Learning Lexicalized Reordering Models from Reordering Graphs

Jinsong Su, Yang Liu, Yajuan Lü, Haitao Mi, Qun Liu
Key Laboratory of Intelligent Information Processing
Institute of Computing Technology
Chinese Academy of Sciences
P.O. Box 2704, Beijing 100190, China
{sujinsong,yliu,lvyajuan,htmi,liuqun}@ict.ac.cn

Abstract

Lexicalized reordering models play a crucial role in phrase-based translation systems. They are usually learned from the word-aligned bilingual corpus by examining the reordering relations of adjacent phrases. Instead of just checking whether there is one phrase adjacent to a given phrase, we argue that it is important to take the number of adjacent phrases into account for better estimations of reordering models. We propose to use a structure named *reordering graph*, which represents all phrase segmentations of a sentence pair, to learn lexicalized reordering models efficiently. Experimental results on the NIST Chinese-English test sets show that our approach significantly outperforms the baseline method.

1 Introduction

Phrase-based translation systems (Koehn et al., 2003; Och and Ney, 2004) prove to be the state-of-the-art as they have delivered translation performance in recent machine translation evaluations. 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.

The early phrase-based paradigm (Koehn et al., 2003) applies a simple distance-based distortion penalty to model the phrase movements. More recently, many researchers have presented lexicalized reordering models that take advantage of lexical information to predict reordering (Tillmann, 2004; Xiong et al., 2006; Zens and Ney, 2006; Koehn et al., 2007; Galley and Manning, 2008). These models are learned from a word-aligned corpus to predict three orientations of a phrase pair with respect to the previous bilingual phrase: monotone (M), swap (S), and discontinuous (D). Take the bilingual phrase *bp* in Figure 1(a) for example. The word-based reordering model (Koehn et al., 2007) analyzes the word alignments at positions $(s-1, u-1)$ and $(s-1, v+1)$. The orientation of *bp* is set to D because the position $(s-1, v+1)$ contains no word alignment. The phrase-based reordering model (Tillmann, 2004) determines the presence of the adjacent bilingual phrase located in position $(s-1, v+1)$ and then treats the orientation of *bp* as S. Given no constraint on maximum phrase length, the hierarchical phrase reordering model (Galley and Manning, 2008) also analyzes the adjacent bilingual phrases for *bp* and identifies its orientation as S.

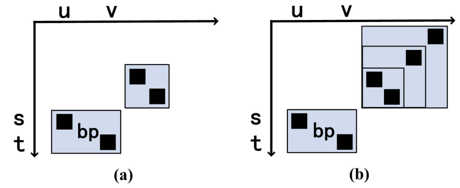

Figure 1: Occurrence of a swap with different numbers of adjacent bilingual phrases: only one phrase in (a) and three phrases in (b). Black squares denote word alignments and gray rectangles denote bilingual phrases. [s,t] indicates the target-side span of bilingual phrase *bp* and [u,v] represents the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

However, given a bilingual phrase, the above-mentioned models just consider the presence of an adjacent bilingual phrase rather than the number of adjacent bilingual phrases. See the examples in Fig-

