Lang Resources & Evaluation
DOI 10.1007/s10579-013-9217-4

ORIGINAL PAPER

A unified approach for effectively integrating source-side syntactic reordering rules into phrase-based translation

Jiajun Zhang · Chengqing Zong

Abstract Phrase-based translation models, with sequences of words (phrases) as translation units, achieve state-of-the-art translation performance. However, phrase reordering is a major challenge for this model. Recently, researchers have focused on utilizing syntax to improve phrase reordering. In adding syntactic knowledge into phrase reordering model, using handcrafted or probabilistic syntactic rules to reorder the source-language approximating the target-language word order has been successful in improving translation quality. However, it suffers from propagating the pre-ordering errors to the later translation step (e.g. decoding). In this paper, we propose a novel framework to uniformly represent the handcrafted and probabilistic syntactic rules and integrate them more effectively into phrase-based translation. In the translation phase, for a source sentence to be translated, handcrafted or probabilistic syntactic rules are first acquired from the source parse tree prior to translation, and then instead of reordering the source sentence directly, we input these rules into the decoder and design a new algorithm to apply these rules during decoding. In order to attach more importance to the syntactic rules and distinguish reordering between syntactic and non-syntactic unit reordering, we propose to design respectively a syntactic reordering model and a non-syntactic reordering model. The syntactic rules will guide phrase reordering in decoding within the syntactic reordering model. Extensive experiments on Chinese-to-English translation show that our approach, whether incorporating handcrafted or probabilistic syntactic rules, significantly outperforms the previous methods.

J. Zhang (✉) · C. Zong
National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing 100190, China
e-mail: jjzhang@nlpr.ia.ac.cn

C. Zong
e-mail: cqzong@nlpr.ia.ac.cn

Published online: 10 February 2013

Keywords Handcrafted syntactic rules · Probabilistic syntactic rules · Effective integration · Phrase-based translation

1 Introduction

Given a source sentence f, statistical machine translation (SMT) searches through all the target sentences e and finds the one with the highest probability:

$$e' = \arg\max_{e} P(e|f) \tag{1}$$

Brown et al. (1990, 1993) proposed first a word-based SMT model which treats each single word as a translation unit. Over the last decade, SMT models that adopt phrases[1] as translation units have dominated the research area. This model is the well-known phrase-based translation model (Koehn et al. 2003, 2007; Och and Ney 2004), which can be formulated as follows:

$$\begin{aligned} e' &= \arg\max_{e} P(e|f) \\ &= \arg\max_{e,f_1^K} P\left(e, f_1^K | f\right) \\ &= \arg\max_{e,e_1^K f_1^K} P\left(f_1^K | f\right) \times P\left(e_1^K | f_1^K, f\right) \times P\left(e | e_1^K, f_1^K, f\right) \end{aligned} \tag{2}$$

In phrase-based SMT, $P\left(f_1^K | f\right)$ first divides the source sentence into K phrases, then $P\left(e_1^K | f_1^K, f\right)$ translates K source phrases into K target phrases, and finally $P\left(e | e_1^K, f_1^K, f\right)$ permutes the K target phrases yielding the target translation. Usually, a language model $P(e)$ is also employed to measure the grammaticality of the target translation. Conventionally, phrase-based SMT assumes that $P\left(f_1^K | f\right)$ follows a uniform distribution. Thus, we mainly focus on the translation model $P\left(e_1^K | f_1^K, f\right)$ and the phrase reordering model $P\left(e | e_1^K, f_1^K, f\right)$. Compared with the translation model, phrase reordering is a bigger challenge. The recent years have witnessed great progress of the phrase reordering model: from the distortion model (Koehn et al. 2003), to the constraint model (Zens et al. 2004), and then to the lexicalized phrase reordering model (Tillman and Zhang 2005; Xiong et al. 2006; Koehn et al. 2007). However, these models are usually criticized for their lack of both deep syntactic knowledge and the ability to handle long-distance phrase reordering. Therefore, more and more researchers concentrate on syntactic approaches to improve phrase reordering model.

