A Joint Rule Selection Model for Hierarchical Phrase-based Translation*

Lei Cui[†], **Dongdong Zhang**[‡], **Mu Li**[‡], **Ming Zhou**[‡], and **Tiejun Zhao**[†]

[†]School of Computer Science and Technology
Harbin Institute of Technology, Harbin, China
{cuilei,tjzhao}@mtlab.hit.edu.cn
[‡]Microsoft Research Asia, Beijing, China
{dozhang,muli,mingzhou}@microsoft.com

Abstract

In hierarchical phrase-based SMT systems, statistical models are integrated to guide the hierarchical rule selection for better translation performance. Previous work mainly focused on the selection of either the source side of a hierarchical rule or the target side of a hierarchical rule rather than considering both of them simultaneously. This paper presents a joint model to predict the selection of hierarchical rules. The proposed model is estimated based on four sub-models where the rich context knowledge from both source and target sides is leveraged. Our method can be easily incorporated into the practical SMT systems with the log-linear model framework. The experimental results show that our method can yield significant improvements in performance.

1 Introduction

Hierarchical phrase-based model has strong expression capabilities of translation knowledge. It can not only maintain the strength of phrase translation in traditional phrase-based models (Koehn et al., 2003; Xiong et al., 2006), but also characterize the complicated long distance reordering similar to syntactic based statistical machine translation (SMT) models (Yamada and Knight, 2001; Quirk et al., 2005; Galley et al., 2006; Liu et al., 2006; Marcu et al., 2006; Mi et al., 2008; Shen et al., 2008).

In hierarchical phrase-based SMT systems, due to the flexibility of rule matching, a huge number of hierarchical rules could be automatically learnt from bilingual training corpus (Chiang, 2005). SMT decoders are forced to face the challenge of proper rule selection for hypothesis generation, including both source-side rule selection and target-side rule selection where the source-side rule determines what part of source words to be translated and the target-side rule provides one of the candidate translations of the source-side rule. Improper rule selections may result in poor translations.

There is some related work about the hierarchical rule selection. In the original work (Chiang, 2005), the target-side rule selection is analogous to the model in traditional phrase-based SMT system such as Pharaoh (Koehn et al., 2003). Extending this work, (He et al., 2008; Liu et al., 2008) integrate rich context information of non-terminals to predict the target-side rule selection. Different from the above work where the probability distribution of source-side rule selection is uniform, (Setiawan et al., 2009) proposes to select source-side rules based on the captured function words which often play an important role in word reordering. There is also some work considering to involve more rich contexts to guide the source-side rule selection. (Marton and Resnik, 2008; Xiong et al., 2009) explore the source syntactic information to reward exact matching structure rules or punish crossing structure rules.

All the previous work mainly focused on either source-side rule selection task or target-side rule selection task rather than both of them together. The separation of these two tasks, however, weakens the high interrelation between them. In this paper, we propose to integrate both source-side and target-side rule selection in a unified model. The intuition is that the joint selection of source-side and target-side rules is more reliable as it conducts the search in a larger space than the single selection task does. It is expected that these two kinds of selection can help and affect each other, which may potentially lead to better hierarchical rule selections with a relative global optimum instead of a local optimum that might be reached in the pre-

*This work was finished while the first author visited Microsoft Research Asia as an intern.

Proceedings of the ACL 2010 Conference Short Papers, pages 6–11,
Uppsala, Sweden, 11-16 July 2010.

vious methods. Our proposed joint probability model is factored into four sub-models that can be further classified into source-side and target-side rule selection models or context-based and context-free selection models. The context-based models explore rich context features from both source and target sides, including function words, part-of-speech (POS) tags, syntactic structure information and so on. Our model can be easily incorporated as an independent feature into the practical hierarchical phrase-based systems with the log-linear model framework. The experimental results indicate our method can improve the system performance significantly.

2 Hierarchical Rule Selection Model

Following (Chiang, 2005), $\langle \alpha, \gamma \rangle$ is used to represent a synchronous context free grammar (SCFG) rule extracted from the training corpus, where α and γ are the source-side and target-side rule respectively. Let C be the context of $\langle \alpha, \gamma \rangle$. Formally, our joint probability model of hierarchical rule selection is described as follows:

$$P(\alpha, \gamma | C) = P(\alpha | C) P(\gamma | \alpha, C) \quad (1)$$

We decompose the joint probability model into two sub-models based on the Bayes formulation, where the first sub-model is *source-side rule selection model* and the second one is the *target-side rule selection model.*

For the source-side rule selection model, we further compute it by the interpolation of two sub-models:

$$\theta P_s(\alpha) + (1 - \theta) P_s(\alpha | C) \quad (2)$$

where $P_s(\alpha)$ is the *context-free source model* (CFSM) and $P_s(\alpha|C)$ is the *context-based source model* (CBSM), θ is the interpolation weight that can be optimized over the development data.