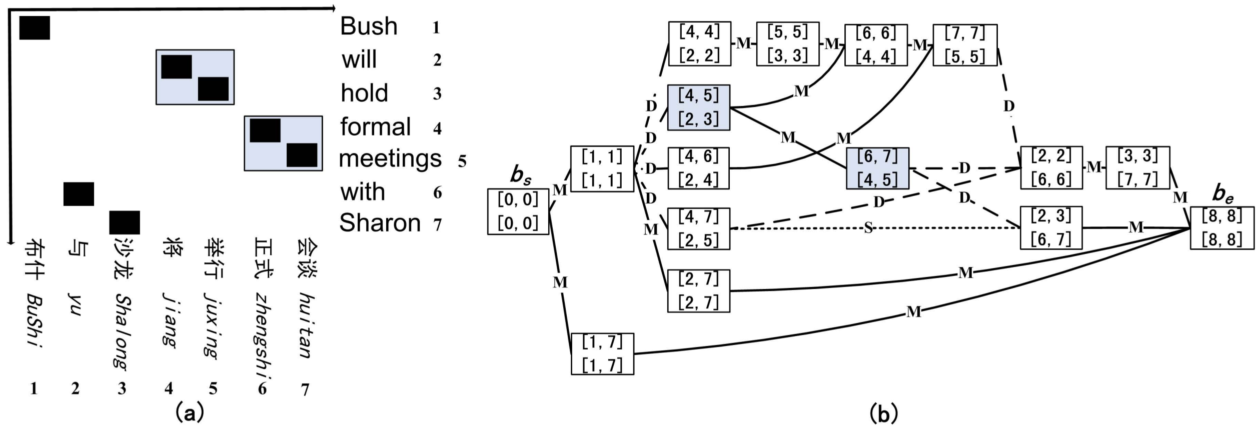

Figure 2: (a) A parallel Chinese-English sentence pair and (b) its corresponding reordering graph. In (b), we denote each bilingual phrase with a rectangle, where the upper and bottom numbers in the brackets represent the source and target spans of this bilingual phrase respectively. M = monotone (solid lines), S = swap (dotted line), and D = discontinuous (segmented lines). The bilingual phrases marked in the gray constitute a reordering example.

ure 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. Lexicalized reordering models do not distinguish different numbers of adjacent phrase pairs, and just give *bp* the same count in the swap orientation.

In this paper, we propose a novel method to better estimate the reordering probabilities with the consideration of varying numbers of adjacent bilingual phrases. Our method uses reordering graphs to represent all phrase segmentations of parallel sentence pairs, and then gets the fractional counts of bilingual phrases for orientations from reordering graphs in an inside-outside fashion. Experimental results indicate that our method achieves significant improvements over the traditional lexicalized reordering model (Koehn et al., 2007).

This paper is organized as follows: in Section 2, we first give a brief introduction to the traditional lexicalized reordering model. Then we introduce our method to estimate the reordering probabilities from reordering graphs. The experimental results are reported in Section 3. Finally, we end with a conclusion and future work in Section 4.

2 Estimation of Reordering Probabilities Based on Reordering Graph

In this section, we first describe the traditional lexicalized reordering model, and then illustrate how to construct reordering graphs to estimate the reordering probabilities.

2.1 Lexicalized Reordering Model

Given a phrase pair $bp = (\overline{e}_i, \overline{f}_{a_i})$, where a_i defines that the source phrase $\overline{f}_{a_i}$ is aligned to the target phrase $\overline{e}_i$, the traditional lexicalized reordering model computes the reordering count of *bp* in the orientation o based on the word alignments of boundary words. Specifically, the model collects bilingual phrases and distinguishes their orientations with respect to the previous bilingual phrase into three categories:

$$o = \begin{cases} M & a_i - a_{i-1} = 1 \\ S & a_i - a_{i-1} = -1 \\ D & |a_i - a_{i-1}| \neq 1 \end{cases} \qquad (1)$$

Using the relative-frequency approach, the reordering probability regarding *bp* is

$$p(o|bp) = \frac{Count(o, bp)}{\sum_{o'} Count(o', bp)} \qquad (2)$$

2.2 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 bilingual phrases using the way of (Och, 2003). Then, the adjacent bilingual phrases are linked according to the target-side order. Some bilingual phrases, which have no adjacent bilingual phrases because of maximum length limitation, are linked to the nearest bilingual phrases in the target-side order.

Shown in Figure 2(b), the reordering graph for the parallel sentence pair (Figure 2(a)) can be represented as an undirected graph, where each rectangle corresponds to a phrase pair, each link is the orientation relationship between adjacent bilingual phrases, and two distinguished rectangles b_s and b_e indicate the beginning and ending of the parallel sentence pair, respectively. With the reordering graph, we can obtain all reordering examples containing the given bilingual phrase. For example, the bilingual phrase ⟨*zhengshi huitan*, formal meetings⟩ (see Figure 2(a)), corresponding to the rectangle labeled with the source span [6,7] and the target span [4,5], is in a monotone order with one previous phrase and in a discontinuous order with two subsequent phrases (see Figure 2(b)).

2.3 Estimation of Reordering Probabilities

We estimate the reordering probabilities from reordering graphs. 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 orientations can be calculated by utilizing an algorithm in the inside-outside fashion.

