An Orientation Model for Hierarchical Phrase-based Translation

Xinyan Xiao, Jinsong Su, Yang Liu, Qun Liu, and Shouxun Lin
Key Laboratory of Intelligent Information Processing, Institute of Computing Technology
Chinese Academy of Sciences, Beijing, China
{*xiaoxinyan,sujinsong,yliu,liuqun,sxlin*}*@ict.ac.cn*

***Abstract*—The hierarchical phrase-based (HPB) translation exploits the power of grammar to perform long distance reorderings, without specifying nonterminal orientations against adjacent blocks or considering the lexical information covered by nonterminals. In this paper, we borrow from phrase-based system the idea of orientation model to enhance the reordering ability of HPB translation. We distinguish three orientations (monotone, swap, discontinuous) of a nonterminal based on the alignment of grammar, and select the appropriate orientation of nonterminal using lexical information covered by it. By incorporating the orientation model, our approach significantly outperforms a standard HPB system up to 1.02 BLEU on large scale NIST Chinese-English translation task, and 0.51 BLEU on WMT German-English translation task.**

I. INTRODUCTION

The orientation model [1], [2], [3] has greatly improved the phrase reordering performance, and becomes a necessary component of phrase-based systems.[1] The orientation model specifies the phrase orientations and estimates the orientation probabilities conditioned on the phrases. In Figure 1, the phrase in bold rectangle ("with Sharon") swaps its position with previous phrase. Besides, the prepositions "*yu*" in Chinese side and "with" in English side strongly imply the swapping orientation. Therefore, it is accurate to condition an orientation probability on the phrase itself.

HPB translation [4] and so as other syntax-based models, on the other hand, exploit the power of grammar to capture both short and long distance reordering. The order of a phrase changes as the order of a nonterminal which covers it changes. In HPB, the swapping orientation of the phrase in bold rectangle in Figure 1 can be accomplished by a rewrite rule:

$$X \Rightarrow \langle X \textit{ juxing le huitan},\ \text{held talks } X \rangle \qquad (1)$$

Unlike the orientation model in phrase-based system which predicts a phrase orientation by the lexical information of the phrase itself, HPB translation reorders a phrase without considering the information of the phrase, which offers accurate cues for the reordering orientation [5], [6]. In HPB system, since a phrase is generalized by a nonterminal, the system fails to consider the phrase information covered by a nonterminal when reordering.

[1]Such model is often called lexicalized reordering in phrase-based system. However, we think orientation is a more accurate representation, since it predicts the phrase orientation in practice. Therefore, to distinct from lexicalized reordering of HPB system, we call it orientation model in our paper.

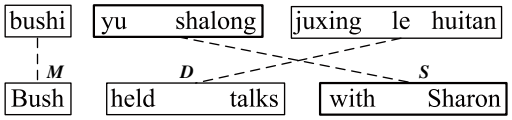


Figure 1. Phase orientations for a Chinese-to-English sentence pair. The three orientations are monotone(*M*), swap(*S*), and discontinuous (*D*).

Since the orientation model and HPB translation model phrase reordering in different ways, we believe that HPB can also benefit from specifying orientation. Consequently, we propose an orientation model for HPB translation, which distinguishes three types of nonterminal orientations including monotone(*M*), swap(*S*), and discontinuous (*D*) based on the alignment of grammar, and selects an appropriate orientation using the lexical information covered by nonterminals. In the experiments, our approach outperforms a standard HPB system up to 1.02 BLEU on large scale NIST Chinese-English translation task, and 0.51 BLEU on WMT German-English translation task. The results confirms that the HPB translation is really enhanced through incorporating the orientation model. Note that, the orientation model is defined on grammar, thus is straight to be extended to other linguistically syntax-based systems.

