Maximum Entropy Based Lexical 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. The hierarchical phrase-based (HPB) model on the basis of a synchronous context-free grammar (SCFG) is prominent in solving global reorderings. However, the HPB model is inadequate to supervise the reordering process so that sometimes positions of different lexicons are switched due to the incorrect SCFG rules. In this paper, we consider the order of two lexicons as a classification problem and propose a novel lexical reordering model based on a maximum entropy classifier. Our model employs the word alignment and translation during the decoding process. 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.

Keywords: hierarchical phrase-based model, reordering, maximum entropy model

1 Introduction

Reordering is a big challenge for statistical machine translation (SMT). The hierarchical phrase-based (HPB) translation model (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. Arbitrary reorderings frequently come up during the decoding phase, worsening the translation quality, such as the example shown in Figure 1(a). The non-terminal "X_2" is reordered with "X_1" and "*anshui rongye*", but the punctuation "、" indicates that the phrase should be translated monotonously.

This kind of reordering error frequently occurs when target phrases contain similar lexicons, which include both terminals and non-terminals, with different permutations, e.g., the rules in Figure 1(b). We believe that there are mainly two reasons that the HPB model can not distinguish such ambiguous rules properly.

- The extraction of SCFG rules is merely based on the word alignment information. Thus any form of rules could be obtained. As shown in Figure 1, the same Chinese part "X_1 *anshui rongye* 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. The order of non-terminals should depend on their concrete translations and contexts. However, the decoder does not take into account those factors when reorderings happen.
- There are not enough features to evaluate the correctness of word order for SCFG rules. Conventionally, 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 word order. Although language model evaluates the fluency of target

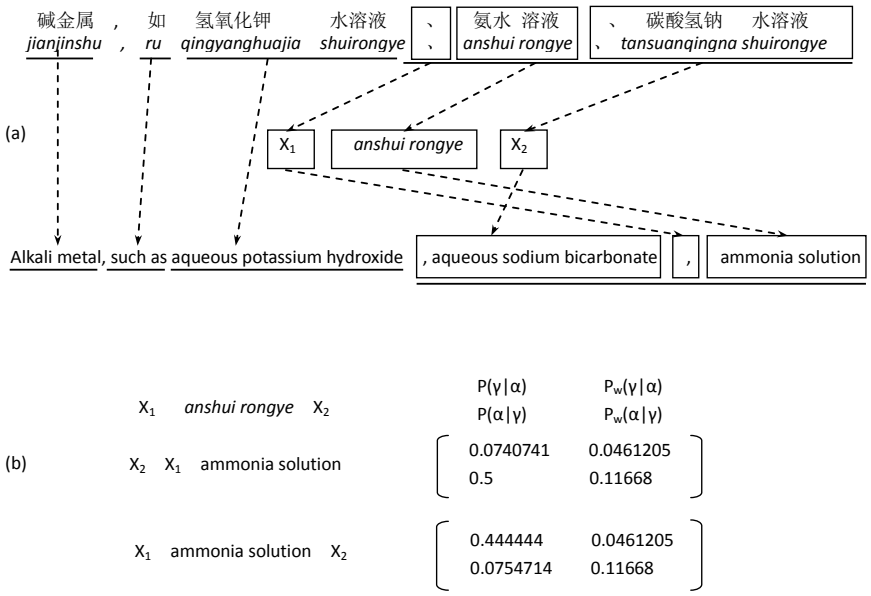

Figure 1: Incorrect reordering.

string, it considers the target words only. As shown in Figure 1(b), the two SCFG rules have different probabilities given by $P(\gamma|\alpha)$, $P(\alpha|\gamma)$ and language model, but they do not yield correct result.

Various methods have been proposed in order to solve the reordering problem for HPB model. Some of them focus on the sentence rewriting in the preprocessing stage (Xia and McCord, 2004; Collins *et al.*, 2005; Wang *et al.*, 2007; Tromble and Elsner, 2009; Du and Way, 2010). The main idea of those studies is to rewrite source language in order to make the source word sequence close to the target language before training and testing by various syntactic methodologies. 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. Moreover, offline methods require syntax analysis which does not always produce convincing result on certain languages. The parsing error will result in wrongly reordered sentences so that the translations cannot be correct.

