Extending the Hierarchical Phrase Based Model with Maximum Entropy Based BTG

Zhongjun He Yao Meng Hao Yu
Fujitsu R&D Center CO., LTD.
15/F., Tower A, Ocean International Center, 56 Dongsihuan Zhong Rd.
Chaoyang District, Beijing, 100025, China
{hezhongjun, mengyao, yu}@cn.fujitsu.com

Abstract

In the hierarchical phrase based (HPB) translation model, in addition to hierarchical phrase pairs extracted from bi-text, *glue* rules are used to perform serial combination of phrases. However, this basic method for combining phrases is not sufficient for phrase reordering. In this paper, we extend the HPB model with maximum entropy based bracketing transduction grammar (BTG), which provides content-dependent combination of neighboring phrases in two ways: serial or inverse. Experimental results show that the extended HPB system achieves absolute improvements of 0.9∼1.8 BLEU points over the baseline for large-scale translation tasks.

1 Introduction

The hierarchical phrase based (HPB) model (Chiang, 2005), built on weighted synchronous context free grammar (SCFG), provides a powerful mechanism to capture both short and long distance phrase reorderings for statistical machine translation (SMT). It utilizes two types of rules:

- Translation rules are learned from word-aligned bilingual corpus. A translation rule can be either a *phrasal* rule consisting of words or a *hierarchical* rule consisting of both words and variables. During decoding, phrasal rules perform lexical translation, while hierarchical rules perform both lexical translation and phrase reordering.
- Two glue rules are defined to serially combine neighboring phrases. Glue rules provide a mechanism for finishing translation in cases where no translation rules are available for a source span.

However, one disadvantage of the HPB model is that the glue rules only provide monotone combinations of phrases. In some cases, however, the order of phrases maybe inverted. Therefore, we need an additional glue rule to perform *inverse* combinations of phrases. It is appropriate to use the bracketing transduction grammar (BTG) (Wu, 1996), which provides two options for combining phrases: serial or inverse. Xiong et al. (2006) and Zens and Ney (2006) presented a discriminative phrase reordering model based on BTG. They regarded phrase reordering as a two-class classification problem and built a content-dependent model under a maximum entropy (ME) framework. The approach yielded significant improvements on phrase reordering over conventional phrase-based SMT systems.

In this paper, we extend the HPB model by using BTG rules instead of the monotone glue rules. Analogous to Xiong et al. (2006), we built an ME based classifier to predict whether the neighboring phrases combined serially or inversely.

The extended HPB approach can be viewed as a combination of HPB translation and ME based BTG translation. Compared with previous methods of system combination (e.g. (Rosti et al., 2007; He et al., 2008)), the basic difference is that while conventional methods combined translation results after the decoding of different models, our method combines translation models during decoding.

Liu et al. (2009) presented a framework for a joint decoding method to combine multiple transla-

tion models. They combined the HPB model and the tree-to-string model (Liu et al., 2006). Since these two models are quite different, e.g. the HPB model is formally syntax-based while the tree-to-string model is linguistically syntax-based, it is difficult to combine them to form a joint decoder. Liu et al. (2009) utilized a hypergraph structure to store partial derivations and a modified MERT (Och, 2003) algorithm for training. They reported an absolute improvement of 1.5 BLEU points on a small corpus (6.9M + 8.9M words) for Chinese-to-English translation.

It is more straightforward to combine the HPB model and the ME based BTG model. On the one hand, the translation grammar is similar as both of the two models are based on CFG. Thus, to combine them together we need only to add into the HPB model an *inverted* rule, which combines phrases in an inverse order. On the other hand, both use a CKY algorithm for decoding, therefore, making it unnecessary to modify the decoding algorithm. Furthermore, all the feature weights of the translation model are tuned together by MERT without any modification. Large-scale data experiments (77M+81M) showed absolute improvements of 0.9$\sim$1.8 BLEU points for different test sets.

