# Effective Use of Discontinuous Phrases for Hierarchical Phrase-based Translation

**Wei Wei**
Digital Media Content Technology Research Center, Institute of Automation,
Chinese Academic of Sciences.
weiwei@hitic.ia.ac.cn

**Bo Xu**
Digital Media Content Technology Research Center, Institute of Automation,
Chinese Academic of Sciences.
xubo@hitic.ia.ac.cn

## Abstract

Hierarchical phrase-based (HPB) models have shown strong capability in generalization and reordering. However, they are heavily dependent on continuous phrases and are difficult for modeling natural linguistic discontinuities directly. In this paper, we propose a novel approach for integrating discontinuous phrases into the Chinese-to-English HPB system. We focus on the extraction method of discontinuous phrases which retrieves various linguistic information missed in the HPB model, such as set phrases and long-distance reordering of adverbials, etc. After being transformed into the similar form to HPB rules, the translation rules with discontinuities can be seamlessly integrated into the CKY decoder. Experimental results show that the proposed approach for incorporating the linguistic discontinuities achieves statistically significant improvements over the traditional HPB system.

## 1. Introduction

Hierarchical phrase-based (HPB) translation has emerged as one of the dominant current approaches to statistical machine translation (SMT), which combines the ideas of syntax-based translation and phrase-based translation. Based on the binary synchronous context-free grammar (2-SCFG), the HPB system (Chiang, 2005) has better generalization capability and can capture long distance reordering. However, due to the limitation of phrasal continuity, HPB model cannot extract the frequent patterns in Chinese-to-English translation, such as "一…就→as soon as"(Figure 1(a)), where words in the source(Chinese) phrases may be separated by gaps. In addition, recent work (Wellington *et al.*, 2006; Søgaard and Kuhn, 2009; Søgaard and Wu, 2009) has questioned the empirical adequacy of 2-SCFG systems, which are unable to generate translation units with certain types of alignments independently. Galley and Manning (2010) pointed out that discontinuous phrases can account for these missed patterns and proposed a generalization of conventional phrase-based decoding to handle discontinuities in both source and target phrases, which yields significant improvements over Joshua(Li *et al.*, 2009)[1], a state-of-the-art HPB system.

Some other attempts to exploit phrasal discontinuities are also based on phrase-based SMT. Simard *et al.* (2005) presented an extension to Moses that allows one-word gaps in source and target phrases. This makes decoding simpler, but fixed-size discontinuous phrases are less general and will increase data sparseness. Cancedda *et al.* (2007) extended Simard's work by using flexible phrases that may contain gaps of variable lengths. It should be noted that these attempts with discontinuous phrases were mainly carried out using the left-right SMT decoder which is ineffective in allowing phrasal discontinuities. Although they tried to extend the linear decoding to support phrases with gaps, the linguistic patterns within the hierarchical structures of discontinuous phrases are difficult to be utilized fully. Additionally, He and Zong (2008) proposed a generalized reordering model for phrase-based SMT which developed a CKY style decoder to combine continuous and discontinuous phrases. However, phrasal discontinuities are only used to improve the generalization capability in their phrase-based SMT while other benefits such as long-distance reordering are not fully exploited.

In this paper, we propose a novel method to improve 2-SCFG through integrating

[1] It also significantly outperformed the conventional phrase-based MT(Moses)

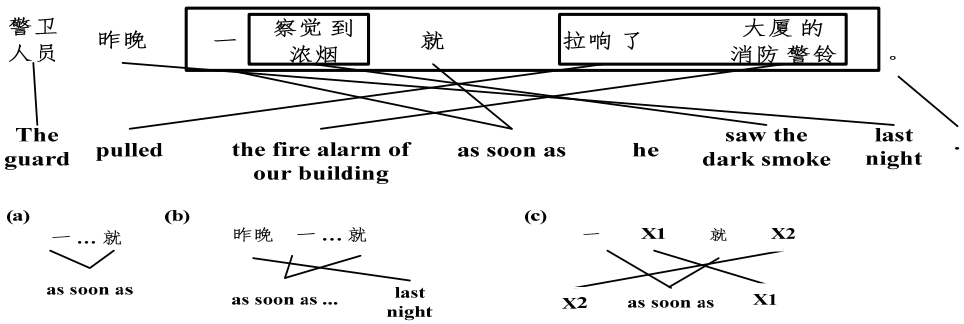


Figure 1: an example of Chinese-to-English sentence pair and some useful translation rules within it. (a) and (b) are the flexible patterns missed in HPB model: *set phrases* and *long-distance reordering of time adverbial*. (c) is the HPB rule which imposes hard hierarchical constraints