CFSM is the probability of source-side rule selection that can be estimated based on maximum likelihood estimation (MLE) method:

$$P_s(\alpha) = \frac{\sum_{\gamma} Count(\langle \alpha, \gamma \rangle)}{Count(\alpha)} \quad (3)$$

where the numerator is the total count of bilingual rule pairs with the same source-side rule that are extracted based on the extraction algorithm in (Chiang, 2005), and the denominator is the total amount of source-side rule patterns contained in the monolingual source side of the training corpus. CFSM is used to capture how likely the source-side rule is linguistically motivated or has the corresponding target-side counterpart.

For CBSM, it can be naturally viewed as a classification problem where each distinct source-side rule is a single class. However, considering the huge number of classes may cause serious data sparseness problem and thereby degrade the classification accuracy, we approximate CBSM by a binary classification problem which can be solved by the maximum entropy (ME) approach (Berger et al., 1996) as follows:

$$\begin{aligned} P_s(\alpha|C) &\approx P_s(\upsilon|\alpha, C) \\ &= \frac{exp[\sum_i \lambda_i h_i(\upsilon, \alpha, C)]}{\sum_{\upsilon'} exp[\sum_i \lambda_i h_i(\upsilon', \alpha, C)]} \end{aligned} \quad (4)$$

where $\upsilon \in \{0, 1\}$ is the indicator whether the source-side rule is applied during decoding, $\upsilon = 1$ when the source-side rule is applied, otherwise $\upsilon = 0$; h_i is a feature function, λ_i is the weight of h_i. CBSM estimates the probability of the source-side rule being selected according to the rich context information coming from the surface strings and sub-phrases that will be reduced to non-terminals during decoding.

Analogously, we decompose the target-side rule selection model by the interpolation approach as well:

$$\varphi P_t(\gamma) + (1 - \varphi) P_t(\gamma | \alpha, C) \quad (5)$$

where $P_t(\gamma)$ is the *context-free target model* (CFTM) and $P_t(\gamma|\alpha, C)$ is the *context-based target model* (CBTM), φ is the interpolation weight that can be optimized over the development data.

In the similar way, we compute CFTM by the MLE approach and estimate CBTM by the ME approach. CFTM computes how likely the target-side rule is linguistically motivated, while CBTM predicts how likely the target-side rule is applied according to the clues from the rich context information.

3 Model Training of CBSM and CBTM

3.1 The acquisition of training instances

CBSM and CBTM are trained by ME approach for the binary classification, where a training instance consists of a label and the context related to SCFG rules. The context is divided into source context

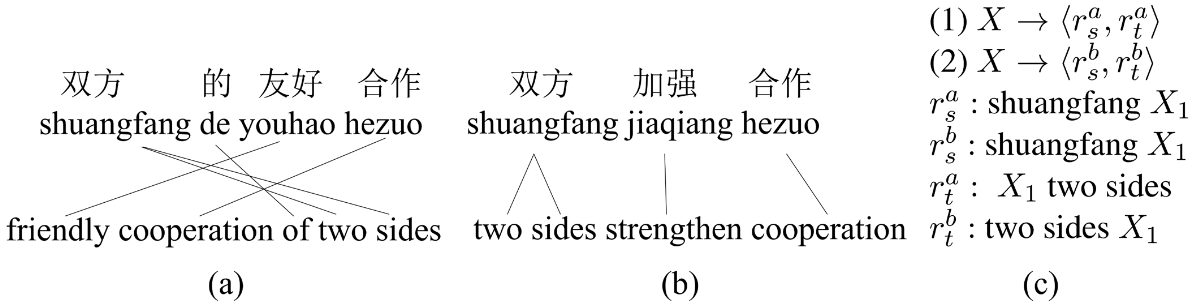
双方 的 友好 合作
shuangfang de youhao hezuo
friendly cooperation of two sides
(a)
双方 加强 合作
shuangfang jiaqiang hezuo
two sides strengthen cooperation
(b)
(1) $X \to \langle r_s^a, r_t^a \rangle$
(2) $X \to \langle r_s^b, r_t^b \rangle$
r_s^a : shuangfang X_1
r_s^b : shuangfang X_1
r_t^a : X_1 two sides
r_t^b : two sides X_1
(c)

