Graph-based Lexicalized Reordering Models for Statistical Machine Translation

SU Jinsong[1], LIU Yang[2], LIU Qun[3], DONG Huailin[1]

[1] Xiamen University, Xiamen 361005, P. R. China
[2] Tsinghua University, Beijing 100084, P. R. China
[3] Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100084, P. R. China

Abstract: Lexicalized reordering models are very important components of phrase-based translation systems. By examining the reordering relationships between adjacent phrases, conventional methods learn these models from the word aligned bilingual corpus, while ignoring the effect of the number of adjacent bilingual phrases. In this paper, we propose a method to take the number of adjacent phrases into account for better estimation of reordering models. Instead of just checking whether there is one phrase adjacent to a given phrase, our method firstly uses a compact structure named reordering graph to represent all phrase segmentations of a parallel sentence, then the effect of the adjacent phrase number can be quantified in a forward-backward fashion, and finally incorporated into the estimation of reordering models. Experimental results on the NIST Chinese-English and WMT French-Spanish data sets show that our approach significantly outperforms the baseline method.

Key words: natural language processing; statistical machine translation; lexicalized reordering model; reordering graph

I. INTRODUCTION

Phrase-based translation systems [1-2] prove to be the state-of-the-art as they have delivered translation performance in recent machine translation evaluations. Compared with word-based translation systems, phrase-based translation systems extend translation unit from word to phrase, thus reducing the ambiguity and reordering at the lexical level. While excelling at memorizing local translation and reordering, phrase-based systems have difficulties in modeling permutations among phrases. As a result, it is important to develop effective reordering models to capture such non-local reordering.

In the earliest phrase-based translation system, the reordering model is a simple one in which the reordering probability is set manually depending on the translation language pairs. Then, the distance-based distortion penalty [1] is applied to better model the phrase movements in translation. Specifically, using this model, the decoder chooses to translate monotonically unless there is sufficient support for a jump from other features, for example, language model. Obviously, without the consideration of the specific content of phrases, phrase movements have not been well solved in the early phrase-based paradigm.

To deal with this problem, many researchers have presented lexicalized reordering models [3-7] that take advantage of lexical information to predict reordering. Instead of entirely ignoring the information of phras-

*This paper is an extended version of (Su et al., 2010) at ACL 2010

es, these models are learned from the word-aligned corpus to predict the orientations of a phrase pair with respect to the adjacent bilingual phrases.

In the popular lexicalized reordering models, there are generally three kinds of relationships between the selected phrase pair and its previous or following one: monotone (*M*), swap (*S*), and discontinuous (*D*). For example, in Figure 1(a), the word-based reordering model [3] analyzes the word alignments at positions (s-1, u-1) and (s-1, v+1), and sets the orientation of *bp* to *D* because the position (s-1, v+1) contains no word alignment. The phrase-based reordering models [4-6] determine the presence of the adjacent bilingual phrase located in position (s-1, v+1) and then treat the orientation of *bp* as *S*. Given no constraint on maximum phrase length, the hierarchical phrase reordering model [7] also analyzes the adjacent bilingual phrases for *bp* and identifies its orientation as *S*.

However, the above-mentioned models just consider the presence of an adjacent bilingual phrase and ignore the effect of the number of adjacent bilingual phrases on the reordering probability estimation. Here we continue with the examples in Figure 1 for illustration. In Figure 1(a), *bp* is in a swap order with only one bilingual phrase. In Figure 1(b), *bp* swaps with three bilingual phrases. Conventional lexicalized reordering models do not distinguish different numbers of adjacent phrase pairs, and just give *bp* the same count in the swap orientation. Intuitively, the more adjacent phrase pairs swap with *bp*, the larger probability *bp* has in the swap orientation. So we believe the conventional reordering model can be improved by distinguishing different numbers of adjacent phrase pairs.

In this paper, we propose a novel method to better estimate the reordering probability with the consideration of varying numbers of adjacent bilingual phrases. Our approach represents all phrase segmentations of a parallel sentence pair with one compact structure named reordering graph. Then, from the reordering graph, the fractional counts of bilingual phrases in different orientations can be efficiently calculated in a forward-backward fashion. To investigate the effectiveness of our method, we integrate the lexicalized reordering model based on reordering graphs into two famous phrase-based translation systems: MOSES [3] and maximal entropy-based bracketing transduction grammar (MEBTG) [6] model. The final experimental results indicate that our method achieves significant improvements over the conventional ones.

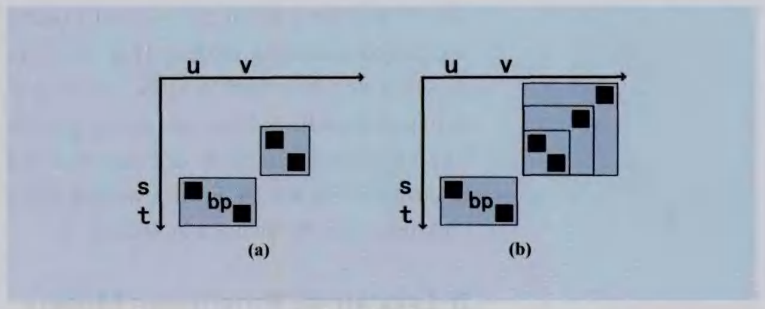

Figure 1: *Occurrence of a swap with different numbers of adjacent bilingual phrases: only one phrase in (a) and three phrases in (b). Black squares denote word alignments and gray rectangles denote bilingual phrases. [s,t] indicates the target-side span of bilingual phrase bp and [u,v] represents the source-side span of bilingual phrase bp.*

In this paper, we propose a method to take the number of adjacent phrases into account for better estimation of reordering models.

