Reordering with Source Language Collocations

Zhanyi Liu[1,2], Haifeng Wang[2], Hua Wu[2], Ting Liu[1], Sheng Li[1]
[1]Harbin Institute of Technology, Harbin, China
[2]Baidu Inc., Beijing, China
{liuzhanyi, wanghaifeng, wu_hua}@baidu.com
{tliu, lisheng}@hit.edu.cn

Abstract

This paper proposes a novel reordering model for statistical machine translation (SMT) by means of modeling the translation orders of the source language collocations. The model is learned from a word-aligned bilingual corpus where the collocated words in source sentences are automatically detected. During decoding, the model is employed to softly constrain the translation orders of the source language collocations, so as to constrain the translation orders of those source phrases containing these collocated words. The experimental results show that the proposed method significantly improves the translation quality, achieving the absolute improvements of 1.1~1.4 BLEU score over the baseline methods.

1 Introduction

Reordering for SMT is first proposed in IBM models (Brown et al., 1993), usually called *IBM constraint* model, where the movement of words during translation is modeled. Soon after, Wu (1997) proposed an ITG (Inversion Transduction Grammar) model for SMT, called *ITG constraint* model, where the reordering of words or phrases is constrained to two kinds: straight and inverted. In order to further improve the reordering performance, many structure-based methods are proposed, including the reordering model in hierarchical phrase-based SMT systems (Chiang, 2005) and syntax-based SMT systems (Zhang et al., 2007; Marton and Resnik, 2008; Ge, 2010; Visweswariah et al., 2010). Although the sentence structure has been taken into consideration, these methods don't explicitly make use of the strong correlations between words, such as collocations, which can effectively indicate reordering in the target language.

In this paper, we propose a novel method to improve the reordering for SMT by estimating the reordering score of the source-language collocations (source collocations for short in this paper). Given a bilingual corpus, the collocations in the source sentence are first detected automatically using a monolingual word alignment (MWA) method without employing additional resources (Liu et al., 2009), and then the reordering model based on the detected collocations is learned from the word-aligned bilingual corpus. The source collocation based reordering model is integrated into SMT systems as an additional feature to softly constrain the translation orders of the source collocations in the sentence to be translated, so as to constrain the translation orders of those source phrases containing these collocated words.

This method has two advantages: (1) it can automatically detect and leverage collocated words in a sentence, including long-distance collocated words; (2) such a reordering model can be integrated into any SMT systems without resorting to any additional resources.

We implemented the proposed reordering model in a phrase-based SMT system, and the evaluation results show that our method significantly improves translation quality. As compared to the baseline systems, an absolute improvement of 1.1~1.4 BLEU score is achieved.

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 1036–1044,
Portland, Oregon, June 19-24, 2011.

The paper is organized as follows: In section 2, we describe the motivation to use source collocations for reordering, and briefly introduces the collocation extraction method. In section 3, we present our reordering model. And then we describe the experimental results in section 4 and 5. In section 6, we describe the related work. Lastly, we conclude in section 7.

2 Collocation

A collocation is generally composed of a group of words that occur together more often than by chance. Collocations effectively reveal the strong association among words in a sentence and are widely employed in a variety of NLP tasks (Mckeown and Radey, 2000).

Given two words in a collocation, they can be translated in the same order as in the source language, or in the inverted order. We name the first case as *straight*, and the second *inverted*. Based on the observation that some collocations tend to have fixed translation orders such as "金融 jin-rong 'financial' 危机 wei-ji 'crisis'" (financial crisis) whose English translation order is usually straight, and "法律 fa-lv 'law' 范围 fan-wei 'scope'" (scope of law) whose English translation order is generally inverted, some methods have been proposed to improve the reordering model for SMT based on the collocated words crossing the neighboring components (Xiong et al., 2006). We further notice that some words are translated in different orders when they are collocated with different words. For instance, when "潮流 chao-liu 'trend'" is collocated with "时代 shi-dai 'times'", they are often translated into the "trend of times"; when collocated with "历史 li-shi 'history'", the translation usually becomes the "historical trend". Thus, if we can automatically detect the collocations in the sentence to be translated and their orders in the target language, the reordering information of the collocations could be used to constrain the reordering of phrases during decoding. Therefore, in this paper, we propose to improve the reordering model for SMT by estimating the reordering score based on the translation orders of the source collocations.

