A Lexicalized Reordering Model for Hierarchical Phrase-based Translation

Hailong Cao[1], Dongdong Zhang[2], Mu Li[2], Ming Zhou[2] and Tiejun Zhao[1]
[1]Harbin Institute of Technology, Harbin, P.R. China
[2]Microsoft Research Asia, Beijing, P.R. China
{hailong, tjzhao}@mtlab.hit.edu.cn
{Dongdong.Zhang, muli, mingzhou}@microsoft.com

Abstract

Lexicalized reordering model plays a central role in phrase-based statistical machine translation systems. The reordering model specifies the orientation for each phrase and calculates its probability conditioned on the phrase. In this paper, we describe the necessity and the challenge of introducing such a reordering model for hierarchical phrase-based translation. To deal with the challenge, we propose a novel lexicalized reordering model which is built directly on synchronous rules. For each target phrase contained in a rule, we calculate its orientation probability conditioned on the rule. We test our model on both small and large scale data. On NIST machine translation test sets, our reordering model achieved a 0.6-1.2 BLEU point improvements for Chinese-English translation over a strong baseline hierarchical phrase-based system.

1 Introduction

In statistical machine translation, the problem of reordering source language into the word order of the target language remains a central research topic. Statistical phrase-based translation models (Och and Ney, 2004; Koehn et al., 2003) are good at local reordering, or the reordering of words within the phrase, since the order is specified by phrasal translations. However, phrase-based models remain weak at long-distance reordering, or the reordering of the phrases. To improve the reordering of the phrases, two types of models have been developed.

The first one is lexicalized reordering models (Tillman, 2004; Huang et al., 2005; Al-Onaizan and Papineni, 2006; Nagata et al., 2006; Xiong et al., 2006; Zens and Ney, 2006; Koehn et al., 2007; Galley and Manning, 2008; Cherry et al., 2012) which predict reordering by taking advantage of lexical information. The model in (Koehn et al., 2007) distinguishes three orientations with respect to the previous and the next phrase—monotone (*M*), swap (*S*) and discontinuous (*D*). For example, we can extract a phrase pair "xiayou ||| the lower reach of" whose orientations with respect to the previous and the next phrase are *D* and *S* respectively, as shown in Figure 1. Such a model is simple and effective, and has become a standard component of phrase-based systems such as MOSES.

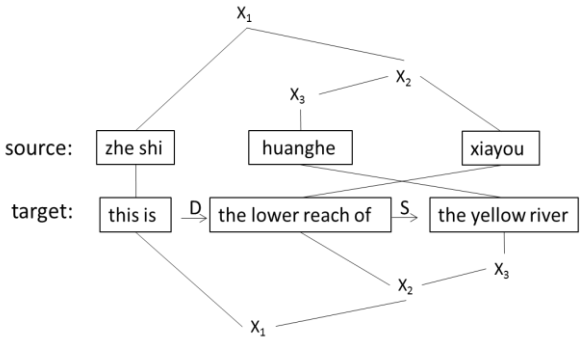

Figure 1. Phrase orientations for Chinese-English translation.

The other is a hierarchical phrase-based (HPB) translation model (Chiang, 2007) based on synchronous grammar. In the HPB model, a synchronous grammar rule may contain both terminals (words) and nonterminals (sub-phrases). The order of terminals and nonterminal are specified by the rule. For

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 1144–1153, Dublin, Ireland, August 23-29 2014.

example, the translation rule <X xiayou, the lower reach of X > specifies that the translation of sub phrase X before "xiayou" should be put after "the lower reach of".