In adding syntax to improve phrase reordering in phrase-based translation, many research works (Collins et al. (2005); Costa-jussà et al. (2007); Wang et al. (2007); Zhang et al. (2007); Li et al. (2007); Xiong et al. (2008); Badr et al. (2009); Xu et al. (2009); Crego and Yvon (2010); Lee et al. (2010); Visweswariah et al. (2010); Genzel (2010); Du and Way (2010); Wu et al. (2011) and Andreas et al. (2011)) have investigated the use of linguistic knowledge and have empirically proven that

[1] In SMT, *phrase* just denotes a sequence of words rather than a syntactic constituent. When we need to represent a syntactic constituent, we use the term "syntactic phrase".

syntactic rules are very helpful to improve phrase reordering. For example, in Chinese-to-English translation, the Chinese prepositional phrase (PP) preceding the verb phrase (VP) **PP-VP** is translated into English **VP-PP** in most cases. Thus, if a special rule is designed to deal with this case, the translation result can be better.

The popular way of integrating the linguistic information into phrase reordering is to reorder the source sentences with syntactic reordering rules so as to make the input much closer to the target language in word order. Collins et al. (2005); Wang et al. (2007); Badr et al. (2009); Xu et al. (2009) and Lee et al. (2010) used **handcrafted syntactic rules** obtained from source parse trees to directly reorder the input sentences. Li et al. (2007); Elming (2008); and Khalilov and Sima'an (2010, 2011) employed **probabilistic syntactic rules** to get a reordered source sentence, an *n*-best reordered sentence list, or a reordered word lattice for decoding. And Costa-jussà et al. (2007) utilized the statistical machine reordering technique to convert a source sentence into a weighted reordering graph which is adopted as the input of the decoder. The former method using handcrafted syntactic rules depends much on both the author's professional knowledge in linguistics and the performance in parsing technology. The latter approach is more robust to the errors in parsing stage; however it increases the burden of decoding as it has to translate an *n*-best sentence list or large word lattices resulting from merging phrases with different reorderings. Furthermore, it might still produce pre-ordering errors prior to translation because the *n*-best list includes only parts, but not all, of the reordering hypotheses. Even though the word lattice can accommodate the entire syntactic reordering hypothesis, it is hard to exhaustively search the best reordering in the huge reordering space. From the methodological point of view, it should be noted that both methods directly deal with the parse trees to get reordered source sentences. It is pointed out in previous work (Habash 2007) that syntactic pre-reordering does not improve translation if the parse quality is not good enough. It becomes a challenge to use the handcrafted and probabilistic syntactic rules properly and adequately even if the parse quality is not very promising (taking Chinese parsers as an example, the parse accuracy is around 80 % (Levy and Manning 2003)).

We have to admit that, although the syntactic reordering rules contain much noise, many researchers empirically prove that systems incorporating syntactic reordering rules significantly outperform those systems applying distortion-based or lexicalized phrase reordering models. However, due to the parsing errors and the discrepancy between the translation units and the syntactic reordering rules, reordering the source sentences prior to translation may cause many pre-ordering errors. Taking the following sentence as an example (in which (a) gives a Chinese sentence with English translation under each word (b) shows the correct English translation):

(a)	以	巴	和平	成为	这次	会议	的	主题
	israeli	palestinian	peace	become	this	meeting	of	theme
(b)	israeli-palestinian peace becomes the theme of this meeting							
(c)	becomes the theme of this meeting israeli-palestinian peace							

A Chinese parser might mistakenly parse the noun phrase (NP) "$NP(以_{NN}\,(israeli)\,巴_{NN}\,(palestinian)\,和平_{NN}\,(peace))$" in the Chinese sentence into a

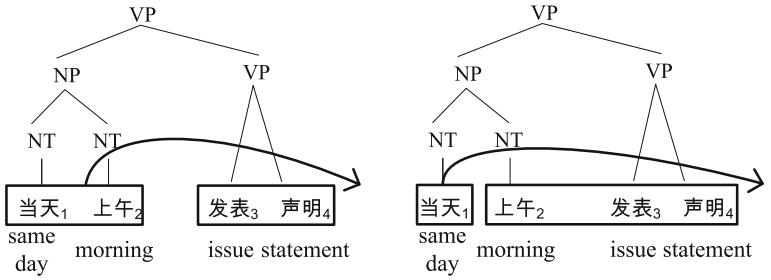

Fig. 1 An example to show which reordering belongs to syntactic phrase reordering and which one is non-syntactic reordering: reordering between spans (*1*, *2*) and (*3*, *4*) on the left is syntactic reordering since each span corresponds to a syntactic phrase; the reordering between spans (*1*, *1*) and (*2*, *4*) on the right belongs to non-syntactic reordering because the phrase corresponding to span (*2*, *4*) is not a syntactic phrase

