Integrating Phrase-based Reordering Features into a Chart-based Decoder for Machine Translation

ThuyLinh Nguyen
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
thuylinh@cs.cmu.edu

Stephan Vogel
Qatar Computing Research Institute
Tornado Tower
Doha, Qatar
svogel@qf.org.qa

Abstract

Hiero translation models have two limitations compared to phrase-based models: 1) Limited hypothesis space; 2) No lexicalized reordering model. We propose an extension of Hiero called Phrasal-Hiero to address Hiero's second problem. Phrasal-Hiero still has the same hypothesis space as the original Hiero but incorporates a phrase-based distance cost feature and lexicalized reodering features into the chart decoder. The work consists of two parts: 1) for each Hiero translation derivation, find its corresponding discontinuous phrase-based path. 2) Extend the chart decoder to incorporate features from the phrase-based path. We achieve significant improvement over both Hiero and phrase-based baselines for Arabic-English, Chinese-English and German-English translation.

1 Introduction

Phrase-based and tree-based translation model are the two main streams in state-of-the-art machine translation. The tree-based translation model, by using a synchronous context-free grammar formalism, can capture longer reordering between source and target language. Yet, tree-based translation often underperforms phrase-based translation in language pairs with short range reordering such as Arabic-English translation (Zollmann et al., 2008; Birch et al., 2009).

We follow Koehn et al. (2003) for our phrase-based system and Chiang (2005) for our Hiero system. In both systems, the translation of a source sentence $\mathbf{f}$ is the target sentence $\mathbf{e}^*$ that maximizes a linear combination of features and weights:

$$\langle \mathbf{e}^*, \mathbf{a}^* \rangle = \underset{\langle \mathbf{e},\mathbf{a} \rangle \in \mathcal{H}(\mathbf{f})}{\operatorname{argmax}} \sum_{m \in M} \lambda_m h_m (\mathbf{e}, \mathbf{f}, \mathbf{a}) . \quad (1)$$

where

- $\mathbf{a}$ is a translation path of $\mathbf{f}$. In the phrase-based system, $\mathbf{a}_{\text{ph}}$ represents a segmentation of $\mathbf{e}$ and $\mathbf{f}$ and a correspondance of phrases. In the Hiero system, $\mathbf{a}_{\text{tr}}$ is a derivation of a parallel parse tree of $\mathbf{f}$ and $\mathbf{e}$, each nonterminal representing a rule in the derivation.
- $\mathcal{H}(\mathbf{f})$ is the hypothesis space of the sentence $\mathbf{f}$. We denote $\mathcal{H}_{\text{ph}}(\mathbf{f})$ as the phrase-based hypothesis space of $\mathbf{f}$ and $\mathcal{H}_{\text{tr}}(\mathbf{f})$ as its tree-based hypothesis space. Galley and Manning (2010) point out that due to the hard constraints of rule combination, the tree-based system does not have the same excessive hypothesis space as the phrase-based system.
- M is the set of feature indexes used in the decoder. Many features are shared between phrase-based and tree-based systems including language model, word count, and translation model features. Phrase-based systems often use a lexical reordering model in addition to the distance cost feature.

The biggest difference in a Hiero system and a phrase-based system is in how the reordering is modeled. In the Hiero system, the reordering decision is encoded in weighted translation rules, determined by nonterminal mappings. For example, the rule $X \rightarrow ne\ X_1\ pas\ ;\ not\ X_1\ : w$ indicates the translation of the phrase between *ne* and *pas* to be after the English word *not* with score w. During decoding, the system parses the source sentence and synchronously generates the target output.

To achieve reordering, the phrase-based system translates source phrases out of order. A reordering distance limit is imposed to avoid search space explosion. Most phrase-based systems are equipped with a distance reordering cost feature to tune the system towards the right amount of reordering, but then also a lexicalized reordering

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 1587–1596,
Sofia, Bulgaria, August 4-9 2013.

model to model the direction of adjacent source phrases reordering as either *monotone*, *swap* or *discontinuous*.

There are two reasons to explain the shortcomings of the current Hiero system:

1. A limited hypothesis space because the synchronous context-free grammar is not applicable to non-projective dependencies.

2. It does not have the expressive lexicalized reordering model and distance cost features of the phrase-based system.

When comparing phrase-based and Hiero translation models, most of previous work on tree-based translation addresses its limited hypothesis space problem. Huck et al. (2012) add new rules into the Hiero system, Carreras and Collins (2009) apply the tree adjoining grammar formalism to allow highly flexible reordering. On the other hand, the Hiero model has the advantage of capturing long distance and structure reordering. Galley and Manning (2010) extend phrase-based translation by allowing gaps within phrases such as $\langle ne \ldots pas, not \rangle$, so the decoder still has the discriminative reordering features of phrase-based, but also uses on average longer phrases. However, these phrase pairs with gaps do not capture structure reordering as do Hiero rules with nonterminal mappings. For example, the rule $\mathrm{X} \rightarrow ne\ \mathrm{X}_1\ pas\ ;\ not\ \mathrm{X}_1$ explicitly places the translation of the phrase between ne and pas behind the English word not through nonterminal X_1. This is important for language pairs with strict reordering. In our Chinese-English experiment, the Hiero system still outperforms the discontinuous phrase-based system.