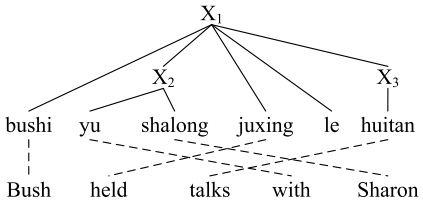
II. AN ORIENTATION MODEL FOR HPB TRANSLATION

HPB translates a source language sentence into a target language sentence by a sequence of synchronous context free grammar (SCFG) rules. The sequence of rules $\{r_i\}$ is called a derivation $\mathbf{d}$. The derivation can also be represented as a synchronous tree. Figure 2 shows the tree representation, where the target side tree is omitted since the grammar is synchronized.

In HPB translation, the orientation of a phrase is identical to the orientation of the nonterminal that covers it. For the derivation tree in Figure 2, the translation of phrase "*yu shalong*" is moved to the end of the sentence. Such movement is exactly the same as the movement of the nonterminal X_2 that covers it. Consequently, we can represent the phrase orientation probability as the nonterminal orientation probability. For the derivation, we set the orientation of X_2 as swap and X_3 as monotone. Note that it is unnecessary to calculate the orientation of root node X_1, since no other phrase outside the root node exists. Therefore, the orientation probability of the derivation is represented as:

$$P(S|f(X_2)) \times P(M|f(X_3)) \qquad (2)$$

978-0-7695-4554-7/11 $26.00 © 2011 IEEE
DOI 10.1109/IALP.2011.43



r_1	$X_1 \Rightarrow \langle$*bushi* X_2 *juxing le* X_3, Bush held X_3 $X_2\rangle$
r_2	$X_2 \Rightarrow \langle$*yu shalong*, with Sharon$\rangle$
r_3	$X_3 \Rightarrow \langle$*huitan*, talks$\rangle$

Figure 2. Derivation tree and correspondent SCFG rules. We also show the word alignment here.

where $f(X)$ is the information about X such as the words covered by X. Notably, such information is contextual for HPB translation.

Formally, we calculate the orientation probability $P_o(\mathbf{d})$ of a derivation $\mathbf{d}$ as the product of orientation probabilities of all nonterminals except the root:

$$P_o(\mathbf{d}) = \prod_{X \in \mathbf{d} \wedge X \neq root} P(o_X | f(X)) \tag{3}$$

where o_X is the orientation of nonterminal X, and $X \in \mathbf{d}$ means the nonterminals in the tree of derivation $\mathbf{d}$.

Furthermore, HPB model only changes the order of a nonterminal with its siblings. Thus, it just needs to compare the relative position of a nonterminal with its siblings in order to decide the orientation. This inspires us to define the orientation based on the alignment of rule, since a rule includes all the siblings of a nonterminal in the right hand side. Such definition implies that the orientation probability can be calculated along with applying rules, thus provides a straight way to incorporate the orientation model into traditional systems. Therefore, we transform equation (3) into equation (5):

$$P_o(\mathbf{d}) = \prod_{X \in \mathbf{d}} \prod_{X' \in child(X)} P(o_{X'} | f(X')) \tag{4}$$

$$= \prod_{r \in \mathbf{d}} \prod_{X \in rhs(r)} P(o_X | f(X)) \tag{5}$$

Equation (4) means that the probability is also equal to the product of orientation probabilities for every child nonterminal X' of every nonterminal X. We can transfer equation (4) into equation (5), because the child nonterminals of a nonterminal X are the same with the nonterminals in the right hand side $rhs(r)$ of the rule r for nonterminal X. For example, X_1 has two children X_2 and X_3, while the right hand side of r_1 also contains two nonterminals X_2 and X_3.

Following Moses, we distinguish three nonterminal orientations: monotone(*M*), swap(*S*), and discontinuous (*D*), and use a bidirectional orientation model. In practice, the bidirectional setting results two similar orientation models: *left* model which refers the orientation to left adjacent blocks and *right* model which refers to right adjacent blocks. In the following sections, we first describe the definition of nonterminal orientation based on the alignment of rule, then introduce the estimation of orientation probabilities based on relative frequency and discriminative approach. Finally, we give some details in decoding.