prepositional phrase (PP) "$PP\left(\text{以}_{P}\left(israeli\right)\text{巴}_{NN}\left(palestinian\right)\text{和平}_{NN}\left(peace\right)\right)$"; and the handcrafted rules[2] will incorrectly reposition this mistakenly recognized prepositional phrase behind its right sibling verb phrase. Thus, this wrong pre-ordering may lead to translation errors and (c) shows the possible bad translation.

According to the analysis above, we know that the source-side syntactic reordering rules are helpful to improve translation quality, but the preprocessing approaches do not take full advantage of the syntactic rules since they reorder the source sentence arbitrarily. Intuitively, the syntactic rules can contribute more if they are applied to help translation inference during decoding together with other information. So it comes to our motivation: instead of using these syntactic rules (handcrafted or probabilistic) to reorder the source sentences arbitrarily, we aim to make full use of the syntactic rules in the decoding stage. To achieve this purpose, we firstly design a unified representation for both handcrafted and probabilistic syntactic reordering rules. Then, we propose two orthogonal reordering models: syntactic reordering model handling only reordering between syntactic phrases and non-syntactic reordering model dealing with other cases (Fig. 1 shows an example of syntactic phrase reordering and non-syntactic phrase reordering). Since it is intuitive that reordering between syntactic phrases is more important than reordering between non-syntactic ones, this design can attach more importance to syntactic reordering model and can facilitate the integration of syntactic reordering rules as well. We tune the syntactic reordering model and the non-syntactic reordering model respectively, and design an approach for enabling the syntactic reordering rules to guide phrase reordering during decoding within syntactic reordering model. Furthermore, we create a feature to reward the syntactic reordering during decoding. As will be shown in our experiments, by utilizing syntactic rules in the decoding stage, we can not only use the correct syntactic rules adequately but also alleviate the pain caused by incorrect syntactic reordering rules with other important model features, such as phrase translation probabilities and the target language model. Moreover, our

[2] The handcrafted rule for this case looks like $NP(DNP(PP)\diamondsuit NP) \rightarrow NP(\diamondsuit NP\ DNP(PP))$ and will be detailed in Sect. 3.1.

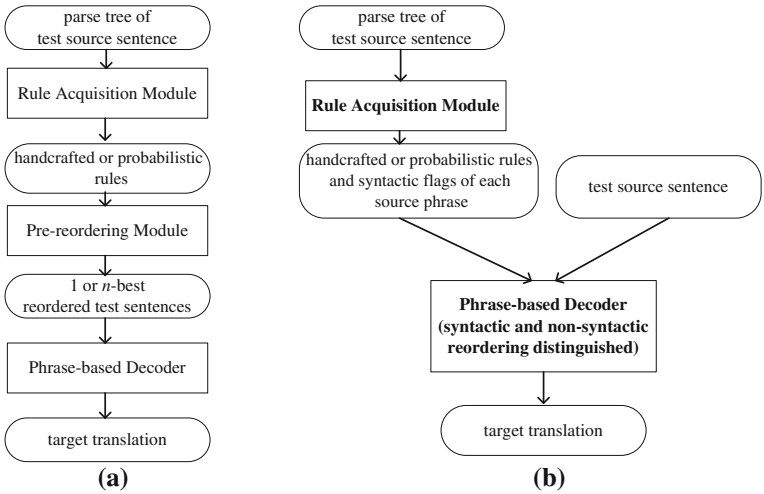

Fig. 2 **a** The translation flowchart of previous pre-ordering methods. **b** Illustrates our translation framework of incorporating handcrafted or probabilistic rules into the decoding stage. We will detail respectively the two key parts which are in *boldface* in Sects. 3 and 4

approach does not increase the time complexity of decoding compared with the baseline.

For a source sentence to be translated, our translation framework can be illustrated in Fig. 2b. Figure 2a corresponds to the previous pre-ordering approaches.

