Lexical-based Reordering Model for Hierarchical Phrase-based Machine Translation

Zhongguang Zheng, Yao Meng, Hao Yu
Fujitsu R&D Center CO., LTD.
15/F, Tower A, Ocean International Center, No.56 Dongsihuan Zhong Rd.
Chaoyang District, Beijing, 100025, China
{zhengzhg, mengyao, yu}@cn.fujitsu.com

Abstract

Reordering is a critical step for statistical machine translation. The hierarchical phrase-based model (HPB) based on a synchronous context-free grammar (SCFG) is prominent in solving global reorderings. However, the model is inadequate to supervise the reordering process so that sometimes phrases or words are reordered due to the wrong choice of translation rule. In order to solve this problem, we propose a novel lexical-based reordering model to evaluate the correctness of word order for each translation rule. Our approach employs the word alignment and translation information during the decoding process without causing too much extra computational consumption. Experimental results on the Chinese-to-English task showed that our method outperformed the baseline system in BLEU score significantly. Moreover, the translation results further proved the effectiveness of our approach.

1 Introduction

Reordering is a big challenge for statistical machine translation (SMT). The hierarchical phrase-based translation model (HPB) (Chiang, 2005), which adopts a synchronous context-free grammar (SCFG), is considered to be prominent in capturing global reorderings. However, the HPB model is weak in controlling the reordering process. Thus arbitrary reorderings frequently come up during the decoding process worsening the translation quality. Figure 1(a) shows an example of an incorrect reordering. The non-terminal "X_2" is reordered

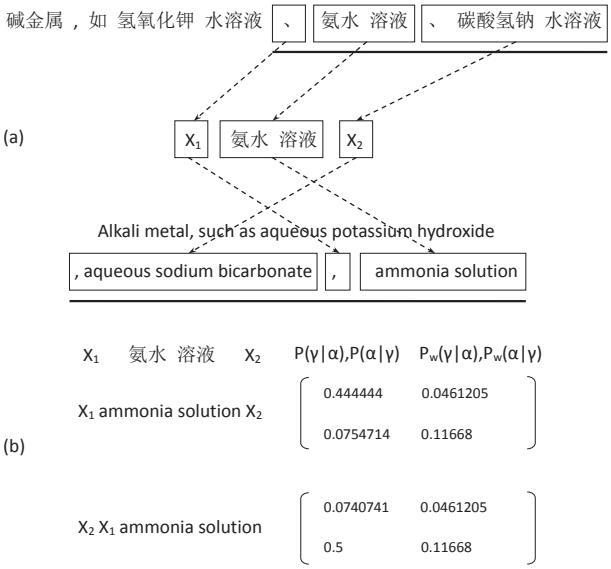

Figure 1: Incorrect reordering.

with "X_1" and "氨水 溶液", but the punctuation "、" indicates that the phrase should be translated monotonously.

As shown in Figure 1(b), although the two rules have different feature weights, the decoder still chose the incorrect one. We believe that there are mainly two reasons for this problem:

- The hierarchical rules are extracted without any linguistic constraints. Thus any form of rules could be obtained. As the example in Figure 1, the same Chinese part "X_1 氨水 溶液 X_2" corresponds to different target rules "X_2 X_1 ammonia solution" and "X_1 ammonia solution

X_2", which contain the same English words but differ in word order.

- There are not enough features to evaluate the correctness of word order. This makes rules in Figure 1(b) ambiguous for the decoder. Generally the HPB model has 8 features (Chiang, 2005), including language model, constituent feature, word penalty, phrase penalty, bi-direction translation weights $P(\gamma|\alpha)$ and $P(\alpha|\gamma)$, bi-direction lexical weights $P_w(\gamma|\alpha)$ and $P_w(\alpha|\gamma)$, but none of them is responsible for the rationality of reorderings. Although language model evaluates the fluency of target string, it considers the target words only.

Various methods have been proposed in order to solve this problem for HPB model. Most of them focused on the preprocessing stage (Xia and McCord, 2004; Collins et al., 2005; Wang et al., 2007; Tromble and Eisner, 2009; Du and Way, 2010). Those methods reordered source language to target language before training and testing by various proposed syntactic rules. Those offline rewriting methods are independent of the decoder. Thus other useful information such as word translation generated in the decoding process cannot be utilized.