discontinuous phrases into HPB system. We focus on the extraction strategy of discontinuous phrases which can flexibly capture the linguistic translation patterns missed in 2-SCFG. Since the generation of discontinuous translation rules in our method follows the similar way as that of continuous HPB rules, the rules can be reasonably fed into CKY decoder after a series of transformations. Different from the previous work, the discontinuous phrases in our system are added into the bottom-up decoder, which naturally handles structural translation model but requires one-to-one correspondence for the gaps in source and target phrases. As a result, we propose a transforming method to seamlessly integrate the translation rules with discontinuities into the CKY decoder. Furthermore, the key to our approach is the observation that linguistic discontinuities can reduce the deficiency of 2-SCFG, hence we only use the discontinuous phrases which can retrieve various linguistic information missed in 2-SCFG. Experimental results show that the proliferation of the added translation rules introduced by discontinuous phrases is well controlled in both training and decoding. Meanwhile, incorporating the translation rules with discontinuities achieves statistically significant improvements over the traditional HPB system.

The rest of this paper is organized as follows: Section 2 introduces the motivation for our approach, especially the analysis of the problems with current HPB model; Section 3 describes the system implementation, including the procedure of discontinuous phrase extraction and integration; Section 4 reports and analyzes experimental results; Section 5 gives conclusion and future work.

## 2. Motivation

The expressive power of 2-SCFG is gained through looking for continuous phrases that contain other continuous phrases and replacing the subphrases with one non-terminal symbol X. However, due to its heavy dependency on phrasal continuities and the computational complexity constraints, HPB model often fails to capture some useful translation patterns, such as long-distance reordering of time adverbials. In this section, we will analyze the problems with current HPB model and propose the solution by integrating discontinuous phrases into HPB system. Firstly, we will give a brief description of HPB model based on 2-SCFG.

### 2.1. HPB model based on 2-SCFG

Chiang (2005) proposed a HPB translation model using 2-SCFG, which is a rewrite system consisting of production rules whose right-hand side is paired (Aho and Ullman, 1969):

$$X \rightarrow (\gamma, \alpha, \sim) \quad (1)$$

Where X is a non-terminal, $\gamma$ and $\alpha$ are strings of terminals and non-terminals. ~ is a one-to-one correspondence for the non-terminals appearing in $\gamma$ and $\alpha$.

The production rules are induced from a bilingual corpus with the help of word alignments. The procedure is as follows:

First, given a word-aligned sentence pair $(f, e, A)$, let $\overline{f}$ and $\overline{e}$ stand for any continuous sequences of $f$ and $e$. Here, $A$ represents the many-to-many word alignments which are induced by running a one-to-many word alignment model, such as GIZA++[2], in both directions and by combining the results based on a heuristic approach (Koehn *et al.*, 2003).Then a rule $(\overline{f}, \overline{e})$ is an initial phrase pair of $(f, e, A)$ iff

$$\forall f_i \in \overline{f} : (i, j) \in A \rightarrow e_j \in \overline{e}$$
$$\forall e_j \in \overline{e} : (i, j) \in A \rightarrow f_i \in \overline{f}$$

[2] http://code.google.com/p/giza-pp/

Second, based on the extracted continuous phrases, production rules are accumulated by computing the "holes" for contiguous phrases (Chiang, 2005):

1. A continuous initial phrase pair $(\bar{f}, \bar{e})$ constitutes a rule
$$X \to (\bar{f}, \bar{e}) \quad (2)$$
2. A rule $X \to (\gamma, \alpha)$ and an initial phrase pair $(\bar{f}, \bar{e})$ such that $\gamma = \gamma_1 \bar{f} \gamma_2$ and $\alpha = \alpha_1 \bar{e} \alpha_2$ constitute a rule
$$X \to (\gamma_1 X_k \gamma_2, \alpha_1 X_k \alpha_2) \quad (3)$$
3. Two glue rules are added so that it prefers combining hierarchical phrases in a serial manner:
$$S \to (SX, SX)$$
$$S \to (X, X) \quad (4)$$

To reduce the rule set size and spurious ambiguity, most HPB systems use some constraints to filter the rule set (Chiang, 2007).

### 2.2. Problems with current HPB model

Even though 2-SCFG allows some phrasal discontinuities, it is heavily dependent on continuous phrases and imposes hard hierarchical constraints (Galley, 2010). For example, the HPB rule in Figure 1(c) can be extracted only if the three long continuous phrases located in the rectangular areas are generated simultaneously. This increases higher demands on the quality of word alignments which is also an unresolved problem in SMT, especially for long phrases. In fact, the discontinuous phrase (Figure 1(a)) can play the similar role as the HPB rule (Figure 1(c)) so that it can alleviate the HPB system's dependence on continuous phrases.