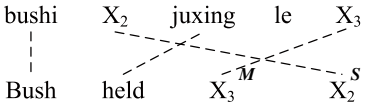


Figure 3. Nonterminal orientations in rule r_1. We set the orientation of X_2 as swap, since it swaps its positions with the maximum left adjacent block $\langle$held X_3, *juxing le* $X_3\rangle$. The orientation of X_3 is monotone comparing with block $\langle$held, *juxing le*$\rangle$.

A. Nonterminal Orientation

As shown before, the nonterminal orientation can be determined by the alignment of rule. Figure 3 shows the rule r_1 with the alignment of both terminals and nonterminals. It is obvious that r_1 is the list of children of node X_1 by contrasting Figure 2 and Figure 3. We set the orientation of X_2 as swap, since it swaps its position with the left adjacent block $\langle$held X_3, *juxing le* $X_3\rangle$. The orientation of X_3 is monotone comparing with block $\langle$held, *juxing le*$\rangle$. Note that, there are multiple choices of adjacent blocks, and we use the one with maximum size.

In order to define the relative orders of nonterminals and their adjacent blocks, we expand the alignment $\mathbf{a}$ in a rule to include both terminal and nonterminal alignments.[2] The **alignment** of rule $\mathbf{a} = \{(i, j)\}$ is a set of links between terminals or between non-terminals. A point (i, j) is a link from rule target side α_i to rule target side γ_j.

Formally, a **block** is a bispan $\langle [s, t], [u, v] \rangle$ of a rule, which can contain both terminals and nonterminals. $[s, t]$ denotes the strings that spans from s to t in the rule target side, and $[u, v]$ denotes the strings that spans from u to v in the rule source side. Furthermore, the bispan must be consistent with the alignment $\mathbf{a}$. The bispan is consistent with the alignment if and only if:

$$\forall (i, j) \in \mathbf{a} : s \leq i \leq t \Leftrightarrow u \leq j \leq v \tag{6}$$

There maybe multiple choices of blocks adjacent to a nonterminal. In such case, we use the one with maximum size, since using larger granularity leads to less discontinuous orientation instances [3]. To decide which block is larger, we first compare their target spans, and if their target sizes are equal we then compare source spans.

We use bidirectional setting as orientation model in Moses. In practice, the bidirectional setting results two similar orientation models: **left** model which refers the

[2]Such alignment of a rule comes from extraction. During rule extraction, we retain the alignment information in the extracted rules. If a rule is observed with more than one set of alignment, we only keep the most frequent one.

orientation to left adjacent blocks and **right** model which refers the orientation to right adjacent blocks. Given the above definitions, we can identify the orientation now. An alignment of a nonterminal X must be one-to-one alignment, which represents as (i_X,j_X). Thus, the left orientation model is defined as following:

- $o = M$, if a block lies at $(i_X - 1, j_X - 1)$.
- $o = S$, if a block lies at $(i_X - 1, j_X + 1)$.
- $o = D$, otherwise.

Similarly, we define the orientation of right direction as:

- $o = M$, if a block lies at $(i_X + 1, j_X + 1)$.
- $o = S$, if a block lies at $(i_X + 1, j_X - 1)$.
- $o = D$, otherwise.

B. Training

The basic element of the orientation model is nonterminal orientation probability $P(o_X|f(X))$. The probability can be calculated by relative frequency estimation or by discriminative training. Both training methods use the same instances extracted from word alignment corpus.

Relative Frequency Estimation: This method estimates $P(o_X|f(X))$ by relative frequency.[3] If $f(X)$ means the words covered by the nonterminal, then it the same with Moses. However, the maximum size of phrases covered by X in HPB system is larger than the maximum size of phrase in phrase-based system. We will face sparseness problem if we directly use the entire phrase. Alternatively, we use *boundary words* covered by X, which are informative for the reordering of phrases [7]. For a nonterminal, there are four boundary words: left boundary word in target side, right boundary word in target side, left boundary word in source side, and right boundary word in source side. Suppose the nonterminal X_1 is not a root node and the orientation is S, then the probability is

$$P(S|f(X_1)) \approx P(S|\text{bush},\ \text{Sharon},\ \textit{bushi},\ \textit{huitan}) \quad (7)$$