While we adopt the idea of the graph-based lexicalized reordering model from our previous work [8], we extend the whole paradigm in this paper, including modeling, evaluation and analysis, in three major aspects. First, we refine the establishment of reordering graphs by excluding invalid derivations. In our previous work, the phrase pairs are linked if they are adjacent in the target side order. However, this implementation leads to some inconsistent derivations. To solve this problem, we distinguish different derivations by keeping the source-side covered state in each node, and exclude the inconsistent derivations from reordering graphs. Secondly, in order to validate the generality and robustness of our approach, we integrate our model into two SMT systems and carry out more experiments across several different data sets. Finally, we investigate the effect of key factor on our method and conduct more in-depth analysis to experimental results.

This paper is organized as follows: in Sec-

tion II, we first give an overview of existing lexicalized reordering models. Then we introduce our novel method to estimate the reordering probabilities from reordering graphs. The experimental results are reported and discussed in Section III. Finally, we end with a conclusion and future work in Section IV.

II. Lexicalized Reordering Models in SMT

Lexicalized reordering models containing phrases with reordering probabilities in different orientations play a crucial role in phrase-based translation systems. To build a standard lexicalized reordering model, conventional studies complement in the two following procedures. Firstly, adjacent bilingual phrases with orientation information are identified as reordering examples. Meanwhile, the corresponding fractional counts are also gathered. After this process is finished, the reordering probabilities of phrases in different orientations are estimated by maximal likelihood estimation or classification model training. Since different models may identify the reordering relationship of the same adjacent phrases as different orientations, we briefly revisit the current three models: word-based model, phrase-based model and hierarchical model. To better illustrate these models, we still take the bilingual sentence shown in Figure 1(a) as example, and depict the orientation of *bp* using different models.

2.1 Word-based lexicalized reordering model

This model is a standard component in the famous phrase-based system MOSES. In this model, the orientation of *bp* is identified by determining the existence of word alignments at positions (s-1, u-1) and (s-1, v+1) in the alignment grid. If the position (s-1, u-1) contains a word alignment while (s-1, v+1) contains no word alignment, the orientation of *bp* is set to *M*. On the contrary, when the position (s-1, u-1) contains no word alignment while (s-1, v+1) contains a word alignment, the orientation of *bp* is set to *S*. In all other cases, the orientation is classified as *D*.

2.2 Phrase-based lexicalized reordering model

Unlike the mentioned-above model which determines the orientations based on specific word alignments, the phrase-based lexicalized reordering model [4-5] analyzes adjacent phrases. Back to the example shown in Figure 1(a), the orientation of *bp* is set to *M* if an adjacent phrase pair covers (s-1, u-1) in the alignment grid. Similarly, if an adjacent phrase pair lies at (s-1, v+1), the orientation of *bp* is set to *S*, and is set to *D* otherwise. Besides, the MEBTG system also adopts a similar phrase-based lexicalized reordering model to predict the relative orders of neighbor phrases. But unlike the above phrase-based model, it considers this problem as a binary classification task: monotone or swap, which is suitable to be solved by maximal entropy model.

2.3 Hierarchical lexicalized reordering model

Different from the previous two models, this model aims at improving non-local reordering and analyzes the alignments beyond adjacent phrases. Continuing the parallel sentence shown in Figure 1(a), we set the orientation of *bp* as *M* if there exists a bilingual block[1] covers the position (s-1, u-1). When a block lies in (s-1, v+1), the orientation of bilingual phrase is *S*, and orientation is *D* otherwise. One thing worth mentioning is that this model only applies to the phrase-based system decoding in the shift-reduce manner.

III. Graph-Based Lexicalized Reordering Model

In this section, we first describe the construction of the reordering graph representing all segmentations of a parallel sentence, and then illustrate how to efficiently learn more accurate reordering probabilities from the reordering graph.

[1]Each block is also a phrase pair but without the limitation of maximum phrase length.