The rest of the paper is organized as follows. In section 2 and 3, we review the hierarchical phrase-based model and the maximum entropy based BTG, respectively. Section 4 describes our approach that extends the HPB model with ME based BTG. In Section 5, we describe the systems that we used for experiments, as well as the experimental results. We analyze the presented method and experimental results in Section 6 and conclude in Section 7.

2 The Hierarchical Phrase-based Model

The hierarchical phrase-based model is based on a weighted SCFG, which consists of the following rewrite rules:

$$X \rightarrow \langle \alpha, \gamma, \sim \rangle \tag{1}$$

where X is a non-terminal symbol, α/γ is a string consisting of source/target terminals and non-terminals. $\sim$ describes a one-to-one correspondence between non-terminals in α and γ.

The greatest advantage of the HPB model is that it utilizes hierarchical phrases, i.e. phrases that contain sub-phrases. Thus the hierarchical phrases capture phrase reordering. For instance, we could extract the following hierarchical rule from a word-aligned bilingual corpus:

$$X \rightarrow \langle \text{zai } X_1 \text{ de } X_2,\ X_2 \text{ in the } X_1 \rangle \tag{2}$$

During decoding, this rule swaps the source phrases covered by X_1 and X_2 on the target side.

Chiang (2007) developed a bottom-up decoder using CKY algorithm. For a source sentence, the decoder produces a translation and a parser tree as a byproduct (Figure 1 (a)). Given a source sentence F_1^J, the goal of decoding is to search the best derivation for $S_{[1,J]}$[1]. The algorithm produces partial derivations for each cell that spans from j_1 to j_2 by using translation rules from bottom to top. This guarantees that when the span $[j_1, j_2]$ is being expanded, all its sub-spans have been already expanded. Finally, we search the best derivation of the span $[1, J]$, from which we can obtain the translation.

However, some spans may not be covered by any translation rules because of data sparseness. Therefore, the decoder faces a risk of being unable to produce the final derivations. For example, if there are no translation rules cover the source span $[i, j]$ in Figure 1 (a), the decoder cannot produce any derivations for $X_{[i,j]}$. To solve this problem, two glue rules were introduced by Chiang (2005):

$$S \rightarrow \langle SX, SX \rangle \tag{3}$$

$$S \rightarrow \langle X, X \rangle \tag{4}$$

The glue rules are used to combine two adjacent phrases serially. Therefore, if there are no translation rules available, the decoder performs monotone translation.

Using translation rules and glue rules, one partial derivation of F_1^J is shown in Figure 2. We noted that the decoder serially combined the translations of four phrase spans $[1, i]$, $[i, k]$, $[k, j]$ and $[j, J]$ on the target side, which indicates that reordering may occur only within these spans and cannot across them.