We address the second problem of the original Hiero decoder by mapping Hiero translation derivations to corresponding phrase-based paths, which not only have the same output but also preserve structure distortion of the Hiero translation. We then include phrase-based features into the Hiero decoder.

A phrase-based translation path is the sequence of phrase-pairs, whose source sides cover the source sentence and whose target sides generate the target sentence from left to right. If we look at the leaves of a Hiero derivation tree, the lexicals also form a segmentation of the source and target sentence, thus also form a discontinuous phrase-based translation path. As an example, let us look at the translation of the French sentence *je ne parle pas le française* into English *i don't speak french* in Figure 1. The Hiero decoder translates the sentence using a derivation of three rules:

- $r_1 = \mathrm{X} \rightarrow parle\ ;\ speak.$
- $r_2 = \mathrm{X} \rightarrow ne\ \mathrm{X}_1\ pas\ ;\ don't\ \mathrm{X}_1.$
- $r_3 = \mathrm{X} \rightarrow$ $Je\ \mathrm{X}_1\ le\ Français\ ;\ I\ \mathrm{X}_1\ french.$

From this Hiero derivation, we have a segmentation of the sentence pairs into phrase pairs according to the word alignments, as shown on the left side of Figure 1. Ordering these phrase pairs according the word sequence on the target side, shown on the right side of Figure 1, we have a phrase-based translation path consisting of four phrase pairs: $(je, i), (ne \ldots pas, not), (parle, speak), (lefrancaise, french)$ that has the same output as the Hiero system. Note that even though the Hiero decoder uses a composition of three rules, the corresponding phrase-based path consists of four phrase pairs. We name this new variant of the Hiero decoder, which uses phrase-based features, Phrasal-Hiero.

Our Phrasal-Hiero addresses the shortcomming of the original Hiero system by incorporating phrase-based features. Let us revisit machine translation's loglinear model combination of features in equation 1. We denote $ph(\mathbf{a})$ as the corresponding phrase-based path of a Hiero derivation $\mathbf{a}$, and $M_{Ph \setminus H}$ as the indexes of phrase-based features currently not applicable to the Hiero decoder. Our Phrasal-Hiero decoder seeks to find the translation, which optimizes:

$$\langle \mathbf{e}^*, \mathbf{a}^* \rangle = \underset{\langle \mathbf{e}, \mathbf{a} \rangle \in \mathcal{H}_{\mathrm{tr}}(\mathbf{f})}{\operatorname{argmax}} \Big(\sum_{m \in M_H} \lambda_m h_m(\mathbf{e}, \mathbf{f}, \mathbf{a}) + \sum_{m' \in M_{Ph \setminus H}} \lambda_{m'} h_{m'}(\mathbf{e}, \mathbf{f}, ph(\mathbf{a})) \Big).$$

We focus on improving the modelling of reordering within Hiero and include discriminative reordering features (Tillmann, 2004) and a distance cost feature, both of which are not modeled in the original Hiero system. Chiang et al. (2008) added structure distortion features into their decoder and showed improvements in their Chinese-English experiment. To our knowledge, Phrasal-Hiero is the first system, which directly integrates phrase-based and Hiero features into one model.

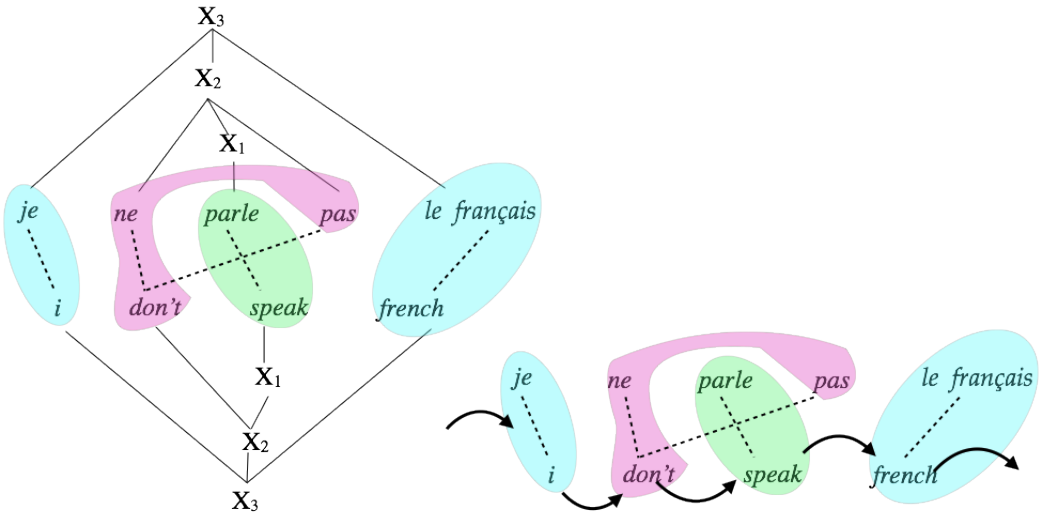

Figure 1: Example of French-English Hiero Translation on the left and its corresponding discontinuous phrase-based translation on the right.