Furthermore, due to the computational complexity, the extraction of 2-SCFG are usually imposed with certain limits (Chiang, 2007), including the length of initial phrases, number of non-terminals plus terminals, etc. These limits may successfully ensure the efficiency of training and decoding, but often fail to capture useful information, especially some non-local reordering patterns in Chinese-to-English translation. For instance, the reordering rule of time adverbial ( "昨晚→last night" in Figure 1(b)) cannot be extracted as HPB rules due to the length constraints (usually 10 words). Although Cai *et al.* (2009) proposed a two-step method to extract useful reordering HPB rules with less limits, the model also lost some useful information because it only extracted reordering patterns in the bilingual corpus. The discontinuous phrase (Figure 1(b)) naturally handles this reordering case.

### 2.3. Generation of discontinuous phrase rules

We present a novel method to overcome the hard constraints of 2-SCFG and to capture the information missed after HPB translation rules filtering. Instead of just using *continuous phrases*, we define the initial phrases as any subset of words of a sentence, i.e., *discontinuous phrases.* The procedure of the additional rule generation is as follows:

First, let $\tilde{f}$ and $\tilde{e}$ stand for any subsequences of $f$ and $e$, including both continuous phrases $(\bar{f}, \bar{e})$ and discontinuous phrases $(f_\diamond, e_\diamond)$. A phrase $(\tilde{f}, \tilde{e})$ is an initial phrase pair of $(f, e, A)$ iff

$$\forall f_i \in \tilde{f} : (i, j) \in A \to e_j \in \tilde{e}$$
$$\forall e_j \in \tilde{e} : (i, j) \in A \to f_i \in \tilde{f}$$

Then, each discontinuous phrase $(f_\diamond, e_\diamond)$ should include a sequence of words and gaps (indicated by the symbol $\diamond$): the gap here acts as a placeholder for the sequence of unspecified words. To avoid redundancy, phrases may not begin or end with a gap. Thus a discontinuous initial phrase pair $(f_\diamond, e_\diamond)$ can be rewritten as $(\bar{\gamma}_1 \diamond \bar{\gamma}_2, \bar{\alpha}_1 \diamond \bar{\alpha}_2)$, where $\diamond$ represents the gaps in source or target discontinuous phrases, $\bar{\gamma}$ and $\bar{\alpha}$ stand for continuous sequence of words in $f_\diamond$ and $e_\diamond$ .The set of rules based on discontinuous phrases must satisfy:

4. A discontinuous initial phrase pair $(f_\diamond, e_\diamond)$ constitutes a rule
$$X \to (\bar{\gamma}_1 \diamond \bar{\gamma}_2, \bar{\alpha}_1 \diamond \bar{\alpha}_2) \quad (5)$$

The notable difference between rules (3) and (5) lies in non-terminal symbols $X_k$ and the gap symbol $\diamond$ . $X_k$ , where $k$ is the index of non-terminals not used in $\gamma$ and $\alpha$ , built a one-to-one correspondence for the continuous subphrase in $(\bar{f}, \bar{e})$ . On the other hand, $\diamond$ collects the related gap information in $f_\diamond$ and $e_\diamond$ , and has no requirement for strict alignment between subphrases. Thus, the discontinuous phrases with gaps impose more flexible hierarchical constraints and can better account

for various linguistic phenomena such as set phrases and long-distance reordering.

## 3. System implementation

### 3.1. Discontinuous phrase extraction

If arbitrary number of words and gaps may be rules, the amount of source discontinuous phrases that cover each sentence is exponential in the sentence length. This is especially problematic for training and decoding. In order to make the balance between efficiency and accuracy, we should impose some limitations on discontinuous phrase extraction. Given the characteristics of the HPB model, which allows some phrasal discontinuities but imposes hard hierarchical constraints and fails to capture some useful linguistic patterns, the aim of integrating discontinuous phrases into our system is to reduce the deficiency of translation model based on 2-SCFG. Therefore, we propose a reasonable approach for filtering the discontinuous phrases $(f_{\Diamond}, e_{\Diamond})$ according to the following constraints:

(a). Either $f_{\Diamond}$ or $e_{\Diamond}$ has no more than one gap, thus the gap can be categorized as $\Diamond \in \{\Diamond_f, \Diamond_e\}$.

(b). The sequence of unspecified source words, which is represented by the gap $\Diamond_f$, should align with the sequence of words outside the target phrase on one direction (that means on the left side of $\bar{\alpha}_1$ or the right side of $\bar{\alpha}_2$), and similar to the alignment of unspecified target words $\Diamond_e$.

(c). $(f_{\Diamond}, e_{\Diamond})$ is limited to a length of 5 words plus gaps on each side.