One problem with the HPB model is that the application of a rule is independent of the actual sub phrase. For example, the rule <X xiayou, the lower reach of X > will always swap the translation of X and "xiayou", no matter what is covered by X. This is an over-generalization problem. Much work has been done to solve this issue. For example, Zollmann and Venugopal (2006) annotate non-terminals by syntactic categories. He et al. (2008) proposes maximum entropy models which combine rich context information for selecting translation rules during decoding. Huang et al. (2010) automatically induce a set of latent syntactic categories to annotate nonterminals. These works alleviate the over-generalization problem by considering the content of X. In this paper, we try to solve it from an alternative view by modeling whether the phrases covered by X prefer the order specified by the rule. This has led us to borrow the lexicalized reordering model from the phrase-based model for the HPB model. We propose a novel lexicalized reordering model for hierarchical phrase-based translation and achieved a 0.6-1.2 BLEU point improvements for Chinese-English translation over a strong HPB baseline system.

2 Related work

In this section, we briefly review two types of related work which are a nonterminal-based lexicalized reordering models and a path-based lexicalized reordering model. Both of them calculate the orientation for HPB translation.

2.1 Nonterminal-based lexicalized reordering models

Xiao et al. (2011) proposed an orientation model for HPB translation. The orientation probability of a derivation is calculated as the product of orientation probabilities of all nonterminals except the root. In order to define the relative orders of nonterminals and their adjacent phrase, they expand the alignment in a rule to include both terminals and nonterminals. There may be multiple ways to segment a rule into phrases; they use the maximum adjacent phrase similar to Galley and Manning (2008). They significantly outperformed the HPB system on both Chinese-English and German-English translation.

Xiao et al. (2011) use the boundary word feature of nonterminals without considering their internal structure. For example, in Figure 1, suppose nonterminal X_1 is not the root node and the orientation probability of X_1 will condition on "zhe, xiayou, this, river".

In this paper, we will consider how the words covered by the nonterminal X_1 are reordered. Rather than using "xiayou" as a feature to determine the orientation of X_1 with respect to the next phrase, we think the immediately translated source word "huanghe" could be more informative through it is not on the boundary of X_1 , since "huanghe" is the exact starting point from where we search for the next phrase to translate.

Huck et al. (2013) proposed a very effective phrase orientation model for HPB translation. The model is also based on nonterminal. They extracted phrase orientation probabilities from word-aligned training data for use with hierarchical phrase inventories, and scored orientations in hierarchical decoding.

2.2 Path-based lexicalized reordering model

The most recent related work is Nguyen and Vogel (2013). They map a HPB derivation into a discontinuous phrase-based translation path in the following two steps:

1) Represent each rule as a sequence of phrase pairs and non-terminals.
2) The rules' sequences are used to find the corresponding phrase-based path of a HPB derivation and calculate the phrase-based reordering features.

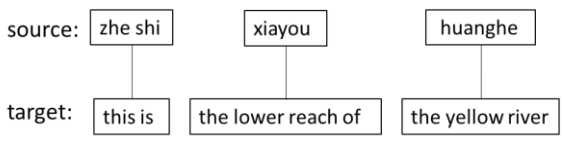

Figure 2. The phrase-based path of the derivation in Figure 1.

A phrase-based path is the sequence of phrase pairs, whose source sides covers the source sentences and whose target sides generated the target sentences from left to right. For example, the phrase-based path of the derivation in Figure 1 is shown in Figure 2.

The phrase-based reordering features for the above phrase-based path are:

$\log P_{next}(D \mid <\text{zheshi, this is}>)$, $\log P_{previous}(D \mid <\text{xiayou, the lower reach of}>)$,

$\log P_{next}(S \mid <\text{xiayou, the lower reach of}>)$, $\log P_{previous}(S \mid <\text{huanghe, the yellow river}>)$.

Nguyen and Vogel (2013) achieved significant improvement over both phrase-based and HPB models on three language pairs respectively.

One problem with the above work is that they did not use rules with unaligned source or target phrases. Though this can get faster and better Arabic-English translation, it leads to a 0.49 BLEU point loss for Chinese-English translation.