Shen et al. (2008,2009) proposed a string-to-dependency language model to exploit long-distance word relations during decoding. He et al. (2010) classified SCFG rules into different patterns and built a maximum entropy classifier to select proper translation rules. Hayashi et al. (2010) integrated the method of (Tromble and Eisner, 2009) into the decoder to make this on-line rewriting method as a source language model. Those online methods are involved in the decoding phase as soft constraints to evaluate the word order of translation rules.

This paper proposes an on-line method which is based on the lexical information as a new feature for the HPB model. This feature is used to evaluate the correctness of word order in the decoding process. The remainder of this paper is organized as follows. Section 2 introduces the previous related work. In Section 3, we describe the implementation of the lexical-based reordering model and the integration into the decoder. Experiment on the Chinese-to-English task is shown in Section 4, followed by a discussion in Section 5. The conclusion and future work are presented in Section 6.

2 Previous Related Work

2.1 Online Reordering Methods

Comparing with the offline method, online method is able to utilize various information during decoding. Shen et al. (2008) proposed a string-to-dependency target language model to capture long distance word orders and Shen et al. (2009) extended the work by applying more features such as phrase length distribution and context language model. Shen et al. (2008) also intended to build a dependency language model on the source language, but the result reported a decline with this feature. He et al. (2010) divided hierarchical rules into several fixed patterns. For example, the rule <"X_1 氨水 溶液 X_2", "X_2 X_1 氨水 溶液"> belongs to the pattern <"X_1FX_2","X_2X_1E">. A maximum entropy classifier is applied to select target rules with proper patterns. This method is insensitive to the terminal order within the rules.

Our work is somewhat similar to the word-based reordering model proposed by Hayashi et al. (2010). In order to differ from their work, we name our approach a lexical-based reordering model, and the differences between the two methods are described below.

- Our method does not change the original HPB model. The former research changed HPB model from Equation 2 to

$$X \longrightarrow < \gamma, \gamma^{'}, \alpha, \sim > \quad (1)$$

  where $\gamma^{'}$ is the rewriting string of γ.

- Former research needed to consider the positions of unaligned words after rewriting a source sentence. But there is no such a problem with our model since we do not rewrite sentences.

- During the decoding process, our model employs the target language and word alignment information which are not included in the former research. The translation and alignment information are both helpful to distinguish word order to some extent. For example, Chinese

phrase “A 的 B” can be translated into English phrases “A ’s B” and “B of A”. The order between “A” and “B” is determined by the translation of “的”. Thus if a rule remains the order of “A” and “B”, but translates “的” to “of”, it is probably an incorrect rule. Furthermore, word alignment information is also useful. Recall the example of Figure 1, “氨水” and “、” are ambiguous words for the rewriting method, since both “氨水 、” and “、 氨水” are reasonable phrases that should be translated without reordering. However if we find a rule that reorders “氨水” and “、” according to the word alignments, it is probably incorrect.

- The former research worked on the Japanese-to-English task, while ours works on the Chinese-to-English task.

2.2 Hierarchical Phrase-Based Model

The hierarchical phrase-based model (HPB) (Chiang, 2005), which is based on a synchronous context-free grammar (SCFG), is presented in the form

$$X \longrightarrow < \gamma, \alpha, \sim > \quad (2)$$

where X is a non-terminal, γ and α denote source and target strings, which contain both terminals and non-terminals. $\sim$ is the one-to-one correspondence between terminals and non-terminals in γ and α. Chiang (2005) integrated all the features mentioned in the first section into the log-linear framework (Och and Ney, 2002).

$$P(e|f) \propto \sum_i \lambda_i h_i(\gamma, \alpha) \quad (3)$$

where $h_i(\gamma, \alpha)$ is a feature function and λ_i is the weight of h_i. Based on the deficiencies of HPB discussed in the early section, we intend to complement the log-linear framework with a feature as a soft constraint to measure the correctness of word order for each hierarchical rule.

3 Lexical-based Reordering Model

3.1 Overview of the Model

A score S_{re} is calculated using the lexical-based reordering model for each hierarchical rule r as follows

$$S_{re}(r) = log(\prod_{i,j:1 \le i < j \le n} P_{re}(order_{i,j}|\phi_{i,j})) \quad (4)$$

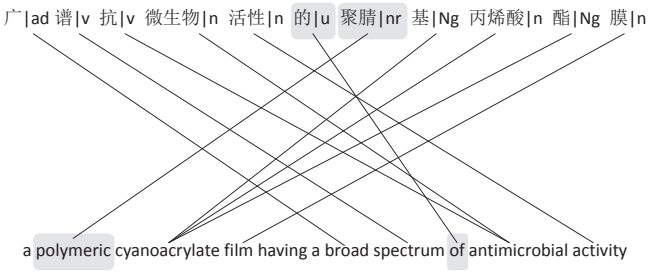

Figure 2: Sentence pair with word alignments.