Following the constraints above, the discontinuous phrase rule (5) can be changed to:

$$X \rightarrow (\bar{\gamma}, \bar{\alpha}_1 \Diamond_e \bar{\alpha}_2) \quad (6)$$

$$X \rightarrow (\bar{\gamma}_1 \Diamond_f \bar{\gamma}_2, \bar{\alpha}) \quad (7)$$

$$X \rightarrow (\bar{\gamma}_1 \Diamond_f \bar{\gamma}_2, \bar{\alpha}_1 \Diamond_e \bar{\alpha}_2) \quad (8)$$

According to constraints (b), $\Diamond_f$ and $\Diamond_e$ can be subdivided into:

$$\Diamond_f \in \{\Diamond_f^R, \Diamond_f^L\} \quad \Diamond_e \in \{\Diamond_e^R, \Diamond_e^L\} \quad (9)$$

where *R* and *L*, the abbreviations of *right* and *left*, are the directions in which the gaps align with the substrings outside the phrases. The discontinuous phrase rules as (6) and (7) are continuous on either source or target side. The rules like (8) are composed of discontinuous phrases which have gaps on both sides.

Based on the many-to-many word alignments between each sentence pair, we extract discontinuous phrases $(\bar{\gamma}, \bar{\alpha}_1 \Diamond_e \bar{\alpha}_2)$ which are continuous on the source side, following the method adopted in traditional phrase-based MT approaches (Zens and Ney, 2003; Och and Ney, 2004). The difference lies in that the target phrase $\bar{\alpha}_1 \Diamond_e \bar{\alpha}_2$ has one gap symbol $\Diamond_e \in \{\Diamond_e^R, \Diamond_e^L\}$, which stands for the sequence of words aligned outside of the source phrase $\bar{\gamma}$. When extracting discontinuous phrases $(\bar{\gamma}_1 \Diamond_f \bar{\gamma}_2, \bar{\alpha})$, we exchange the source and target sentences due to the symmetry of word alignments. After locating the two sets of discontinuous phrases above and all the continuous phrases, $(\bar{\gamma}_1 \Diamond_f \bar{\gamma}_2, \bar{\alpha}_1 \Diamond_e \bar{\alpha}_2)$ can be produced by enumerating different collocations of them, and then be filtered with the constraints (a)、(b) and (c).

### 3.2. Features in the model

Based on the 2-SCFG like (1), each rule in the novel model is associated with a score that is computed via the following log linear formula:

$$w(X \rightarrow \langle \gamma, \alpha, \sim \rangle) = \prod_i \phi_i(\tilde{f}, \tilde{e})^{\lambda_i}$$

where $\phi_i(\tilde{f}, \tilde{e})$ is a feature describing one particular aspect of the rule associated with the source and target phrases $(\tilde{f}, \tilde{e})$, and $\lambda_i$ is the corresponding weight of that feature. Following the HPB model, typical features used in our system are relative-frequency phrase translation probability $P(\tilde{f} \mid \tilde{e})$ and its inverse $P(\tilde{e} \mid \tilde{f})$, lexically weighted phrase translation probability $lex(\tilde{f} \mid \tilde{e})$ and its inverse $lex(\tilde{e} \mid \tilde{f})$. However, our computation of translation probability is different from the one in the previous model, since we need to account for discontinuous phrases with one gap. For the relative-frequency translation probability, the gap symbol $\Diamond$ can be viewed as a special token, and we give a count of one to each discontinuous phrase pair occurrence. For the lexical weight of phrase pair, we check how well its words translate to each other except the gap.

Besides the translation probability features, our system generally employ the phrase length penalty, glue rule and target language model features (Chiang, 2007). We also add a penalty feature in order to allow the model to learn a