Figure 1: Example of training instances in CBSM and CBTM.

and target context. CBSM is trained only based on the source context while CBTM is trained over both the source and the target context. All the training instances are automatically constructed from the bilingual training corpus, which have labels of either positive (i.e., $\upsilon = 1$) or negative (i.e., $\upsilon = 0$). This section explains how the training instances are constructed for the training of CBSM and CBTM.

Let s and t be the source sentence and target sentence, W be the word alignment between them, r_s be a source-side rule that pattern-matches a sub-phrase of s, r_t be the target-side rule pattern-matching a sub-phrase of t and being aligned to r_s based on W, and $C(r)$ be the context features related to the rule r which will be explained in the following section.

For the training of CBSM, if the SCFG rule $\langle r_s, r_t \rangle$ can be extracted based on the rule extraction algorithm in (Chiang, 2005), $\langle \upsilon = 1, C(r_s) \rangle$ is constructed as a positive instance, otherwise $\langle \upsilon = 0, C(r_s) \rangle$ is constructed as a negative instance. For example in Figure 1(a), the context of source-side rule ”X_1 hezuo” that pattern-matches the phrase ”youhao hezuo” produces a positive instance, while the context of ”X_1 youhao” that pattern-matches the source phrase ”de youhao” or ”shuangfang de youhao” will produce a negative instance as there are no corresponding plausible target-side rules that can be extracted legally[1].

For the training of CBTM, given r_s, suppose there is a SCFG rule set $\{\langle r_s, r_t^k \rangle | 1 \leq k \leq n\}$ extracted from multiple distinct sentence pairs in the bilingual training corpus, among which we assume $\langle r_s, r_t^i \rangle$ is extracted from the sentence pair $\langle s, t \rangle$. Then, we construct $\langle \upsilon = 1, C(r_s), C(r_t^i) \rangle$ as a positive instance, while the elements in $\{\langle \upsilon = 0, C(r_s), C(r_t^j) \rangle | j \neq i \wedge 1 \leq j \leq n\}$ are viewed as negative instances since they fail to be applied to the translation from s to t. For example in Figure 1(c), Rule (1) and Rule (2) are two different SCFG rules extracted from Figure 1(a) and Figure 1(b) respectively, where their source-side rules are the same. As Rule (1) cannot be applied to Figure 1(b) for the translation and Rule (2) cannot be applied to Figure 1(a) for the translation either, $\langle \upsilon = 1, C(r_s^a), C(r_t^a) \rangle$ and $\langle \upsilon = 1, C(r_s^b), C(r_t^b) \rangle$ are constructed as positive instances while $\langle \upsilon = 0, C(r_s^a), C(r_t^b) \rangle$ and $\langle \upsilon = 0, C(r_s^b), C(r_t^a) \rangle$ are viewed as negative instances. It is noticed that this instance construction method may lead to a large quantity of negative instances and choke the training procedure. In practice, to limit the size of the training set, the negative instances constructed based on low-frequency target-side rules are pruned.

3.2 Context-based features for ME training

ME approach has the merit of easily combining different features to predict the probability of each class. We incorporate into the ME based model the following informative context-based features to train CBSM and CBTM. These features are carefully designed to reduce the data sparseness problem and some of them are inspired by previous work (He et al., 2008; Gimpel and Smith, 2008; Marton and Resnik, 2008; Chiang et al., 2009; Setiawan et al., 2009; Shen et al., 2009; Xiong et al., 2009):

1. **Function word features**, which indicate whether the hierarchical source-side/target-side rule strings and sub-phrases covered by non-terminals contain function words that are often important clues of predicting syntactic structures.

[1] Because the aligned target words are not contiguous and ”cooperation” is aligned to the word outside the source-side rule.

2. **POS features**, which are POS tags of the boundary source words covered by non-terminals.

3. **Syntactic features**, which are the constituent constraints of hierarchical source-side rules exactly matching or crossing syntactic subtrees.

4. **Rule format features**, which are non-terminal positions and orders in source-side/target-side rules. This feature interacts between source and target components since it shows whether the translation ordering is affected.