Shen *et al.* (2008, 2009) proposed a string-to-dependency language model to capture long-distance word order. He *et al.* (2010) classified SCFG rules into different patterns and built a maximum entropy classifier to select proper translation rules during decoding. Analogously, Cui *et al.* (2010) proposed a joint model for SCFG rule selection. Four sub-models which include context free model and context based model are used to predict proper rules. Both source and target strings are considered simultaneously in the model. Hayashi *et al.* (2010) integrated the method of (Tromble and Elsner, 2009) into the decoder as a source language model. Those online methods are involved in the decoding phase as a soft constraint to bias the decoder toward certain SCFG rules that are considered appropriate. This paper proposes a lexical reordering model based on the maximum entropy model for the HPB model, and we also integrate our model into the decoder to exploit various information during decoding.

The rest of this paper is organized as follows. Section 2 introduces the related work. In Section 3, we describe the implementation of the maximum entropy based lexical 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 useful information during decoding. Shen *et al.* (2008) proposed a string-to-dependency target language model to capture long distance word orders. Furthermore, Shen *et al.* (2009) extended the work by applying more features such as phrase length distribution and context language model. Shen *et al.* (2009) 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) classified SCFG rules into several fixed patterns. For example, the rule $<X_1$ *anshui rongye* X_2, X_2 X_1 ammonia solution$>$ belongs to the pattern $<X_1$FX_2, X_2X_1E$>$. A maximum entropy classifier is trained to select rules according to the patterns on the both source and target sides. This method is insensitive to the terminal order.

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 reordering model, and the differences between the two methods are described below.

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

$$X \longrightarrow < \gamma, \gamma^{'}, \alpha, \sim > \tag{1}$$

  where $\gamma^{'}$ is the rewriting string of γ.
- Former research needs to consider the positions of unaligned words after rewriting a source string. 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 will benefit the word order disambiguation to some extent. For example, Chinese phrase "A 的(*de*) 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 "*de*".

  Furthermore, word alignment information is also useful. Recall the example of Figure 1, "*anshui*" and "、" are ambiguous words for the rewriting method, since both "*anshui* 、" and "、 *anshui*" are reasonable phrases that should be translated without reordering. However, if a rule reorders "*anshui*" and "、" according to the word alignment, it is probably incorrect and should not be used here.
- 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 (HPB) model (Chiang, 2005), which is based on a synchronous context-free grammar (SCFG), is presented in the form

$$X \longrightarrow < \gamma, \alpha, \sim > \tag{2}$$

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)

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

where $h_i(\gamma, \alpha)$ is a feature function and λ_i is the weight of h_i. One merit of the log-linear framework is that we are able to adopt various features into the HPB model conveniently. Hence we intend to complement the HPB model with our proposed method as a new feature.

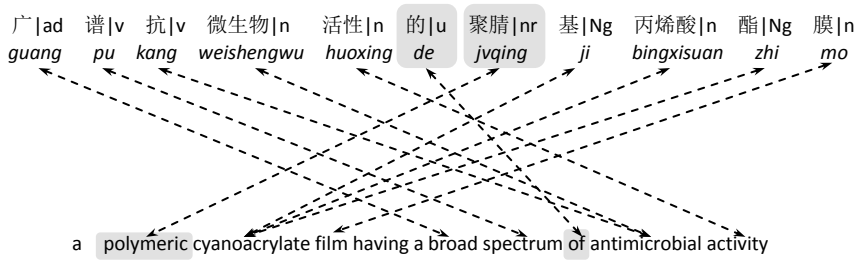

Figure 2: Sentence pair with word alignment.

Table 1: Features extracted from "*de*" and "*jvqing*".

Feature	Feature Value
Feat1	order=0 lw=*de* lp=u lt=of rw=*jvqing* rp=nr rt=polymeric
Feat2	order=0 lw=*de* lp=u lt=of rw=*null* rp=nr rt=polymeric
Feat3	order=0 lw=*null* lp=u lt=of rw=*jvqing* rp=nr rt=polymeric

3 Maximum Entropy Based Lexical Reordering Model

3.1 Overview of the Model

A score S_{re} is calculated using the maximum entropy based lexical reordering model for each SCFG rule

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

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 alignment, $order_{i,j}$ equals "0" when w_i and w_j are reordered, otherwise $order_{i,j}$ equals "1".

3.2 Feature Extraction

The feature extraction is carried out together with SCFG rule extraction. We use GIZA++ (Och and Ney, 2003) to obtain word alignment on the training set. 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 of w_i and w_j as features. Firstly, we employ a non-linguistic constraint to limit the distance between w_i and w_j, which is described as follows.

