2017 International Conference on Computer Network, Electronic and Automation (ICCNEA 2017)

Learning Better Classification-based Reordering Model for Phrase-based Translation

Li Fuxue
NiuTrans Lab School of Computer Science and Engineering,
Northeastern University, China,
YingKou Institute of technology,
e-mail:lifuxue119@163.com

Zhu Jingbo
NiuTrans Lab School of Computer Science and Engineering, Northeastern University, China

Xiao Tong
NiuTrans Lab School of Computer Science and Engineering, Northeastern University, China

***Abstract*—Reordering is of a challenging issue in phrase-based statistical machine translation systems. This paper proposed three techniques to optimize classification-based reordering models for phrase-based translation under the bracket transduction grammar framework. First, a forced decoding technique is adopted to learn reordering samples for maximum entropy model training. Secondly, additional features are learned from the context of two consecutive phrases to enhance the prediction ability of the reordering classifier. Thirdly, the reordering model score is integrated as two feature functions (STRAIGHT and INVERTED) into the log-linear model to improve its discriminative ability. Experimental result demonstrates significant improvements over the baseline in two translation tasks such as Chinese to English and Chinese to Japanese translation.**

Keywords-statistical machine translation; word reordering; log linear model; feature selection

I. INTRODUCTION

The phrase-based translation approach has been a popular and widely used strategy to the statistical machine translation (SMT). In phrase-based statistic machine translation (PBMT), reordering is of a big challenge and a great importance issue, and it is typically handled by two different models such as distortion model and lexicalized reordering model. Distortion models consider the distance of the words or phrases movement (Brown et al., 1993; Koehn et al. 2003). Lexicalized reordering models are proposed to learn phrase orientation base on content (Tillmann, 2004; Koehn et al., 2005; Nagata et al., 2006). In this paper, we focus on lexicalized reordering models for phrase-based translation. Among the lexicalized reordering models, Bracket Transduction Grammar (BTG) restriction is widely used for reordering in SMT (Zens et al., 2004) due to its good tradeoff between efficiency and expressiveness. Under framework of BTG, the reordering task is considered as classification problem and achieves good performance (Abdullah et al., 2014), referred to as the classification-based reordering model (CRM). The maximum entropy classifier is widely adopted by many researchers to implement the CRM (Zens and Ney, 2006; Xiong et al., 2006; Nguyen et al., 2009; Xiang et al., 2011), and is also considered in this work.

In principle, three key issues should be addressed to build effective classification-based reordering models. The first key issue is how to learn reordering samples from bilingual corpus to train the classifier. The traditional way is to learn reordering samples from each sentence pair based on its word alignments. However, it is sensitive to word alignment noise because a word alignment error would result in some incorrect reordering samples and block some desirble reordering samples. To alleviate this problem, this paper presents a forced decoding based approach to learning reordering samples from derivations of each sentence pair instead of word alignments. Secondly, to build a powerful classifier for CRM, e.g. based on maximum entropy model, traditional methods learn classification features only from source and target sides of two consecutive phrases for reordering, e.g., boundary information of both phrases. Since the source-side context of two consecutive phrases can provide more valuable information for reordering, in our work some additional features are learned from the context of two consecutive phrases to enhance the prediction ability of the reordering classifier. Thirdly, reordering model score is typically integrated as one feature function into the log-linear model. Our method considers reordering model score as two feature functions (STRAIGHT and INVERTED) to improve reordering discriminative ability. Experimental results show significant improvements over the baseline in two translation tasks such as Chinese to English and Chinese to Japanese translation.

II. RELATED WORK

A number of approaches have been proposed to address the reordering issue in phrase-based translation. In principle, the reordering approaches can be divided into two categories: pre-reordering and reordering mode at decoding time.

The first category reorders the source language in a preprocessing step before decoding (Nieben and Ney 2001;Collins et al., 2005; Isozaki et al., 2010), this kind of methods aim at arranging source words in a target-like order before decoding. This paper focuses on the reordering model at decoding time.