Discriminative Training: We can also calculate the orientation probabilities by discriminative training [7], [8]. In this way, we can utilize arbitrary features to improve the performance of model. We use maximum entropy model to estimate the probability $P(o_X|f(X))$

$$P(o_X|f(X)) = \frac{\exp[\sum_i \lambda_i h_i(o_X, f(X))]}{\sum_{o'} \exp[\sum_i \lambda_i h_i(o', f(X))]} \quad (8)$$

where h_i is feature function, λ_i is the feature weight of h_i. In addition to the boundary word features, we also use the *linear context* features, which is the surrounding words of nonterminals. Similar to boundary word features, there are four types of linear context features, including left adjacent words and right adjacent words in both source and target side. Such words provide good evidences for the syntax type of a string, and have been widely used in unsupervised parsing.

[3]we use add 0.1 for smoothing the probability.

Extraction: We extract orientation instances from word alignment parallel corpus. The extraction process is similar with phrase extraction. For each bispan, we calculate its maximum adjacent block. Given the adjacent block, we decide the orientations of the bispan. If there is not any adjacent block, then it is a discontinuous orientation. After identifying the orientation, we can extract an instance with the orientation label of the bispan as well as the related words. We limit the source size of a bispan to 10 (same as initial phrase size in HPB system), but do not limit the size of adjacent blocks. Given such instances, we can estimate the probability by relative frequency or by maximum entropy training.

C. Decoding

We incorporate the orientation model into traditional HPB system under the log-linear framework [9]. We assign three distinct features for each orientation category like Moses, rather than using the log probability of the orientation model. This means that the probability of equation (5) is divided into three feature scores for each orientation. For example the probability of swapping orientation is calculated by

$$P_{o=S}(\mathbf{d}) = \prod_{r \in \mathbf{d}} \prod_{X \in rhs(r)} P(o_X = S|f(X)) \quad (9)$$

Considering there are two types of orientation models indeed, this results 6 new features. The feature weights are rescaled by minimum error rate training [10]. In this way, we can optimize the weights of each orientation according to its effect on translation quality in terms of BLEU.

During decoding, we search the best translation using CKY algorithm with cube pruning [4]. Since the nonterminal orientation is decided by the rule, we can calculate the nonterminal orientation before decoding. Note that it's possible that the linear context feature of a nonterminal may be non-local. For example, the right adjacent word of X_2 in target side is unknown when applying the rule r_1. We simplify this problem to only consider those words cover by current rule. In this way, we can use as many lexicalized features as possible while maintain simplicity. OOV may occur during decoding. In such case, we just set the three orientations with equal probability.

III. EXPERIMENTS

Our main experiments work on the Chinese-English translation task. The bilingual training data contains 1.5M sentence pairs with 42.3M Chinese words and 48.2M English words, which come from subsets of LDC data.[4] The monolingual data for training English language model includes the Xinhua portion of the GIGAWORD corpus, which contains 238M English words. We used the NIST evaluation sets of 2002 (MT02) as our development data set, and sets of MT03/MT04/MT05 as test sets.

SCFG rules were extracted as described in Chiang [4] and a 4-gram language model was trained on the monolingual data. We extracted the orientation instances from

[4]including LDC2002E18, LDC2003E07, LDC2003E14, Hansards portion of LDC2004T07, LDC2004T08 and LDC2005T06.

Table I
PERCENTAGE OF ORIENTATION INSTANCES.

Direction	$o = M$	$o = S$	$o = D$
Left	65.2%	7.0%	27.8%
Right	65.3%	8.7%	26.0%

Table II
BLEU SCORE FOR CHINESE-ENGLISH TRANSLATION TASKS. ** MEANS SIGNIFICANTLY BETTER THAN *Baseline* ($p < 0.01$).