In order to verify the effectiveness of our usage of source-side syntactic rules, we have developed two systems: one integrating the handcrafted syntactic rules, and the other incorporating the probabilistic syntactic rules. Compared with the systems using the previous pre-ordering methods, extensive experiments show that our proposed approach (regardless of handling handcrafted rules or probabilistic rules) performs much better.

The rest of this paper is organized as follows: Sect. 2 introduces related work. Section 3 describes the acquisition and unified representation of source-side syntactic reordering rules. Section 4 elaborates the baseline and our adapted translation model, and details the integration algorithm of syntactic reordering rules into the decoding module. In Sect. 5, we report the experiments on middle-scale data and large-scale data, and give a comprehensive analysis as well. Section 6 concludes the paper and discusses the future work.

2 Related work

Adding syntax into phrase reordering model has become a hot topic in the recent years. Chiang (2007)'s hierarchical phrase-based translation system utilized the formal syntax (Synchronous Context Free Grammars, SCFG) to model phrase reordering. However, incorporating the linguistically syntactic information of source language to improve phrase reordering has drawn more and more attention.

Collins et al. (2005) described six types of transformation rules to reorder the German clauses to better match the English word order in German-to-English translation. Wang et al. (2007) analyzed the systematic difference between Chinese and English, and then proposed specific reordering rules for three categories of Chinese phrases: verb phrases, noun phrases, and localizer phrases. Badr et al. (2009) addressed two syntactic structures (Subject-Verb structure and noun phrase structure) and exploited well-defined reordering rules for English-to-Arabic translation. Xu et al. (2009) and Lee et al. (2010) designed syntactic reordering rules for English-to-Japanese translation using dependency structure and constituent structure respectively. They all showed that translation quality can be improved significantly if syntactic reordering rules are adopted to reorder the source sentences prior to translation. However, all the rules in the above methods are handcrafted and they often cause many pre-ordering errors (Wang et al. 2007). In order to improve the robustness, Li et al. (2007) used the weighted reordered *n*-best source sentences as input for the decoder. They utilized the probabilistic rules based on source parse trees in Chinese-to-English translation to determine whether the children of a node should be reordered or not, and finally to obtain a reordered *n*-best list used as input for the decoder. Nevertheless, all these methods are separated from the decoder and reorder the source sentences arbitrarily prior to translation. Once a pre-ordering error happens, it is very difficult to undo this mistake in later translation steps. In our approach, we just retain the syntactic rules instead of using them to reorder the source sentences directly. During decoding, the syntactic rules will serve as a strong informative feature to guide and enhance the phrase reordering within the syntactic reordering model.

Zhang et al. (2007) only allowed reordering between syntactic phrases and enforced the non-syntactic phrases to be translated in straight order. Xiong et al. (2008) proposed a linguistically annotated bracketing transduction grammar model (BTG) for SMT. This method uses some heuristic rules to linguistically annotate each source phrase with the source-side parse tree in decoding and builds a linguistic reordering model besides a conventional reordering model. Xiang et al. (2011) just employed one reordering model but enriched each phrase for reordering with multiple syntactic features. Crego and Yvon (2010) proposed a linguistically-informed bilingual n-gram language model to tackle mid-range reorder problem in SMT. All these approaches acquired and applied the syntactic rules during the decoding stage; however they increased the decoding time to a large extent since they have to compute the syntactic information for each phrase during decoding. Our work differs from the four described above in three ways. First, when translating a test sentence, we obtain the corresponding syntactic rules prior to translation instead of during the decoding stage and thus alleviate the decoding complexity. Second, we distinguish syntactic phrase reordering from non-syntactic phrase reordering because we believe they play different roles in translation. To our best knowledge, this idea is not considered in previous works. Third, we add a feature to reward the syntactic reordering so as to attach more importance to syntactic phrase reordering.

The underlying philosophy of this paper is to use soft syntactic constraints to guide phrase reordering during decoding. The main idea of (Cherry 2008, Marton and Resnik 2008) is under the same philosophy although their focus is on translation

boundary rather than phrase reordering. Cherry (2008) imposed soft syntactic constraints on translation boundaries in phrase-based SMT (Koehn et al. 2007) based on the source-side dependency parse trees and proposed a counting feature to penalize the hypothesis violating the syntactic boundaries. Marton and Resnik (2008) extended the hierarchical phrase-based SMT (Chiang 2007) with a number of counting features which are accumulated whenever the translation hypothesis violates the source-side constituent boundaries. Different from their works, our paper concentrates on designing and applying soft syntactic constraints to help phrase reordering in phrase-based decoding.