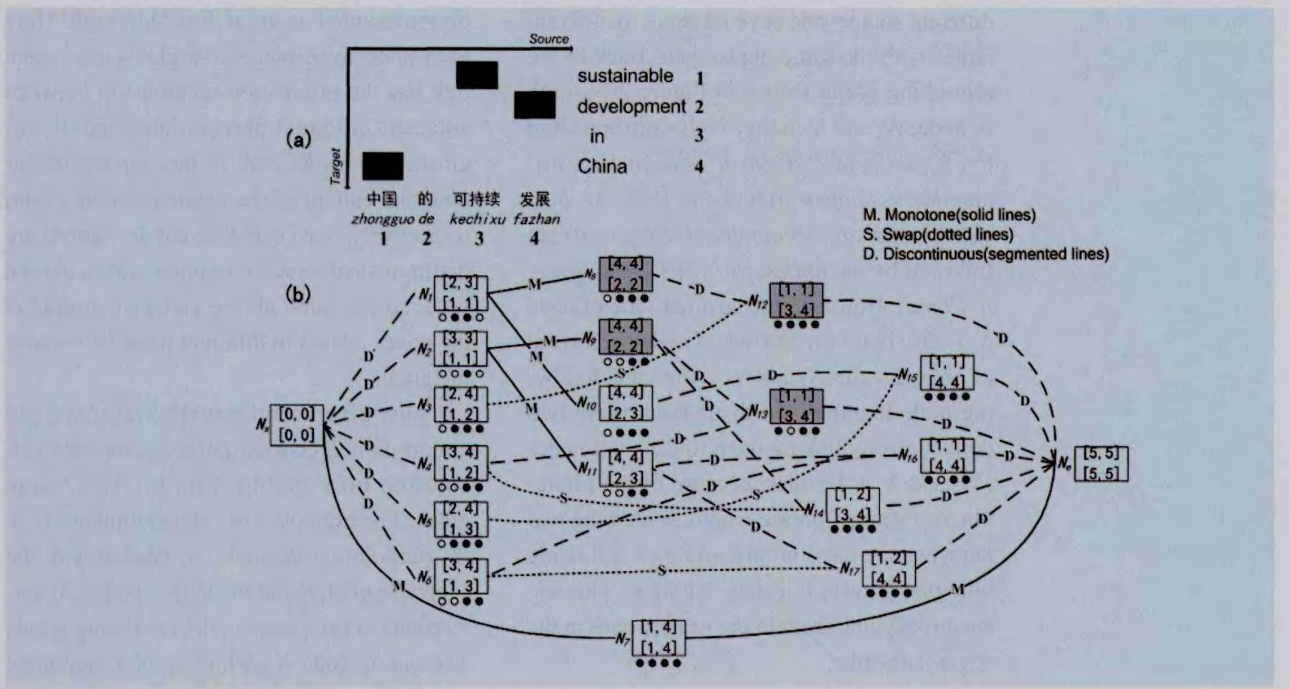

Fig.2: *(a) A parallel Chinese-English sentence pair and (b) its corresponding reordering graph. In (b), we denote each bilingual phrase with a rectangle (node), where the upper and bottom numbers in the brackets represent the source and target spans of this bilingual phrase, respectively. Besides, the dot sequence under the node denotes the source-side covered state of the derivation, where the black ones represent the corresponding source-side words have been included. Note that the null-aligned words may lead to different derivations, so the same bilingual phrase may occur in different nodes.*

3.1 Reordering graph

For a parallel sentence pair, its reordering graph indicates all possible translation derivations consisting of the extracted bilingual phrases. To construct a reordering graph, we first extract the consistent bilingual phrases using the conventional method [9]. Given a parallel sentence with word alignment, the extracted bilingual phrase is said to be consistent with the alignment if and only if: (1) there must be at least one word inside one phrase aligned to a word inside the other phrase; (2) no words inside one phrase can be aligned to a word outside the other phrase. Then, in terms of the relative order on the target side, the adjacent bilingual phrases are linked from left to right, forming nodes related to a bilingual phrase respectively.

The main drawback of the method above, however, is that it may lead to some inconsistent derivations, causing negative influence on the final estimation of reordering probabilities. This is because, in the phrase-based SMT systems, the source-side null-aligned words can be extracted by attaching to any neighboring words, and thus these words may be included by adjacent phrase pairs, which result in some inconsistent derivations. For example, in Figure 2, although the bilingual phrases (中国 */zhongguo* 的 */de*, *in China*) and (的 */de* 可持续 */kechixu* 发展 */fazhan*, *sustainable development*) are adjacent to each other in the target-side order, they are unable to form a consistent derivation because the source-side word ' 的 */de*' appears in both phrases.

To avoid this inconsistency, we distinguish different derivations by keeping the source-side covered state in each node. Here each covered state indicates which source-side words have been included by the partial derivation. Unlike our previous graph-based model [8] in which the adjacent phrase pairs are linked based solely on the target-side order, we respectively link the preceding nodes with

different source-side covered states to different nodes with the same phrase pair. Back to the reordering graph shown in Figure 2(b), both of nodes N_3 and N_4 respectively corresponding to (的 */de* 可持续 */kechixu* 发展 */fazhan, sustainable development*) and (可 持 续 */kechixu* 发 展 */fazhan, sustainable development*) are followed by the phrase pair (中国 */zhongguo, in China*). However, the covered state of node N_3 is different from the one of node N_4. Instead of linking nodes N_3 and N_4 to the same following node together, we distinguish these two derivations by linking them to different nodes (N_{12} and N_{13}). Besides, because of the limitation of maximal phrase length, some bilingual phrases may have no preceding or following bilingual phrase. For these bilingual phrases, we directly link them to the nearest ones in the target-side order.