Constraint 1. Non-linguistic 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 selection of word pairs in case of useless information. We also adopt linguistic rules to capture collocations that reveal word order explicitly when they violate the non-linguistic 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 "*de*" structures for Chinese to English translation and their experiment reported a significant improvement. Linguistic knowledge acquisition generally requires language analysis tools such as dependency parser. However, the "*de*" structure is relatively easier to capture without parsing the sentence in that word "*de*" is closely related to its context words. Therefore,

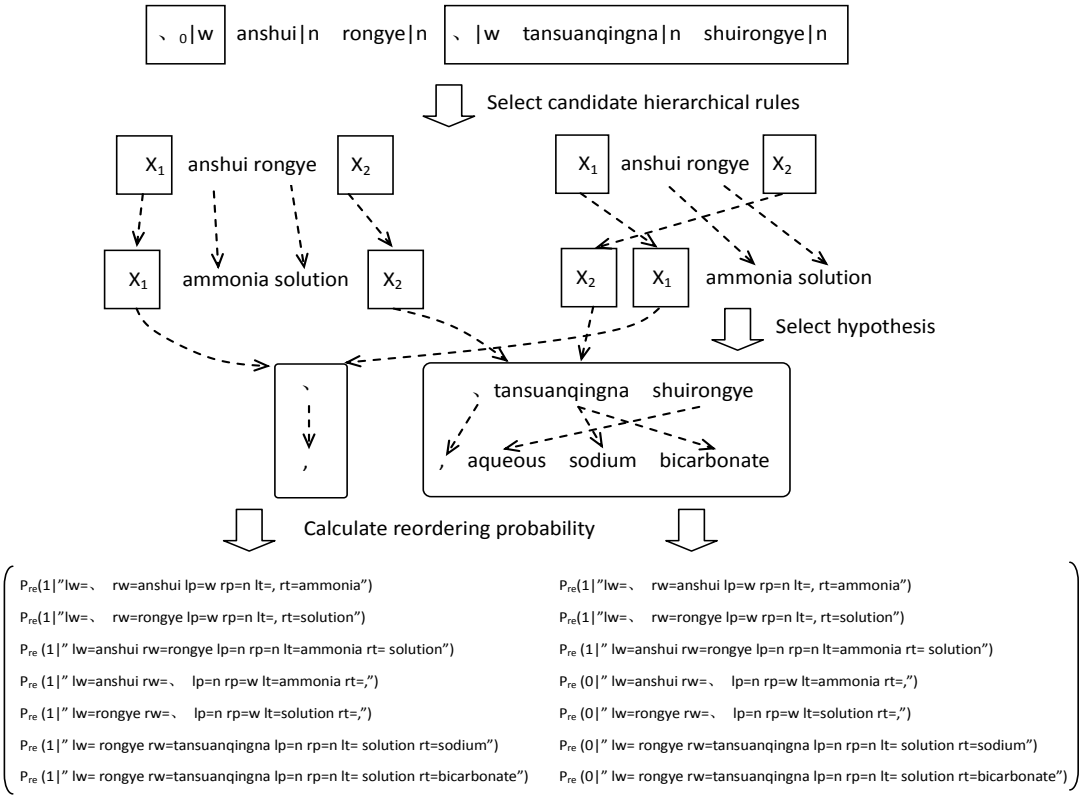

Figure 3: The process of computing reordering probabilities for SCFG rules.

we especially propose a linguistic constraint based on the Chinese word "*de*" when the word pair violates the non-linguistic constraint.

Constraint 2. Linguistic constraint.

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

The linguistic constraint is simple but effective. 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 "*de*" and "*jvqing*", which indicates a reordering, by only considering the non-linguistic constraint. In this case, the linguistic constraint is helpful to complement the lost collocations. Table 1 lists all the features extracted from word pair "*de*" and "*jvqing*", where "w", "p" and "t" stand for word, POS and translation, respectively. "*null*" denotes empty word. Note that there is a default precondition that w_i and w_j must both have alignments.

3.3 Model Training

There is a simple way to calculate the reordering probability using maximum likelihood estimation (MLE) method, which is in the form

$$P_{re}\left(order_{i,j}|\phi_{i,j}\right) = \frac{count\left(order_{i,j}, \phi_{i,j}\right)}{count\left(\phi_{i,j}\right)} \tag{5}$$

where $count\left(*\right)$ is the occurrence of $*$ in the training set, and $\phi_{i,j}$ is a set of features. However, MLE method faces the severe data sparsity.