3 Acquisition and representation of syntactic rules

Without loss of generality, we use Chinese-to-English translation as a case study in this paper. However, our approach is also suitable for other language pairs. Whether incorporating handcrafted or probabilistic syntactic reordering rules in decoding, acquiring these rules is our first task. We first detail the acquisition methods of the frequently-used handcrafted and probabilistic rules, and then propose a unified representation for the two categories of the rules. Note that we use the Penn Chinese Treebank guidelines (Xue et al. 2005) to represent all the syntactic reordering rules. Table 1 provides a list of Chinese Treebank phrase tags for reference.

3.1 Handcrafted rule acquisition

The handcrafted syntactic rules are not trained by any generative or discriminative model and thus should reflect the true structural difference between the language pair Chinese and English. Wang et al. (2007) described three kinds of handcrafted rules for Chinese-to-English which have proven to be reasonable. Here, we revisit and summarize these specific rules.

- **Verb Phrases** If there is a node in the Chinese parse tree labeled as verb phrase VP, we have three rules to reorder its children.

Table 1 Some Penn Chinese Treebank phrase tags borrowed from Wang et al. (2007)

ADJP	Adjective phrase
CLP	Classifier phrase
CP	Clause headed by complementizer
DNP	Defective noun phrase formed by "XP+DEG"
DP	Determiner phrase
DVP	Phrase formed by "XP+DEV"
IP	Inflectional phrase headed by INFL (I)
LCP	Localizer phrase formed by "XP+LC"
NP	Noun phrase
PP	Prepositional phrase
QP	Quantifier phrase
VP	Verb phrase

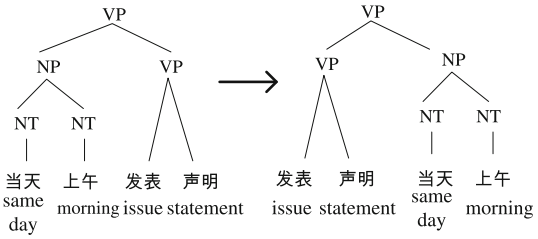

Fig. 3 A handcrafted reordering rule for verb phrase

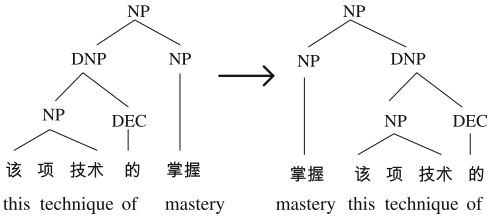

Fig. 4 A handcrafted reordering rule for noun phrase

1. *VP*(*PP*◇*VP*) → *VP*(◇*VP PP*)[3] and *VP*(*LCP*◇*VP* → *VP*(◇*VP LCP*) indicate that either the prepositional phrase PP or the localizer phrase LCP under the parent VP needs to be repositioned after the sibling VP.
2. *VP*(*NP*(*NT*)◇*VP*) → *VP*(◇*VP NP*(*NT*)) means a preverbal noun phrase NP containing at least one temporal noun NT should be repositioned after the sibling VP.
3. *VP*(*QP*◇*VP*) → *VP*(◇*VP QP*) states that the quantifier phrase QP below a parent VP will be repositioned after the sibling VP.

Figure 3 shows an example of a handcrafted reordering rule for the verb phrase.

- **Noun Phrases** When we find a noun phrase NP node in the Chinese parse tree, four rules are considered.
  1. *NP*(*DNP*(*PP*|*LCP*)◇*NP*) → *NP*(◇*NP DNP*(*PP*|*LCP*)) indicates that the defective noun phrase DNP is repositioned after the last sibling NP if the child DNP has a child PP or LCP.
  2. *NP*(*DNP*(!*PN*) ◇*NP*) → *NP*(◇*NP DNP*(!*PN*)) denotes that if a parent NP has a child DNP which in turn has a child NP that is not a pronoun PN, then the DNP should be repositioned after the last sibling NP.
  3. *NP*(*CP*◇*NP*) → *NP*(◇*NP CP*) means that the child complementizer phrase CP will be repositioned after its sibling NP.
  4. *CP*(*IP DEC*) → *CP*(*DEC IP*) says that if the CP in rule (3) is formed by "IP+DEC" (inflectional phrase followed by a particle DEC), we have to swap these two nodes.