The second category estimates phrase movement with reordering models at decoding time. In distortion models, IBM models 1 and 2 define the distortion parameters in accordance with the word positions in the sentence pair instead of actual words at those positions (Brown et al., 1993). Models 4 and 5 limit this by replacing absolute word positions with relative word positions (Brown et al., 1993). Lexicalized reordering models introduce reordering probabilities conditioned on the words of each phrase pair, and they distinguish three orientations with respect to the previous phrase pair (Tillmann, 2004; Koehn et al., 2005; Nagata et al., 2006).

Tillman (2004) considers the position of each phrase as a class, and Koehn et al. (2005) extend the classes to any arbitrary number. Galley and Manning (2008) extended the lexicalized reordering model to tackle long-distance reordering. These reordering models learn local orientations with probabilities for each bilingual phrase from training data. However, since reordering is related to concrete phrases, the data sparseness problem may be introduced. Under the restriction of BTG, some researchers had posed the phrase movement problem as a classification problem. Zens and Ney (2006) introduced a maximum entropy classifier for phrase reordering. Xiong et al. (2006) proposed a maximum entropy model to predicate reordering of neighbour blocks (i.e. phrase pairs), and considered straight or inverted orientations. Nguyen et al. (2009) applied a maximum entropy model to learn orientations identified by the hierarchical reordering model. Xiang et al. (2011) introduced a smoothed prior probability to maximum entropy model, and used multiple features based on syntactic parsing to improve reordering in PBMT. Alrajeh and Niranjan (2014) posed phrase movements as a classification problem, and explored a generative learning approach named Bayesian naive Bayes to dealing with phrase reordering. Recently, neural reordering model (Li P et al., 2014) is also adopted to deal the reordering issue and it could address the data sparseness problem.

III. Classification-based Reordering Model for PBMT

Phrase-based SMT systems move from using words as translation units to using phrases, it has been widely used and achieves the state-of-the-art performance. However, reordering is still a crucial issue for PBMT. Many researchers proposed lexicalized reordering models to address this issue (Tillmann, 2004; Koehn et al., 2005; Nagata et al., 2006). In prinpicle, lexicalized reordering models learn local orientations with probabilities for each bilingual phrase from training data. To alleviate the data sparness problem of lexicaliezd reordering, a kind of models which treat the reordering issue as classification problem are proposed under the BTG framework (Zens et al., 2004).

BTG is employed firstly in statistical machine translation in (Wu, 1996). Under the framework of BTG, three rules are adopted to generate the translations:

(1) A→ [A1, A2];

(2) A→＜A1, A2＞;

(3) A→ x / y;

where rule (1) merges two consecutive blocks into a larger blocks in the straight order, rule (2) does the same work in the inverted order and rule (3) translates phrase y into target phrase x and generates a block A.

The maximum entropy-based approach (so called MaxEnt) is widely used to implement classification-based lexicalized reordering models by many researchers (Zens and Ney, 2006; Xiong et al., 2006; Nguyen et al., 2009; Xiang et al., 2011), which is defined as:

$$\Omega = f(o, A_1, A_2) \tag{1}$$

$$o \in (straight, invered) \tag{2}$$

$$\Omega = p_\theta(o \mid A_1, A_2) = \frac{\exp(\sum_i \theta_i h_i(o, A_1, A_2))}{\sum_o \exp(\sum_i \theta_i h_i(o, A_1, A_2))} \tag{3}$$

where o∈(straight, inverted) indicates phrase orientation, h ∈{0,1}is the ith classification feature and θ i is weight of the ith feature.

IV. Learning Better CRM

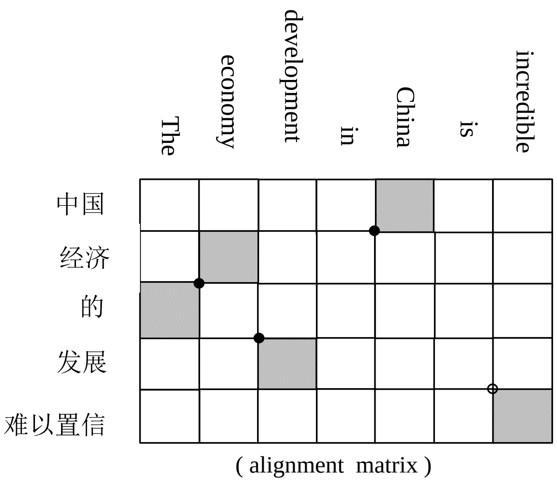
This section presents three optimization techniques to improve classification-based reordering models for PBMT, involving reordering sample generation for training, feature selection for classification and reordering feature functions for decoding. We will discuss these techniques in details as follows.

