A Linguistically Annotated Reordering Model for BTG-based Statistical Machine Translation

Deyi Xiong, Min Zhang, Aiti Aw and Haizhou Li
Human Language Technology
Institute for Infocomm Research
21 Heng Mui Keng Terrace, Singapore 119613
{dyxiong, mzhang, aaiti, hli}@i2r.a-star.edu.sg

Abstract

In this paper, we propose a linguistically annotated reordering model for BTG-based statistical machine translation. The model incorporates linguistic knowledge to predict orders for both syntactic and non-syntactic phrases. The linguistic knowledge is automatically learned from source-side parse trees through an annotation algorithm. We empirically demonstrate that the proposed model leads to a significant improvement of 1.55% in the BLEU score over the baseline reordering model on the NIST MT-05 Chinese-to-English translation task.

1 Introduction

In recent years, Bracketing Transduction Grammar (BTG) proposed by (Wu, 1997) has been widely used in statistical machine translation (SMT). However, the original BTG does not provide an effective mechanism to predict the most appropriate orders between two neighboring phrases. To address this problem, Xiong et al. (2006) enhance the BTG with a maximum entropy (MaxEnt) based reordering model which uses boundary words of bilingual phrases as features. Although this model outperforms previous unlexicalized models, it does not utilize any linguistically syntactic features, which have proven useful for phrase reordering (Wang et al., 2007). Zhang et al. (2007) integrates source-side syntactic knowledge into a phrase reordering model based on BTG-style rules. However, one limitation of this method is that it only reorders syntactic phrases because linguistic knowledge from parse trees is only carried by syntactic phrases as far as reordering is concerned, while non-syntactic phrases are combined monotonously with a flat reordering score.

In this paper, we propose a linguistically annotated reordering model for BTG-based SMT, which is a significant extension to the work mentioned above. The new model annotates each BTG node with linguistic knowledge by projecting source-side parse trees onto the corresponding binary trees generated by BTG so that syntactic features can be used for phrase reordering. Different from (Zhang et al., 2007), our annotation algorithm is able to label both syntactic and non-syntactic phrases. This enables our model to reorder any phrases, not limited to syntactic phrases. In addition, other linguistic information such as head words, is also used to improve reordering.

The rest of the paper is organized as follows. Section 2 briefly describes our baseline system while Section 3 introduces the linguistically annotated reordering model. Section 4 reports the experiments on a Chinese-to-English translation task. We conclude in Section 5.

2 Baseline SMT System

The baseline system is a phrase-based system which uses the BTG lexical rules ($A \rightarrow x/y$) to translate source phrase x into target phrase y and the BTG merging rules ($A \rightarrow [A, A]|\langle A, A\rangle$) to combine two neighboring phrases with a straight or inverted order. The BTG lexical rules are weighted with several features, such as phrase translation, word penalty and language models, in a log-linear form. For the merging rules, a MaxEnt-based reordering model using boundary words of neighboring phrases as features is used to predict the merging order, similar to (Xiong et al., 2006). We call this reordering model

Proceedings of ACL-08: HLT, Short Papers (Companion Volume), pages 149–152,
Columbus, Ohio, USA, June 2008.

boundary words based reordering model (BWR). In this paper, we propose to incorporate a linguistically annotated reordering model into the log-linear translation model, so as to strengthen the BWR's phrase reordering ability. We train all the model scaling factors on the development set to maximize the BLEU score. A CKY-style decoder is developed to generate the best BTG binary tree for each input sentence, which yields the best translation.

3 Linguistically Annotated Reordering Model

The linguistically annotated reordering model (LAR) is a MaxEnt-based classification model which predicts the phrase order $o \in \{inverted, straight\}$ during the application of merging rules to combine their left and right neighboring phrases A_l and A_r into a larger phrase A. [1] The model can be formulated as

$$LAR = \frac{exp(\sum_i \theta_i h_i(o, A_l, A_r, A))}{\sum_{o'} exp(\sum_i \theta_i h_i(o', A_l, A_r, A))} \quad (1)$$

where the functions $h_i \in \{0, 1\}$ are reordering features and θ_i are weights of these features. We define the features as linguistic elements which are annotated for each BTG node through an annotation algorithm, which comprise (1) head word hw, (2) the part-of-speech (POS) tag ht of head word and (3) syntactic label sl.

Each merging rule involves 3 nodes (A, A_l, A_r) and each node has 3 linguistic elements (hw, ht, sl). Therefore, the model has 9 features in total. Taking the left node A_l as an example, the model could use its head word w as feature as follows

$$h_i(o, A, A_l, A_r) = \begin{cases} 1, & A_l.hw = w, o = straight \\ 0, & otherwise \end{cases}$$

3.1 Annotation Algorithm

There are two steps to annotate a phrase or a BTG node using source-side parse tree information: (1) determining the span on the source side which is exactly covered by the node or the phrase, then (2) annotating the span according to the source-side parse tree. If the span is exactly covered by a single subtree in the source-side parse tree, it is called **syntactic span**, otherwise it is **non-syntactic span**. One of the challenges in this annotation algorithm is that phrases (BTG nodes) are not always covering syntactic span, in other words, they are not always aligned to all constituent nodes in the source-side tree. To solve this problem, we use heuristic rules to generate pseudo head word and **composite label** which consists of syntactic labels of three relevant constituents for the non-syntactic span. In this way, our annotation algorithm is capable of labelling both syntactic and non-syntactic phrases and therefore providing linguistic information for any phrase reordering.