  Figure 4 gives an example of handcrafted reordering rule for the noun phrase.

[3] ◇ denotes a placeholder which indicates other syntactic nodes, in this example between PP and VP.

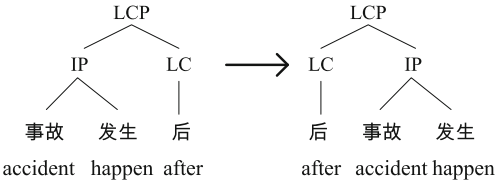

Fig. 5 A handcrafted reordering rule for localizer phrase

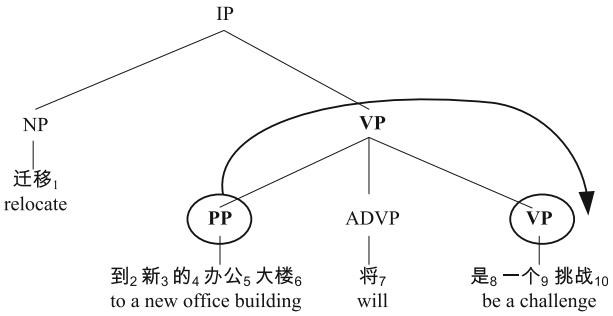

Fig. 6 The simplified Chinese parse tree of the example sentence where the leaves are Chinese words with their indices and corresponding English translation

- **Localizers** We have one rule for the node localizer phrase LCP: *LCP*(◇*LC*) → *LCP*(*LC*◇) denoting the child localizer LC node will be repositioned before its left sibling under a parent LCP node. Figure 5 shows an example of handcrafted reordering rule for localizer phrase.

More details about handcrafted rules can be found in (Wang et al. 2007). Here, we give a real example from our test data. We know that all the possible handcrafted rules belonging to the three categories above can be extracted if the parse tree of the source sentence is given.

The example is shown below and its parse tree is illustrated in Fig. 6. In view of the parse tree obtained, we can apply the handcrafted rule *VP*(*PP*◇*VP*) → *VP*(◇*VP PP*) in which PP is (到$_2$ 新$_3$ 的$_4$ 办公$_5$ 大楼$_6$) and VP corresponds to (是$_8$ 一个$_9$ 挑战$_{10}$).

Chinese: 迁移*(relocate)*到*(to)*新*(new)*的*(of)*办公*(office)*大楼*(building)*将*(will)*是*(be)* 一个*(a)*挑战*(challenge)*

English reference: *relocation to a new office building will be a challenge*

Note that if we apply the pre-ordering approach to reorder the input sentence and reposition the PP after the VP, we obtain the reordered source sentence "*(*迁移*)(*将 是 一个 挑战*)(*到 新 的 办公 大楼*)*" and might obtain the bad translation "*relocation will be a challenge to a new office building*". It is because the Chinese sentence is parsed incorrectly and the first part "迁移*(relocate)*到*(to)*新*(new)*的*(of)*办公*(office)*大楼*(building)*" is not recognized correctly as a clause.

3.2 Probabilistic rule acquisition

For the probabilistic rules, we use an approach similar with Li et al. (2007) to extract rules and learn their probabilities. Li et al. (2007) are only concerned with the nodes with two or three children, and predict a probability for each permutation of the children. We turn to a different strategy. For the nodes with two children, we design a rule to determine if they should be swapped. For the nodes with more than two children, we first search the head node (VP or NP), and if it exists, we design a rule to decide whether any preceding modifier node should be repositioned after the head node. The second rule is based on the phenomenon that the modifiers before VP or NP in Chinese usually appear after VP or NP in English. The two rules can be formalized as:

$$P : N^L \diamondsuit N^R \Rightarrow \begin{cases} N^L \diamondsuit N^R & straight \\ \diamondsuit N^R N^L & inverted \end{cases} \tag{3}$$

in which $\diamondsuit$ is NULL if the parent node P has two children (left node N^L and right node N^R), or is a placeholder denoting other nodes between the modifier node N^L and the head node N^R if P has more than two children. The inverted case in (3) means that the node N^L will be repositioned after the nodes $\diamondsuit$ and N^R.