In general, the collocations can be automatically identified based on syntactic information such as dependency trees (Lin, 1998). However these methods may suffer from parsing errors. Moreover, for many languages, no valid dependency parser exists. Liu et al. (2009) proposed to automatically detect the collocated words in a sentence with the MWA method. The advantage of this method lies in that it can identify the collocated words in a sentence without additional resources. In this paper, we employ MWA Model l~3 described in Liu et al. (2009) to detect collocations in sentences, which are shown in Eq. (1)~(3).

$$p_{\text{MWA Model 1}}(A \mid S) \propto \prod_{j=1}^{l} t(w_j \mid w_{c_j}) \tag{1}$$

$$p_{\text{MWA Model 2}}(A \mid S) \propto \prod_{j=1}^{l} t(w_j \mid w_{c_j}) \cdot d(j \mid c_j, l) \tag{2}$$

$$p_{\text{MWA Model 3}}(A \mid S) \propto \prod_{i=1}^{l} n(\phi_i \mid w_i) \cdot \prod_{j=1}^{l} t(w_j \mid w_{c_j}) \cdot d(j \mid c_j, l) \tag{3}$$

Where $S = w_1^l$ is a monolingual sentence; ϕ_i denotes the number of words collocating with w_i; $A = \{(i, c_i) \mid i \in [1,l] \,\&\, c_i \neq i\}$ denotes the potentially collocated words in S.

The MWA models measure the collocated words under different constraints. MWA Model 1 only models word collocation probabilities $t(w_j \mid w_{c_j})$. MWA Model 2 additionally employs position collocation probabilities $d(j \mid c_j, l)$. Besides the features in MWA Model 2, MWA Model 3 also considers fertility probabilities $n(\phi_i \mid w_i)$.

Given a sentence, the optimal collocated words can be obtained according to Eq. (4).

$$A^* = \arg\max_{A} p_{\text{MWA Model } i}(A \mid S) \tag{4}$$

Given a monolingual word aligned corpus, the collocation probabilities can be estimated as follows.

$$r(w_i, w_j) = \frac{p(w_i \mid w_j) + p(w_j \mid w_i)}{2} \tag{5}$$

Where, $p(w_i \mid w_j) = \dfrac{count(w_i, w_j)}{\sum_{w'} count(w', w_j)}$; (w_i, w_j) denotes the collocated words in the corpus and $count(w_i, w_j)$ denotes the co-occurrence frequency.

3 Reordering Model with Source Language Collocations

In this section, we first describe how to estimate the orientation probabilities for a given collocation, and then describe the estimation of the reordering score during translation. Finally, we describe the integration of the reordering model into the SMT system.

3.1 Reordering probability estimation

Given a source collocation (f_i, f_j) and its corresponding translations (e_{a_i}, e_{a_j}) in a bilingual sentence pair, the reordering orientation of the collocation can be defined as in Eq. (6).

$$o_{i,j,a_i,a_j} = \begin{cases} \text{straight if } i < j \,\&\, a_i < a_j \text{ or } i > j \,\&\, a_i > a_j \\ \text{inverted if } i > j \,\&\, a_i < a_j \text{ or } i < j \,\&\, a_i > a_j \end{cases} \quad (6)$$

In our method, only those collocated words in source language that are aligned to different target words, are taken into consideration, and those being aligned to the same target word are ignored.

Given a word-aligned bilingual corpus where the collocations in source sentences are detected, the probabilities of the translation orientation of collocations in the source language can be estimated, as follows:

$$p(o = \text{straight} \mid f_i, f_j) = \frac{count(o = \text{straight}, f_i, f_j)}{\sum_{o'} count(o', f_i, f_j)} \quad (7)$$

$$p(o = \text{inverted} \mid f_i, f_j) = \frac{count(o = \text{inverted}, f_i, f_j)}{\sum_{o'} count(o', f_i, f_j)} \quad (8)$$

Here, $count(o, f_i, f_j)$ is collected according to the algorithm in Figure 1.