A. *Reordering Sample Generation for Training*

The first step is to learn reordering samples to train the MaxEnt classifier used by CRM. In traditional method, the reodering samples are learned from bilingual sentence pairs based on word alignments. Given a bilingual sentence pair with its word alignments, we can get the alignment matrix as shown in Figure 1. There are some vertexes shared between two blocks which have four directions: top-left, top-right, bottom-left and bottom-right. The top-right and bottom-left link blocks with the straight order, so we call them INVERTED links. Similarly, we call the top-left and bottom-right STRAIGHT links since they link blocks with the inverted order. For example, in Figure 1, the order of “经济 -economy” (Block1) and “的-the” (Block2) is INVERTED, and the order of “经济 的-the economy” and “发展-development” is STRAIGHT. Actually the traditional approach is sensitive to the word alignments, because word alignments errors would result in incorrect reordering training samples and block some desirable reordering samples extracted. For example, the word alignment error [“的”-“the”] introduces some incorrect reordering samples, e.g., {“经济-Economy”, “的-the”, INVERTED}.

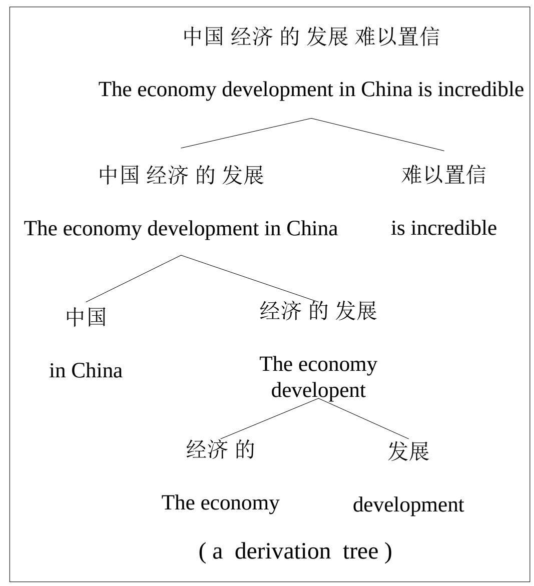
To alleviate this problem, this paper adopts a phrase-based forced decoding approach to learning reordering samples from derivation tree (or forest) of each bilingual sentence pair, as shown in the right side of Figure 1. The phrase-based forced decoding technique is different from the typical phrase-based decoding method, in which the derivation of each translation hypothesis must yield the same target sentence during the phrase-based decoding process. In other words, a derivation hypothesis different from the given target sentence could not survive during the phrase-based forced decoding process.

In the CYK decoding process, the words in segmented source sentence are treated as the basic unit, referred to as cell. For each cell that spans from i to j on the source side, the derivations in cell (i, j) was generated by merging derivations from any two neighbouring sub-cells. For each cell (i, j), k is defined as i < k < j. There would be two sub-cells: cell(i, k) and cell (k, j), we can combine the two cells by the straight and inverted rules, and the application of two rules will generate new translation hypotheses, then we drop the derivations which are not yield the target sentence. When the whole source sentence is covered, the decoding process is finished, we can trace back the path of the derivation to learn the details of how to derive the target sentence (translation reference).



Block1	Block2	Type
经济 economy	的 The	INVERTED
经济的 The economy	发展 development	STRAIGNT
经济 的 发展 The economy evelopment	中国 in China	INVERTED

Reordering samples base on word alignments



Block1	Block2	Type
经济 的 The economy	发展 development	STRAIGHT
中国 in China	经济 的 发展 The economy development	INVERTED

Reordering samples base on forced decoding

Figure 1. Alignment matrix and parts of the reordering samples base on word alignments