Rules	Alignments	Phrase pairs & nonterminals
$r_1 = \text{X} \rightarrow parle\ ;\ speak.$	0-0	$(parle\ ;\ speak)$
$r_2 = \text{X} \rightarrow ne\ \text{X}_1\ pas\ ;\ don't\ \text{X}_1.$	0-0 1-1 2-0	$(ne \ldots pas\ ;\ don't)\ ;\ \text{X}_1$
$r_3 = \text{X} \rightarrow Je\ \text{X}_1\ le\ Francais\ ;\ I\ \text{X}_1\ French$	0-0 1-1 3-2	$(Je\ ;I)\ ;\ \text{X}_1\ ;\ (le\ Francais;\ french)$
$r_4 = \text{X} \rightarrow je\ \text{X}_1\ le\ \text{X}_2\ ;\ i\ \text{X}_1\ \text{X}_2$	0-0 1-1 3-2	Not Applicable

Table 1: Rules and their sequences of phrase pairs and nonterminals

Previous work has attempted to weaken the context free assumption of the synchronous context free grammar formalism, for example using syntactic non-terminals (Zollmann and Venugopal, 2006). Our approach can be viewed as applying soft context constraint to make the probability of substituting a nonterminal by a subtree depending on the corresponding phrase-based reordering features.

In the next section, we explain the model in detail.

2 Phrasal-Hiero Model

Phrasal-Hiero maps a Hiero derivation into a discontinuous phrase-based translation path by the following two steps:

1. Training: Represent each rule as a sequence of phrase pairs and nonterminals.

2. Decoding: Use the rules' sequences of phrase pairs and nonterminals to find the corresponding phrase-based path of a Hiero derivation and calculate its feature scores.

2.1 Map Rule to A Sequence of Phrase Pairs and Nonterminals

We segment the rules' lexical items into phrase pairs. These phrase pairs will be part of the phrase-based translation path in the decoding step. The rules' nonterminals are also preserved in the sequence, during the decoding they will be substituted by other rules' phrase pairs. We now explain how to map a rule to a sequence of phrase pairs and nonterminals.

Let $r = \text{X} \rightarrow s_0 X_1 s_1 \ldots X_k s_k\ ;\ t_0 X_{\alpha(1)} t_1 \ldots X_{\alpha(k)} t_k$ be a rule of k nonterminals, $\alpha(.)$ defines the sequence of nonterminals on the target. s_i or t_i , $i = 0 \ldots k$ are phrases between nonterminals, they can be empty because nonterminals can be at the border of the rule or two nonterminals are adjacent. For example the rule $\text{X} \rightarrow ne\ \text{X}_1\ pas\ ;\ not\ \text{X}_1$ has $k = 1$, $s_0 = ne$, $s_1 = pas$, $t_0 = not$, t_1 is an empty phrase because the target X_1 is at the rightmost position.

Phrasal-Hiero retains both nonterminals and lexical alignments of Hiero rules instead of only nonterminal mappings as in (Chiang, 2005). A

rule's lexical alignment is the most frequent one in the training data. We use the lexical alignments of a rule to decide how source phrases and target phrases are connected. In the rule r, a source phrase s_i is connected to a target phrase $t_{i'}$ if at least one word in s_i aligns to a target word in $t_{i'}$. In the rule $\mathrm{X} \rightarrow Je\ \mathrm{X}_1\ le\ Français\ ;\ I\ \mathrm{X}_1\ french$ extract from sentence pair in Figure 1, the phrase *le Français* connects to the phrase *french* because the French word *Français* aligns with the English word *french* even though *le* is unaligned.

We then group the source phrases and target phrases into phrase pairs such that only phrases that are connected to each other are in the same phrase pair. So phrase pairs still preserve the lexical dependency of the rule. Phrase pairs and nonterminals are then ordered according to the target side of the rule. Table 1 shows an example of rules, alignments and their sequences of phrase pairs and nonterminals on the last column.

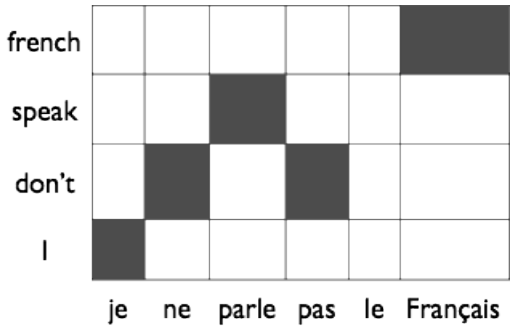

Figure 2: Alignment of a sentence pair.