LW	RW	LT	RT	LP	RP	Order
的	聚腈	of	polymeric	u	nr	0
的	聚腈	of	polymeric	u	--	0
的	聚腈	of	polymeric	--	nr	0
的	聚腈	of	polymeric	u	nr	0
的	聚腈	of	--	u	nr	0
的	聚腈	--	polymeric	u	nr	0
的	--	of	polymeric	u	nr	0
--	聚腈	of	polymeric	u	nr	0

Table 1: A part of features extracted from “的” and “聚腈”.

where i and j are subscripts of the source words in rule r. $\phi_{i,j}$ is a set of features extracted from w_i and w_j. $order_{i,j}$ presents the position relationship between w_i and w_j. According to the word alignments, $order_{i,j}$ equals “0” in the process of reordering, otherwise $order_{i,j}$ equals “1”.

3.2 Feature Extraction and Model Training

The features are extracted from the training set, where the hierarchical rules also come from. We use GIZA++ (Och and Ney, 2003) to obtain word alignments. Given a word aligned sentence pair $\langle f, e \rangle$, where $f = \{w_0, ..., w_n\}$, we select translations, part of speech (POS) tags and the order relationship for w_i and w_j according to the following constraints.

Constraint 1. Common constraint.

- w_i and w_j must be in the same *initial phrase pair* defined by Chiang (2005).

- $|j - i| \leq Threshold\_Word\_Scope$.

$Threshold\_Word\_Scope$ is an empirical threshold used to avoid arbitrary selections of word pairs, which may contain useless information. We also exploit linguistic rules to capture collocations that reveal the word order explicitly in case they violate the common constraint.

There are many linguistic phenomena between Chinese and English that indicate the word order explicitly, even though they often violate the $initial\ phrase\ pair$ constraint. Du and Way (2010) studied the reorderings of "的" structures for Chinese to English translation and reported significant improvement. Though it is a tough job to acquire linguistic knowledge without language analysis tools, the "的" structure is relatively easier to capture. Therefore, we especially propose a linguistic constraint based on the Chinese word "的" when the word pair violates the common constraint.

Constraint 2. Linguistic constraint.

- w_i or w_j is Chinese auxiliary word "的".
- $|j - i| \leq Threshold\_Word\_Scope$.

Figure 2 shows a sentence pair with word alignments. If confine an initial phrase pair to the maximum coverage of 10 source words, we will lose the collocations such as "的" and "聚腈", which indicates a reordering, by only considering the common constraint. In this case the linguistic constraint is applied to capture such collocations. The features extracted from words "的" and "聚腈" are listed in Figure 2, where "W", "P" and "T" stand for word, POS and translation, respectively. Note that there is a default precondition that w_i and w_j must both have alignments.

After extracting all the features from the training set, we use the maximum likelihood estimation method (MLE) to obtain the lexical-based reordering model.

$$P_{re}\left(order|\phi\right) = \frac{count\left(order, \phi\right)}{count\left(\phi\right)} \tag{5}$$

where $count\left(*\right)$ is the occurrence of $*$ int the training set.

Data		Sentence	Word
Training	Ch	100 k	3.7 M
	En	100 k	4.4 M
Development Set	Ch	1.0 k	37.5 k
	En	1.0 k	33.8 k
Test Set	Ch	1.0 k	38.8 k
	En	1.0 k	34.2 k

Table 2: Information of our data sets.

System ID	Threshold_Word_Scope	BLEU
sys1	2	30.05%
sys2	3	30.32%
sys3	4	29.76%

Table 3: Experiments on different threshold settings.

3.3 Integration into the Decoder

Given a source sentence $Sent$, candidate rules are first selected from the rule table. For a candidate rule r, all the source words it covers are easy obtained according to the rule span. Since the word translations and alignments of non-terminals are also known beforehand, we are able to extract features for every two source words and calculate the reordering score of r according to Equation 4. In this way, each candidate rule will get a reordering score, which will help the decoder to choose the right rules in cube pruning.