To alleviate this problem, this paper adopts a phrase-based forced decoding approach to learning reordering samples from derivation tree (or forest) of each bilingual sentence pair, as shown in the right side of Figure 1. The phrase-based forced decoding technique is different from the typical phrase-based decoding method, in which the derivation of each translation hypothesis must yield the same target sentence during the phrase-based decoding process. In other words, a derivation hypothesis different from the given target sentence could not survive during the phrase-based forced decoding process. In the CYK decoding process, the words in segmented source sentence are treated as the basic unit, referred to as cell. For each cell that spans from i to j on the source side, the derivations in cell (i, j) was generated by merging derivations from any two neighbouring sub-cells. For each cell (i, j), k is defined as i < k < j. There would be two sub-cells: cell(i, k) and cell (k, j), we can combine the two cells by the straight and inverted rules, and the application of two rules will generate new translation hypotheses, then we drop the derivations which are not yield the target sentence. When the whole source sentence is covered, the decoding process is finished, we can trace back the path of the derivation to learn the details of how to derive the target sentence (translation reference).

The process of a source sentence which is decoded successfully by forced decoding will form a tree structure, referred to as the derivation tree, as shown Figure 1. The phrase in the node which has two chilidren nodes can be composed by the combination of two phrases in its children nodes. The algorithm of learning reordering samples based on forced decoding is summarized into six steps as follows:

1. Extract translation rules needed for a specific phrase-based SMT paradigm M from bilingual training corpus C;

2. Perform minimum error rate training (MERT) on a development data set to obtain a set of optimized feature weights;

3. For each {s,t}∈C, translate s into accurate t based on M with translation rules learned in step 1 and feature weights optimized in step 2;

4. For each {s,t}∈C, save the derivation forest produced in step 3 as TreeSet.

5. For each derivation tree belongs to {s,t}, Traversing Treei produced in step4 and extracting the reordering samples from the combination of two phrases in children node.

6. Combine the reordering samples belongs to each sentence pair {s,t}, and remove the duplicate reordering samples.

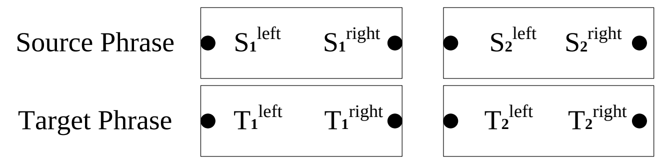
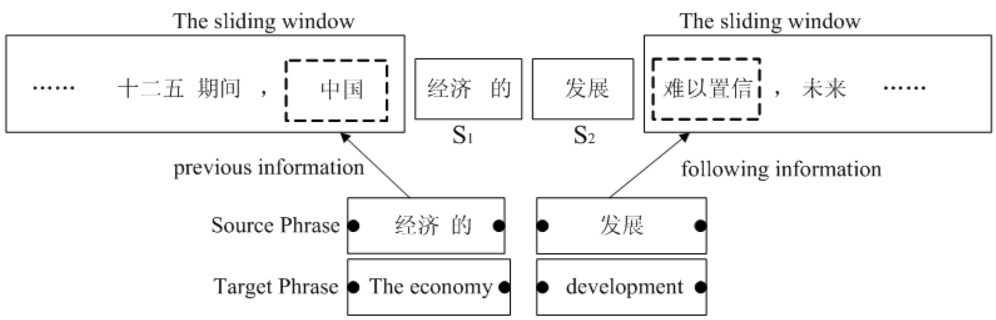


Figure 2. Boundary words (black dots) in the two neighboring phrases



Feature	S_1.left	S_1.right	S_2.left	S_2.right	T_1.left	T_1.right	T_2.left	T_2. right	S_0.right	S_3.left
Feature value	经济	的	发展	发展	The	economy	development	development	中国	难以置信

Figure 3. Boundary features (the solid frame) and contextual features (the dotted frame) for the classifier when setting the sliding window K = 1

Take the derivation tree shown in Figure 1 as example, the node with phrase "中国 经济 的 发展—The economy development in China" can be generated by the combination of "中国—in China" and "经济 的 发展—The economy development" in the inverse order, and we can learn the reordering samples from this combination. In other words, the forced decoding based method learns the reordering samples from the combination of the two phrases, which represents the details of how to generate the derivation. Therefore the quality of reordering samples is much higher than that of traditional methods. In fact, there may be multiple ways to decode a source sentence to target reference by forced decoding technique. In other words, there are several ways to derive the generating tree, referred to as the derivation forest.