Table 2: Information of our data sets.

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

Table 3: Experiment results of different settings of $Threshold\_Word\_Scope$.

$Threshold\_Word\_Scope$	MLE	ME
2	30.05%	30.17%
3	30.32%	30.66%
4	29.76%	29.96%

On the other hand, the order of two words is either monotone or inverse which means it belongs to a binary classification task. The maximum entropy (ME) classifier is suitable for solving this kind of problem and has been successfully applied in many previous studies. The ME model we trained is in the form

$$p_{re}(o|w_l, w_r) = \frac{\exp(\sum_i \lambda_i h_i(o, w_l, w_r))}{\sum_o \exp(\sum_i \lambda_i h_i(o, w_l, w_r))} \quad (6)$$

where o denotes the order of two source words w_l and w_r, h_i is a feature function and λ_i is the weight of h_i.

3.4 Integrating Our Model into the Decoder

Given a source sentence s, candidate SCFG rules are first selected from the rule table. For one candidate rule r, all the source words it covers are easy obtained according to the rule span. Since word alignments are also known beforehand, it is easily to extract features described in 3.2 for each two terminals in r. As to non-terminals, we extract features from their boundary terminals in each hypothesis during the cube-pruning. The reordering probability of r is then calculated using Equation 4. Here, again, we limit the computational scope with the following constraints.

Constraint 3. Suppose w_i and w_j are source words containing either terminals or non-terminals of r. The subscripts denote the word positions in r. We calculate reordering probability of that word pair only when

- Neither w_i nor w_j aligns to empty word.
- There is at most one non-terminal between w_i and w_j.
- $\left|j^{'} - i^{'}\right| \leq Threshold\_Word\_Scope$, where $j^{'}$ and $i^{'}$ denote the original positions of w_i and w_j in s.

Figure 3 depicts the process of distinguishing the two rules of Figure 1(b) by our method with the setting $Threshold\_Word\_Scope = 2$. The incorrect rule will be assigned with a lower order 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

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

System ID	Language Model	BLEU
Baseline system	4-gram	29.74%
Baseline 5-gram	5-gram	30.37%
Baseline 6-gram	6-gram	30.27%
MLE	4-gram	30.32%*
ME-no-de	4-gram	30.28%*
ME-all	4-gram	30.66%*

Table 5: The ambiguous SCFG rules.

	Source Rules	Ambiguous Target Rules	
Sentence 1	(X_1	(X_1	X_1 (
Sentence 2	X_1 texing	X_1 characteristic	characteristic X_1

hand, since the language of patent text is well organized and constrained in expression, our model would be more suitable for this kind of data.

Our data set is a part of NTCIR-9 Patent Machine Translation Task (PatentMT) Document Data provided by NTCIR-9 Workshop[1]. We selected 100,000 sentence pairs randomly from the whole data set as our training set, and then 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. The English tokenization was implemented using our own script.

GIZA++ (Och and Ney, 2003) was run in both translation directions to obtain the word alignment on the training set, and then 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 three language models from 4-gram to 6-gram on the target portion of the training set. Considering that long distance language models may influence the word order on the target string, it makes sense to compare our model with those long distance language models.

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. An open source Maximum Entropy Toolkit (Zhang, 2004) was employed to train our lexical model.

We implemented two systems based on maximum likelihood estimation (MLE) and maximum entropy (ME) model, respectively. Our experiments were carried out in two steps.

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

From the results, the ME based systems produce better results than the MLE based systems. This indicates that the ME classifier is more suitable for our approach. The 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 a reasonable result that we could not get enough features in a smaller scope and may obtain too much noise in a larger scope.