1: Annotator (span $s = \langle i, j \rangle$, source-side parse tree t)
2: **if** s is a syntactic span **then**
3: Find the subtree c in t which exactly covers s
4: $s.\{\ \} := \{c.hw, c.ht, c.sl\}$
5: **else**
6: Find the smallest subtree c^* subsuming s in t
7: **if** $c^*.hw \in s$ **then**
8: $s.hw := c^*.hw$ and $s.ht := c^*.ht$
9: **else**
10: Find the word $w \in s$ which is nearest to $c^*.hw$
11: $s.hw := w$ and $s.ht := w.t$ /*$w.t$ is the POS tag of w*/
12: **end if**
13: Find the left boundary node ln of s in c^*
14: Find the right boundary node rn of s in c^*
15: $s.sl := ln.sl\text{-}c^*.sl\text{-}rn.sl$
16: **end if**

Figure 1: The Annotation Algorithm.

The annotation algorithm is shown in Fig. 1. For a syntactic span, the annotation is trivial. Annotation elements directly come from the subtree that covers the span exactly. For a non-syntactic span, the process is much complicated. Firstly, we need to locate the smallest subtree c^* subsuming the span (line 6). Secondly, we try to identify the head word/tag of the span (line 7-12) by using its head word directly if it is within the span. Otherwise, the word within the span which is nearest to hw will be assigned as the head word of the span. Finally, we determine the composite label of the span (line 13-15), which is formulated as L-C-R. L/R means the syntactic label of the left/right **boundary node** of s which is the highest leftmost/rightmost sub-node of c^* not overlapping the span. If there is no such boundary node

[1] Each phrase is also a node in the BTG tree generated by the decoder.

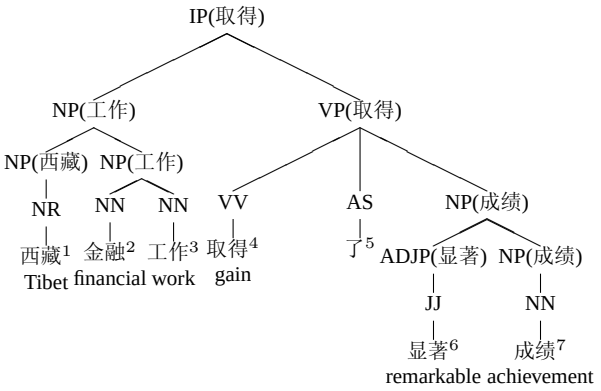

Figure 2: A syntactic parse tree with head word annotated for each internal node. The superscripts of leaf nodes denote their surface positions from left to right.

$span$	hw	ht	sl
$\langle 1, 2\rangle$	金融	NN	NULL-NP-NN
$\langle 2, 3\rangle$	工作	NN	NP
$\langle 2, 4\rangle$	取得	VV	NP-IP-NP
$\langle 3, 4\rangle$	取得	VV	NP-IP-NP

Table 1: Annotation samples according to the tree shown in Fig. 2. hw/ht represents the head word/tag, respectively. sl means the syntactic label.

(the span s is exactly aligned to the left/right boundary of c^*), L/R will be set to NULL. C is the label of c^*. L, R and C together define the external syntactic context of s.

Fig. 2 shows a syntactic parse tree for a Chinese sentence, with head word annotated for each internal node. Some sample annotations are given in Table 1.

3.2 Training and Decoding

Training an LAR model takes three steps. Firstly, we extract annotated reordering examples from source-side parsed, word-aligned bilingual data using the annotation algorithm and the reordering example extraction algorithm of (Xiong et al., 2006). We then generate features using linguistic elements of these examples and finally estimate feature weights. This training process flexibly learns rich syntactic reordering information without explicitly constructing BTG tree or forest for each sentence pair.

During decoding, each input source sentence is firstly parsed to obtain its syntactic tree. Then the CKY-style decoder tries to generate the best BTG tree using the lexical and merging rules. When two neighboring nodes are merged in a specific order, the two embedded reordering models, BWR and LAR, evaluate this merging independently with individual scores. The former uses boundary words as features while the latter uses the linguistic elements as features, annotated on the BTG nodes through the annotation algorithm according to the source-side parse tree.

4 Experiments

All experiments in this section were carried out on the Chinese-to-English translation task of the NIST MT-05. The baseline system and the new system with the LAR model were trained on the FBIS corpus. We removed 15,250 sentences, for which the Chinese parser (Xiong et al., 2005) failed to produce syntactic parse trees. The parser was trained on the Penn Chinese Treebank with a F1 score of 79.4%. The remaining FBIS corpus (224,165 sentence pairs) was used to obtain standard bilingual phrases for the systems.