Figure 1 shows a derivation tree of the generating forest and parts of reordering samples extracted from the generating forest. From the reordering samples extracted by two methods, we discover that the incorrect reordering samples extracted base on word alignments is discarded in this forced decoding based approach.

B. Classification Features

In traditional classification-based reordering model, the maximum entropy classifier generally considers phrase boundary words of reordering examples as features. It can be illustrated as shown in Figure 2. Sleft imeans the most left word in source phrase Si and Sright imeans the most right word in source phrase Si; Tleft imeans the most left word in target phrase Ti and Tright imeans the most right word in Ti. Figure 2 shows the eight boundary words (bold dots) of two consecutive phrases {S1, S2} and their corresponding target phrases {T1, T2}. Boundary words of the source phrases and target phrases are selected as eight features to build the classifier. As shown in Figure 3, these eight features are listed within solid frames. Since a target phrase T2 only contains one word "development", two boundary word features (T2.left and T2.right) of T2 are the same "development". In other words, the left-most word is the same as the right-most word, and the rule is also applied to source phrases.

In traditional method, only boundary information (i.e., in the form of eight features) is considered. In our opinion, the source-side context of both two consecutive phrase pairs in the source sentence can also provide more valuable information for reordering. Therefore, in our approach, the contextual information in source sentence is considered to predict the order of two consecutive phrases. Along this line of thinking, we can choose the contextual of the source phrases as additional features. First, the sliding window K is defined as the phrase number that we extend in two directions from the current phrase in source sentence, theoretically the max value of K is can be set to the distance from the beginning of the source sentence to current phrase position.

In fact, the bigger value of K is set1, the sparseness problem of data is more serious, especially for the maximum entropy classifier. In this paper, for illustrating simplicity without loss of generality, we set K = 1, therefore, the source word before the first phrase and the source word after the second phrase are adopted to be two additional features, i.e., S0.right and S3.left in Figure 3. Notice that if the first phrase is at the beginning of the source sentence, the S0.right feature will be set to “<s>”, and if the second phrase is at the end of the source sentence, the S3.left feature will be set to “</s>”. Compared to traditional eight features used in traditional methods, two additional features S0.right and S3.left are used by our method, as shown in Figure 3 with dotted frames. S0.right represents the right-most word of the phrases which is previous source phrase S1, and S3.left represents the left-most word of the phrases which is after the source phrase S2.

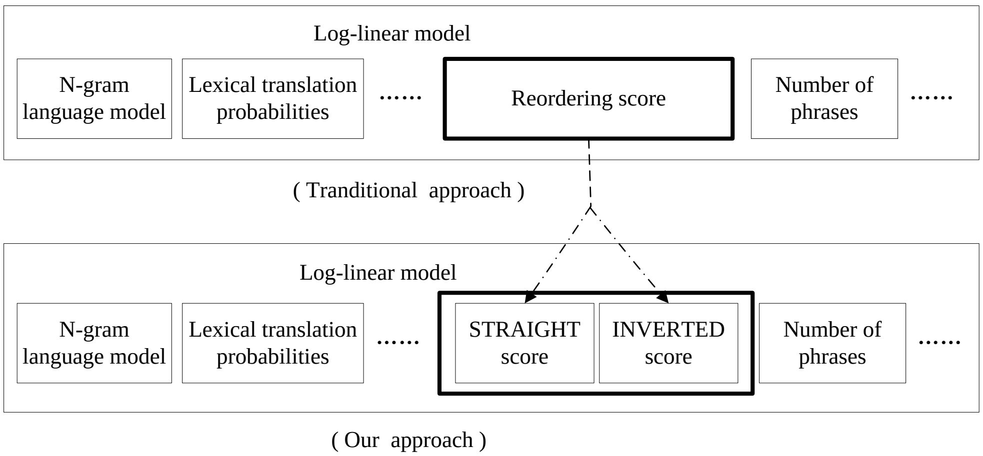


Figure 4. Reordering feature functions for decoding in two approachs

C. *Reordering Feature Functions for Decoding*

In statistic machine translation, all sub-models are trained separately and combined under the assumption that they are independent of each other in the log-linear model, the associated weightsλcan be tuned using minimum error rate training (MERT) (Och 2003). Base on the reordering samples and classification features, we can train a MaxEnt classifier to get the feature weights defined in section 3, and the reordering score is caculated by formular (3).