As described above, our method distinguishes different derivations based on source-side covered state. However, the execution efficiency of our method will become poor when dealing with the long parallel sentence[2] with many null-aligned words. To solve this problem, we split long parallel sentence into short ones, each of which is used to construct a reordering graph. To be specific, for each parallel sentence, we first identify the commas, semicolons and colons, which are translated into themselves and belong to *rift points* proposed by Berge et al. [10], to form a set of split points. According to the split points, each parallel sentence can be spitted into a sequence of bilingual segments, where the reordering relationship of adjacent segments is always monotone. Meanwhile, it should be noted that the utilization of segments results in the loss of a few bilingual rules[3]. In order to reduce the discarded rules, here we keep combining the adjacent segments into larger one in a left-to-right fashion, until the generated segment becomes a long sub parallel sentence. Finally, we establish reordering graphs to learning reordering probabilities based on the generated bilingual segments.

Shown in Figure 2(b), the reordering graph for the parallel sentence pair (Figure 2(a)) can be represented as an undirected graph. Here each node corresponds to a phrase pair, each link has the orientation relationship between adjacent bilingual phrases, and two distinguished nodes *Ns* and *Ne* indicate the beginning and ending of the parallel sentence pair, respectively. Note that different derivations are distinguished based on source-side covered state, so the same phrase pair may appear in the nodes related to different paths of reordering graph.

Thus, given a bilingual phrase, we can obtain its neighboring phrase pairs with orientation information from the reordering graph. For example, the bilingual phrase (发 展 */fazhan, development in*) labeled with the source span [4,4] and the target span [2,3] corresponds to two nodes in the reordering graph. The one is node N_{10} which is in a monotone order with one previous phrase (node N_1) and in a discontinuous order with one subsequent phrase (nodes N_{15}); the other is node N_{11} which is in a monotone order with one previous phrase (node N_2) and in a discontinuous order with two subsequent phrases (nodes N_{16} and N_{17}).

3.2 Estimation of reordering probabilities

With the reordering graphs, we can learn more accurate reordering probabilities from them. Given a parallel sentence pair, there are many translation derivations corresponding to different paths in its reordering graph. Assuming all derivations have a uniform probability, the fractional counts of bilingual phrases for the orientations can be calculated by utilizing an algorithm in the forward-backward fashion.

Specifically, for the given phrase pair *bp* in the ordering graph, we first introduce two notations to represent the numbers of paths going through its related node N_{bp}:

1) $\alpha(N_{bp})$ is the number of paths from node N_s to N_{bp}, and it can be iteratively computed as $\alpha(N_{bp}) = \sum_{bp'} \alpha(N_{bp'})$, where bp' is one of the previous phrase pairs of *bp* and $\alpha(N_s) = 1$.

2) $\beta(N_{bp})$ denotes the number of paths

[2]In our work, we identify the parallel sentence containing more than 50 words and 12 null-aligned words in either side as the long parallel sentence.

[3]Statistically, in two experimental data sets described later, the number of the lost bilingual rules does not exceed 0.4% of whole bilingual rules.

from node N_e to N_{bp}, and it is simply $\beta(N_{bp}) = \sum_{bp''} \beta(N_{bp''})$, where bp'' is one of the subsequent phrase pairs of bp and $\beta(N_e) = 1$.

Continuing with the reordering graph shown in Figure 2(b), we show the final α and β values of all nodes in Table I.

After the calculation of the α and β values of all nodes, we assign different weights to the reordering examples. Inspired by the parsing literature on pruning [12], the fractional count of the reordering example (o, bp', bp) is

$$Count(o, bp', bp) = \sum_{N_{bp'}} \sum_{N_{bn}} \frac{\alpha(N_{bp'}) \cdot \beta(N_{bp})}{\beta(N_S)} \quad (1)$$

where the numerator indicates the number of paths containing the reordering example (o, bp', bp) and the denominator is the total number of paths in the reordering graph. Returning to the example in Figure 2, the reordering example consisting of (发 展 */fazhan*, *development*) and (中国 */zhongguo*, *in China*) marked in the gray color, appear in two derivations and its fractional count in the D orientation is $(\alpha(N_8) * \beta(N_{12}) + \alpha(N_9) * \beta(N_{13}))/\beta(N_e) = (1*1+1*1)/13 = 2/13$.

Next, we incorporate the fractional count of reordering examples into the establishment of lexicalized reordering models. In MOSES system, we apply the maximal likelihood estimation to calculate the final reordering probabilities of phrases. Formally, the fractional count of bp in the orientation o is calculated as described below:

$$Count(o, bp) = \sum_{bp'} Count(o, bp', bp) \quad (2)$$

In the implementation of conventional MEBTG system [6], the count of the reordering example (o, bp', bp) is set to its frequency number in training corpus. Here, we replace this count with its fractional count *Count*(o, bp', bp) in reordering graph for better estimation of maximal entropy based reordering model.