Note that, we do not calculate reordering probabilities for all the word pairs of r. We use a constraint to confine the calculation scope so as to avoid noisy computational results.

Constraint 3. Suppose w_i and w_j are source words of r and the subscript denotes the word position in r. We compute their reordering probability only when

- Neither w_i nor w_j aligns to empty word.
- There is at most one word of w_i and w_j belonging to a non-terminal span.
- $\left|j^{'} - i^{'}\right| < Threshold\_Word\_Scope$, where

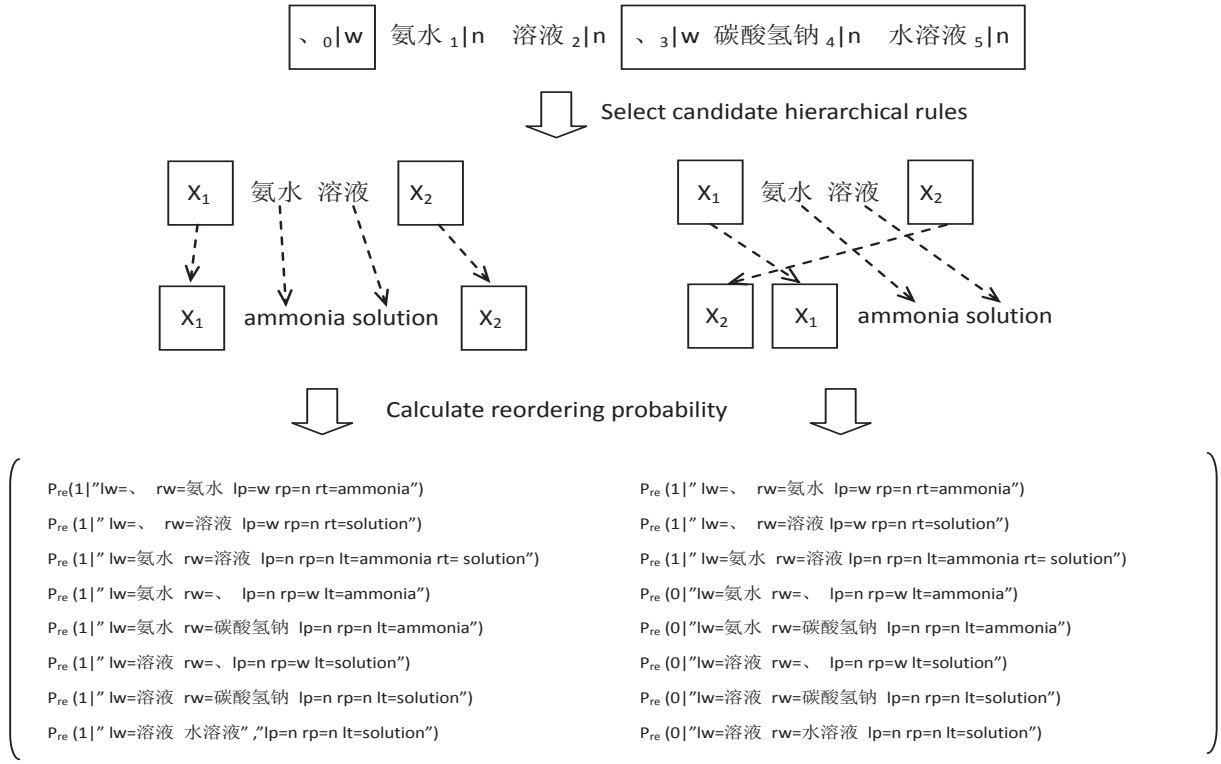

Figure 3: The process of computing reordering probabilities for HPB rules.

$j^{'}$ and $i^{'}$ denote the original positions of w_i and w_j in $Sent$.

Figure 3 depicts the process of distinguishing the two rules of Figure 1(b) by our method, where $Threshold\_Word\_Scope = 3$. The incorrect one is assigned with a lower reordering probability so that it is probably ignored in the cube pruning.

4 Experiment

4.1 Data Set

We conducted experiments on Chinese-to-English patent translation. On the one hand, word order is different between Chinese and English, thus it is a sensible testbed for our model. On the other hand, since the language of patent literature is well organized and constrained in expressions, our model would be more suitable for this kind of data.