There are Hiero rules in which one of its source phrases or target phrases is not aligned. For example in the rule $r_4 = \mathrm{X} \rightarrow je\ \mathrm{X}_1\ le\ \mathrm{X}_2\ ;\ i\ \mathrm{X}_1\ \mathrm{X}_2$ extracted from the sentence pair in Figure 2, the phrase *le* is not aligned. In our Arabic-English experiment, rules without nonaligned phrases account for only 48.54% of the total rules. We compared the baseline Hiero translation from the full set of rules and the translation from only rules without nonaligned phrases. The later translation is faster and Table 2 [1] shows that it outperforms the translation with the whole set of rules. We therefore decided to not use rules with nonaligned phrases in Phrasal-Hiero.

It is important to note that there are different ways to use all the rules and map rules with unaligned phrases into a sequence of phrase pairs.

[1]The dataset and experiment setting description are in section 4.

Test set	MT04	MT05	MT09
All rules	48.17	47.85	42.37
Phrasal Hiero	48.52	47.78	42.8

Table 2: Arabic-English pilot experiment. Compare BLEU scores of translation using all extracted rules (the first row) and translation using only rules without nonaligned subphrases (the second row).

For example, adding these unaligned phrases to the previous phrase pair i.e. the rule r_4 has one discontinuous phrase pair $(je \ldots le, i)$ or treat these unaligned phrases as deletion/insertion phrases. We started the work with Arabic-English translation and decided not to use rules with nonaligned phrases in Phrasal-Hiero. In the experiment section, we will discuss the impact of removing rules with nonaligned sub-phrases in our German-English and Chinese-English experiments.

2.2 Training: Lexicalized Reordering Table

Phrasal-Hiero needs a phrase-based lexicalized reordering table to calculate the features. The lexicalized reordering table could be from a discontinuous phrase-based system. To guarantee the lexicalized reordering table to cover all phrase pairs of the rule table, we extract phrase-pairs and their reordering directions during rule extraction.

Let $(\mathbf{s}, \mathbf{t})$ be a sentence pair in the training data and $r = \mathrm{X} \rightarrow s_0 X_1 s_1 \ldots X_k s_k\ ;\ t_0 X_1 t_1 \ldots X_k t_k$ be a rule extracted from the sentence. The lexical phrase pair corresponding to the rule r is $ph = (s_0 \ldots s_1 \ldots s_k, t_0 \ldots t_1 \ldots t_k)$, with nonterminals are replaced by the gaps. Because the nonterminal could be at the border of the rule, the lexical phrase pair might have smaller coverage than the rule. For example, the training sentence pair in Figure 2 generates the rule $r2 = \mathrm{X} \rightarrow ne\ \mathrm{X}_1\ pas\ ;\ don't\ \mathrm{X}_1$ spanning $(1 \ldots 3, 1 \ldots 2)$ but its lexical phrase pair $(ne \ldots pas, not)$ only spans $(1 \ldots 3, 1 \ldots 1)$.

Also, two different rules can have the same lexical phrase pairs. In Phrasal-Hiero, each lexical phrase pair is only generated once for a sentence. Look at the example of the training sentence pair in Figure 2, the rule $\mathrm{X} \rightarrow je\ ;\ I$ spanning $(0 \ldots 1, 0 \ldots 1)$ and the rule $\mathrm{X} \rightarrow je\ \mathrm{X}_1\ ;\ I\ \mathrm{X}_1$ spanning $(0 \ldots 3, 0 \ldots 2)$ are both sharing the same lexical phrase pair (je, i) spanning $(0 \ldots 1, 0 \ldots 1)$. But Phrasal-Hiero only gen-

erates (je, i) once for the sentence. Phrase pairs are generated together with phrase-based reordering orientations to build lexicalized reordering table.

3 Decoding

Chiang (2007) applied bottom up chart parsing to parse the source sentence and project on the target side for the best translation. Each chart cell $[\mathrm{X}, i, j, r]$ indicates a subtree with rule r at the root covers the translation of the i-th word upto the j-th word of the source sentence. We extend the chart parsing, mapping the subtree to the equivalent discontinuous phrase-based path and includes phrase-based features to the log-linear model.

In Phrasal-Hiero, each chart cell $[\mathrm{X}, i, j, r]$ also stores the first phrase pair and the last phrase pair of the phrase-based translation path covered the i-th to the j-th word of the source sentence. These two phrase pairs are the back pointers to calculate reordering features of later larger spans. Because the distance cost feature and phrase-based discriminative reordering feature calculation are both only required the source coverage of two adjacent phrase pairs, we explain here the distance cost calculation.

We will again use three rules r_1, r_2, r_3 in Table 1 and the translation *je ne parle pas le français* into *I don't speak French* to present the technique. Table 3 shows the distance cost calculation.

First, when the rule r has only terminals, the rule's sequence of phrase pairs and nonterminals consists of only a phrase pair. No calculation is needed, the first phrase pair and the last phrase pair are the same. The chart cell $\mathrm{X}_1 : 2 \ldots 2$ in Table 3 shows the translation with the rule $r_1 = \mathrm{X} \rightarrow parle\ ;\ speak$. The first phrase pair and the last phrase pair point to the phrase $(parle, speak)$ spanning $2 \ldots 2$ of the source sentence.

