, and V. J. D. Pietra, “A maximum entropy approach to natural language processing”, Computational Linguistics, 1996, pp. 39–69.

[9] S. Niessen, and H. Ney, “Statistical machine translation with scarce resources using morpho-syntactic information”, Computational Linguistics, 2004, pp. 181–204.

[10] D. Klein, R. Levy, C. Manning, T. Grenager, G. Andrew, M. C. Marneffe, B. MacCartney, A. Rafferty, H. Tseng, P. Chang, W. Maier, and J. Finkel, “The Stanford Parser: A statistical parser”, http://nlp.stanford.edu/software/lex-parser.shtml, accessed January 2009