For the two kinds of probabilistic rules, we adopt a maximum entropy (MaxEnt) model to estimate the probabilities of *straight* and *inverted*. In the training example extraction, our algorithm uses the Chinese parse tree and the word alignment between Chinese and English as input. If the English sides aligned to the two Chinese nodes that we are interested in (N^L and N^R) have an empty intersection, a training example can be extracted.

The rich features we employ for MaxEnt training and prediction include three different levels from shallow to deep:

1. lexicalized evidence: the leftmost/rightmost word of N^L and N^R, and the word immediately before/after the leftmost/rightmost word of N^L/N^R;
2. the part-of-speech evidence: the part-of speech of lexicalized words used in (1);
3. the syntactic tag evidence: the combined phrase tags of N^L, N^R and their parent in the form of $N^L + N^R + P$.

For example in Fig. 6, we can extract a training instance, namely $N^L = PP(\text{到}_2\ \text{新}_3\ \text{的}_4\ \text{办公}_5\ \text{大楼}_6)$ and $N^R = VP(\text{是}_8\ \text{一个}_9\ \text{挑战}_{10})$. The specific features about this rule are listed in Table 2.

Given the parse tree of a test source sentence, we first extract all the probabilistic rules. Meanwhile, we predict their reordering probabilities with the trained MaxEnt

Table 2 The specific features for a rule, "l/r" denotes leftmost/rightmost, "w" means word, "p" indicates part-of-speech, and "b/a" means before/after

lw of N^l	rw of N^l	lp of N^l	rp of N^l	lw of N^r	rw of N^r	lp of N^r	rp of N^r	bw of N^l	aw of N^r	Tag of rule
到	大楼	P	NN	是	挑战	VV	NN	迁移	NULL	PP-VP-VP

model. For the pre-ordering approach, these probabilistic rules are employed to produce an n-best reordered source sentences as the input of the decoder. Alternatively, in our approach, we will design an algorithm to apply these rules to guide phrase reordering in the decoding stage.

3.3 Unified representation for handcrafted and probabilistic rules

Let us first review the forms of the handcrafted and probabilistic syntactic rules. The handcrafted syntactic rules have forms like $VP(PP \diamondsuit VP) \rightarrow VP(\diamondsuit VP\ PP)$, $CP(IP\ DEC) \rightarrow CP(DEC\ IP)$, and $LCP(\diamondsuit LC) \rightarrow LCP(LC \diamondsuit)$. It should be noted that $\diamondsuit$ in the last rule cannot be NULL and we regard it as a special node. Therefore, all the handcrafted rules are binary relations between two nodes. The same relationship holds in the probabilistic syntactic rules in the forms $\langle N^L \diamondsuit N^R \rightarrow N^L \diamondsuit N^R, P(s)\rangle$ and $\langle N^L \diamondsuit N^R \rightarrow \diamondsuit N^R\ N^L, P(i)\rangle$ where $P(s)$ and $P(i)$ denote probabilities of straight and inverted respectively. It is obvious and easy to change the handcrafted rule into an equivalent probabilistic format. For example, $VP(PP \diamondsuit VP) \rightarrow VP(\diamondsuit VP\ PP)$ is equivalent to $\langle PP \diamondsuit VP \rightarrow \diamondsuit VP\ PP, 1.0\rangle$. Thus, we can see that the handcrafted rules are a special case of probabilistic rules, and the only difference lies in that handcrafted rules only has the inverted format.