Our data set is a part of NTCIR-9 Patent Machine Translation Task Document Data provided by NTCIR Workshop 9[1]. We selected 100,000 sentence pairs randomly from the whole data set as the training set, and divided the original development set into our development set and test set respectively. Table 2 shows the information of our data sets in detail.

4.2 Experimental Setup

Our experiments were on Chinese-to-English patent translation. Chinese word segmentation and POS tagging was implemented using an in-house Chinese word segmentation toolkit.

GIZA++ (Och and Ney, 2003) was run in both translation directions to obtain the word alignments, and the alignment result was refined by "grow-diag-final" method (Koehn, 2003).

For the language model, we used the SRI Language Modeling Toolkit (SRILM) (Stolcke, 2002) to train a 4-gram model on the target portion of the training set.

We used the minimum error rate training algorithm (MERT) (Och, 2003) for tuning the feature weights of the log-linear model, and adopted BLEU (Papineni et al., 2002) as our evaluation metric.

The experiments were carried out in two steps.

[1]http://research.nii.ac.jp/ntcir/index-en.html

System ID	BLEU
Baseline system	29.74%
LBR-no-trans	**30.00%**
LBR-no-de	**30.23%***
LBR-all	**30.32%***

Table 4: Experiment results on all the systems. "*" denotes significant better than the baseline system at $p < 0.01$.

	Source Phrase	Ambiguous Target Rules	
Sentence 1	X_1 特性 $_0$	characteristics$_0$ X_1	X_1 characteristics$_0$
Sentence 2	X_1 画面 $_0$ X_2	X_2 X_1 picture$_0$	X_1 picture$_0$ X_2
	X_1 强 $_0$ X_2	strong$_0$ X_2 X_1	X_1 strong$_0$ X_2

Table 5: The ambiguous hierarchical rules.

Firstly, in order to find out the most effective model settings, we tested different values on the threshold $Threshold\_Word\_Scope$. The results are shown in table 3.

From the results, best performance is achieved by setting $Threshold\_Word\_Scope = 3$, which was adopted as the final setting in the rest of the experiments. It is reasonable that we could not get enough features in a smaller scope and may obtain too much noise in a larger scope.

To evaluate the effectiveness of our model, we conducted experiments on four systems:

- Baseline: an in-house hierarchical phrase-based machine translation system (Chiang, 2007).
- LBR-no-trans: integrated with the lexical-based reordering model, but did not adopt translations as feature.
- LBR-no-de: integrated with the lexical-based reordering model, but was trained with the features merely satisfying the common constraint.
- LBR-all: fully integrated with our proposed model into the baseline system.

4.3 Results

The experimental results are shown in table 4. We can observe that systems integrated with the lexical-based reordering model all outperformed the baseline system. The improvements of "LBR-no-de" and "LBR-all" are statistically significant at $p < 0.01$ according to the significant test method described in (Koehn, 2004).

As applying more features to the model, the BLEU score rises accordingly. This proves that the linguistic constraint and the translation feature are effective.

We compared the translation results between the baseline system and "LBR-all" to check out the actual influence of our model. Figure 4 shows some examples of the effectiveness of our method. When analyzing the translation results, we find that many reordering mistakes are caused by the ambiguous hierarchical rules which are similar with those in Figure 1. Table 5 lists the ambiguous rules which occur in the examples of Figure 4.

5 Discussion

The experiment result confirms us that the application of linguistic knowledge is beneficial. And the examples in Figure 4 show that our method is effective in judging local word orders. In our experiment we also tried to use preposition and verb as an alternative linguistic constraint to capture more reordering relationships, since such collocations frequently trigger reorderings. For example, Chinese phrase "用|p A 覆盖|v B" always corresponds to English phrase "cover|v B with|p A". And those collocations are prevalent in the training corpous.

However the result turned out a decline on BLEU score. We believe that there are mainly two reasons for this

- Sometimes the preposition and verb are far from each other so that they exceed the coverage of one hierarchical rule, e.g. 10 words. Thus they are split into two rules. Our model can not calculate reordering score between two rules.