3.2 Reordering model

Given a sentence $F = f_1^l$ to be translated, the collocations are first detected using the algorithm described in Eq. (4). Then the reordering score is estimated according to the reordering probability weighted by the collocation probability of the collocated words. Formally, for a generated translation candidate T, the reordering score is calculated as follows.

$$P_O(F,T) = \sum_{(i,c_i)} r(f_i, f_{c_i}) \log p(o_{i,c_i,a_i,a_{c_i}} \mid f_i, f_{c_i}) \quad (9)$$

Input: A word-aligned bilingual corpus where the source collocations are detected
Initialization: $count(o, f_i, f_j) = 0$
for each sentence pair <*F*, *E*> in the corpus **do**
  for each collocated word pair (f_i, f_{c_i}) in *F* **do**
    if $i < c_i \,\&\, a_i < a_{c_i}$ or $i > c_i \,\&\, a_i > a_{c_i}$ **then**
      $count(o = straight, f_i, f_{c_i})$++
    if $i < c_i \,\&\, a_i > a_{c_i}$ or $i > c_i \,\&\, a_i < a_{c_i}$ **then**
      $count(o = inverted, f_i, f_{c_i})$++
Output: $count(o, f_i, f_j)$

Figure 1. Algorithm of estimating reordering frequency

Here, $r(f_i, f_{c_i})$ denotes the collocation probability of f_i and f_{c_i} as shown in Eq. (5).

In addition to the detected collocated words in the sentence, we also consider other possible word pairs whose collocation probabilities are higher than a given threshold. Thus, the reordering score is further improved according to Eq. (10).

$$P_O(F,T) = \alpha \cdot \sum_{(i,c_i)} r(f_i, f_{c_i}) \log p(o_{i,c_i,a_i,a_{c_i}} \mid f_i, f_{c_i}) + \beta \cdot \sum_{\substack{(i,j) \notin \{(i,c_i)\} \\ \& \, r(f_i,f_j) > \theta}} r(f_i, f_j) \log p(o_{i,j,a_i,a_j} \mid f_i, f_j) \} \quad (10)$$

Where α and β are two interpolation weights. θ is the threshold of collocation probability. The weights and the threshold can be tuned using a development set.

3.3 Integrated into SMT system

The SMT systems generally employ the log-linear model to integrate various features (Chiang, 2005; Koehn et al., 2007). Given an input sentence *F*, the final translation *E** with the highest score is chosen from candidates, as in Eq. (11).

$$E^* = \arg\max_E \{ \sum_{m=1}^{M} \lambda_m h_m(E,F) \} \quad (11)$$

Where $h_m(E, F)$ (m=1,...,M) denotes features. λ_m is a feature weight.

Our reordering model can be integrated into the system as one feature as shown in (10).

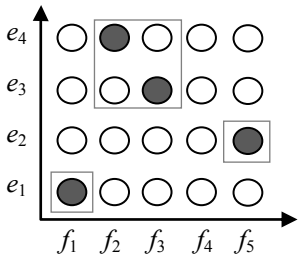


Figure 2. An example for reordering

4 Evaluation of Our Method

4.1 Implementation

We implemented our method in a phrase-based SMT system (Koehn et al., 2007). Based on the GIZA++ package (Och and Ney, 2003), we implemented a MWA tool for collocation detection. Thus, given a sentence to be translated, we first identify the collocations in the sentence, and then estimate the reordering score according to the translation hypothesis. For a translation option to be expanded, the reordering score inside this source phrase is calculated according to their translation orders of the collocations in the corresponding target phrase. The reordering score crossing the current translation option and the covered parts can be calculated according to the relative position of the collocated words. If the source phrase matched by the current translation option is behind the covered parts in the source sentence, then $\log p(o = \text{staight} \mid ...)$ is used, otherwise $\log p(o = \text{inverted} \mid ...)$. For example, in Figure 2, the current translation option is ($f_2\, f_3 \rightarrow e_3\, e_4$). The collocations related to this translation option are (f_1, f_3) , (f_2, f_3) , (f_3, f_5) . The reordering scores can be estimated as follows:

$$r(f_1, f_3) \log p(o = \text{straight} \mid f_1, f_3)$$
$$r(f_2, f_3) \log p(o = \text{inverted} \mid f_2, f_3)$$
$$r(f_3, f_5) \log p(o = \text{inverted} \mid f_3, f_5)$$

In order to improve the performance of the decoder, we design a heuristic function to estimate the future score, as shown in Figure 3. For any uncovered word and its collocates in the input sentence, if the collocate is uncovered, then the higher reordering probability is used. If the collocate has been covered, then the reordering orientation can be determined according to the relative positions of the words and the corresponding reordering probability is employed.