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

Src1	weishengyongpin[*hygiene article*] , tebieshi[*particular*] **niaobu[*diaper*] (yinger[*infant*]** 、 shijin[*incontinent*] chengren[*adult*]) 、 nvxing[*feminine*] weishegnyongpin [*hygiene article*] ,
BL	Hygiene products, in particular **(baby diapers**, incontinent adult), feminine hygiene products,
MEL	Hygiene products, particular **diapers (baby**, incontinent adult), feminine hygiene products,
RF	Sanitary articles, in particular **diapers (infant**, incontinent adult), feminine hygiene products,
Src2	kepeizhi[*configurable*] canshu[*parameter*] 375 **(liru[*such as*] shujv[*data*] leixing[*category*] , shujv[*data*] dingxiang[*orientation*], he[*and*] shujv[*data*] texing[*characteristic*])** ,
BL	configurable parameters 375 **characteristics (e. g., data type, orientation and data)**,
MEL	configurable parameters 375 **(e. g., data type, data orientation and data characteristics),**
RF	configurable parameters 375 **(such as data categories, data orientations and data characteristics)**,
Src3	canzhao[*refer*] tu[*figure*] 7 , **yonghu[*user*] dui[to] caidan[*menu*] xuanxiang[*option*] 607 de[*'s*] xuanze[*selection*]** (tu[*figure*] 6)
BL	Referring now to Figure 7, **the user selects to menu options 607** (Figure 6)
MEL	Referring to FIG.7, **user selection of the menu options 607** (Figure 6)
RF	Referring to FIG. 7, **the user's selection of menu option 607** (FIG. 6)

Figure 4: The actual influence of our method on translation results. Src: The source sentence. BL: Translation of Baseline system. MEL: Translation of ME-all system. RF: Reference.

Then, to evaluate the effectiveness of our model, we conducted experiments on six systems.

- Baseline system: an in-house hierarchical phrase-based machine translation system (Chiang, 2007) with 4-gram language model.
- Baseline 5-gram: baseline system using 5-gram language model.
- Baseline 6-gram: baseline system using 6-gram language model.
- MLE: proposed method using maximum likelihood estimation.
- ME-no-de: proposed method using maximum entropy model, but was trained with the features merely satisfying the non-linguistic constraint.
- ME-all: proposed method using maximum entropy model using all sorts of features.

Note that all the MLE and ME systems are trained with 4-gram language model only.

4.3 Results

The experimental results are shown in Table 4. We can observe that all the proposed methods outperformed the baseline system. The improvements are all statistically significant at $p < 0.01$ according to the significant test method described in (Koehn, 2004). The fact that our methods yield better results than the long distance n-gram language model illustrates that information on the source side is useful to judge the word order. Moreover, as applying the simple linguistic constraint, the BLEU score rises accordingly. We can observe that the ME-all system outperformed ME-no-de systems at the significance $p < 0.05$. This proves that the linguistic constraint is effective.

We compared the translation results between the baseline system and ME-all system to investigate the actual influence of our method. Figure 4 shows some examples. From the first two sentences we can see that incorrect reorderings occur in the *BL* system due to the ambiguous SCFG rules, which are listed in Table 5, while *MEL* system is able to produce correct results. The third sentence demonstrates that *MEL* system is prone to produce more proper result than *BL* system since translation is considered when measuring the word order.

5 Discussion

The experiment result confirms us that the application of linguistic knowledge is beneficial. In our experiment we also tried to exploit more linguistic knowledge from training set, such as collocations of preposition and verb (pv). We used pv as an alternative linguistic constraint to capture more reordering relationships, since such collocations frequently trigger reorderings. For example, Chinese phrase "*yong(with)*$_1$|p A *fugai(cover)*$_2$|v B" always corresponds to English phrase "$cover_2$|v B $with_1$|p A". And those pv collocations are prevalent in the training set.

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 conventionally. Thus they are split into different SCFG rules. Our model can not calculate reordering scores between two rules.
- We did not apply any kind of parser to analyze the Chinese sentence. 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 must be involved.

6 Conclusion and Future Work

In this paper we proposed a maximum entropy based lexical reordering model for the hierarchical phrase-based translation model. Our model employs useful features, e.g., word alignment and translation, during the decoding process to measure the correctness of word order for SCFG rules. The experimental results showed that our method outperformed baseline system significantly on the Chinese-to-English patent translation.

Although we only use a simple linguistic constraint, the experimental result shows a significant improvement. This is a positive signal that our model will become much stronger by exploiting more sophisticated linguistic knowledge. Furthermore, the maximum entropy model makes it convenient to incorporate various features. Thus in the future, we will apply language analysis tools to extract more beneficial features to improve our model.

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. 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. 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 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 WeiJing 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.

Le Zhang. 2004. Maximum Entropy Modeling Toolkit for Python and C++. available at http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html

Lei Cui, Dongdong Zhang, Mu Li, Ming Zhou, and Tiejun Zhao. 2010. A Joint Rule Selection Model for Hierarchical Phrase-based Translation. In *Proceedings of the ACL 2010 Conference Short Papers,* pages 6-11.

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.