| Discontinuous phrase rules | Transformed rules |
|---|---|
| $X \rightarrow (\bar{\gamma}, \bar{\alpha}_1 \Diamond_e^L \bar{\alpha}_2)$<br>$X \rightarrow (\bar{\gamma}, \bar{\alpha}_1 \Diamond_e^R \bar{\alpha}_2)$ | $X \rightarrow (X_1 \circ \bar{\gamma}, \bar{\alpha}_1 X_1 \bar{\alpha}_2)$<br>$X \rightarrow (\bar{\gamma} \circ X_1, \bar{\alpha}_1 X_1 \bar{\alpha}_2)$ |
| | $S \rightarrow (\bar{\gamma}_{\alpha_1} \bar{\gamma}_{\alpha_2} \circ X, \bar{\alpha}_1 X \bar{\alpha}_2)$<br>$S \rightarrow (\bar{\gamma}_{\alpha_2} \bar{\gamma}_{\alpha_1} \circ X, \bar{\alpha}_1 X \bar{\alpha}_2)$ |
| $X \rightarrow (\bar{\gamma}_1 \Diamond_f^L \bar{\gamma}_2, \bar{\alpha})$<br>$X \rightarrow (\bar{\gamma}_1 \Diamond_f^R \bar{\gamma}_2, \bar{\alpha})$ | $X \rightarrow (\bar{\gamma}_1 X_1 \bar{\gamma}_2, X_1 \circ \bar{\alpha})$<br>$X \rightarrow (\bar{\gamma}_1 X_1 \bar{\gamma}_2, \bar{\alpha} \circ X_1)$ |
| $X \rightarrow (\bar{\gamma}_1 \Diamond_f^L \bar{\gamma}_2, \bar{\alpha}_1 \Diamond_e^L \bar{\alpha}_2)$<br>$X \rightarrow (\bar{\gamma}_1 \Diamond_f^L \bar{\gamma}_2, \bar{\alpha}_1 \Diamond_e^R \bar{\alpha}_2)$<br>$X \rightarrow (\bar{\gamma}_1 \Diamond_f^R \bar{\gamma}_2, \bar{\alpha}_1 \Diamond_e^L \bar{\alpha}_2)$<br>$X \rightarrow (\bar{\gamma}_1 \Diamond_f^R \bar{\gamma}_2, \bar{\alpha}_1 \Diamond_e^R \bar{\alpha}_2)$ | $X \rightarrow (X_1 \circ \bar{\gamma}_1 X_2 \bar{\gamma}_2, X_2 \circ \bar{\alpha}_1 X_1 \bar{\alpha}_2)$<br>$X \rightarrow (\bar{\gamma}_1 X_1 \bar{\gamma}_2 \circ X_2, X_1 \circ \bar{\alpha}_1 X_2 \bar{\alpha}_2)$<br>$X \rightarrow (X_1 \circ \bar{\gamma}_1 X_2 \bar{\gamma}_2, \bar{\alpha}_1 X_1 \bar{\alpha}_2 \circ X_2)$<br>$X \rightarrow (\bar{\gamma}_1 X_1 \bar{\gamma}_2 \circ X_2, \bar{\alpha}_1 X_2 \bar{\alpha}_2 \circ X_1)$ |

Table 1: Corresponding relationship between Discontinuous phrase rules and Transformed rules

preference for continuous or discontinuous phrases.

### 3.3. Integrating discontinuous phrase rules into CKY decoding

The decoder of HPB system is built upon the parser style algorithm, such as CKY, which naturally handles structural translation rules with gaps. Moreover, the chart-based CKY decoder requires one-to-one correspondence for the gaps in source and target phrases. In the ordinary HPB rules, the non-terminal symbol $X_k$ builds the correspondence for continuous subphrases. However, the discontinuous phrase rules ((6), (7) and (8)) use the symbol $\Diamond \in \{\Diamond_f, \Diamond_e\}$ to record the gap information, which has no requirement for strict alignment. Thus, we propose a solution to add the proper correspondence related with the gap symbols in order to integrate the discontinuous phrase rules into CKY decoder.

According to the procedure of discontinuous phrase extraction (Section 3.1), the gap symbol represents the sequence of unspecified words, which are aligned with the substrings on the left or right side of corresponding phrases. Thus the discontinuous phrase rules can be transformed into the rules which have the similar form of 2-SCFG with the non-terminal $X$ for one-to-one correspondence. In order to reserve the flexibility, we introduce the symbol $\circ$ to match any number of words, including NONE. Moreover, we try to add phrasal discontinuities and reordering into the glue rules when transforming the discontinuous phrase rules like (6). The corresponding relationship is shown in table 1.

Since there exist some overlaps, we integrate the transformed rules into translation system only if no matching HPB rules can be found. The procedure can be described as follows:

First，we collect the transformed rules in each sentence span simultaneously with the analysis of HPB rules related with the same span.

Then the collected transformed rules can be divided into two parts: one can be matched with certain HPB rules; the other contains the lost knowledge in the HPB rule set. Thus it is necessary to incorporate the latter set as complementary rules. Due to the placeholder symbol $\circ$ in each transformed rule, we should treat the problem as fuzzy matching and have to integrate more than one corresponding HPB rules in decoding.

Finally，for the transformed glue rules (starting with the symbol $S$ in Table 1), we record the alignment between $\bar{\gamma}_\alpha$ and $\bar{\alpha}$ so that they can provide discontinuity and reordering information to help the HPB decoder connect two adjacent source phrases together, which play the similar role as glue rules.

## 4. Experiments and results

### 4.1. Corpus

Our experiments were made on two Chinese-to-English translation tasks: IWSLT-07 (dialogue domain) and NIST-06 (news domain).