IV. EXPERIMENT

To investigate the effectiveness of our approach, we evaluate our method on the Chinese-to- English and English-to-Spanish translation tasks. After a brief description of the experimental setup, we report the experimental results and discuss the effects of various

Table I ***srcSpan*** = *source span, and* ***tgtSpan*** = *target span. The α and β values of the nodes shown in Figure 2(b).*

node	*srcSpan*	*tgtSpan*	α	β
N_s	[0, 0]	[0, 0]	1	13
N_1	[2,3]	[1,1]	1	2
N_2	[3,3]	[1,1]	1	4
N_3	[2,4]	[1,2]	1	1
N_4	[3,4]	[1,2]	1	2
N_5	[2,4]	[1,3]	1	1
N_6	[3,4]	[1,3]	1	2
N_7	[1,4]	[1,4]	1	1
N_8	[4,4]	[2,2]	1	1
N_9	[4,4]	[2,2]	1	2
N_{10}	[4,4]	[2,3]	1	1
N_{11}	[4,4]	[2,3]	1	2
N_{12}	[1,1]	[3,4]	2	1
N_{13}	[1,1]	[3,4]	2	1
N_{14}	[1,2]	[3,4]	2	1
N_{15}	[1,1]	[4,4]	2	1
N_{16}	[1,1]	[4,4]	2	1
N_{17}	[1,2]	[4,4]	2	1
N_e	[5,5]	[5,5]	13	1

factors on the proposed method.

4.1 Experiment setup

To comprehensively investigate the generality of our method, we carry out experiments on two data sets. In the experiment with the first data set, our training corpus comes from LDC[4], which mainly consists of the FBIS corpus and the Hansards part of LDC2004T07 corpus. The 2002 NIST MT evaluation test data is used as the development set and the 2003, 2004, 2005 NIST MT test data are the test sets. Besides, we use SRILM Toolkits [13] to train a 4-gram language model on the Xinhua portion of Gigaword corpus (181.1M words). In the experiment with the second data set, the training data come from the French-Spanish part of Europarl corpus[5]. We use the same in-domain development set and test set provided by the shared task of NAACL/HLT 2006 Workshop on SMT[6]. Each sentence in the sets is with single reference. Likewise, we directly use the 3-gram language model which is also provided by the shared task. Table II and III show the statistics of various data sets.

As for the processing of various data sets, we firstly use the open-source toolkit ICTCLAS [14] to segment the Chinese sentences of LDC data, and use the token script provided by the shared task to split the data in other languages. Then, GIZA++ [15] and the heuristics "grow-diag-final-and" are used to generate a word-aligned corpus. Finally, different methods are applied to extract phrase pairs and reordering examples, depending on the SMT system.

Our method is generic for SMT systems with lexicalized reordering model, so we carry out experiments using two translation systems: one is a famous open-source translation system MOSES; the other is the MEBTG system which is based on the weighted synchronous context free grammar without explicit linguistic syntactic knowledge. Both systems are widely applied and deliver good performance in various machine translation evaluations.

In the experiment of MOSES system, we investigate the effect of our method under two conditions: the maximal length of extract phrase pairs is set to 5 or 7. As we all know, there are six word-based lexical reordering models used in MOSES, while we only focus on the msd-fe and msd-bidirectional-fe reordering models which are widely applied in MOSES. The former has three features representing the probabilities of bilingual phrases in three orientations: monotone, swap, and discontinuous. If the latter is used, then the number of features doubles: one for each direction.

To build MEBTG system, we also try different maximal lengths to extract phrase pairs: 5 or 7, meanwhile, we respectively collect reordering examples using the conventional method and our approach. Following (Xiong et al., 2006) [6] which has indicated the effectiveness of boundary words in reordering

Table II *Data sets of the LDC experiment.*

Data Sets	Source	Sentence Pairs	Source Words	Target Words
Training Set	LDC	1,039,140	25,232,267	29,047,725
Development Set	NIST02	878	23,181	94,774
Test Set	NIST03	919	24,997	101,129
	NIST04	1,788	50,392	210,703
	NIST05	1,082	30,602	123,474

Table III *Data sets of the WMT experiment.*

Data Sets	Source	Sentence Pairs	Source Words	Target Words
Training Set	WMT	1,653,130	48,930,228	46,229,158
Development Set	NAACL/HLT 2006 Workshop	2000	63,265	60,233
Test Set	NAACL/HLT 2006 Workshop	2000	65,097	61,774

[4]Available at: http://www.ldc.upenn.edu/.

[5]Available at: http://www.statmt.org/.

[6]Available at http://www.statmt.org/wmt06/shared-task/baseline.html.

models, we also use four boundary words to capture the phrase movement in our model: the last source and target words of left bilingual phrase, the last source and target words of right bilingual phrase. Then, we adopt the maximal entropy tool [16] developed by Zhang to train reordering model with the following parameters: iteration number i=100 and Gaussian prior g=1.0.

During decoding, we set the same pruning parameters for two systems. To be specific, we set ttable-limit = 20 to keep translation candidates for each source phrase, stack-size = 30 to prune hypotheses for each span. In exception to the reordering probabilities, we use the same features as the baseline systems in the comparative experiments. For the weights of various features, we perform minimum-error-rate training [17] to maximize the BLEU score on the development set. The translation quality is evaluated by case-insensitive BLEU-4 metric [18]. Finally, we conduct paired bootstrap sampling [19] to test the significance in BLEU score differences.