Another problem with path-based model is: there are many forms of HPB rules which we cannot map into a reasonable sequence of phrase pairs and non-terminals. We will show this with an example derivation shown in Figure 3. The main difference between Figure 3 and Figure 1 is there is such a rule <fangzhi X, prevent X from> that a source phrase "fangzhi" is aligned with a discontinuous target phrase "prevent…from". This makes it hard to find the corresponding phrase-based path because we do not know what is the right order of "fangzhi ||| prevent…from" and "daozei ||| the thieves" in the discontinuous phrase-based path. We face the following dilemmas:

- If "fangzhi ||| prevent…from" goes first, then the discontinuous phrase-based path is as shown in Figure 4(a). On such a path, we will consider the orientation of "the thieves" with respect to "breaking in". This is unreasonable because "the thieves" and "breaking in" are not adjacent in the target side. It does not satisfy the definition of the phrase-based reordering model which predicts the orientation with respect to previous or next adjacent target phrase.
- If "daozei ||| the thieves" goes first, then the discontinuous phrase-based path is as shown in Figure 4(b). This is unreasonable because "The policeman" and "the thieves" are not adjacent on the target side.

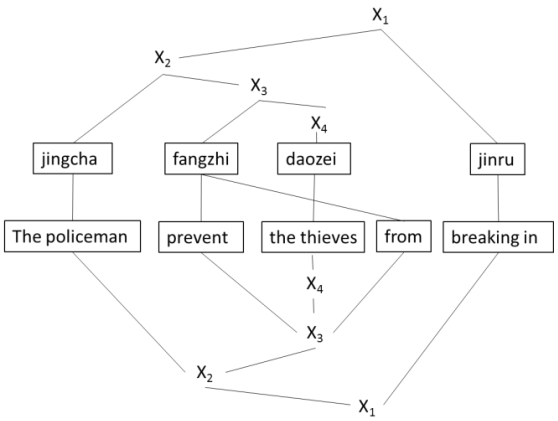

Figure 3. Example of Chinese-English translation and its derivation.

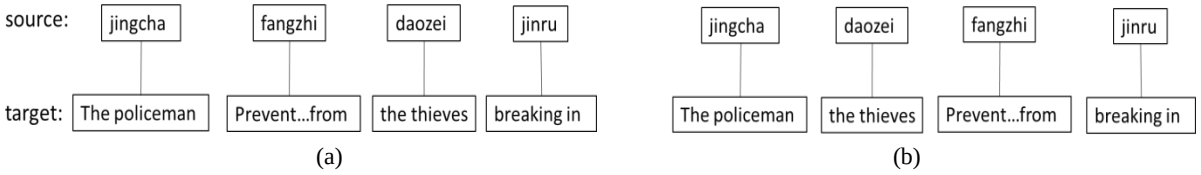

Figure 4. Two discontinuous phrase-based path candidates of the HPB derivation.

From the above example, we can see that if a target phrase is aligned to a discontinuous target phrase in a HPB rule, then it is hard to find a reasonable path whose target sides can generate the target sentence from left to right.

3 Our lexicalized reordering model

Rather than mapping a HPB derivation into a discontinuous phrase-based path and applying reordering model built on phrases, we propose a lexicalized reordering model which is built directly on HPB rules. For each target phrase contained in a HPB rule, we calculate its orientation probability conditioned on the rule. For the example derivation in Figure 3, we represent it by the structure shown in the following figure:

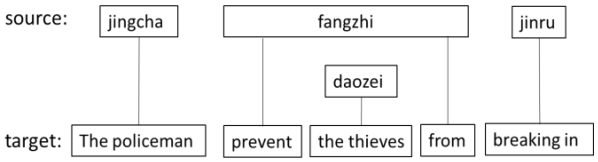

Figure 5. Our representation of the HPB derivation in Figure 3.