5. **Length features**, which are the length of sub-phrases covered by source non-terminals.

4 Experiments

4.1 Experiment setting

We implement a hierarchical phrase-based system similar to the Hiero (Chiang, 2005) and evaluate our method on the Chinese-to-English translation task. Our bilingual training data comes from FBIS corpus, which consists of around 160K sentence pairs where the source data is parsed by the Berkeley parser (Petrov and Klein, 2007). The ME training toolkit, developed by (Zhang, 2006), is used to train our CBSM and CBTM. The training size of constructed positive instances for both CBSM and CBTM is 4.68M, while the training size of constructed negative instances is 3.74M and 3.03M respectively. Following (Setiawan et al., 2009), we identify function words as the 128 most frequent words in the corpus. The interpolation weights are set to $\theta = 0.75$ and $\varphi = 0.70$. The 5-gram language model is trained over the English portion of FBIS corpus plus Xinhua portion of the Gigaword corpus. The development data is from NIST 2005 evaluation data and the test data is from NIST 2006 and NIST 2008 evaluation data. The evaluation metric is the case-insensitive BLEU4 (Papineni et al., 2002). Statistical significance in BLEU score differences is tested by paired bootstrap re-sampling (Koehn, 2004).

4.2 Comparison with related work

Our baseline is the implemented Hiero-like SMT system where only the standard features are employed and the performance is state-of-the-art. We compare our method with the baseline and some typical approaches listed in Table 1 where XP+ denotes the approach in (Marton and Resnik, 2008) and TOFW (topological ordering of function words) stands for the method in (Setiawan et al., 2009). As (Xiong et al., 2009)'s work is based on phrasal SMT system with bracketing transduction grammar rules (Wu, 1997) and (Shen et al., 2009)'s work is based on the string-to-dependency SMT model, we do not implement these two related work due to their different models from ours. We also do not compare with (He et al., 2008)'s work due to its less practicability of integrating numerous sub-models.

Methods	NIST 2006	NIST 2008
Baseline	0.3025	0.2200
XP+	0.3061	0.2254
TOFW	0.3089	0.2253
Our method	0.3141	0.2318

Table 1: Comparison results, our method is significantly better than the baseline, as well as the other two approaches ($\boldsymbol{p} < 0.01$)

As shown in Table 1, all the methods outperform the baseline because they have extra models to guide the hierarchical rule selection in some ways which might lead to better translation. Apparently, our method also performs better than the other two approaches, indicating that our method is more effective in the hierarchical rule selection as both source-side and target-side rules are selected together.

4.3 Effect of sub-models

Due to the space limitation, we analyze the effect of sub-models upon the system performance, rather than that of ME features, part of which have been investigated in previous related work.

Settings	NIST 2006	NIST 2008
Baseline	0.3025	0.2200
Baseline+CFSM	0.3092^*	0.2266^*
Baseline+CBSM	0.3077^*	0.2247^*
Baseline+CFTM	0.3076^*	0.2286^*
Baseline+CBTM	0.3060	0.2255^*
Baseline+CFSM+CFTM	0.3109^*	0.2289^*
Baseline+CFSM+CBSM	0.3104^*	0.2282^*
Baseline+CFTM+CBTM	0.3099^*	0.2299^*
Baseline+all sub-models	0.3141^*	0.2318^*

Table 2: Sub-model effect upon the performance, *: significantly better than baseline ($\boldsymbol{p} < 0.01$)

As shown in Table 2, when sub-models are integrated as independent features, the performance is improved compared to the baseline, which shows that each of the sub-models can improve the hierarchical rule selection. It is noticeable that the performance of the source-side rule selection model is comparable with that of the target-side rule selection model. Although CFSM and CFTM perform only slightly better than the others among the individual sub-models, the best performance is achieved when all the sub-models are integrated.

5 Conclusion

Hierarchical rule selection is an important and complicated task for hierarchical phrase-based SMT system. We propose a joint probability model for the hierarchical rule selection and the experimental results prove the effectiveness of our approach.

In the future work, we will explore more useful features and test our method over the large scale training corpus. A challenge might exist when running the ME training toolkit over a big size of training instances from the large scale training data.

Acknowledgments

We are especially grateful to the anonymous reviewers for their insightful comments. We also thank Hendra Setiawan, Yuval Marton, Chi-Ho Li, Shujie Liu and Nan Duan for helpful discussions.