4.2 Experimental results

4.2.1 Effect of models based on reordering graphs

Our first experiment investigates the effect of the graph-based lexicalized reordering models in two systems. We compare the performance of system using the conventional reordering models with the proposed graph-based models.

Table IV gives the experimental results in the LDC data set. Here we mainly focus on the average BLEU scores on the three test sets.

When the maximal phrase length is set as 5, the average BLEU scores of three baseline systems are 31.85, 32.63 and 32.54, respectively. Then, we replace the conventional models with the graph-based ones, and find that our new graph-based reordering model performs better than other models including baseline and old graph-based one. More specifically, the average BLEU scores using new graph-based method are 32.60, 33.15 and

Table IV. *Experimental results in the LDC data set.* ***RG-old*** *= Graph-based lexicalized reordering model represented in (Su et al., 2010) [8],* ***RG-new*** *= Graph-based lexicalized reordering model represented in this paper. "**MSD-FE RM Model**" and "**MSD-BI-FE RM Model**" denote msd-fe and msd-bi-fe reordering models, respectively. "**Max Bp Length**" is the maximal length of the extracted bilingual phrase. "**Avg**" is the average BLEU score on the three test sets. * or **: significantly better than baseline ($p < 0.05$ or $p < 0.01$).*

System	Method	MT-03	MT-04	MT-05	AVG
MOSES	Baseline	32.06	32.45	31.05	31.85
MSD-FE RM Model	RG-old	32.41	33.70**	31.61**	32.57
Max Bp Length = 5	RG-new	32.36	33.77**	31.68**	32.60
MOSES	Baseline	33.06	33.36	31.47	32.63
MSD-BI-FE RM Model	RG-old	32.98	33.93**	31.91*	32.94
Max Bp Length = 5	RG-new	32.85	34.51**	32.08**	33.15
	Baseline	31.96	33.69	31.97	32.54
MEBTG	RG-old	32.39*	34.02	32.06	32.82
Max Bp Length = 5	RG-new	33.11**	34.35**	32.20	33.22
MOSES	Baseline	33.04	32.41	30.50	31.98
MSD-FE RM Model	RG-old	33.30	32.95*	30.87	32.37
Max Bp Length = 7	RG-new	33.22	33.05**	31.20**	32.49
MOSES	Baseline	33.01	33.52	31.40	32.64
MSD-BI-FE RM Model	RG-old	33.83**	33.44	31.89*	33.05
Max Bp Length = 7	RG-new	33.97**	33.95*	32.04**	33.32
	Baseline	32.34	33.90	31.83	32.69
MEBTG	RG-old	32.82*	34.29*	32.41**	33.17
Max Bp Length = 7	RG-new	33.00**	34.22	32.32*	33.18

33.22, achieving absolute improvements of 0.75, 0.52 and 0.68 over baseline systems, respectively.

When we extract bilingual phrase with maximal length 7 to build SMT systems, the average BLEU scores of three baseline systems are 31.98, 32.64 and 32.69, respectively. By using our two graph-based methods to build lexi-calized reordering models, all systems perform better than the corresponding baseline. Overall, the new graph-based method obtains slight improvements than the old one for three systems. To be specific, the average BLEU scores by our new graph-based method are 32.49, 33.32 and 33.18, achieving absolute improvements of 0.51, 0.68 and 0.49 on three test sets, respectively.

The experimental results of the WMT data set are shown in Table IIV. These results are similar to the ones of previous experiment.

When the maximal phrase length is set as 5, the BLEU scores of three baseline systems are 51.62, 51.90 and 52.33, respectively. Using our two methods to incorporate the effect of the adjacent phrase number into reordering models, we bring different levels of improvements to system performance. Especially, when we use the new graph-based method, the MOSES system with msd-fe and msd-bi-fe reordering models achieve better performance. On the test set, the BLEU scores of three systems under this condition are 52.20, 52.43 and 52.73, respectively, achieving 0.58, 0.53 and 0.40 improvements than the baseline, all of which are significant by using paired bootstrap sampling.

When the maximal length of phrase pairs increases to 7, our new graph-based method also yields best performance. Specifically, the BLEU scores acquired by three SMT systems with new graph-based lexicalized models are 58.72, 58.78 and 60.86, respectively. These scores respectively obtain and 0.76, 0.57, and 0.43 points higher than the corresponding baseline, the scores of which are 57.96, 58.21 and 60.43, respectively.

From the above experimental results, we know that all systems with graph-based reordering models perform better than the baseline systems. This finding strongly proves that it is helpful to learn more accurate reordering probabilities by distinguishing the number of adjacent bilingual phrases. Besides, in most cases, the lexicalized reordering model based on reordering graphs seems to be more effective in MOSES than MEBTG system. The underlying reason is that the conventional reordering model utilized by MOSES is a word-based one, which is simpler than the phrase-based model, thus it can be improved more significantly by our method.