Different from Figure 4(a) and Figure 4(b) which contain a discontinuous phrase "prevent…from", we represent "prevent…from" as two individual target phrases: "prevent" and "from". Instead of considering the orientation of "prevent…from", we consider the orientation of "prevent" and "from" respectively. For example, we will consider the orientation of "prevent" with respect the previous phrase "the policeman" $O_{previous}(\text{prevent})$, and the orientation of "prevent" with respect the next phrase "the thieves" $O_{next}(\text{prevent})$. The probabilities of both $O_{previous}(\text{prevent})$ and $O_{next}(\text{prevent})$ are conditioned on the rule <fangzhi X, prevent X from>.

In Figure 5, every two neighboring target phrases are adjacent in the original target side. In this way, we can borrow the phrase-based reordering model which calculates the orientation with respect to previous and next adjacent phrase.

More formally, we represent a HPB rule in the general form of:

$$r = < s_0 \mathrm{X}_1 s_1 \mathrm{X}_2 s_2 ... \mathrm{X}_n s_n, t_0 \mathrm{X}_1 t_1 \mathrm{X}_2 t_2 ... \mathrm{X}_n t_n, \alpha >$$

where *n* is the number of nonterminals. $s_i, i = 1...n$, is the source phrase which is a continuous source word sequences. $t_i, i = 1...n$, is the target phrase which is a continuous target word sequences. We use α to represent the alignment of words and nonterminals in the rule. Note that s_i or t_i can be empty if there are adjacent nonterminals or there is nonterminal on the boundary. The lexicalized reordering probability of rule *r* is defined as the product of each target phrase's orientation probabilities conditioned on the rule *r*:

$$\prod_{i=0}^{n} P_{previous}(O_{previous}(t_i) \mid r, i) P_{next}(O_{next}(t_i) \mid r, i)$$

In the above equation, each probability is conditioned on the whole rule. In this way, we avoid the problem of mapping a HPB derivation into a discontinuous phrase-based path. There are two advantages for our reordering model:

- It is compatible with HPB rules which contain unaligned phrases.
- It is compatible with HPB rules in which a source phrase is aligned to a discontinuous target phrase.

Actually, our model is compatible with any kind of HPB rules since it is defined on the general form of rule.

Now we describe how to define $O_{previous}(t_i)$ and $O_{next}(t_i)$ in the model. Suppose t_i contains k_i target words and we write t_i as $t_i = w_{i(1)} w_{i(2)} ... w_{i(k_i\text{-}1)} w_{i(k_i)}$. Then we define:

$$O_{previous}(t_i) = O_{previous}(w_{i(1)}) = O(w_{i(1)\text{-}1}, w_{i(1)}), \qquad O_{next}(t_i) = O_{next}(w_{i(k_i)}) = O(w_{i(k_i)}, w_{i(k_i)+1})$$

where $O(w_j, w_{j+1})$ is the orientation of two adjacent target words and is determined as follows:

$$\begin{aligned}
&\text{If } (\ rm(w_j)+1 = lm(w_{j+1})\ ) && O(w_j, w_{j+1}) = M;\\
&\text{Else if } (\ rm(w_{j+1})+1 = lm(w_j)\ ) && O(w_j, w_{j+1}) = S;\\
&\text{Else} && O(w_j, w_{j+1}) = D;
\end{aligned}$$

$rm(w)$ is the position of the right most source word aligned to target word *w*; $lm(w)$ is the position of the left most source word aligned to target word *w*.

Above is our lexicalized reordering model which is built upon HPB rules. We complete its description using an example. For the rule <fangzhi X, prevent X from>, *n*=1, $t_0 = \text{“prevent”}$ and $t_1 = \text{“from”}$, the lexicalized reordering probability is:

$$P_{previous}(O_{previous}(\text{prevent})|\text{<fangzhi X, prevent X from>},0) \cdot P_{next}(O_{next}(\text{prevent})|\text{<fangzhi X, prevent X from>},0)$$
$$\cdot P_{previous}(O_{previous}(\text{from})|\text{<fangzhi X, prevent X from>},1) \cdot P_{next}(O_{next}(\text{from})|\text{<fangzhi X, prevent X from>},1)$$

Note that we calculate the orientation of plain phrase pairs in the same way as for HPB rules. We can represent a phrase pair in the form of $r = < s_0, t_0, \alpha >$, which is a rule that does not contain any nonterminal. Then we can apply our above model which is general enough to cover both HPB rules and plain phrase pairs.