As we know, in traditional approach, the reordering is a sub-model which is in log-linear model, and the reordering score is used as a feature function. However, one feature function cannot indicate two phrase orientations. Therefore we define two feature functions to indicate two orientations. In this approach, we treat the reordering scores as two feature functions, STRAIGHT and INVERTED respectively.

The motivation behind this method is very simple: we want to depict the reordering model accurately in more dimensions to improve the discrimitive ability of the model. Taking the sentence mentioned in section 4.2 as an example, while the combination of two phrases which are “经济的” and “发展”, the order of the consecutive phrases is predicted by the (maximum entropy) ME model to be STRAIGTH, then the reordering score is added to the STRAGTH score; The order of the combination “中国” and “经济 的 发展” is INVERTED, then the reordering score is added to the INVERTED score. The decoding algrithim repeats this operation to caculate STRAIGHT score and INVERTED score. After the whole source sentence is decoded, there are two reordering scores such as STRANGIT score and INVERTED score, they are integrated into the log-linear model and treated as two feature functions. The details can be illustrated by Figure 4.

V. EVALUATION

In this section, we compare the typical and our proposed methods within a phrase-based SMT system by experiments on the NIST Chinese to English translation tasks and Chinese to Japanses translation tasks.

A. *Settings*

The open source NiuTrans system (Xiao et al., 2012) was selected to build the baseline system. Our training corpus consists of 2 million sentences pairs in Chinese-English (Ch-En) task and 1.8 million sentences in Chinese-to-Japanese (Ch-Ja) task. Development data in Ch-En task is the NIST evaluation sets of mt04, and test data is the NIST evaluation sets of mt05 and mt06 2000 sentences are selected as development data and another 1200 sentences are selected as test data in Ch-Ja task. The base feature set used for all systems is similar to that used in (Marcu et al. 2006), including 14 base features in total such as 5-gram langusage model, bidirectional lexical and phrase-based translation probabilities. All features were combined log-linearly and their weights were estimated by performing minimum error rate training (MERT) (Och 2003).

B. Result

Observed from table 1, method 1, method 2 and method 3 show better performances than baseline with the increase of BLEU points on development set and test set in both Ch-En task and Ch-Ja task. Method 4 which integrates three methods synchronously shows significant improvements than baseline.

The improvements in method1 could be illustrated as follows. In traditional method, the reordering samples were learned base on word alignments, in other words, and it only consideres word alignment in current bilingual sentence pair, so it’s sensitive to the words alignment mistakes. Method 1 can alleviate this problem, it learns reordering samples from derivations of each bilingual sentence pair, and the derivations repesent the detais of how to generate the translation refference, therefore the quality of reordering samples is much higher than that of traditional method.

TABLE I. IBM-BLEU4 (%) SCORE OF OUR METHOD ON DEVELOPMENT SET AND TWO TEST SETS ON TWO TASKS, * INDICATES SIGNIFICANTLY BETTER ON TEST PERFORMANCE AT THE P=0.05 LEVEL, COMPARE TO THE BASELINE METHOD.

		Ch-En		Ch-Ja	
Method	Description	Dev	Test	Dev	Test
Baseline	Baseline	39.83	33.27	30.11	25.40
Method1	Learning samples base on forced decoding	40.11 (+0.28)	33.58 (+0.31)	30.43 (+0.32)	25.67 (+0.27)
Method2	Boundary features and contextual features	39.94 (+0.11)	33.36 (+0.09)	30.24 (+0.13)	25.52 (+0.12)
Method3	STRAIGHT score and INVERTED score	40.03 (+0.2)	33.20 (+0.13)	30.40 (+0.29)	25.63 (+0.23)
Method4	using method1, method2 and method3 synchronously	40.34* (+0.53)	33.79* (+0.51)	30.74* (+0.63)	25.97* (+0.57)