Input: Input sentence $F = f_1^L$
Initialization: *Score* = 0
for each uncovered word f_i **do**
  for each word f_j ($j = c_i$ or $r(f_{i,} f_j) > \theta$) **do**
    if f_j is covered **then**
      if $i > j$ **then**
        $Score \mathrel{+}= r(f_{i,} f_j) \log p(o = \text{straight} \mid f_i, f_j)$
      else
        $Score \mathrel{+}= r(f_{i,} f_j) \log p(o = \text{inverted} \mid f_i, f_j)$
    else
      $Score \mathrel{+}= \arg\max_o r(f_{i,} f_j) \log p(o \mid f_i, f_j)$
Output: *Score*

Figure 3. Heuristic function for estimating future score

4.2 Settings

We use the FBIS corpus (LDC2003E14) to train a Chinese-to-English phrase-based translation model. And the SRI language modeling toolkit (Stolcke, 2002) is used to train a 5-gram language model on the English sentences of FBIS corpus.

We used the NIST evaluation set of 2002 as the development set to tune the feature weights of the SMT system and the interpolation parameters, based on the minimum error rate training method (Och, 2003), and the NIST evaluation sets of 2004 and 2008 (MT04 and MT08) as the test sets.

We use BLEU (Papineni et al., 2002) as evaluation metrics. We also calculate the statistical significance differences between our methods and the baseline method by using the paired bootstrap resample method (Koehn, 2004).

4.3 Translation results

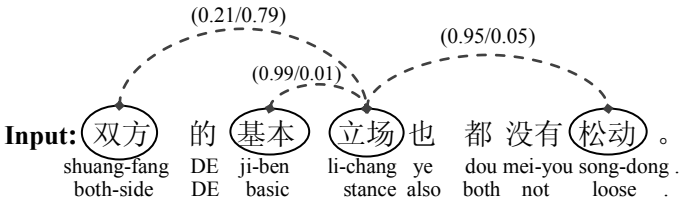
We compare the proposed method with various reordering methods in previous work.

Monotone model: no reordering model is used.

Distortion based reordering (DBR) model: a distortion based reordering method (Al-Onaizan & Papineni, 2006). In this method, the distortion cost is defined in terms of words, rather than phrases. This method considers *outbound*, *inbound*, and *pairwise* distortions that

Reorder models		MT04	MT08
Monotone model		26.99	18.30
DBR model		26.64	17.83
MSDR model (Baseline)		28.77	18.42
MSDR+	DBR model	28.91	18.58
	SCBR Model 1	29.21	19.28
	SCBR Model 2	29.44	19.36
	SCBR Model 3	29.50	19.44
	SCBR models (1+2)	29.65	19.57
	SCBR models (1+2+3)	29.75	19.61

Table 1. Translation results on various reordering models



T1: The two sides are also the basic stand of not relaxed.
T2: The basic stance of the two sides have not relaxed.
Reference: The basic stances of both sides did not move.

Figure 4. Translation example. (*/*) denotes ($p_{straight}$ / $p_{inverted}$)

are directly estimated by simple counting over alignments in the word-aligned bilingual corpus. This method is similar to our proposed method. But our method considers the translation order of the collocated words.

msd-bidirectional-fe reordering (MSDR or Baseline) model: it is one of the reordering models in Moses. It considers three different orientation types (*monotone*, *swap*, and *discontinuous*) on both source phrases and target phrases. And the translation orders of both the next phrase and the previous phrase in respect to the current phrase are modeled.

Source collocation based reordering (SCBR) model: our proposed method. We investigate three reordering models based on the corresponding MWA models and their combinations. In SCBR Model *i* (*i*=1~3), we use MWA Model *i* as described in section 2 to obtain the collocated words and estimate the reordering probabilities according to section 3.

The experiential results are shown in Table 1. The DBR model suffers from serious data sparseness. For example, the reordering cases in the trained pairwise distortion model only covered 32~38% of those in the test sets. So its performance is worse than that of the monotone model. The MSDR model achieves higher BLEU scores than the monotone model and the DBR model. Our models further improve the translation quality, achieving better performance than the combination of MSDR model and DBR model. The results in Table 1 show that "MSDR + SCBR Model 3" performs the best among the SCBR models. This is because, as compared to MWA Model 1 and 2, MWA Model 3 takes more information into consideration, including not only the co-occurrence information of lexical tokens and the position of words, but also the fertility of words in a sentence. And when the three SCBR models are combined, the performance of the SMT system is further improved. As compared to other reordering models, our models achieve an absolute improvement of 0.98~1.19 BLEU score on the test sets, which are statistically significant ($p < 0.05$).