**IWSLT-07.** We performed translation experiments on the Basic Traveling Expression Corpus (BTEC) for the Chinese-English task. Both the bilingual training data and the 4-gram language model (LM) training data are restricted to the supplied corpus, which contain 39,950 sentence pairs, 350K Chinese words and 382K English words. IWSLT-07 test set consists of 489 sentences, which includes 3,287 Chinese words. We used the IWSLT-05 test set consisting of 506 sentence pairs as development set.

**NIST-06.** The bilingual training corpus comes from Linguistic Data Consortium (LDC)[3], which consists of 3.4M sentence pairs with 64M/70M words of Chinese/English. The LM training corpus is from the English side of the parallel data as well as the English Gigaword corpus[4], which consists of 11.3M sentences. Our test set is 2006 NIST MT Evaluation test set (1664 sentences), and our development set is 2005 NIST MT Evaluation test set (1084 sentences).

[3] LDC2002E18, LDC2002T01, LDC2003E07, LDC2003E14, LDC2003T17, LDC2004T07, LDC2004T08, LDC2005T06, LDC2005T10, LDC2005T34, LDC2006T04, LDC2007T09

[4] LDC2007T07

| | HPB | TDP | UTDP |
|---|---|---|---|
| Training | 1929.6K | 589.1K | |
| Decoding | 100.2K | 34.7K | 9.2K |

Table 2: The number of translation rules on IWSLT-07 task

| | HPB | TDP | UTDP |
|---|---|---|---|
| Training | 146.4M | 36.5M | |
| Decoding | 13.5M | 4.2M | 1.4M |

Table 3: The number of translation rules on MT NIST-06 task

### 4.2. Setup

We obtained the word alignments using the way of Koehn *et al*. (2003). After running GIZA++ in both directions, we applied the "grow-diag-final" refinement rule on the intersection alignments for each sentence pair, and extracted continuous phrases of length at most 10 words on both sides together with their internal alignments. Simultaneously, we can extract discontinuous phrases according to the method in Section 3.1.

In the baseline system, we only produced the ordinary hierarchical rules following the same constraints as in Chiang (2007). In the contrast experiments, we added the discontinuous phrase rules to the baseline. Our 4-gram language model was trained on English corpus using the SRILM toolkit (Stolcke, 2002) with modified Kneser-Ney smoothing methods. The base feature set for all systems is similar to the set used in Chiang (2007). Additionally, we added a penalty feature in order to learn a preference for continuous or discontinuous phrases. All the features were combined into a standard log-linear model, which was trained using minimum error-rate training (Och, 2003) to maximize the BLEU-4 (Papineni *et al*., 2002) score on the development sets.

### 4.3. Rules Comparison

The statistics of the hierarchical phrase rules (**HPB rules**) and the transformed discontinuous phrase rules (**TDP rules**) are listed in Table 2 and 3. In order to reflect the differences of rule distributions in the process of training and decoding, we counted the number of rules in the two procedures, respectively. Since some of the **TDP** rules may overlap with the ordinary **HPB** rules in certain sentence spans, we also gave the number of the actual use of the TDP rules (**UTDP rules**) to show how many discontinuous phrases rules are incorporated into decoding.

We could see that the number of UTDP rules, which are considered as the additional information, is acceptable compared with the number of HPB rules used in decoding. In IWSLT-07 translation task, the UTDP rule set accounts for about 9.1% of the HPB rules and 26.5% of the whole TDP rule set. For NIST-06 task, the ratios of UTDP rules are higher (10.4% and 33.3%, respectively). The main reason is that the sentences in news domain are longer so that the flexibility of TDP rules will play a more important role. The results also confirm that the proliferation of UTDP rules can be well controlled so that it is feasible to integrate discontinuous phrases into CKY decoder following the way described in this paper.

| System | **IWSLT-07** | | **NIST-06** | |
|---|---|---|---|---|
| | BLEU | METEOR | BLEU | METEOR |
| **Baseline** | 31.78 | 56.61 | 30.83 | 53.78 |
| **+UTDP** | **32.35**** | **57.68**** | **31.75**** | **54.97**** |

Table 4: BLEU [%] and METEOR[%] scores in experiments(**: significance at the 0.01 level)

| Condition | BLEU | METEOR |
|---|---|---|
| Baseline | 30.83 | 53.78 |
| $+\Diamond_{tgt}(\Diamond_{tgt}S+\Diamond_{tgt}X)$ | 31.57 | 54.64 |
| only $+\Diamond_{tgt}S$ | **31.48** | **54.46** |
| $+\Diamond_{src}$ | 31.08 | 54.14 |
| $+\Diamond_{src+tgt}$ | 31.19 | 54.09 |
| ALL($+\Diamond_{tgt}+\Diamond_{src}+\Diamond_{src+tgt}$) | 31.75 | 54.97 |

Table 5: BLEU [%] and METEOR[%] scores on NIST06 after integrating each category of UTDP rules

### 4.4. Results and Analysis

To test the effect of discontinuous phrase rules when integrating into CKY decoder, we ran various kinds of translation experiments and used two evaluation metrics[5]: BLEU and METEOR (Banerjee and Lavie, 2005). Statistical significance of difference from the baseline scores was measured by using paired bootstrap re-sampling (Koehn, 2004).