4 Training and decoding

The training of our model is similar to the reordering model of Moses. During the standard phrase pair extraction and rule extraction, besides the nonterminal alignment in rules, we also keep the lexical alignments and orientations. If a phrase pair or a rule is observed with more than one set of alignment, we only keep the most frequent one and only count the orientations corresponding to the most frequent alignment.

Following Moses, we use relative frequency and add 0.5 smoothing technique to estimate the orientation probability based on all samples collected from the training corpus. Generally, given a rule *r* with *n* target phrases, we estimated the reordering probability for each t_i as follows:

$$P_{previous}(O_{previous}(t_i)|r,i) = \frac{0.5 + \#(O_{previous}(t_i),\ r)}{1.5 + \#(r)}, \qquad P_{next}(O_{next}(t_i)\,|\,r,i) = \frac{0.5 + \#(O_{next}(t_i), r)}{1.5 + \#(r)}$$

For each parallel sentences pair, we add a start and an end mark on both sides. They are aligned respectively.

Our phrase pairs and rules are extracted from word aligned parallel sentences. There are many phrase pairs and rules which contain unaligned target or source words. How to deal with them is quite important for our reordering model. We will describe how to process them in the following two subsections.

4.1 The processing of unaligned target words

Our main principle for processing an unaligned word is to: skip it and use the nearest aligned word. For example in Figure 3, the orientation of “prevent” with respect to the next phrase is determined by:

$$O_{next}(\text{prevent}) = O(\text{prevent}, \text{the})$$

If the target word “the” is unaligned and “thieves” is aligned with “daozei”, we will define:

$$O_{next}(\text{prevent}) = O(\text{prevent, the}) = O(\text{prevent, thieves}) = M$$

Similarly, in Figure 1, the orientation of “the lower reach of” with respect with “the yellow river” is determined by *O*(of, the). Suppose both “of” and “the” are unaligned and there are alignments for “reach-xiayou” and “yellow-huanghe”, we will have:

$$O(\text{of, the}) = O(\text{reach, yellow}) = S$$

We believe this orientation is consistent with our intuitions.

More formally, before we determine the orientation of two adjacent target words $O(w_p, w_q)$, we apply the following processing procedure:

While (target word w_p is unaligned) p--;

While (target word w_q is unaligned) q++;

If all words in a target phrase t_i are unaligned, we do not need to consider its orientation since t_i does not trigger any movement along the source words at all. Actually, it will be skipped when we determine the orientation of the previous and next aligned target phrases. (See also the decoding algorithm in Section 4.3)

4.2 The processing of unaligned source words

The processing of Section 4.1 can guarantee that the orientation is determined based on two aligned target words, namely w_p and w_q, which must be continuous or separated by unaligned target words.

Now we introduce the processing of unaligned source words. Before we determine the orientation of two target words $O(w_p, w_q)$, we apply the following procedure to modify the position index of the left most source word aligned to w_p and w_q respectively:

While (the $(lm(w_p)\text{-}1)^{\text{th}}$ source word is unaligned) $lm(w_p)$--;

While (the $(lm(w_q)\text{-}1)^{th}$ source word is unaligned) $lm(w_q)$--;

For the example shown in the Figure 6, initially we have $rm(w_1) = 1$ and $lm(w_4) = 4$. Since the source words w_3 and w_2 are unaligned, our procedure will modify the value of $lm(w_4)$ from 4 to 2. Finally, since $rm(w_1) + 1 = lm(w_4)$, the orientation of the two phrases marked by rectangular boxes in Figure 6 is:

$$O(w_2, w_3) = O(w_1, w_4) = M$$

Again, we believe this result is consistent with our intuition.