Figure 4 shows an example: T1 is generated by the baseline system and T2 is generated by the system where the SCBR models (1+2+3)[1] are used.

[1] In the remainder of this paper, "SCBR models" means the combination of the SCBR models (1+2+3) unless it is explicitly explained.

Reordering models		MT04	MT08
MSDR model		28.77	18.42
MSDR+	DBR model	28.91	18.58
	CBR model	28.96	18.77
	WCBR model	29.15	19.10
	WCBR+SCBR models	29.87	19.83

Table 2. Translation results of co-occurrence based reordering models

	CBR model	SCBR Model3
Consecutive words	77.9%	73.5%
Interrupted words	74.1%	87.8%
Total	74.3%	84.9%

Table 3. Precisions of the reordering models on the development set

The input sentence contains three collocations. The collocation (基本, 立场) is included in the same phrase and translated together as a whole. Thus its translation is correct in both translations. For the other two long-distance collocations (双方, 立场) and (立场, 松动), their translation orders are not correctly handled by the reordering model in the baseline system. For the collocation (双方, 立场), since the SCBR models indicate $p(o$=straight|双方, 立场$) < p(o$=inverted|双方, 立场$)$, the system finally generates the translation T2 by constraining their translation order with the proposed model.

5 Collocations vs. Co-occurring Words

We compared our method with the method that models the reordering orientations based on co-occurring words in the source sentences, rather than the collocations.

5.1 Co-occurrence based reordering model

We use the similar algorithm described in section 3 to train the co-occurrence based reordering (CBR) model, except that the probability of the reordering orientation is estimated on the co-occurring words and the relative distance. Given an input sentence and a translation candidate, the reordering score is estimated as shown in Eq. (12).

$$P_O(F,T) = \sum_{(i,j)} \log p(o_{i,j,a_i,a_j} \mid f_i, f_j, \Delta_{i-j}) \quad (12)$$

Here, Δ_{i-j} is the relative distance of two words in the source sentence.

We also construct the weighted co-occurrence based reordering (WCBR) model. In this model, the probability of the reordering orientation is additionally weighted by the pointwise mutual information[2] score of the two words (Manning and Schütze, 1999), which is estimated as shown in Eq. (13).

$$P_O(F,T) = \sum_{(i,j)} s_{\mathrm{MI}}(f_i, f_j) \log p(o_{i,j,a_i,a_j} \mid f_i, f_j, \Delta_{i-j}) \quad (13)$$

5.2 Translation results

Table 2 shows the translation results. It can be seen that the performance of the SMT system is improved by integrating the CBR model. The performance of the CBR model is also better than that of the DBR model. It is because the former is trained based on all co-occurring aligned words, while the latter only considers the adjacent aligned words. When the WCBR model is used, the translation quality is further improved. However, its performance is still inferior to that of the SCBR models, indicating that our method (SCBR models) of modeling the translation orders of source collocations is more effective. Furthermore, we combine the weighted co-occurrence based model and our method, which outperform all the other models.

5.3 Result analysis

Precision of prediction

First of all, we investigate the performance of the reordering models by calculating precisions of the translation orders predicted by the reordering models. Based on the source sentences and reference translations of the development set, where the source words and target words are automatically aligned by the bilingual word alignment method, we construct the reference translation orders for two words. Against the references, we calculate three kinds of precisions as follows:

$$P_{\mathrm{CW}} = \frac{|\{|i-j|=1 \;\&\; o_{i,j} = o_{i,j,a_i,a_j}\}|}{|\{o_{i,j} \parallel i-j|=1\}|} \quad (14)$$

[2] For occurring words extraction, the window size is set to [-6, +6].

$$P_{\mathrm{IW}} = \frac{|\{|i-j|>1 \;\&\; o_{i,j} = o_{i,j,a_i,a_j}\}|}{|\{o_{i,j} \parallel i-j|>1\}|} \quad (15)$$

$$P_{\mathrm{total}} = \frac{|\{o_{i,j} = o_{i,j,a_i,a_j}\}|}{|\{o_{i,j}\}|} \quad (16)$$