When the translation rule's right hand side has nonterminals, the nonterminals in the sequence belong to smaller chart cells that we already found phrase-based paths and calculated their features before. The decoder then substitute these paths into the rule's sequence of phrase pairs and nonterminals to form the complete path for the current span.

We now demonstrate finding the phrase based path and calculate distance cost of the chart cell X_2 spanning $1 \ldots 3$. The next phrase pair of $(ne \ldots pas, don't)$ is the first phrase pair of the chart cell X_1 which is $(parle, speak)$. The distance cost of these two phrase pairs according to discontinuous phrase-based model is $|2 - 3 - 1| = 2$. The distance cost of the whole chart cell X_2 also includes the cost of the translation path covered by chart cell X_1 which is 0, therefore the distance cost for X_2 is $2 + \mathrm{dist}(X_1) = 2$. We then update the first phrase pair and the last phrase pair of cell X_2. The first phrase pair of X_2 is $(ne \ldots pas, don't)$, the last phrase pair is also the last phrase pair of cell X_1 which is $(parle, speak)$.

Similarly, finding the phrase-based path and calculate its distortion features in the chart cell X_3 include calculate the feature values for moving from the phrase pair (je, I) to the first phrase pair of chart cell X_2 and also from last phrase pair of chart cell X_2 to the phrase pair $(le\ français e, french)$.

4 Experiment Results

In all experiments we use phrase-orientation lexicalized reordering (Galley and Manning, 2008)[2] which models monotone, swap, discontinuous orientations from both reordering with previous phrase pair and with the next phrase pair. There are total six features in lexicalized reordering model.

We will report the impact of integrating phrase-based features into Hiero systems for three language pairs: Arabic-English, Chinese-English and German-English.

4.1 System Setup

We are using the following three baselines:

- Phrase-based without lexicalized reodering features. (PB+nolex)
- Phrase-based with lexicalized reordering features.(PB+lex)
- Hiero system with all rules extracted from training data. (Hiero)

We use Moses phrase-based and chart decoder (Koehn et al., 2007) for the baselines. The score difference between PB+nolex and PB+lex results indicates the impact of lexicalized reordering features on phrase-based system. In Phrasal-Hiero we

[2]Galley and Manning (2008) introduce three orientation models for lexicalized reordering: word-based, phrase-based and hierarchical orientation model. We apply phrase-based orientation in all experiment using lexicalized reordering.

Chart Cell	Rule's phrase pairs & NTs	Distance	First Phrase Pair	Last Phrase Pair
$X_1 : 2 \dots 2$	$(parle, speak)$	$\emptyset$	$2 \dots 2\,(parle, speak)$	
$X_2 : 1 \dots 3$	$(ne \dots pas, don't)$; X_1	$2 + \text{dist}\,(X_1)$ $= 2$	$1 \dots 3$ $(ne \dots pas, don't)$	$2 \dots 2\,(parle, speak)$
$X_3 : 0 \dots 5$	$(Je\,;I)$; X_2 ; $(le\ Français; french)$	$0 + \text{dist}\,(X_2)$ $+1 = 3$	$0 \dots 0\ \ (je, I)$	$4 \dots 5$ $(le\ Français; french)$

Table 3: Phrasal-Hiero Decoding Example: Calculate distance cost feature for the translation in Figure 1.

will compare if these improvements still carry on into Hiero systems.

The original Hiero system with all rules extracted from training data (Hiero) is the most relevant baseline. We will evaluate the difference between this Hiero baseline and our Phrasal-Hiero.

To implement Phrasal-Hiero, we extented Moses chart decoder (Koehn et al., 2007) to include distance-based reordering as well as the lexicalized phrase orientation reordering model. We will report the following results for Phrasal-Hiero:

- Hiero translation results on the subset of rules without unaligned phrases. (we denote this in the table scores as P.H.)
- Phrasal-Hiero with phrase-based distance cost feature (P.H.+dist).
- Phrasal-Hiero with phrase-based lexicalized reordering features(P.H.+lex).
- Phrasal-Hiero with distance cost and lexicalized reordering features(P.H.+dist+lex).

4.2 Arabic-English Results

The Arabic-English system was trained from 264K sentence pairs with true case English. The Arabic is in ATB morphology format. The language model is the interpolation of 5-gram language models built from news corpora of the NIST 2012 evaluation. We tuned the parameters on the MT06 NIST test set (1664 sentences) and report the BLEU scores on three unseen test sets: MT04 (1353 sentences), MT05 (1056 sentences) and MT09 (1313 sentences). All test sets have four references per each sentence.

The results are in Table 4. The three rows in the first block are the baseline scores. Phrase-based with lexicalized reordering features(PB+lex) shows significant improvement on all test sets over the simple phrase-based system without lexicalized reordering (PB+nolex). On average the improvement is 1.07 BLEU score (45.66 versus 46.73). We make the same observation as Zollmann et al. (2008), i.e, that the Hiero baseline system underperforms compared to the phrase-based system with lexicalized phrase-based reordering for Arabic-English in all test sets, on average by about 0.60 BLEU points (46.13 versus 46.73). This is because Arabic language has relative free reordering, but mostly short distance, which is better captured by discriminative reordering features.