The first experiment (**+UTDP**) is to compare the results of the integration of transformed rules with the results of the baseline system. Note that the UTDP rules are added only if there are no matching HPB rules in the sentence span. The experiments' results are shown in Table 4. We could see that the performance of adding UTDP rules in both evaluation tasks is significantly better than the baseline system. Moreover, the improvements on large data track are more noticeable (while adding UTDP rules improves performance by 0.92 BLEU points and 1.19

[5] All the models are tuned on BLEU ( case-insensitive) and evaluated on BLEU and METEOR

| | |
|---|---|
| **Sentence 1** | 梅丽安***是***海湾国家巴林王室***的成员***。 |
| Baseline | Meriam is the Gulf state of Bahrain members of the royal family. |
| +UTDP rules | Meriam ***is a member of*** the Gulf state of Bahrain's royal family. |
| Reference | Meriam is a member of the royal family of the Gulf country of Bahrain. |
| **Sentence 2** | 布希***明年二月二十二日将飞往***布鲁塞尔与欧盟及北大西洋公约组织领袖会面 。 |
| Baseline | Bush will fly to Brussels on February 22 next year and the European Union and NATO leaders. |
| +UTDP rules | Bush ***will fly to*** Brussels with the EU and NATO leaders ***on February 22 next year***. |
| Reference | Bush will fly to Brussels to meet with EU and NATO leaders on February 22 next year. |
| **Sentence 3** | ***通过立法遏止***“台独”分子的***分裂行径是***海外侨胞和台湾海峡两岸同胞***的共同意志***。 |
| Baseline | Through legislation to curb Taiwan independence elements of secession activities is the common aspiration of the overseas Chinese and Taiwan compatriots on both sides of the strait. |
| +UTDP rules | ***To curb separatist acts*** of Taiwan independence ***through legislation is the common will of*** overseas Chinese and Taiwan compatriots on both sides of the strait. |
| Reference | To suppress the divisive action of "Taiwanese Separatists" through legislation is the common will of overseas Chinese and compatriots across the Taiwan Strait |

Table 6: Actual translation results produced by HPB and our systems

METEOR points on NIST-06). It is probably because translating complex sentences in news domain needs more additional rules to capture linguistic translation patterns, such as long distance reordering, set phrases and so on. It reflects from another aspect that our method with discontinuous phrases is able to retrieve various linguistic information which is complementary to that given by the traditional HPB systems.

We then divided the UTDP rules into three categories like rules (6), (7), (8), and study the effects of different kinds of UTDP rules on NIST-06 set. The evaluation scores after integrating each category into the baseline system are shown in the Table 5. The left column shows various conditions of using the rules, for example the symbol $+\Diamond_{tgt}$ represents the integration of the rules which only have the gap on the target side. From the results on the right column, we could conclude that each kind of the UTDP rules made their own contributions to the improvement of translation result.

Further analysis of Table 5 shows that our system allowing phrasal discontinuities on the target side ($+\Diamond_{tgt}$) has the best performance. In order to have a more comprehensive understanding, we subdivided this kind of UTDP rules into $\Diamond_{tgt}X$ and $\Diamond_{tgt}S$ following the transformed relationship (row 1 in Figure 1), where $X$ and $S$ represent different start symbols of the transformed rules. It is interesting to notice that our system only integrating the transformed glue rules (only $+\Diamond_{tgt}S$) performs almost as well as the system that allows all the target side discontinuities ($\Diamond_{tgt}S+\Diamond_{tgt}X$). For instance, while adding $\Diamond_{tgt}S$ improves the system performance by 0.65 BLEU points and 0.68 METEOR points, further enabling $\Diamond_{tgt}X$ only raises the performance by a mere 0.09 BLEU points and 0.18 METEOR points. Although the amount of $\Diamond_{tgt}S$ rules is about 6.8%[6] of all the UTDP rules, their effects cannot be overlooked because the rules introduce additional knowledge into the huge amount of glue rules which simply connect translations of two adjacent blocks in monotonic order. In HPB system, glue rules are often used as the only rules to connect two spans when there are no matching hierarchical rules. However, this half-measure neglect the possibilities of phrasal reordering and discontinuity thus may degrade the translation performance. Moreover, the number of the glue rules used in translating NIST-06 set accounts for about 25.9% of all the rules used in HPB system[7], and they play an important role in decoding. Therefore it is an efficient way to improve the translation quality through integrating discontinuous phrases to help the ordinary glue rules.