Here, $o_{i,j}$ denotes the translation order of (f_i, f_j) predicted by the reordering models. If $p(o=\text{straight}\,|\,f_i,f_j) > p(o=\text{inverted}\,|f_i,f_j)$, then $o_{i,j}=\text{straight}$, else if $p(o=\text{straight}\,|\,f_i,f_j) < p(o=\text{inverted}\,|f_i,f_j)$, then $o_{i,j}=\text{inverted}$. o_{i,j,a_i,a_j} denotes the translation order derived from the word alignments. If $o_{i,j}=o_{i,j,a_i,a_j}$, then the predicted translation order is correct, otherwise wrong. P_{CW} and P_{IW} denote the precisions calculated on the consecutive words and the interrupted words in the source sentences, respectively. P_{total} denotes the precision on both cases. Here, the CBR model and SCBR Model 3 are compared. The results are shown in Table 3.

From the results in Table 3, it can be seen that the CBR model has a higher precision on the consecutive words than the SCBR model, but lower precisions on the interrupted words. It is mainly because the CBR model introduces more noise when the relative distance of words is set to a large number, while the MWA method can effectively detect the long-distance collocations in sentences (Liu et al., 2009). This explains why the combination of the two models can obtain the highest BLEU score as shown in Table 2. On the whole, the SCBR Model 3 achieves higher precision than the CBR model.

Effect of the reordering model

Then we evaluate the reordering results of the generated translations in the test sets. Using the above method, we construct the reference translation orders of collocations in the test sets. For a given word pair in a source sentence, if the translation order in the generated translation is the same as that in the reference translations, then it is correct, otherwise wrong.

We compare the translations of the baseline method, the co-occurrence based method, and our method (SCBR models). The precisions calculated on both kinds of words are shown in Table 4. From the results, it can be seen that our method achieves higher precisions than both the baseline and the method modeling the translation orders of the co-occurring words. It indicates that the proposed method effectively constrains the reordering of source words during decoding and improves the translation quality.

Test sets	Baseline (MSDR)	MSDR+ WCBR	MSDR+ SCBR
MT04	78.9%	80.8%	82.5%
MT08	80.7%	83.8%	85.0%

Table 4. Precisions (total) of the reordering models on the test sets

6 Related Work

Reordering was first proposed in the IBM models (Brown et al., 1993), later was named *IBM constraint* by Berger et al. (1996). This model treats the source word sequence as a coverage set that is processed sequentially and a source token is covered when it is translated into a new target token. In 1997, another model called *ITG constraint* was presented, in which the reordering order can be hierarchically modeled as *straight* or *inverted* for two nodes in a binary branching structure (Wu, 1997). Although the *ITG constraint* allows more flexible reordering during decoding, Zens and Ney (2003) showed that the *IBM constraint* results in higher BLEU scores. Our method models the reordering of collocated words in sentences instead of all words in IBM models or two neighboring blocks in ITG models.

For phrase-based SMT models, Koehn et al. (2003) linearly modeled the distance of phrase movements, which results in poor global reordering. More methods are proposed to explicitly model the movements of phrases (Tillmann, 2004; Koehn et al., 2005) or to directly predict the orientations of phrases (Tillmann and Zhang, 2005; Zens and Ney, 2006), conditioned on current source phrase or target phrase. Hierarchical phrase-based SMT methods employ SCFG bilingual translation model and allow flexible reordering (Chiang, 2005). However, these methods ignored the correlations among words in the source language or in the target language. In our method, we automatically detect the collocated words in sentences and

their translation orders in the target languages, which are used to constrain the ordering models with the estimated reordering (straight or inverted) score. Moreover, our method allows flexible reordering by considering both consecutive words and interrupted words.

In order to further improve translation results, many researchers employed syntax-based reordering methods (Zhang et al., 2007; Marton and Resnik, 2008; Ge, 2010; Visweswariah et al., 2010). However these methods are subject to parsing errors to a large extent. Our method directly obtains collocation information without resorting to any linguistic knowledge or tools, therefore is suitable for any language pairs.