TABLE II. THE COMPARATION ON THE NUMBER OF REORDERING SAMPLES EXTRACED BY TWO METHODS

Method	Number of STRAIGHT samples	Number of INVERTED samples	STRAIGHT / INVERTED
Base on word alignments (WA)	14.78 million	1.7 million	8.7 : 1
Base on forced decoding (FD)	10.58 million	2.46 million	4.3 : 1

In another view, this method considers the whole phrase table and chooses the phrase with the maximum model score when generating the translation hypothesis, so the word alignment mistakes in current sentence pair affect the training samples little in some extent. Comparing with baseline, method 2 considers both the information of the bilingual (source and target) phrases and the context of the two phrases in the source sentence, therefore the classifier could capture more contextual information and enhance the reordering prediction ability. Method 3 utlizes two feature functions to indicate the orientation during decoding, and show better performance than baseline.

In our approach, the reordering samples are extracted base on forced decoding, therefore the success rate of decoding influences the number of reordering samples. Table 2 lists the number of reordering samples by different method in Chinese to English bilingual sentence pairs (taking 1 million sentences as an example). In our experiments, when the beam size is set to 60, 24% of the bilingual sentence pairs fail to be decoded in the process of the forced decoding. In this case, for these failed sentence pairs, we adopt the results of traditional method for them. As shown in Table2, the number of samples base on WA is larger than that of FD, the ratio of STRAIGHT and INVERTED number reaches 8.7:1 and the ration of STRAIGHT and INVERTED numbers base on FD is 4.3:1, which is more preferable to the classifier, the distribution of the reordering samples is better than that of traditional method. The number of FD INVERTED reordering samples is larger than that of WA, the reason is that reordering samples are extracted from multiple derivation trees in a sentence pair.

To show the influence of our approach on translation compared with baseline, we present some examples which are listed in Table 3. Obviously, the translation result by our approach is better than that of baseline. In fact, the reordering model in this work influences the translation results which can be shown in two conditions. Firstly, during the decoding process, the reordering model in this work influences the selection of translation hypotheses and what we can see is the better translation result than baseline; Secondly, comparing with baseline, this model optimizes the phrases order in translation hypotheses and uses the same translation hypothesis with baseline, but better translation result shown for us.

VI. CONCLUSION

This paper presents three optimization techniques to improve classification-based reordering methods for PBMT, involving reordering sample generation for classifier training, feature selection for classification and reordering feature functions for decoding. To our best knowledge, we are the first to apply forced decoding technique to generate training samples on reordering and treat the reordering score as two feature functions into log-linear model. Experimental results show that the work in this paper improves the baseline system significantly. In future work, we can make research on extending the sliding window defined in this paper to capture more contextual information and utilize other models to improve the reordering for PBMT.

TABLE III. THREE EXAMLES OF TEST BY TRADITIONAL METHOD AND OUR APPROACH, PHRASE WHICH REPRESENTS THE VARIABLE POSITION IN DIFFERENT POSITION IS MARKED IN BOLD

Case 1	Chinese	同时 , 将 海军 新 装备 武器 试验 与 部队 科技 练兵 相结合 , 缩短 了 海军 新 装备 形成 战斗力 的 时间
	Baseline	Meanwhile, naval weapons testing of new equipment with the combination of the science and technology training of troops, to shorten the time new equipment to form combat the navy
	Our approach	Meanwhile, the combination of naval weapons testing of new equipment and the science and technology training of troops, to shorten the time new equipment to form combat the navy
Case 2	Chinese	组委会 和 国际 联盟 在 同一天 作出 了 对 她 禁赛 两年 的 处罚 决定
	Baseline	The organizing committee and the international union banned for two years a decision on the penalty made on the same day to her
	Our approach	The organizing committee and the international union banned for two years a decision on the penalty made to her on the same day
Case 3	Chinese	几天前，孩子模仿电视自杀了。
	Baseline	数日前、子供は模倣テレビが自殺です。
	Our approach	数日前、子供はテレビを模倣して自殺します。
Case 4	Chinese	水产厅资源管理部的负责人就该海域的情况进行了说明。
	Baseline	水産庁資源管理部の責任者が海域の状況を説明し、説明を行った。
	Our aproach	水産庁資源管理部の担当者は、当該海域の状況をこう説明する。

REFERENCES

[1] Li, P., Liu, Y., Sun, M., Izuha, T., & Zhang, D. (2014). A Neural Reordering Model for Phrase-based Translation. In COLING (pp. 1897-1907).