[1]We use S to represent the start of a sentence

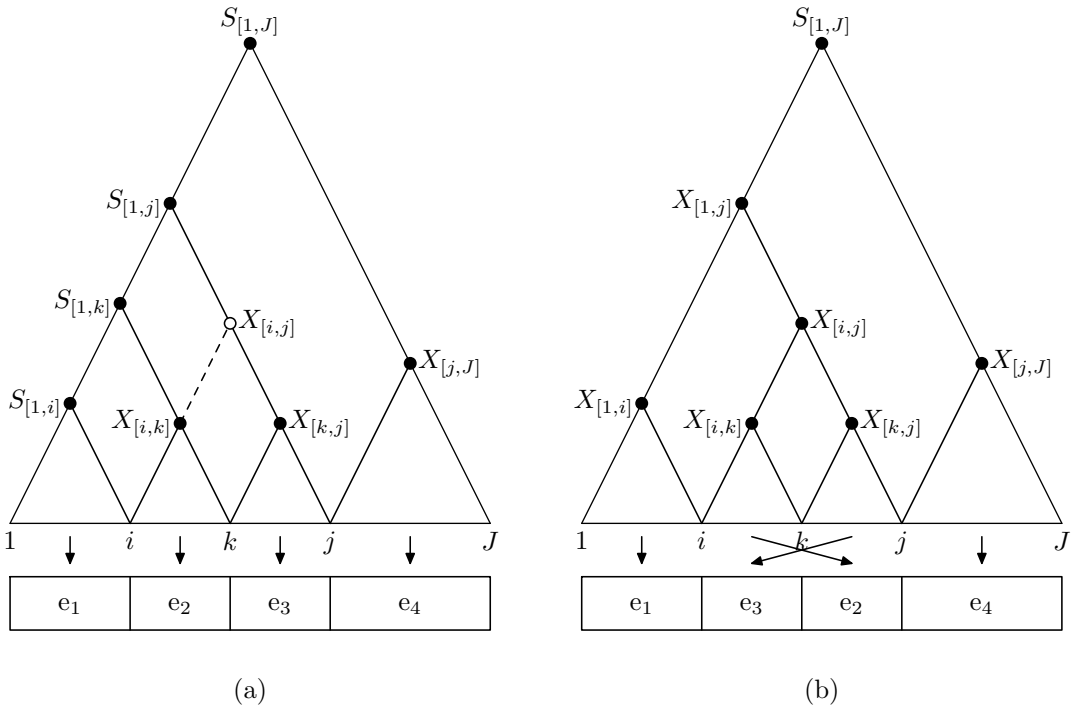

Figure 1: Illustration for the HPB translation (a) and extended HPB translation (b) with BTG. The extended HPB translation provides an option to swap two adjacent source spans [i,k] and [k,j] when no translation rules cover them.

$$
\begin{aligned}
S_{[1,J]} &\Rightarrow \langle S_{[1,j]}X_{[j,J]}, S_{[1,j]}X_{[j,J]}\rangle \\
&\Rightarrow \langle S_{[1,k]}X_{[k,j]}X_{[j,J]}, S_{[1,k]}X_{[k,j]}X_{[j,J]}\rangle \\
&\Rightarrow \langle S_{[1,i]}X_{[i,k]}X_{[k,j]}X_{[j,J]}, \\
&\quad S_{[1,i]}X_{[i,k]}X_{[k,j]}X_{[j,J]}\rangle \\
&\Rightarrow \langle X_{[1,i]}X_{[i,k]}X_{[k,j]}X_{[j,J]}, \\
&\quad X_{[1,i]}X_{[i,k]}X_{[k,j]}X_{[j,J]}\rangle
\end{aligned}
$$

Figure 2: The derivation of Figure 1 (a).

$$
\begin{aligned}
S_{[1,J]} &\Rightarrow \langle X_{[1,J]}, X_{[1,J]}\rangle \\
&\Rightarrow \langle X_{[1,j]}X_{[j,J]}, X_{[1,j]}X_{[j,J]}\rangle \\
&\Rightarrow \langle X_{[1,i]}X_{[i,j]}X_{[j,J]}, X_{[1,i]}X_{[i,j]}X_{[j,J]}\rangle \\
&\Rightarrow \langle X_{[1,i]}\mathbf{X_{[i,k]}}\mathbf{X_{[k,j]}}X_{[j,J]}, \\
&\quad X_{[1,i]}\mathbf{X_{[k,j]}}\mathbf{X_{[i,k]}}X_{[j,J]}\rangle
\end{aligned}
$$

Figure 3: The derivation of Figure 1 (b).

3 Maximum Entropy based BTG

BTG (Wu, 1996) consists of three types of rules:

$$X \rightarrow \langle \tilde{f}, \tilde{e}\rangle \tag{5}$$

$$X \rightarrow \langle X_1X_2, X_1X_2\rangle \tag{6}$$

$$X \rightarrow \langle X_1X_2, X_2X_1\rangle \tag{7}$$

Rule 5 is a phrasal rule that translates a source phrase $\tilde{f}$ into a target phrase $\tilde{e}$. Rule 6 merges two consecutive phrases in monotone order, while Rule 7 merges them in inverse order. During decoding, the decoder first uses Rule 5 to produce phrase translations, then iteratively uses Rules 6 and 7 to merge two phrases into a larger phrase until the whole sentence is covered.