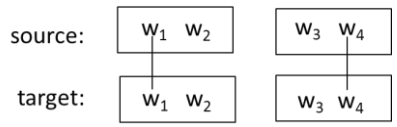

Figure 6. An example of phrases contain unaligned words

Note that during decoding, both the unaligned source and target words are also processed in the same way as in the training step. This makes our lexicalized reordering model consistent.

4.3 Decoding

Now we introduce how to integrate our reordering model into the HPB system during the standard CYK bottom-up decoding.

During decoding, if we just apply a plain phrase, we do not need to consider the orientation at once. It will be triggered when the phrase is used to compose a larger translation hypothesis together with other phrases or rules.

We need to calculate the reordering features whenever we apply a HPB rule or a glue rule during the CYK decoding. Generally, given a rule $r = < s_0 \mathrm{X}_1 s_1 \mathrm{X}_2 s_2 ... \mathrm{X}_n s_n, t_0 \mathrm{X}_1 t_1 \mathrm{X}_2 t_2 ... \mathrm{X}_n t_n, \alpha >$ defined in section 3, we calculate the reordering probability for the span covered by *r* with algorithm 1. In the algorithm, LL(X) represents the lowest rule which covers the left most word of X; LR(X) is the lowest

rule which covers the right most word of X; Both LL(X) and LR(X) can be found by traversing the derivation tree top to down recursively. LI(r) is the index of the last target phrase of rule *r*.

As in the example shown in Figure 3, for the rule $r_{2=}$<X_2 jinru, X_2 breaking in>, the orientation of X_2 and "breaking in" is:

$$O = O_{previous}(\text{breaking in}) = O(\text{from, breaking}) = D$$

The right most target word of X_2 is "from", the lowest rule covering "from" is $r_{3=}$<fangzhi X_4, prevent X_4 from> and the index of the last target phrase of r_3 is 1. So the reordering probability is:

$$prob = P_{previous}(D|r_1,0) \cdot P_{next}(D|r_3,1)$$

Note that, for readability, we use the product of probabilities to demonstrate the decoding process. Actually in practice, we use a linear model which sums the weighted log probabilities.

```
prob=1;
for (int i=1; i<=n; i++)
{
  if ( t_{i-1} is not empty and contains aligned words)
  {
      O = O_next(t_{i-1});
      prob* = P_next(O|r, i-1);
      prob* = P_previous(O|LL(X_i),0);
  }
  if ( t_i is not empty and contains aligned words)
  {
      O = O_previous(t_i);
      prob* = P_next(O|LR(X_i), LI(LR(X_i)));
      prob* = P_previous(O | r, i);
  }
  else if (i<n)
  {
      // X_i and X_{i+1} are continuous
      //or all words between them is unaligned
      rule r_p = LR(X_i);
      rule r_q = LL(X_{i+1});
      t = the first phrase of r_q;
      O = O_previous(t);
      prob* = P_next(O | r_p, LI(r_p))
      prob* = P_previous(O | r_q, 0);
      i++;
  }
}
```

Algorithm 1. Calculating the reordering probability for a span covered by a rule: $r = < s_0 X_1 s_1 X_2 s_2 ... X_n s_n, t_0 X_1 t_1 X_2 t_2 ... X_n t_n, \alpha >$.

As shown in Algorithm 1, the reordering probability depends on the lowest rules which cover the left/right most word. Therefore, we keep the lowest rules which cover the left/right most word for each partial translation. If two partial translations are same in everything but differ in the lowest rule, we need to keep both of them, rather than only keep the one with higher score. This will increase the complexity of the searching.

4.4 Discussion

Orientation can be determined based on word, phrase and hierarchical phrase (Galley and Manning, 2008). What we adopt in this paper is word based orientation. It is based on the following considerations:

- Our baseline is a HPB system, which can capture hierarchical orientation. We use word based orientation with the aim to complement the HPB system.
- Word based orientation is consistent during training and decoding; phrase based orientation is prone to inconsistent between training and decoding.