Src	在 一 个 实施 例 中 , 每 个 预定 义 集合 393 可 不仅 包含 可 配置 参数 375 (***例如 数据 类型 、 数据 定向 和 数据 特性***) , 还 包含 一 个 或 多 个 隐藏 配置 参数 377 。
BL	In one embodiment, each set of predefined 393 may not only comprises a configurable parameters 375 ***characteristics (e. g., data type, orientation and data***), and also comprises one or more hidden configuration parameter 377.
LR	In one embodiment, each predefined set 393 can not only include a configurable parameters 375 (***e. g., data type, data orientation and data characteristics***), further comprising one or more hidden configuration parameter 377.
RF	In one embodiment, each of predefined sets 393 may include not only configurable parameters 375 (***such as data categories, data orientations and data characteristics***), but also one or more hidden configuration parameters 377.
Src	使用 MPEG 或 VC1 运动 向量 来 确定 场景 为 3 个 等级 : ***静止 画面 (1) 、 弱 运动 (2) 、 强 谐 运动 (3)***
BL	Using MPEG or VC1 motion vector to determine a scene are three classes: ***(1), still picture weak moves the strong flip (2), (3),***
LR	Use MPEG or VC1 motion vector to determine the scene is 3 level: ***still picture (1), weak motion (2), strong flip motion (3)***,
RF	Using MPEG or VCl motion vectors to determine scenes: 3 levels: ***still picture (1), low motion (2), strong harmonic motion (3)***;

Figure 4: The actual influence of our method on translation results. Src: The source sentence. BL: Baseline translation. LR: The result of our method. RF: Reference.

- We did not apply any parser to analyze the Chinese sentences. It is too ambiguous to capture collocations only referring to the POS tags. As a result, too much noise was obtained when training our model.

Therefore, though linguistic knowledge is beneficial, if we want to employ more useful linguistic rules, language analysis toolkit should be involved.

6 Conclusion and Future Work

In this paper we proposed a lexical-based reordering model. This model employs useful information, such as word alignments and translations, during the decoding process to measure the correctness of word order. The experimental results showed that our method outperformed baseline system significantly on the Chinese-to-English patent translation.

Though our model is trained using the maximum likelihood estimation method in this work, it would be profitable to apply other methods to train the model, such as the maximum entropy model. Since the order of two words can be considered as a binary classification problem, we could use a maximum entropy classifier, which is trained with the features we proposed, to calculate the reordering probabilities. And language analysis toolkit should also be applied to exploit more useful linguistic constraints just like we discussed in Section 5.

References

Andreas Stolcke. 2002. SRIM - An Extensible Language Modeling Toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP),* pages 901-904.

Chao Wang, Michael Collins and Philipp Koehn. 2007. Learning linear ordering problems for better translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning,* pages 737-745.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association of Computational Linguistics,* pages 263-270.

David Chiang. 2007. Hierarchical phrase-based translation. In *Computational Linguistics,* pages 201-228.

Fei Xia and Michael McCord. 2004. Improving a Statistical MT System with Automatically learned Rewrite Patterns In *Proceedings of the 18th ICON,* page 508-514.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting of Computational Linguistics (ACL),* pages 295-302.

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

Franz Josef Och and Hermann Ney. 2003. A system comparison of various statistical alignment models. *Computational Linguistics,* pages 19-51.

Jinhua Du and Andy Way. 2010. The Impact of Source-Side Syntactic Reordering on Hierarchical Phrase-based SMT. *2010 European Association for Machine Translation.*

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 the 23rd International Conference on Computational Linguistics (Coling 2010),* pages 439-446.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhou. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics,* pages 311-318.

Libin Shen, Jinxi Xu and Ralph Weischedel. 2008. A New String-to-Dependency Machine Translation Algorithm with a Target Dependency Language Model. In *Proceedings of ACL 08,* pages 577-585.

Libin Shen, Jinxi Xu, Bing Zhang, Spyros Matsoukas and Ralph Weischedel. 2009. Effective Use of Linguistic and Contextual Information for Statistical Machine Translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing,* pages 72-80.

Michael Collins, Fhilipp Koehn and Ivona Kucerova. 2005. Clause Restructuring for Statistical Machine Translation. In *Proceedings of the 43rd Association of Computational Linguistics,* page 531-540.

Philipp Koehn, Franz Josef Och and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 40th Annual Meeting of HLT-NAACL 2003,* pages 127-133.

Philipp Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of EMNLP 2004,* pages 901-904.

Roy Tromble and Jason Elsner. 2009. Learning linear ordering problems for better translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing,* pages 1007-1016.

Zhongjun He, Yao Meng and Hao Yu. 2010. Maximum entropy based phrase reordering for hierarchical phrase-based translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing,* pages 555-563.