Table V *Experimental results in the WMT data set.*

System	Method	Test
MOSES	Baseline	51.62
MSD-FE RM Model	RG-old	52.07*
Max Bp Length = 5	RG-new	52.20**
MOSES	Baseline	51.90
MSD-BI-FE RM Model	RG-old	52.14
Max Bp Length = 5	RG-new	52.43**
	Baseline	52.33
MEBTG	RG-old	52.80*
Max Bp Length = 5	RG-new	52.73*
	Baseline	57.96
MOSES	RG-old	58.55**
MSD-FE RM Model	RG-new	58.72**
Max Bp Length = 7	Baseline	58.21
MOSES	RG-old	58.53
MSD-BI-FE RM Model	RG-new	58.78*
Max Bp Length = 7	Baseline	60.43
	RG-old	60.87*
MEBTG	RG-new	60.86*
Max Bp Length = 7		

4.2.2 Effect of maximal phrase length

In the experiments above, we construct reordering graphs using all the extracted bilingual phrases, thus the derivation number of reordering graph is limited by the maximal length of extracted phase pairs. On one hand, a much lower setting of the maximal phrase length may make many possible derivations be excluded from reordering graphs, resulting in the low coverage of bilingual phrases and inaccurate reordering probabilities. On the other hand, a much larger setting of the maximal phrase length may be counterproductive,

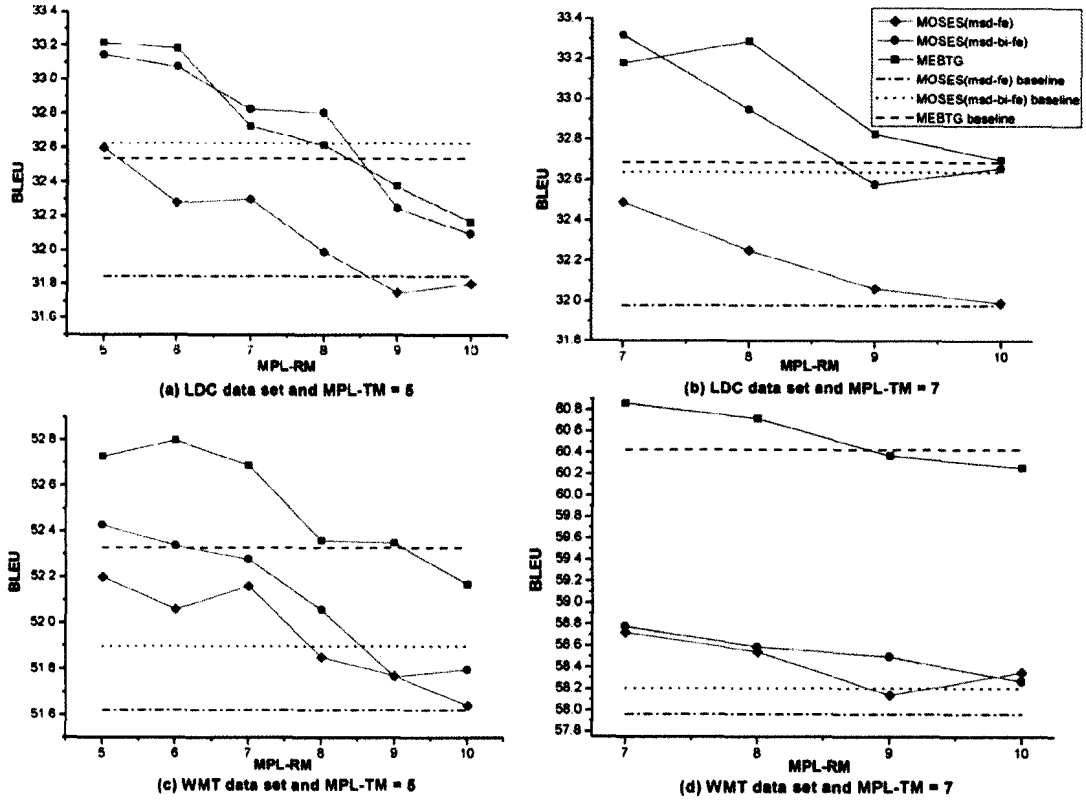

Figure 3: *The changes of BLEU score with the increment of the maximal phrase length of reordering models.* **MPL-TM** = *maximal phrase length of translation models.* **MPL-RM** = *maximal phrase length of reordering models.*

because reordering graphs may contain large number of derivations, which are simply assumed to have a uniform probability in our method. Intuitively, every derivation of reordering graph should have unequal probability, thus this complementation may cause negative effect on reordering models. Therefore, the maximal phrase length has a great effect on the graph-based lexical reordering model, and it makes sense to investigate the effect of maximal phrase length on our method.

In this experiment, we enlarge the maximal phrase length to enable reordering graphs to contain more derivations, and study the effect of maximal phrase length on reordering models by the change of BLEU score. Note that larger phrase length allows us to extract more phrase pairs. To ensure a fair comparison, we keep only the bilingual phrase pairs whose length is not exceed 5 or 7 in the final phrase table depending on the specific setting of translation model.

To avoid confusion, we rename the parameter of maximal phrase length depending on the specific models: one is **maximal phrase length of translation model (MPL-TM)**, which is set as 5 or 7 and is used to limit the maximal length of extracted bilingual phrase in translation model; the other is **maximal phrase length of reordering model (MPL-RM)**, which is greater than or equal to MPL-TM and is used to limit the maximal length of bilingual phrase within reordering graph. For example, in Figure 3(a), the MPL-TM is set as 5, and we try different MPL-RMs to build reordering graphs: from 5 to 10 with an increment of 1 each time.