We extracted 2.8M reordering examples from these sentences. From these examples, we generated 114.8K reordering features for the BWR model using the right boundary words of phrases and 85K features for the LAR model using linguistic annotations. We ran the MaxEnt toolkit (Zhang, 2004) to tune reordering feature weights with iteration number being set to 100 and Gaussian prior to 1 to avoid overfitting.

We built our four-gram language model using Xinhua section of the English Gigaword corpus (181.1M words) with the SRILM toolkit (Stolcke, 2002). For the efficiency of minimum-error-rate training (Och, 2003), we built our development set (580 sentences) using sentences not exceeding 50 characters from the NIST MT-02 evaluation test data.

4.1 Results

We compared various reordering configurations in the baseline system and new system. The baseline system only has BWR as the reordering model, while the new system employs two reordering models: BWR and LAR. For the linguistically annotated reordering model LAR, we augment its feature pool incrementally: firstly using only single labels

[2](SL) as features (132 features in total), then constructing composite labels for non-syntactic phrases (+BNL) (6.7K features), and finally introducing head words and their POS tags into the feature pool (+BNL+HWT) (85K features). This series of experiments demonstrate the impact and degree of contribution made by each feature for reordering. We also conducted experiments to investigate the effect of restricting reordering to syntactic phrases in the new system using the best reordering feature set (SL+BNL+HWT) for LAR. The experimental results (case-sensitive BLEU scores together with confidence intervals) are presented in Table 2, from which we have the following observations:

(1) The LAR model improves the performance statistically significantly. Even we only use the baseline feature set SL with only 132 features for the LAR, the BLEU score improves from 0.2497 to 0.2588. This is because most of the frequent reordering patterns between Chinese and English have been captured using syntactic labels. For example, the pre-verbal modifier PP in Chinese is translated into post-verbal counterpart in English. This reordering can be described by a rule with an inverted order: $VP \rightarrow \langle PP, VP \rangle$, and captured by our syntactic reordering features.

(2) Context information, provided by labels of boundary nodes (BNL) and head word/tag pairs (HWT), also improves phrase reordering. Producing composite labels for non-syntactic BTG nodes (+BNL) and integrating head word/tag pairs into the LAR as reordering features (+BNL+HWT) are both effective, indicating that context information complements syntactic label for capturing reordering patterns.

(3) Restricting phrase reordering to syntactic phrases is harmful. The BLEU score plummets from 0.2652 to 0.2512.

Reordering Configuration	BLEU (%)
BWR	24.97 ± 0.90
BWR + LAR (SL)	25.88 ± 0.95
BWR + LAR (+BNL)	26.27 ± 0.98
BWR + LAR (+BNL+HWT)	26.52 ± 0.96
Only allowed SPs reordering	25.12 ± 0.87

Table 2: The effect of the linguistically annotated reordering model. BWR denotes the boundary word based reordering model while LAR denotes the linguistically annotated reordering model. (SL) is the baseline feature set, (+BNL) and (+BNL+HWT) are extended feature sets for the LAR. SP means *syntactic phrase*.

5 Conclusion

In this paper, we have presented a linguistically annotated reordering model to effectively integrate linguistic knowledge into phrase reordering by merging source-side parse trees with BTG binary trees. Our experimental results show that, on the NIST MT-05 task of Chinese-to-English translation, the proposed reordering model leads to BLEU improvement of 1.55%. We believe that our linguistically annotated reordering model can be further improved by using better annotation which transfers more knowledge (morphological, syntactic or semantic) to the model.

[2]For non-syntactic node, we only use the single label C, without constructing composite label L-C-R.

References

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of ACL 2003*.

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

Chao Wang, Michael Collins and Philipp Koehn. 2007. Chinese Syntactic Reordering for Statistical Machine Translation. In *Proceedings of EMNLP-CoNLL 2007*.

Dekai Wu. 1997. Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora. *Computational Linguistics*, 23(3):377-403.

Deyi Xiong, Shuanglong Li, Qun Liu, Shouxun Lin, Yueliang Qian. 2005. Parsing the Penn Chinese Treebank with Semantic Knowledge. In *Proceedings of IJCNLP*, Jeju Island, Korea.

Deyi Xiong, Qun Liu and Shouxun Lin. 2006. Maximum Entropy Based Phrase Reordering Model for Statistical Machine Translation. In *Proceedings of ACL-COLING 2006*.

Dongdong Zhang, Mu Li, Chi-Ho Li and Ming Zhou. 2007. Phrase Reordering Model Integrating Syntactic Knowledge for SMT. In *Proceedings of EMNLP-CoNLL 2007*.

Le Zhang. 2004. Maximum Entropy Modeling Tooklkit for Python and C++. Available at http://homepages.inf.ed.ac.uk/s0450736 /maxent_toolkit.html.