BTG is adept at phrase reordering. However, one disadvantage is that the phrase reordering is content-independent. To overcome this problem, Xiong et al. (2006) improved it with a maximum entropy framework. They regarded phrase reordering as a two-class classification problem, and built an ME based classifier to predict the order of two adjacent

Rule Type	HPB	BTG
Phrasal rule	Yes	Yes
Hierarchical rule	Yes	No
Monotone glue rule	Yes	Yes
Inverted glue rule	No	Yes

Table 1: Comparison of Rule Types in the HPB model and the BTG model.

phrases:

$$P(o|X_1, X_2) = \frac{exp(\sum_i \lambda_i h_i(o, X_1, X_2))}{\sum_o exp(\sum_i \lambda_i h_i(o, X_1, X_2)} \quad (8)$$

where, $o \in \{monotone, inverted\}$ is the order of X_1 and X_2 , $h_i(o, X_1, X_2)$ is a feature function and λ_i is the weight of h_i. Xiong et al. (2006) defined feature functions on boundary words of phrases, e.g. the first word of a phrase. They reported significant improvements using ME based BTG for phrase reordering.

4 Extended HPB Translation

4.1 Extending the Glue Rules with BTG

Table 1 provides a comparison between HPB translation and BTG translation. Both have phrasal rule and monotone glue rule. This allows them to perform a standard phrase-based translation. To improve phrase reordering, the HPB model uses hierarchical rules, which consist of both terminals and non-terminals, while the BTG model uses an inverted glue rule to combine phrases in inverse order. Furthermore, the ME based BTG provides a mechanism for content-dependent phrase reordering.

We believe that the HPB model can benefit from BTG in the following aspects:

- In the HPB model, the most important rule is the hierarchical rule since it captures phrase reordering. However, it heavily increases the rule table size and makes training and decoding slow. In addition, it gives rise to a spurious ambiguity problem. Therefore, Chiang (2005) introduced some constraints. One of them was to prohibit adjacent nonterminals on the source side, such that the source side cannot contain "X_1X_2". The inverted glue rule in BTG can solve this problem.

HPB	Extended HPB
$S \rightarrow \langle X, X \rangle$	$S \rightarrow \langle X, X \rangle$
$S \rightarrow \langle SX, SX \rangle$	$X \rightarrow \langle X_1X_2, X_1X_2 \rangle$
-	$X \rightarrow \langle X_1X_2, X_2X_1 \rangle$

Table 2: Extending the glue rules in the HPB model with BTG.

- The HPB model only provides a monotone glue rule to merge phrases. As shown in Figure 1 (a), it will produce a left-heavy parser tree. The inverted glue rule will provide another option to merge phrases inversely, allowing it to produce a more balanced parser tree.

We extended the HPB model by replacing its glue rules with BTG rules, as shown in Table 2. With this extension, the parser tree of the source sentence in Figure 1 (a) is more balanced, as shown in (b). Figure 3 shows a partial derivation. We observed that the translations of source phrases f_i^k and f_k^j were reversed on the target side.

4.2 The Extended Translation Model

Both the HPB model and the ME based BTG model were built within the standard log-linear framework (Och and Ney, 2002):

$$Pr(e|f) \propto \sum_i \lambda_i h_i(\alpha, \gamma) \quad (9)$$

where $h_i(\alpha, \gamma)$ is a feature function and λ_i is the weight of h_i. The HPB model has the following features: translation probabilities $p(\gamma|\alpha)$ and $p(\alpha|\gamma)$, lexical weights $p_w(\gamma|\alpha)$ and $p_w(\alpha|\gamma)$, word penalty, phrase penalty, glue rule penalty, and a target n-gram language model.