References

Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. 1996. *A Maximum Entropy Approach to Natural Language Processing*. *Computational Linguistics*, 22(1): pages 39-72.

David Chiang. 2005. *A Hierarchical Phrase-Based Model for Statistical Machine Translation.* In *Proc. ACL*, pages 263-270.

David Chiang, Kevin Knight, and Wei Wang. 2009. *11,001 New Features for Statistical Machine Translation.* In *Proc. HLT-NAACL*, pages 218-226.

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. 2006. *Scalable Inference and Training of Context-Rich Syntactic Translation Models.* In *Proc. ACL-Coling*, pages 961-968.

Kevin Gimpel and Noah A. Smith. 2008. *Rich Source-Side Context for Statistical Machine Translation.* In *Proc. the Third Workshop on Statistical Machine Translation*, pages 9-17.

Zhongjun He, Qun Liu, and Shouxun Lin. 2008. *Improving Statistical Machine Translation using Lexicalized Rule Selection.* In *Proc. Coling*, pages 321-328.

Philipp Koehn. 2004. *Statistical Significance Tests for Machine Translation Evaluation.* In *Proc. EMNLP*.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. *Statistical Phrase-Based Translation.* In *Proc. HLT-NAACL*, pages 127-133.

Qun Liu, Zhongjun He, Yang Liu, and Shouxun Lin. 2008. *Maximum Entropy based Rule Selection Model for Syntax-based Statistical Machine Translation.* In *Proc. EMNLP*, pages 89-97.

Yang Liu, Yun Huang, Qun Liu, and Shouxun Lin. 2007. *Forest-to-String Statistical Translation Rules.* In *Proc. ACL*, pages 704-711.

Yang Liu, Qun Liu, and Shouxun Lin. 2006. *Tree-to-String Alignment Template for Statistical Machine Translation.* In *Proc. ACL-Coling*, pages 609-616.

Daniel Marcu, Wei Wang, Abdessamad Echihabi, and Kevin Knight. 2006. *SPMT: Statistical Machine Translation with Syntactified Target Language Phrases.* In *Proc. EMNLP*, pages 44-52.

Yuval Marton and Philip Resnik. 2008. *Soft Syntactic Constraints for Hierarchical Phrased-Based Translation.* In *Proc. ACL*, pages 1003-1011.

Haitao Mi, Liang Huang, and Qun Liu. 2008. *Forest-Based Translation.* In *Proc. ACL*, pages 192-199.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. *Bleu: a Method for Automatic Evaluation of Machine Translation.* In *Proc. ACL*, pages 311-318.

Slav Petrov and Dan Klein. 2007. *Improved Inference for Unlexicalized Parsing.* In *Proc. HLT-NAACL*, pages 404-411.

Chris Quirk, Arul Menezes, and Colin Cherry. 2005. *Dependency Treelet Translation: Syntactically Informed Phrasal SMT.* In *Proc. ACL*, pages 271-279.

Libin Shen, Jinxi Xu, and Ralph Weischedel. 2008. *A New String-to-Dependency Machine Translation Algorithm with a Target Dependency Language Model.* In *Proc. ACL*, pages 577-585.

Libin Shen, Jinxi Xu, Bing Zhang, Spyros Matsoukas, and Ralph Weischedel. 2009. *Effective Use of Linguistic and Contextual Information for Statistical Machine Translation.* In *Proc. EMNLP*, pages 72-80.

Hendra Setiawan, Min Yen Kan, Haizhou Li, and Philip Resnik. 2009. *Topological Ordering of Function Words in Hierarchical Phrase-based Translation.* In *Proc. ACL*, pages 324-332.

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

Deyi Xiong, Qun Liu, and Shouxun Lin. 2006. *Maximum Entropy Based Phrase Reordering Model for Statistical Machine Translation.* In *Proc. ACL-Coling*, pages 521-528.

Deyi Xiong, Min Zhang, Aiti Aw, and Haizhou Li. 2009. *A Syntax-Driven Bracketing Model for Phrase-Based Translation.* In *Proc. ACL*, pages 315-323.

Kenji Yamada and Kevin Knight. 2001. *A Syntax-based Statistical Translation Model*. In *Proc. ACL*, pages 523-530.

Le Zhang. 2006. *Maximum entropy modeling toolkit for python and c++*. available at `http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html`.