	MT04	MT05	MT09	Avg.
PB+nolex	47.40	46.83	42.75	45.66
PB+lex	48.62	48.07	43.51	46.73
Hiero	48.17	47.85	42.37	46.13
P.H. (48.54% rules)	48.52	47.78	42.80	46.37
P.H.+dist	48.46	47.92	42.62	46.33
P.H. +lex	48.70	48.59	**43.84**	47.04
P.H +lex+dist	**49.35**	**49.07**	43.40	**47.27**
Improv. over PB+lex	0.73	1.00	0.34	0.54
Improv. over P.H.	0.83	1.29	1.04	0.90
Improv. over Hiero	1.18	1.22	1.47	1.14

Table 4: Arabic-English true case translation scores in BLEU metric. The three rows in the first block are the baseline scores. The next four rows in the second block are Phrasal-Hiero scores, the best scores are in boldface. The three rows in the last block are the Phrasal-Hiero improvements.

The next four rows in the second block of Table 4 show Phrasal-Hiero results. The P.H. line is the result of Hiero experiment on only a subset of rules without nonaligned phrases. As mentioned in section 2.1, Phrasal-Hiero only uses 48.54% of the rules but achieves as good or even better performance (on average 0.24 BLEU points better) compared to the original Hiero system using the full set of rules.

We do not benefit from adding only the

distance-based reordering feature (P.H+dist) to the Arabic-English experiment but get significant improvements when adding the six features of the lexicalized reordering (P.H+lex). Table 4 shows that the P.H.+lex system gains on average 0.67 BLEU points (47.04 versus 46.37). Even though the baseline Hiero underperforms phrase-based system with lexicalized reordering(P.B+lex), the P.H.+lex system already outperforms P.B+lex in all test sets (on average 47.04 versus 46.73).

Adding both distance cost and lexicalized reordering features (P.H.+dist+lex) performs the best. On average P.H.+dist+lex improves 0.90 BLEU points over P.H. without new phrase-based features and 1.14 BLEU score over the baseline Hiero system. Note that Hiero rules already have lexical context in the reordering, but adding phrase-based lexicalized reordering features to the system still gives us about as much improvement as the phrase-based system gets from lexicalized reordering features, here 1.07 BLEU points. And our best Phrasal-Hiero significantly improves over the best phrase-based baseline by 0.54 BLEU points. This shows that the underperformance of the Hiero system is due to its lack of lexicalized reordering features rather than a limited hypothesis space.

4.3 Chinese-English Results

The Chinese-English system was trained on FBIS corpora of 384K sentence pairs, the English corpus is lower case. The language model is the trigram SRI language model built from Xinhua corpus of 180 millions words. We tuned the parameters on MT06 NIST test set of 1664 sentences and report the results of MT04, MT05 and MT08 unseen test sets. The results are in Table 5.

We also make the same observation as Zollmann et al. (2008) on the baselines for Chinese-English translation. Even though the phrase-based system benefits from lexicalized reordering, PB+lex on average outperforms PB+nolex by 1.16 BLEU points (25.87 versus 27.03), it is the Hiero system that has the best baseline scores across all test sets, with and average of 27.70 BLEU points.

Phrasal Hiero scores are given in the second block of Table 5. It uses 84.19% of the total training rules, but unlike the Arabic-English system, using a subset of the rules costs Phrasal-Hiero on all test sets and on average it loses 0.49 BLEU points (27.21 versus 27.70). Similar to Chiang et al. (2008) in their Chinese-English experiment, we benefit by adding the distance cost feature. PH.+dist outperforms P.H. on all test sets. We have better improvements when adding the six features of the lexicalized reordering model: P.H.+lex on average has 28.05 BLEU points, i.e. gains 0.84 over P.H.. The P.H.+lex system is even better than the best Hiero baseline using the whole set of rules.

	MT04	MT05	MT08	Avg.
PB+nolex	29.99	26.4	21.23	25.87
PB+lex	31.03	27.57	22.41	27.03
Hiero	32.49	28.06	22.57	27.70
P.H. (84.19% rules)	31.83	27.66	22.16	27.21
P.H.+dist	32.18	28.25	22.46	27.63
P.H.+lex	32.55	28.51	23.08	28.05
P.H+lex+dist	**33.06**	**28.78**	**23.23**	**28.35**
Improv. over PB+lex	2.03	1.21	0.82	1.32
Improv. over P.H.	1.23	1.12	1.07	1.14
Improv. over Hiero	0.57	0.72	0.66	0.65

Table 5: Chinese-English lower case translation scores in BLEU metric.

We again get the best translation when adding both the distance cost feature and the lexicalized reordering features. The P.H+dist+lex has the best score across all the test sets and on average gains 1.14 BLEU points over P.H. So adding phrase-based features to the Hiero system yields nearly the same improvement as adding lexicalized reordering features to the phrase-based system. This shows that a strong Chinese-English Hiero system still benefits from phrase-based features. Further more, the P.H+dist+lex also outperforms the Hiero baseline using all rules from training data.