When extending the glue rules with ME based BTG, we modify the features of the log-linear model as follows:

- An ME based reordering feature was added to predict the order of neighboring phrases:

  $$h_{mebtg}(o|X_1, X_2) = \sum P(o|X_1, X_2) \quad (10)$$

  To train an ME classifier, we used two kinds of content information from the adjacent phrases: the boundary words (the first and last word) and their part-of-speech tags.

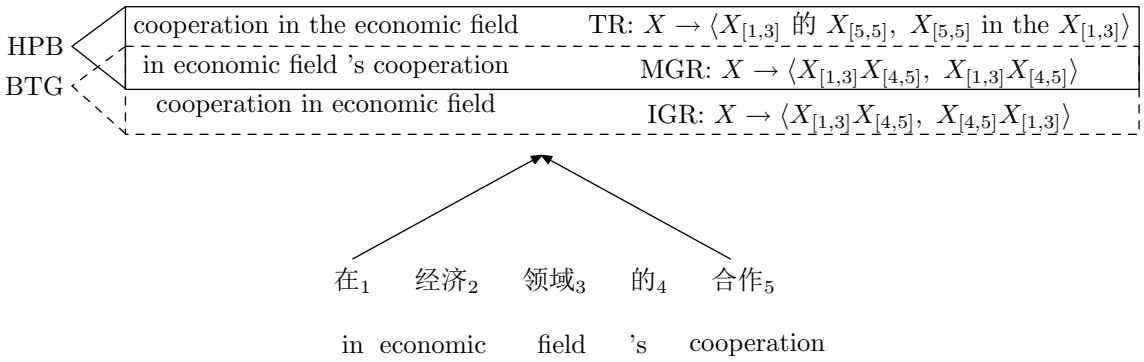

Figure 4: The extended HPB translation. Subscript of X is the source span it covers. TR: translation rule, MGR: monotone glue rule, IGR: inverted glue rule. Solid box contains translations generated by the HPB model using TR and MGR. Dashed box contains translations generated by the ME based BTG model using MGR and IGR.

- We split the "glue rule penalty" into two features: the monotone glue rule number and the inverted glue rule number. These features reflect the preference of the decoder for using monotone or inverted glue rules.

The advantage of our extension method is that the weights of the new features can be tuned together with the other features by MERT algorithm without any modification.

4.3 Decoding

We developed a decoder for the extended HPB model with a CKY algorithm, which is used for both the HPB model (Chiang, 2005) and the ME based BTG model (Xiong et al., 2006). The difference is that, in our decoder, the translations for a source span may be derived from three types of rules: the translation rule (including the hierarchical rule and phrasal rule), monotone glue rule and inverted glue rule. See Figure 4 for illustration. Given a source sentence, the decoder will search for the best derivation and generate a final translation.

The extended decoder is flexible: if we prohibit the usage of the inverted glue rule, the decoder performs HPB translation, while if we prohibit the usage of hierarchical rules (or there are no valid hierarchical rules), the decoder performs ME based BTG translation.

5 Experiments

5.1 Systems

We developed an extended HPB SMT system, which provides four mechanisms for translation:

- **HPB**: the system based on the original HPB model by prohibiting the inverted glue rule;
- **MEBTG**: the system based on ME BTG by prohibiting the hierarchical rules;
- **HPB+BTG**: the system is a combination of the HPB model and the BTG model, without using an ME classifier for content-dependent reordering;
- **ExtHPB**: the system is a combination of the HPB model and the MEBTG model, as described in Section 3.

5.2 Experimental Setup

We carried out experiments on Chinese-to-English translation. The training data contains 77M Chinese words and 81M English words. These data come from 17 corpora: LDC2002E18, LDC2002L27, LDC2002T01, LDC2003E07, LDC2003E14, LDC2004T07, LDC2005E83, LDC2005T06, LDC2005T10, LDC2005T34, LDC2006E24, LDC2006E26, LDC2006E34, LDC2006E86, LDC2006E92, LDC2006E93, LDC2004T08 (HK_News, HK_Hansards).