System	MT03	MT04	MT05
Baseline	34.59	35.54	33.11
+o_{rq}	35.12	36.19**	33.99**
+o_{me}	35.34**	36.56**	34.01**
Impr.	+0.75	+1.02	+0.90

word aligned bilingual corpus. We used minimum error rate training [10] for optimizing the feature weights. Case insensitive NIST BLEU4 was used to measure translation performance.

A. Result on Chinese-English Task

We extract 158.5 million orientation instances from the bilingual data. Note that the total number of instances of left direction is same as the number of right direction, since for each bispan we extract an instance for left direction and an instance for the right direction either. Table I shows the percentage of each orientation. Although the swapping orientations account for only 7.0% and 8.7% respectively, they are important for the probability estimation.

Table II shows the result on the Chinese-English translation task. Enhanced by the orientation model, our approach significantly outperforms the baseline system. When using frequency estimation (+o_{rq}), the improvement ranges from 0.53 to 0.88 point. If the orientation model is learned by discriminative training (+o_{me}), the improvement increase to range from 0.75 to 1.02. The discriminative training is slightly higher than relative frequency estimation. This may results from the fact that discriminative method uses more lexical information. Therefore, we only compare the result between +o_{me} and baseline in the following section. The result indicates that the HPB system does benefit from the orientation model.

B. Result on German-English Task

We also compare the performances of baseline system and +o_{me} system on a different language pair which is German-English. The bilingual data we used is Europarl V6 German-English corpus, including 1.6M sentence pairs. We use the English part of the bilingual corpus to train the language model. The development set is newstest 2008 and test set is newstest 2009 (WMT09).

Tested on the WMT09. +o_{me} system (18.21) outperforms the baseline system (17.70) by 0.51 point, which is less than the improvement on Chinese to English. The reason may be that German is closer to English than Chinese to English. Therefore, there are less swapping orientations (account for 4.0% during extraction) between German and English. This leaves a tight room for the orientation model to show its effect.

IV. CONCLUSION AND FUTURE WORK

We propose an orientation model for HPB translation, and confirm that the HPB system can also benefit from orientation model which has been wildly used in traditional phrase-based system.

Our method directly defines the nonterminal orientation based on the grammar, therefore, it's straight to be extended to other linguistic syntax-based systems including STSG-based translation and string-dependency. We believe that linguistic syntax-based systems can also benefit from specifying the nonterminal orientation and using lexical information of nonterminals. Another direction is to improve the discriminative training of orientation model with more features, such as length features and linguistical features.

REFERENCES

[1] C. Tillman, "A unigram orientation model for statistical machine translation," in *Proc. HLT-NAACL 2004*, 2004.

[2] 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, "Moses: Open source toolkit for statistical machine translation," in *Proc. ACL 2007 (demonstration session)*, 2007.

[3] M. Galley and C. D. Manning, "A simple and effective hierarchical phrase reordering model," in *Proc. EMNLP 2008*, 2008.

[4] D. Chiang, "Hierarchical phrase-based translation," *Computational Linguistics*, vol. 33, no. 2, pp. 201–228, 2007.

[5] Z. He, Q. Liu, and S. Lin, "Improving statistical machine translation using lexicalized rule selection," in *Proc. EMNLP 2008*, 2008.

[6] Z. He, Y. Meng, and H. Yu, "Maximum entropy based phrase reordering for hierarchical phrase-based translation," in *Proc. EMNLP 2010*, 2010.

[7] D. Xiong, Q. Liu, and S. Lin, "Maximum entropy based phrase reordering model for statistical machine translation," in *Proc. ACL 2006*, 2006.

[8] R. Zens and H. Ney, "Discriminative reordering models for statistical machine translation," in *Proc. WSMT 2006*, 2006.

[9] F. J. Och and H. Ney, "Discriminative training and maximum entropy models for statistical machine translation," in *Proc. ACL 2002*, 2002.

[10] F. J. Och, "Minimum error rate training in statistical machine translation," in *Proc. ACL 2003*, 2003.