Table 6 gives some comparisons of the translation results between HPB and our system. The first two examples show the efficiency of the UTDP rules to accurately translate the discontinuous phrases either on source or target side. The last one is an example of more complex sentence which needs the rules to capture both long distance reordering and discontinuity. The actual translation results show that the UTDP rules can capture certain linguistic translation patterns that ordinary HPB rules fail to describe.

## 5. Conclusion and future work

In this paper, we proposed a novel approach for integrating discontinuous phrases into the

[6] The number of $\Diamond_{tgt}S$ rules we used when translating NIST-06 set is about 95K(the number of UDTP is 1.4M )

[7] When translating NIST-06 set, HPB system used 2,350,914 2-SCFG rules and 821,709 glue rules in total.

Chinese-to-English HPB system. Because the integration of discontinuous phrases can reduce the deficiency of 2-SCFG, our system significantly outperformed the conventional HPB system. We found that it is an efficient way to improve the translation performance through integrating the transformed glue rules with discontinuity and reordering. Recently, we have successfully built the distributed HPB translation platform based on parallel CKY decoding and large-scale distributed language model, which proves that it is feasible to integrate much more complex knowledge into machine translation.

In future work, we plan to realize some more efficient methods of extracting discontinuous phrases, such as the online solution using suffix array for pattern matching (Lopez, 2007). In order to integrate more useful information, we will also consider extending the coverage of discontinuous phrases based on our distributed HPB translation system.

## References


Alfred V. Aho and Jeffrey D. Ullman. 1969. Syntax directed translations and the pushdown assembler. *Journal of Computer and System Sciences*, 3:37–56.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of ACL*, pages 101–104.

Shu Cai, Yajuan L and Qun Liu. 2009. Improved Reordering Rules for Hierarchical Phrase-based Translation. In *Proceedings of International Conference on Asian Language Processing*, pages 65-70.

Nicola Cancedda, Marc Dymetman, Éric Gaussier, Cyril Goutte. 2007. An Elastic-Phrase Model for Statistical Machine Translation. In *Jounées de l'ATALA (Assocation pour le Traitement Automatique des Langues).*

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of ACL* , pages 263–270.

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

Michel Galley and Christopher D. Manning. 2010. Accurate Non-Hierarchical Phrase-Based Translation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (HLT-NAACL)*, pages 966–974.

Yanqing He, Chengqing Zong. 2008. A Generalized Reordering Model for Phrase-Based Statistical Machine Translation. In *Proceedings of the 8th Conference of the Association for Machine Translation in the Americas (AMTA)*, pages117-124.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of NAACL*, pages 48–54.

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

Zhifei Li, Chris Callison-Burch, Chris Dyer, Juri Ganitkevitch, Sanjeev Khudanpur, Lane Schwartz, Wren N. G. Thornton, JonathanWeese, and Omar F. Zaidan.2009. Joshua: an open source toolkit for parsing based MT. In *Proceedings of the Workshop on Statistical Machine Translation (WMT09).*

Adam Lopez. 2007. Hierarchical phrase-based translation with suffix arrays. In *Proceedings of EMNLP-CONLL*, pages 976–985.

Franz Josef Och. 2003. Minimum error rate training for statistical machine translation. In *Proceedings of ACL*, pages 160-167.

Franz Josef Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417-449.

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

Michel Simard, Nicola Cancedda, Bruno Cavestro, Marc Dymetman, Eric Gaussier, Cyril Goutte, Kenji Yamada, Philippe Langlais, and Arne Mauser. 2005. Translating with non-contiguous phrases. In *Proceedings of HLT-EMNLP*, pages 755–762.

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

Anders Søgaard and Jonas Kuhn. 2009. Empirical lower bounds on alignment error rates in syntax-based machine translation. In *Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation (SSST-3) at NAACL-HLT*, pages 19–27.

Anders Søgaard and Dekai Wu. 2009. Empirical lower bounds on translation unit error rate for the full class of inversion transduction grammars. In *Proceedings of the 11th International Conference on Parsing Technologies( IWPT)*, pages 33–36.

Benjamin Wellington, Sonjia Waxmonsky, and I. Dan Melamed. 2006. Empirical lower bounds on the complexity of translational equivalence. In *Proceedings of COLING-ACL*, pages 977–984

Richard Zens and Hermann Ney. 2003. A comparative study on reordering constraints in statistical machine translation. In *Proceedings of ACL*, pages 144–151.