To obtain word alignments, we first ran GIZA++ (Och and Ney, 2000) in both translation directions

Rule Type	Number	Percentage
Hierarchical Rules	127M	78.4%
Phrasal Rules	35M	21.6%

Table 3: Statistical information for translation rules extracted from the training corpus.

Reordering Type	Number	Percentage
Monotone	63M	95.7%
Inverted	2.8M	4.3%

Table 4: Statistical information for reordering examples extracted from the training corpus.

and then refined the results using the "grow-diag-final" method (Koehn et al., 2003). For the language model, we used the SRI Language Modeling Toolkit (Stolcke, 2002) to train two 4-gram models on the Xinhua portion of GigaWord corpora and the English side of the training corpus. We tuned our systems on NIST MT03 and tested on MT06 and MT08. The evaluation metric was BLEU-4 (Papineni et al., 2002).

5.3 Training

To extract translation rules from the word-aligned bilingual corpus, as described in Chiang (2005), we limited the initial phrase length to 10 and constrained the rules to have at most two nonterminals, prohibiting them from being adjacent on the source side. Table 3 shows the statistical information for the translation rules extracted from the training corpus. We extracted 162M translation rules. Among them, 35M were phrasal rules, accounting for 21.6%. The other 127M were hierarchical rules, accounting for 78.4%. Hierarchical rules were effective at capturing phrase reorderings, however, they caused a large size rule table.

Using the algorithm described in Xiong et al. (2006), we extracted 65.8M reordering instances from the training corpus. Table 3 shows the statistical information. Most instances were monotone, accounting for 95.7%. While instances of inverted reordering accounted for 4.3%. Although the number of inverted reorderings was small, they were important for phrase ordering. The ME classifier was trained by a toolkit (Zhang, 2004).

Systems	06G	06N	08
HPB	14.19	33.93	25.85
MEBTG	13.86	32.61	24.76
HPB+BTG	13.01	32.32	23.41
ExtHPB	15.09	35.72	27.34

Table 5: BLEU percentage scores (case-insensitive) on the test data. G=GALE set, N=NIST set. Note that the GALE set only has one reference for each source sentence.

5.4 Results

Table 5 shows the BLEU scores of the four systems on MT06 (GALE set and NIST set) and MT08. From the table, we made the following observations:

- The MEBTG system yielded a lower BLEU score than the HPB system. The reason is that there are no hierarchical rules in BTG. The hierarchical rule can be regarded as half-lexicalized since it contains both variables and words. The words in a hierarchical rule can guide rule matching and phrase reordering.

- The HPB+BTG system performed worse than the HPB system. The reason is that the phrase reordering of BTG is content-independent. Such an arbitrary reordering is harmful for HPB translation.

- The ExtHPB system outperformed both the HPB system and the MEBTG system, indicating that the extended model benefits from both the HPB model and the MEBTG model. The ExtHPB system achieved significant improvements ($p < 0.01$) over the HPB system on all test sets, with absolute increases of BLEU scores ranging from 0.9 (on MT06G) to 1.8 (on MT06N) percentage points. On the other hand, the ExtHPB outperformed the HPB+BTG, which indicates that content-dependent phrase reordering is important for SMT.

6 Analysis

The experimental results consistently demonstrated that the extended HPB model achieved significant gains on BLEU scores. However, the BLEU score

system	Number	Percentage (%)			
		PR	HR	MGR	IGR
HPB	26,578	39.7	34.2	26.1	-
MEBTG	46,501	48.2	-	40.6	4.7
ExtHPB	40,523	47.4	11.0	37.1	4.5

Table 6: Statistical information for grammar rules on 1-best outputs for MT08. PR=phrasal rule, HR=hierarchical rule, MGR=monotone glue rule, IGR=inverted glue rule.

is not sufficient to provide detailed insight into the nature of improvements. Therefore, we propose to further study in detail what happens after combining the HPB and MEBTG models.