Figure 3 shows the average BLEU scores of systems under four conditions. From these scores, we can see that our approach outperforms the conventional method in most cases, although the performances of SMT systems

are a bit unstable. However, when the MPL-RM is greater than 8, the performances of the systems with our new graph-based reordering models degrade significantly. For this experimental result, we speculate that with the further increment of phrase length, more long bilingual phrases which are actually rare in training corpus are used to establish reordering graphs, leading to more bias in the estimation of reordering probabilities.

V. Conclusion and Future Work

In this paper, we propose a method to improve the reordering model by considering the effect of the number of adjacent bilingual phrases on the reordering probabilities estimation. Experimental results on Chinese-to- English and English-to-Spanish tasks demonstra- te the effectiveness of our method.

In future, we will continue our research in the following aspects. First, our method is also general to other lexicalized reordering models, thus we plan to apply this method to the hierarchical phrase reordering model [7]. Second, we simply assume that all derivations have equal probabilities in our approach. This assumption has a negative effect on reordering model, so how to further improve the reordering model by distinguishing the derivations with different probabilities will become another study emphasis in further research. Finally, the state-of-the-art SMT has focused on the application of syntax information, so we will focus on how to integrate syntax information into our graph-based reordering model in future work.

Acknowledgements

The authors would like to thank the reviewers for their detailed reviews and constructive comments, which have helped improve the quality of this paper. This work was supported by the National Natural Science Foundation of China (No. 61303082), the Research Fund for the Doctoral Program of Higher Education of China (No. 20120121120046).

References

[1] Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. [C] Proceedings of HLT-NAACL 2003, pages 127–133.

[2] Franz Joseph Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. [J] Computational Linguistics, 30(4), pages 417–449.

[3] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris 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. [C] Proceedings of ACL 2007, Demonstration Session, pages 177–180.

[4] Christoph Tillmann. 2004. A unigram orientation model for statistical machine translation. [C] Proceedings of HLTACL 2004, pages 101–104.

[5] Richard Zens and Hermann Ney. 2006. Discriminative reordering models for statistical machine translation. [C] Proceedings of Workshop on Statistical Machine Translation 2006, pages 521–528.

[6] Deyi Xiong, Qun Liu, and Shouxun Lin. 2006. Maximum entropy based phrase reordering model for statistical machine translation. [C] Proceedings of ACL 2006, pages 521–528.

[7] Michel Galley and Christopher D. Manning. 2008. A simple and effective hierarchical phrase reordering model. [C] Proceedings of EMNLP 2008, pages 848–856.

[8] Jinsong Su, Yang Liu, Yajuan Lv, Haitao Mi, and Qun Liu. 2010. Learning lexicalized reordering models from reordering graphs. [C] Proceedings of ACL 2010, pages 12–16.

[9] Franz Joseph Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. [J] Computational Linguistics, 29(1), pages 19–51.

[10] A. L. Berger, S. A. D. Pietra, and V. J. D. Pietra. 1996. A maximum entropy approach to natural language processing. [J] Computational Linguistics, 22(1), pages 39–71.

[11] Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and maxent discriminative reranking. [C] Proceedings of ACL 2005, pages 173–180.

[12] Liang Huang. 2008. Forest reranking: Discriminative parsing with non-local features. [C] Proceedings of ACL 2008, pages 586–594.

[13] Andreas Stolcke. 2002. Srilm - an extensible language modeling toolkit. [C] Proceedings of ICSLP 2002, pages 901–904.

[14] Hua-Ping Zhang, Hong-Kui Yu, De-Yi Xiong and Qun Liu. 2003. HHMM-based Chinese Lexical Analyzer ICTCLAS, Proceedings of 2nd SigHan Workshop, pages 184-187.

[15] Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment model. [C] Proceedings of ACL 2000, pages 440–447.
[16] Le Zhang. 2004. Maximum entropy modeling toolkit for python and c++.
[17] Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. [C] Proceedings of ACL 2003, pages 160–167.
[18] Kishore Papineni, Salim Roukos, Todd Ward, and Weijing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. [C] Proceedings of ACL 2002, pages 311–318.
[19] Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. [C] Proceedings of EMNLP 2004, pages 388–395.

Biographies

SU Jinsong, received the Ph.D degree in computer application technology from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2011. He is now an assistant professor in the Software School of Xiamen University. His research interests include statistical machine translation and metaphor computing.

LIU Yang, received the Ph.D degree in computer application technology from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2007. He is now an associate professor in Department of Computer Science, Tsinghua University. His research interests include statistical machine translation and social computing.

LIU Qun, professor with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. His current research interests include computational linguistics, statistical and corpus-based natural language processing, including statistical machine translation, segmentation and parsing.

DONG Huailin, received the B.S degree in mathematics from Xiamen University, Xiamen, China, in 1980. He is now a professor in the Software School of Xiamen University. His research interests include data mining and software engineering.