Galley and Manning (2008) has pointed out an inconsistency in Moses between training and decoding. Here we would like to note that phrase based orientation depends on phrase segmentation. For example, in Figure 1, the orientation of phrase "this is" with respect to next phrase could be either:

- D, if we think the next phrase is "the lower reach of " which is what Figure 1 shows.
- or S, if the next phrase is "the lower reach of the yellow river" which can compose a legal phrase pair with "huanghe xiayou" according to the standard phrase pair extraction algorithm.

The decision to adopt word-based orientation makes our work similar with Hayashi et al. (2010) who proposed a word-based reordering model for HPB system. The difference between our work and Hayashi et al. (2010) is: they adopt the reordering model proposed by Tromble and Eisner (2009) for the preprocessing approach, while we borrow the idea of lexicalized reordering models which are originally proposed for phrase-based machine translation.

5 Experiments

5.1 Experimental settings

Our baseline system is re-implementation of Hiero, a hierarchical phrase-based system (Chiang, 2007). Besides the standard features of a HPB model, there are six reordering features in our reordering model which are M, S and D with respect to the previous and next phrase respectively. They are integrated into the log-linear model of the HPB system. The Minimum Error Rate Training (MERT) (Och, 2003) algorithm is adopted to tune feature weights for translation systems.

We test our reordering model on a Chinese-English translation task. The NIST evaluation set MT06 was used as our development set to tune the feature weights, and the test data are MT04, MT 05 and MT08. We first conduct experiments by using the FBIS parallel training corpus, and then further test the effect of our method on a large scale parallel training corpus.

Word alignment is performed by GIZA++ (Och and Ney, 2000) in both directions with the default setting. The language model is a 4-gram model trained with the Xinhua portion of LDC English Gigaword Version 3.0 and the English part of the bilingual training data. Translation performances are measured with case-insensitive BLEU4 score (Papineni et al., 2002).

5.2 Experimental results on FBIS corpus

We first conduct experiments by using the FBIS parallel corpus to train the model of both the baseline and our lexicalized reordering model. After pre-processing, the statistics of FBIS corpus is shown in table 1.

	#sentences	#words
Chinese	128832	3016570
English	128832	3922816

Table 1. The statistics of FBIS corpus

Table 2 summarizes the translation performance. The first row shows the results of baseline HPB system, and the second row shows the results when we integrated our lexicalized reordering model (LRM). We get 1.2, 0.8 and 0.7 BLEU point improvements over the baseline HPB system on three test sets respectively.

	MT04	MT05	MT08
HPB	33.53	32.97	25.08
HPB+LRM	34.71	33.77	25.84

Table 2. Translation performance on the FBIS corpus.

5.3 Experimental results on large scale corpus

To further test the effect of our reordering model, we use a large scale corpus released by LDC. The catalog number of them is LDC2003E07, LDC2003E14, LDC2005T06, LDC2005T10, LDC2005E83, LDC2006E26, LDC2006E34, LDC2006E85 and LDC2006E92. There are 498K sentence pairs, 12.1M Chinese words and 13.8M English words. Table 3 summarizes the translation performance on the large scale of corpus.

	MT04	MT05	MT08
HPB	38.72	37.59	29.03
HPB+LRM	39.81	38.24	29.63

Table 3. Translation performance on a large scale parallel corpus.

Our model is still effective when we train the translation system on large scale data. We get 1.1, 0.7 and 0.6 BLEU point improvements over the baseline HPB system on three test sets respectively.

6 Conclusion and future work

We proposed a novel lexicalized reordering model for hierarchical phrase based machine translation. The model is compatible with any kind of HPB rules no matter how complex the alignments are. We tested our reordering model on both small and large scale data. On NIST machine translation test sets, our reordering model achieved a 0.6-1.2 BLEU point improvements for Chinese-English translation over a strong baseline hierarchical phrase-based system.

In future work, we will further test our model on other language pairs and compare it with other reordering models for HPB translation.