In addition, a few models employed the collocation information to improve the performance of the ITG constraints (Xiong et al., 2006). Xiong et al. used the consecutive co-occurring words as collocation information to constrain the reordering, which did not lead to higher translation quality in their experiments. In our method, we first detect both consecutive and interrupted collocated words in the source sentence, and then estimated the reordering score of these collocated words, which are used to softly constrain the reordering of source phrases.

7 Conclusions

We presented a novel model to improve SMT by means of modeling the translation orders of source collocations. The model was learned from a word-aligned bilingual corpus where the potentially collocated words in source sentences were automatically detected by the MWA method. During decoding, the model is employed to softly constrain the translation orders of the source language collocations. Since we only model the reordering of collocated words, our methods can partially alleviate the data sparseness encountered by other methods directly modeling the reordering based on source phrases or target phrases. In addition, this kind of reordering information can be integrated into any SMT systems without resorting to any additional resources.

The experimental results show that the proposed method significantly improves the translation quality of a phrase based SMT system, achieving an absolute improvement of 1.1~1.4 BLEU score over the baseline methods.

References

Yaser Al-Onaizan and Kishore Papineni. 2006. Distortion Models for Statistical Machine Translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pp. 529-536.

Adam L. Berger, Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Andrew S. Kehler, and Robert L. Mercer. 1996. Language Translation Apparatus and Method of Using Context-Based Translation Models. *United States Patent, Patent Number 5510981*, April.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert. L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter estimation. *Computational Linguistics*, 19(2): 263-311.

David Chiang. 2005. A Hierarchical Phrase-based Model for Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pp. 263-270.

Niyu Ge. 2010. A Direct Syntax-Driven Reordering Model for Phrase-Based Machine Translation. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, pp. 849-857.

Philipp Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 388-395.

Philipp Koehn, Franz Joseph Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics*, pp. 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 *Proceedings of the 45th Annual Meeting of the ACL, Poster and Demonstration Sessions*, pp. 177-180.

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. 2005. Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation. In *Proceedings of International Workshop on Spoken Language Translation*.

Dekang Lin. 1998. Extracting Collocations from Text Corpora. In *Proceedings of the 1st Workshop on Computational Terminology*, pp. 57-63.

Zhanyi Liu, Haifeng Wang, Hua Wu, and Sheng Li. 2009. Collocation Extraction Using Monolingual Word Alignment Method. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp. 487-495.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*, Cambridge, MA; London, U.K.: Bradford Book & MIT Press.

Yuval Marton and Philip Resnik. 2008. Soft Syntactic Constraints for Hierarchical Phrased-based Translation. In *Proceedings of the 46st Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 1003-1011.

Kathleen R. McKeown and Dragomir R. Radev. 2000. Collocations. In Robert Dale, Hermann Moisl, and Harold Somers (Ed.), *A Handbook of Natural Language Processing*, pp. 507-523.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pp. 160-167.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1) : 19-51.

Kishore Papineni, Salim Roukos, Todd Ward, and Weijing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311-318.

Andreas Stolcke. 2002. SRILM - An Extensible Language Modeling Toolkit. In *Proceedings for the International Conference on Spoken Language Processing*, pp. 901-904.

Christoph Tillmann. 2004. A Unigram Orientation Model for Statistical Machine Translation. In *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics*, pp. 101-104.

Christoph Tillmann and Tong Zhang. 2005. A Localized Prediction Model for Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pp. 557-564.

Karthik Visweswariah, Jiri Navratil, Jeffrey Sorensen, Vijil Chenthamarakshan, and Nanda Kambhatla. 2010. Syntax Based Reordering with Automatically Derived Rules for Improved Statistical Machine Translation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pp. 1119-1127.

Dekai Wu. 1997. Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora. *Computational Linguistics*, 23(3):377-403.

Deyi Xiong, Qun Liu, and Shouxun Lin. 2006. Maximum Entropy Based Phrase Reordering Model for Statistical Machine Translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pp. 521-528.

Richard Zens and Herman Ney. 2003. A Comparative Study on Reordering Constraints in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pp. 192-202.

Richard Zens and Herman Ney. 2006. Discriminative Reordering Models for Statistical Machine Translation. In *Proceedings of the Workshop on Statistical Machine Translation*, pp. 55-63.

Dongdong Zhang, Mu Li, Chi-Ho Li, and Ming Zhou. 2007. Phrase Reordering Model Integrating Syntactic Knowledge for SMT. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 533-540.