Given a phrase pair *bp* in the reordering graph, we denote the number of paths from b_s to *bp* with $\alpha(bp)$. It can be computed in an iterative way $\alpha(bp) = \sum_{bp'} \alpha(bp')$, where *bp′* is one of the previous bilingual phrases of *bp* and $\alpha(b_s)$=1. In a similar way, the number of paths from b_e to *bp*, notated as $\beta(bp)$, is simply $\beta(bp) = \sum_{bp''} \beta(bp'')$, where *bp″* is one of the subsequent bilingual phrases of *bp* and $\beta(b_e)$=1. Here, we show the α and β values of all bilingual phrases of Figure 2 in Table 1. Especially, for the reordering example consisting of the bilingual phrases bp_1=⟨*jiang juxing*, will hold⟩ and bp_2=⟨*zhengshi huitan*, formal meetings⟩, marked in the gray color in Figure 2, the α and β values can be calculated: $\alpha(bp_1)$ = 1, $\beta(bp_2)$ = 1+1 = 2, $\beta(b_s)$ = 8+1 = 9.

Inspired by the parsing literature on pruning (Charniak and Johnson, 2005; Huang, 2008), the fractional count of (o, bp', bp) is

src span	*trg span*	α	β
[0, 0]	[0, 0]	1	9
[1, 1]	[1, 1]	1	8
[1, 7]	[1, 7]	1	1
[4, 4]	[2, 2]	1	1
[4, 5]	[2, 3]	1	3
[4, 6]	[2, 4]	1	1
[4, 7]	[2, 5]	1	2
[2, 7]	[2, 7]	1	1
[5, 5]	[3, 3]	1	1
[6, 6]	[4, 4]	2	1
[6, 7]	[4, 5]	1	2
[7, 7]	[5, 5]	3	1
[2, 2]	[6, 6]	5	1
[2, 3]	[6, 7]	2	1
[3, 3]	[7, 7]	5	1
[8, 8]	[8, 8]	9	1

Table 1: The α and β values of the bilingual phrases shown in Figure 2.

$$Count(o, bp', bp) = \frac{\alpha(bp') \cdot \beta(bp)}{\beta(b_s)} \tag{3}$$

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. Continuing with the reordering example described above, we obtain its fractional count using the formula (3): $Count(M, bp_1, bp_2) = (1 \times 2)/9 = 2/9$.

Then, the fractional count of *bp* in the orientation o is calculated as described below:

$$Count(o, bp) = \sum_{bp'} Count(o, bp', bp) \tag{4}$$

For example, we compute the fractional count of bp_2 in the monotone orientation by the formula (4): $Count(M, bp_2) = 2/9$.

As described in the lexicalized reordering model (Section 2.1), we apply the formula (2) to calculate the final reordering probabilities.

3 Experiments

We conduct experiments to investigate the effectiveness of our method on the **msd-fe** reordering model and the **msd-bidirectional-fe** reordering model. These two models are widely applied in phrase-based system (Koehn et al., 2007). The msd-fe reordering model has three features, which represent the probabilities of bilingual phrases in three orientations: monotone, swap, or discontinuous. If a msd-bidirectional-fe model is used, then the number of features doubles: one for each direction.

3.1 Experiment Setup

Two different sizes of training corpora are used in our experiments: one is a small-scale corpus that comes from FBIS corpus consisting of 239K bilingual sentence pairs, the other is a large-scale corpus that includes 1.55M bilingual sentence pairs from LDC. 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. We choose the MOSES[1] (Koehn et al., 2007) as the experimental decoder. GIZA++ (Och and Ney, 2003) and the heuristics "grow-diag-final-and" are used to generate a word-aligned corpus, where we extract bilingual phrases with maximum length 7. We use SRILM Toolkits (Stolcke, 2002) to train a 4-gram language model on the Xinhua portion of Gigaword corpus.

In exception to the reordering probabilities, we use the same features in the comparative experiments. During decoding, we set ttable-limit = 20, stack = 100, and perform minimum-error-rate training (Och, 2003) to tune various feature weights. The translation quality is evaluated by case-insensitive BLEU-4 metric (Papineni et al., 2002). Finally, we conduct paired bootstrap sampling (Koehn, 2004) to test the significance in BLEU scores differences.