Acknowledgments

We thank anonymous reviewers for insightful comments. The work of Hailong Cao is sponsored by Microsoft Research Asia Star Track Visiting Young Faculty Program. The work of HIT is also funded by the project of National Natural Science Foundation of China (No. 61173073) and International Science & Technology Cooperation Program of China (No. 2014DFA11350).

Reference

Yaser Al-Onaizan and Kishore Papineni. 2006. Distortion Models for Statistical Machine Translation. In Proceedings of *ACL*.

Colin Cherry, Robert C. Moore and Chris Quirk. 2012. On Hierarchical Re-ordering and Permutation Parsing for Phrase-based Decoding. In Proceedings of *NAACL Workshop on SMT*.

David Chiang. 2007. Hierarchical Phrase-based Translation. Computational Linguistics, 33(2):201–228.

Michel Galley and Christopher D. Manning. 2008. A Simple and Effective Hierarchical Phrase Reordering Model. In Proceedings of *EMNLP*.

Katsuhiko Hayashi, Hajime Tsukada, Katsuhito Sudoh, Kevin Duh and Seiichi Yamamoto. 2010. Hierarchical Phrase-based Machine Translation with Word-based Reordering Model. In Proceedings of *COLING*.

Zhongjun He, Qun Liu, Shouxun Lin. 2008. Improving Statistical Machine Translation using Lexicalized Rule Selection. In Proceedings of *COLING*.

Liang Huang, Hao Zhang and Daniel Gildea. 2005. Machine Translation as Lexicalized Parsing with Hooks. In Proceedings of *IWPT*.

Zhongqiang Huang, Martin Čmejrek, and Bowen Zhou. 2010. Soft Syntactic Constraints for Hierarchical Phrase-based Translation Using Latent Syntactic Distributions. In Proceedings of *EMNLP*.

Matthias Huck, Joern Wuebker, Felix Rietig, and Hermann Ney. 2013. A Phrase Orientation Model for Hierarchical Machine Translation. In Proceedings of *ACL Workshop on SMT*.

Philipp Koehn, Hieu Hoang, Alexandra Birch Mayne, Christopher Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst.. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *ACL* demonstration session.

Masaaki Nagata, Kuniko Saito, Kazuhide Yamamoto and Kazuteru Ohashi. 2006. A Clustered Global Phrase Reordering Model for Statistical Machine Translation. In Proceedings of *ACL*.

Thuylinh Nguyen and Stephan Vogel. 2013. Integrating Phrase-based Reordering Features into Chart-based Decoder for Machine Translation. In Proceedings of *ACL*.

Franz Josef Och and Hermann Ney. 2000. Improved Statistical Alignment Models. In Proceedings of *ACL*.

Franz Josef Och and Hermann Ney. 2004. The Alignment Template Approach to Statistical Machine Translation. Computational Linguistics, 30(4):417–449.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In Proceedings of *ACL*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In Proceedings of *ACL*.

Christoph Tillmann. 2004. A Unigram Orientation Model for Statistical Machine Translation. In Proceedings of *HLT-NAACL*.

Roy Tromble, Jason Eisner. 2009. Learning Linear Ordering Problems for Better Translation. In Proceedings of *EMNLP*.

Xinyan Xiao, Jinsong Su, Yang Liu, Qun Liu, and Shouxun Lin. 2011. An Orientation Model for Hierarchical Phrase-based Translation. In Proceedings of *IALP*.

Deyi Xiong, Qun Liu, and Shouxun Lin. 2006. Maximum Entropy Based Phrase Reordering Model for Statistical Machine Translation. In Proceedings of *ACL*.

Richard Zens and Hermann Ney. 2006. Discriminative Reordering Models for Statistical Machine Translation. In Proceedings of Workshop on *SMT*.

Andreas Zollmann and Ashish Venugopal. 2006. Syntax Augmented Machine Translation via Chart Parsing. In Proceedings of *NAACL Workshop on SMT*.