6.1 Grammar rules

The usage of grammar rules can reflect the preferences of a decoder. We classify grammar rules into four types: phrasal rule, hierarchical rule, monotone glue rule and inverted glue rule. We counted the number of each type of rule used during decoding for three systems on 1-best outputs for MT08 test set.

The statistical information is shown in Table 6. The HPB system and the MEBTG system used the least (26,578) and greatest (46,501) number of rules, respectively. One of the possible reason is that the hierarchical rule can perform both lexical translation and phrase reordering, since it consists of both words and variables. Therefore, a source span translated by a hierarchical rule in HPB may be translated by several phrasal rules and glue rules (monotone and inverted) in MEBTG. After combination, the rule number of the extended system falls between HPB and MEBTG. Furthermore, the percentage of hierarchical rules decreased in the ExtHPB system. This indicates that some phrase reorderings are performed by the inverted glue rule and phrasal rules instead of hierarchical rules.

Table 6 indicated that ExtHPB used more shorter phrase pairs than HPB, because the rule number for ExtHPB is almost 2 times greater than HPB. We also counted the average phrase length for the three systems, as shown in Table 7. It is observed that the average phrase length for ExtHPB is shorter than HPB and greater than MEBTG. Intuitively, longer

System	Average Len	PR Len	HR Len
HPB	2.26	1.43	3.22
MEBTG	1.34	1.34	-
ExtHPB	1.70	1.30	3.46

Table 7: Average length of phrases used in three systems. PR Len= average phrasal rule length, HR Len= average hierarchial rule length.

phrase pairs can have more fluent impacts for translation quality since they contain more information for local reorderings. However, the phrase number, as a feature in most of the phrase-based SMT models, also impacts translation quality. From Table 6 and 7, we observed that the HPB prefers less phrases in number and longer phrases in length than MEBTG. Although MEBTG used shorter phrases, it produced high quality translation results. Because the ME based phrase reordering model combines various contextual information to guide phrase reorderings on both short and long distance. The extended system ExtHPB combined the advantages of HPB and MEBTG and leveraged on phrase length and number.

6.2 Improving Phrase Reordering

We further compared the outputs of HPB, MEBTG and ExtHPB. We observed that the improvement of translation performance resulted mostly from the improvement of phrase reordering. For example, the translations of a source sentence are as follows [2]:

- **Src:** 这$_1$ 架$_2$ 飞机$_3$ 将$_4$ 在$_5$ 巴基斯坦$_6$ 首都$_7$ 伊斯兰堡$_8$ 作$_9$ 短暂$_{10}$ 停留$_{11}$
- **Ref:** This$_1$ plane$_3$ will$_4$ make$_9$ a brief$_{10}$ stop$_{11}$ in$_5$ the Pakistan$_6$ capital$_7$ of Islamabad$_8$
- **HPB:** The plane in the Pakistani capital of Islamabad for a short stay
- **MEBTG:** The plane will make a short stopover in the Pakistani capital of Islamabad
- **ExtHPB:** The plane will make a short stopover in the Pakistani capital of Islamabad

[2]The co-indexes of the words in the source and reference sentence indicate word alignments.