3.2 Experimental Results

Table 2 shows the results of experiments with the small training corpus. For the msd-fe model, the BLEU scores by our method are 30.51 32.78 and 29.50, achieving absolute improvements of **0.89, 0.66** and **0.62** on the three test sets, respectively. For the msd-bidirectional-fe model, our method obtains BLEU scores of 30.49 32.73 and 29.24, with absolute improvements of **1.11, 0.73** and **0.60** over the baseline method.

model	method	MT-03	MT-04	MT-05
m-f	baseline	29.62	32.12	28.88
	RG	30.51^{**}	32.78^{**}	29.50^{*}
m-b-f	baseline	29.38	32.00	28.64
	RG	30.49^{**}	32.73^{**}	29.24^{*}

Table 2: Experimental results with the **small-scale** corpus. m-f: msd-fe reordering model. m-b-f: msd-bidirectional-fe reordering model. RG: probabilities estimation based on Reordering Graph. * or **: significantly better than baseline ($p < 0.05$ *or* $p < 0.01$).

model	method	MT-03	MT-04	MT-05
m-f	baseline	31.58	32.39	31.49
	RG	32.44^{**}	33.24^{**}	31.64
m-b-f	baseline	32.43	33.07	31.69
	RG	33.29^{**}	34.49^{**}	32.79^{**}

Table 3: Experimental results with the **large-scale** corpus.

Table 3 shows the results of experiments with the large training corpus. In the experiments of the msd-fe model, in exception to the MT-05 test set, our method is superior to the baseline method. The BLEU scores by our method are 32.44, 33.24 and 31.64, which obtain **0.86, 0.85** and **0.15** gains on three test set, respectively. For the msd-bidirectional-fe model, the BLEU scores produced by our approach are 33.29, 34.49 and 32.79 on the three test sets, with **0.86, 1.42** and **1.1** points higher than the baseline method, respectively.

4 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 NIST Chinese-to-English tasks demonstrate the effectiveness of our method.

Our method is also general to other lexicalized reordering models. We plan to apply our method to the complex lexicalized reordering models, for example, the hierarchical reordering model (Galley and Manning, 2008) and the MEBTG reordering model (Xiong et al., 2006). In addition, how to further improve the reordering model by distinguishing the derivations with different probabilities will become another study emphasis in further research.

[1] The phrase-based lexical reordering model (Tillmann, 2004) is also closely related to our model. However, due to the limit of time and space, we only use Moses-style reordering model (Koehn et al., 2007) as our baseline.

Acknowledgement

The authors were supported by National Natural Science Foundation of China, Contracts 60873167 and 60903138. We thank the anonymous reviewers for their insightful comments. We are also grateful to Hongmei Zhao and Shu Cai for their helpful feedback.

References

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proc. of ACL 2005*, pages 173–180.

Michel Galley and Christopher D. Manning. 2008. A simple and effective hierarchical phrase reordering model. In *Proc. of EMNLP 2008*, pages 848–856.

Liang Huang. 2008. Forest reranking: Discriminative parsing with non-local features. In *Proc. of ACL 2008*, pages 586–594.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proc. of HLT-NAACL 2003*, pages 127–133.

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. In *Proc. of ACL 2007, Demonstration Session*, pages 177–180.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proc. of EMNLP 2004*, pages 388–395.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Franz Joseph Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, pages 417–449.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. of ACL 2003*, pages 160–167.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. of ACL 2002*, pages 311–318.

Andreas Stolcke. 2002. Srilm - an extensible language modeling toolkit. In *Proc. of ICSLP 2002*, pages 901–904.

Christoph Tillmann. 2004. A unigram orientation model for statistical machine translation. In *Proc. of HLT-ACL 2004, Short Papers*, pages 101–104.

Deyi Xiong, Qun Liu, and Shouxun Lin. 2006. Maximum entropy based phrase reordering model for statistical machine translation. In *Proc. of ACL 2006*, pages 521–528.

Richard Zens and Hermann Ney. 2006. Discriminvative reordering models for statistical machine translation. In *Proc. of Workshop on Statistical Machine Translation 2006*, pages 521–528.