[2] Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics, 19(2):263--311, June.

[3] Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human

[4] Language Technology - Volume 1, NAACL'03, pages 48--54, Stroudsburg, PA, USA. Association for Computational Linguistics.

[5] Koehn, P., Axelrod, A., Mayne, R. B., Callison-burch, C., Osborne, M., and Talbot, D. (2005). Edinburgh system description for the 2005 iwslt speech translation evaluation. In In Proc. International Workshop on Spoken Language Translation (IWSLT).

[6] C. Tillmann. 2004. A unigram orientation model for statistical machine translation. In Proceedings of HLT-NAACL: Short Papers, pages 101--104.

[7] Nagata, M., Saito, K., Yamamoto, K., & Ohashi, K. (2006, July). A clustered global phrase reordering model for statistical machine translation. InProceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the

Association for Computational Linguistics (pp. 713-720). Association for Computational Linguistics.

[8] D. Xiong, Q. Liu, and S. Lin. 2006. Maximum entropy based phrase reordering model for statistical machine translation. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 521--528, Sydney.

[9] B. Xiang, N. Ge, and A. Ittycheriah. 2011. Improving reordering for statistical machine translation with smoothed priors and syntactic features. In Proceedings of SSST-5, FifthWorkshop on Syntax, emantics and Structure in Statistical Translation, pages 61--69, Portland, Oregon, USA. Association for Computational Linguistics.

[10] R. Zens and H. Ney. 2006. Discriminative reordering models for statistical machine translation. In Proceedings on theWorkshop on Statistical Machine Translation, pages 55--63, New York City, June. Association for Computational Linguistics.

[11] V. Nguyen, A. Shimazu, M. Nguyen, and T. Nguyen. 2009. Improving a lexicalized hierarchical reordering model using maximum entropy. In Proceedings of the Twelfth Machine Translation Summit (MT Summit XII). International Association for Machine Translation.

[12] Zens R, Ney H, Watanabe T, et al. Reordering constraints for phrase-based statistical machine translation[C]//Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics, 2004: 205.

[13] Alrajehab, Abdullah, and Mahesan Niranjanb. 2014.Bayesian Reordering Model with Feature Selection. ACL 2014, 477.

[14] Nieben, S., & Ney, H. (2001). Lehrstuhl Fur Informatik Vi. Morphosyntactic analysis for reordering in statistical machine translation.

[15] Collins, M., Koehn, P., Ku\v{c}erov\'{a}, I. (2005, June). Clause restructuring for statistical machine translation. In Proceedings of the 43rd annual meeting on association for computational linguistics (pp. 531-540). Association for Computational Linguistics.

[16] Zang, S., Zhao, H., Wu, C., & Wang, R. (2015, August). A novel word reordering method for statistical machine translation. In Fuzzy Systems and Knowledge Discovery (FSKD), 2015 12th International Conference on (pp. 843--848). IEEE.

[17] Isozaki, H., Sudoh, K., Tsukada, H., & Duh, K. (2010, July). Head finalization: A simple reordering rule for sov languages. In Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR(pp. 244-251). Association for Computational Linguistics.

[18] Galley, M., & Manning, C. D. (2008, October). A simple and effective hierarchical phrase reordering model. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 848-856). Association for Computational Linguistics.

[19] Wu, D. (1996, June). A polynomial-time algorithm for statistical machine translation. In Proceedings of the 34th annual meeting on Association for Computational Linguistics (pp. 152--158). Association for Computational Linguistics.

[20] Och, F. J. 2003. Minimum error rate training in statistical machine translation. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL), Sapporo, Japan, pp. 160-67

[21] Xiao T, Zhu J, Zhang H, et al. NiuTrans: An open source toolkit for phrase-based and syntax-based machine translation[C] Proceedings of the ACL 2012 System Demonstrations. Association for Computational Linguistics, 2012: 19-24