Src:	他 是 在 海诺宁 抵达 德黑兰 后 发表 这 项 谈话 。
Ref:	He made this remark after Heinonen arrived in Tehran.
HPB:	He is Heinonen arrived in Tehran, issued after the talks.
MEBTG:	He arrived in Tehran on Heinonen said after the talks.
ExtHPB:	He made this remark after Heinonen arrived in Tehran.
Src:	帕蒂尔 是 印度 史 上 第一 位 女性 总统 候选 人 。
Ref:	Patil is the first female presidential candidate in the history of India.
HPB:	Patil is India's first female in the history of presidential candidate.
MEBTG:	Patil is India's first woman president in the history of candidates.
ExtHPB:	Patil is the first female presidential candidate in the history of India.
Src:	北韩 并 未 参加 在 美国 盐湖 城 举行 的 第 十九 届 冬季 奥运 。
Ref:	North Korea did not participate in the 19th Winter Olympics in Salt Lake City in the US.
HPB:	North Korea did not participate in the United States at the 19th Winter Olympics at Salt Lake City.
MEBTG:	North Korea did not participate in the United States at the 19th Salt Lake City Winter Olympics.
ExtHPB:	North Korea did not attend the 19th Winter Olympic Games held in Salt Lake City of the United States.

Table 8: Translation examples.

All systems produced correct lexical translations for the source phrases "在$_5$ 巴基斯坦$_6$ 首都$_7$ 伊斯兰堡$_8$, in the Pakistani capital of Islamabad" and "作$_9$ 短暂$_{10}$ 停留$_{11}$, for a short stay/make a short stopover". However, the HPB system does not have a rule to cover the whole source span $[5, 11]$, thus producing a monotone translation using the monotone glue rule to serially combine sub-spans $[5, 8]$ and $[9, 11]$. The MEBTG system used an inverted rule to swap these two sub-spans on the target side and generated a correct phrase reordering. Analogous to the MEBTG system, the ExtHPB system also used an inverted glue rule to generate a translation.

Table 8 shows more examples generated by the three systems. From these examples, we clearly observed that the extended HPB produced better phrase reorderings than the baseline systems.

7 Conclusions and Future Work

We have presented an extended HPB translation method that incorporates an ME based BTG into the HPB model. We added an inverted glue rule to complement the HPB rules and built an ME classifier to predict reorderings between neighboring phrases. Thus the extended HPB model benefits from both hierarchical rules and content-dependent phrase reordering. Experiments on large-scale translation tasks showed that the extended HPB method achieves significant improvements.

In the future, we will continue this work by combining syntactical information. For example, Zhao and Al-onaizan (2008) extended the HPB model with shallow tree-to-string rules. They disambiguated hierarchical rules by one-level tree with syntactic labels on the source side. We would also like to explore more features for content-dependent phrase reordering.

References

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*, pages 263–270.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, pages 33(2):201–228.

Xiaodong He, Mei Yang, Jianfeng Gao, Patrick Nguyen, and Robert Moore. 2008. Indirect-hmm based hypothesis alignment for combining outputs from machine translation systems. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 98–107.

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

Yang Liu, Qun Liu, and Shouxun Lin. 2006. Tree-to-string alignment template for statistical machine translation. In *Proceedings of the 44th Annual Meeting of*

the Association for Computational Linguistics, pages 609–616.

Yang Liu, Qun Liu, and Shouxun Lin. 2009. Joint decoding with multiple translation models. In *Proceedings of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 576–584.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 440–447.

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 the Association for Computational Linguistics*, 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 for Computational Linguistics*, pages 160–167.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Antti-Veikko Rosti, Spyros Matsoukas, and Richard Schwartz. 2007. Improved word-level system combination for machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 312–319.

Andreas Stolcke. 2002. Srilm – an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken language Processing*, volume 2, pages 901–904.

Dekai Wu. 1996. A polynomial-time algorithm for statistical machine translation. In *Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics*.

Deyi Xiong, Qun Liu, and Shouxun Lin. 2006. Maximum entropy based phrase reordering model for statistical machine translation. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics*, pages 521–528.

Richard Zens and Hermann Ney. 2006. Discriminative reordering models for statistical machine translation. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 55–63.

Le Zhang. 2004. Maximum entropy modeling toolkit for python and c++. available at http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html.

Bing Zhao and Yaser Al-onaizan. 2008. Generalizing local and non-local word-reordering patterns for syntax-based machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 572–581.