4.4 German-English Results

We next consider German-English translation. The systems were trained on 1.8 million sentence pairs using the Europarl corpora. The language model is three-gram SRILM trained from the target side of the training corpora. We use WMT 2010 (2489 sentences) as development set and report scores on WMT 2008 (2051 sentences), WMT 2009 (2525 sentences), WMT 2011 (3003 sentences). All test sets have one reference per test sentence. The results are in Table 6.

WMT test	08	09	11	Avg.
PB+nolex	17.46	17.38	16.76	17.20
PB+lex	18.16	17.85	17.18	17.73
Hiero	18.20	18.23	17.46	17.96
P.H. (80.54% rules)	18.24	18.15	17.39	17.92
P.H. +dist	18.19	17.97	17.41	17.85
P.H. +lex	18.59	18.46	17.69	18.24
P.H.+lex+dist	**18.70**	**18.53**	**17.81**	**18.34**
Improv. over PB+lex	0.54	0.68	0.63	0.61
Improv. over P.H.	0.46	0.38	0.42	0.42
Improv. over Hiero	0.50	0.30	0.35	0.38

Table 6: German-English lower case translation scores in BLEU metric.

The Hiero baseline performs on average 0.26 BLEU points better than the phrase-based system with lexicalized reordering features (PB+lex). The hrasal-Hiero system used 80.54% of the total training rules, but on average the P.H. system has the same performance as the Hiero system using all the rules extracted from training data. Similar to the Arabic-English experiment, Phrasal-Hiero does not benefit from adding the distance cost feature. We do, however, see improvements on all test sets when adding lexicalized reordering features. On average the P.H.+lex results are 0.32 BLEU points higher than the P.H. results. The best scores are achieved with P.H+lex+dist. The German-English translations on average gain 0.38 BLEU score by adding both distance cost and discriminative reordering features.

4.5 Impact of segment rules into phrase pairs

Phrasal Hiero is the first system using rules' lexical alignments. If lexical alignments are not available, we can not divide the rules' lexicals into phrase pairs without losing their dependancies. An alternative approach would be combining all lexicals of a rule into one phrase pair. We run an addition experiment for this approach on Arabic-English dataset. Table 7 shows the examples rules and its new sequence of nonterminals and phrase pairs. The rules r_1 and r_2 have the same sequences as in Table 1. Without segment rules into phrase pairs, the rule r_3 has only one phrase pair: $ph = (Je \ldots le\ Francaise\ ;\ I \ldots french)$ and ph is repeated twice in r_3's sequence of phrase pairs and nonterminals. The new experiment uses the complete set of rules so the rule r_4 is included.

According to the new sequence of phrase pairs and nonterminals, during decoding the rule r_3 has *discontinous* translation directions on both from phrase pair ph to the nonterminal X_1 and from X_1 to ph. But using lexical alignment and divide the rule into phrase pairs as in section 2.1 , the sequence preserves the translation order of r_3 as two *monotone* translations from $(je;\ I)$ to X_1 and from X_1 to $(le\ Francaise\ ;\ french)$.

	Avg
Hiero	46.13
Hiero+lex (no lex. alignments)	46.45 (+0.32)
P.H	46.37
P.H.+lex (with lex. alignments)	47.04 (+0.67)

Table 8: Average of Arabic-English translation scores in BLEU metric. Compare the improvement of using rules' lexical alignments (2nd block) and not using rules' lexical alignments (1st block).

Table 8 compares the two experiments results. The additional experiment is denoted as Hiero+lex in the table. The first block shows an improvement of 0.32 BLEU score when adding discriminated reordering features on Hiero (using the whole set of rules and no rule segmentation). The second block is the impact of adding discriminated reordering features on Phrasal Hiero (using a subset of rules and segment rules into phrase pairs). Here the improvement of P.H+lex over P.H is 0.67 BLEU score. It shows the benefit of segment rules into phrase pairs.

4.6 Rules without unaligned phrases

	A-E	C-E	G-E
Hiero	46.13	27.70	17.96
P.H.	46.36	27.21	17.92
%Rules used	48.54%	84.19%	80.54%
P.H.+lex+dist	47.27	28.35	18.34

Table 9: The impact of using only rules without nonaligned phrases on Phrasal-Hiero results.

Table 9 summarizes the impact of using only rules without nonaligned phrases on Phrasal-

Rules	Phrase pairs & nonterminals
$r_1 = \mathrm{X} \rightarrow parle\ ;\ speak.$	$(parle\ ;\ speak)$
$r_2 = \mathrm{X} \rightarrow ne\ \mathrm{X}_1\ pas\ ;\ don't\ \mathrm{X}_1.$	$(ne \ldots pas\ ;\ don't)\ ;\ \mathrm{X}_1$
$r_3 = \mathrm{X} \rightarrow Je\ \mathrm{X}_1\ le\ Francais\ ;\ I\ \mathrm{X}_1\ French$	$(Je \ldots le\ Francais\ ; I \ldots french)\ ;\ \mathrm{X}_1\ ;$ $(Je \ldots le\ Francais\ ; I \ldots french)$
$r_4 = \mathrm{X} \rightarrow je\ \mathrm{X}_1\ le\ \mathrm{X}_2\ ;\ i\ \mathrm{X}_1\ \mathrm{X}_2$	$(je \ldots le\ ;\ i)\ ;\ \mathrm{X}_1\ ;\ \mathrm{X}_2$

Table 7: Example of translation rules and their sequences of phrase pairs and nonterminals when lexical alignments are not available.

Hiero. Using only rules without nonaligned phrases can get the same performance with translation with full set of rules for Arabic-English and German-English experiments but underperforms for the Chinese-English system. We suggest the difference might come from the linguistic divergences of source and target languages.

Phrasal Hiero includes all lexical rules (rules without nonterminal) therefore it still has the same lexical coverage as the original Hiero system. In the Arabic-English system, the Arabic is in ATB format, therefore most English words should have alignments in the ATB source, rules with nonaligned phrases could be the results of bad alignments or non-informative rules, therefore we could have better performance by using a subset of rules in Phrasal-Hiero.

As Chinese and English are highly divergent, we expect many phrases in one language correctly unaligned in the other language. So leaving out the rules with nonaligned phrases could degrade the system. Even though the current Phrasal-Hiero with extra phrase-based features outperforms the Hiero baseline, future work for Phrasal-Hiero will focus on including all rules extracted from training corpora.

4.7 Discontinuous Phrase-Based

	C-E	G-E
PB+lex	27.03	17.73
PB+lex+gap	27.11	17.55
Hiero	27.70	17.96
P.H.+lex+dist	28.35	18.34

Table 10: Comparing Phrasal-Hiero with translation with gap for Chinese-English and German-English. The numbers are average BLEU scores of all test sets.

We compare Phrasal-Hiero with a discontinuous phrase-based system introduced by Galley and Manning (2010) for Chinese-English and German-English system. Table 10 shows the average results. We used Phrasal decoder (Cer et al., 2010) for phrase-based with gaps (PB+lex+gap) results. While we do not focus on the differences in the toolkits, our Phrasal-Hiero still outperforms the phrase-based with gaps experiments.

Conclusion

We have presented a technique to combine phrase-based features and tree-based features into one model. Adding a distance cost feature, we only get better translation for Chinese-English translation. Phrasal-Hiero benefits from adding discriminative reodering features in all experiment. We achieved the best result when adding both distance cost and lexicalized reordering features. Phrasal-Hiero currently uses only a subset of rules from training data. A future work on the model can include complete rule sets together with word insertion/deletion features for nonaligned phrases.

References

A. Birch, P. Blunsom, and M. Osborne. 2009. A Quantitative Analysis of Reordering Phenomena. In *StatMT '09: Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 197–205.

X. Carreras and M. Collins. 2009. Non-Projective Parsing for Statistical Machine Translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 200–209.

D. Cer, M. Galley, D. Jurafsky, and C. Manning. 2010. Phrasal: A Statistical Machine Translation Toolkit for Exploring New Model Features. In *Proceedings of the NAACL HLT 2010 Demonstration Session*, pages 9–12. Association for Computational Linguistics, June.

D. Chiang, Y. Marton, and P. Resnik. 2008. Online Large-Margin Training of Syntactic and Structural Translation Features. In *Proceedings of the Conference on Empirical Methods in Natural Language*

Processing, pages 224–233. Association for Computational Linguistics.

D. Chiang. 2005. A Hierarchical Phrase-Based Model for Statistical Machine Translation. In *Proc. of ACL*.

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

M. Galley and C. Manning. 2008. A Simple and Effective Hierarchical Phrase Reordering Model. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 847–855, Honolulu, Hawaii, October.

M. Galley and C. D. Manning. 2010. Accurate Non-Hierarchical Phrase-Based Translation. In *Proceedings of NAACL-HLT*, pages 966–974.

M. Huck, S. Peitz, M. Freitag, and H. Ney. 2012. Discriminative Reordering Extensions for Hierarchical Phrase-Based Machine Translation. In *EAMT*, pages 313–320.

P. Koehn, F. J. Och, and D. Marcu. 2003. Statistical Phrase-Based Translation. In *Proc. of HLT-NAACL*, pages 127–133.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL demonstration session*.

C. Tillmann. 2004. A Unigram Orientation Model for Statistical Machine Translation. In *Proceedings of HLT-NAACL: Short Papers*, pages 101–104.

A. Zollmann and A. Venugopal. 2006. Syntax Augmented Machine Translation via Chart Parsing. In *Proc. of NAACL 2006 - Workshop on Statistical Machine Translation*.

A. Zollmann, A. Venugopal, F. Och, and J. Ponte. 2008. A Systematic Comparison of Phrase-Based, Hierarchical and Syntax-Augmented Statistical MT. In *Proceedings of the Conference on Computational Linguistics (COLING)*.