For the sake of convenience, hereafter, we consider only the generalized rule formats $\langle N^L \diamondsuit N^R \rightarrow N^L \diamondsuit N^R, P(s)\rangle$ and $\langle N^L \diamondsuit N^R \rightarrow \diamondsuit N^R N^L, P(i)\rangle$. Since $P(s) + P(i) = 1.0$, we can just use one format to denote these two formats. The unique format is $\langle N^L, N^R, P(i)\rangle$ which means the left node N^L will be repositioned after the right node N^R with the probability $P(i)$. $P(i) = 1.0$ if it is a handcrafted rule, otherwise $P(i)$ is estimated by MaxEnt model. As the unit of phrase-based translation is any word sequence (phrase) but not a parse tree node, we need to make a conversion from tree nodes to source phrases in order to incorporate the syntactic rules. Since each tree node in the test parse tree can be projected to a span on the source sentence, we can easily use spans to denote the tree nodes. Finally, each syntactic rule can be denoted as a triple $\langle span(N^L), span(N^R), P(i)\rangle$ which is a unified representation for handcrafted and probabilistic rules for test source sentences.

To have a better intuition, we use the unified format to represent the handcrafted and probabilistic rules in Fig. 6. Like the Sect. 3.1 illustrates, the handcrafted rule is in the form of $VP(PP \diamondsuit VP) \rightarrow VP(\diamondsuit VP\ PP)$ with PP $(到_2\ 新_3\ 的_4\ 办公_5\ 大楼_6)$ and $VP(是_8\ 一个_9\ 挑战_{10})$. Thus, the unified format is $\langle(2, 6), (8, 10), 1.0\rangle$. Similarly, the probabilistic rule is $\langle(2, 6), (8, 10), 0.6826\rangle$ where 0.6826 is the probability of the inverted case predicted by our trained MaxEnt model.

4 Integrating syntactic reordering rules in decoding

In phrase-based SMT, the system adapting bracketing transduction grammars (BTG) to phrasal translation obtains the state-of-the-art translation performance (Wu 1997; Xiong et al. 2006, 2011; Zhang et al. 2009). This BTG-based model translates a sentence through dynamically handling each source-side span by

merging any two sub-spans, and meanwhile the syntactic reordering rules are about reordering between source-side spans. Therefore, it is very convenient to integrate syntactic reordering rules in the BTG-based model, and accordingly we choose BTG-based phrase-based model as our baseline in this paper. In theory, we can incorporate the syntactic reordering rules in any phrase-based models, such as beam search decoder Moses (Koehn et al. 2007). For Moses, we need to record the history of each partial translation hypothesis in decoding (the history keeps the source-side span sequence generating the current partial translation hypothesis) and check whether it matches the syntactic reordering rules. We leave this for our future work.

4.1 BTG-based phrasal SMT

The BTG-based translation can be viewed as a monolingual parsing process, in which only lexical rules $A \rightarrow (x,\ y)$ and two binary merging rules $A \rightarrow \left[A^l,\ A^r\right]$ and $A \rightarrow \left\langle A^l,\ A^r\right\rangle$ are allowed.

During decoding, the source sentence is first divided into phrases (note again that in phrase-based SMT the phrase means only any sequence of words); then the lexical rule $A \rightarrow (x,\ y)$ translates each source phrase x into the target phrase y and forms a block A. The straight rule $A \rightarrow \left[A^l,\ A^r\right]$ (or the inverted rule $A \rightarrow \left\langle A^l,\ A^r\right\rangle$) continually merges the two smaller neighboring blocks into a bigger one until the whole source sentence is covered. It is natural to adopt a bottom-up CYK (Cocke–Younger–Kasami) algorithm (Younger 1967) for this decoding process. The straight rule $A \rightarrow \left[A^l,\ A^r\right]$ combines the two neighboring blocks into a bigger one by monotonically concatenating the two partial target translations while the inverted rule $A \rightarrow \left\langle A^l,\ A^r\right\rangle$ yields the bigger block by swapping the two partial target translations. The lexical rule plays the same role as phrase pairs (tuples consisting of the source phrase and its target translation) in conventional phrase-based SMT (Koehn et al. 2007). The score of the lexical rule is computed as follows:

$$\Pr\left(r^l\right) = p(y|x)^{\lambda_1} \cdot p(x|y)^{\lambda_2} \cdot p_{lex}(y|x)^{\lambda_3} \cdot p_{lex}(x|y)^{\lambda_4} \tag{4}$$

where the first two factors are bidirectional phrase translation probabilities, $p_{lex}(y|x)$ and $p_{lex}(x|y)$ denote bidirectional lexical translation probabilities. The λs are their corresponding feature weights.

The score of the merging rules is formulated as:

$$\Pr(r^m) = \Omega^{\lambda_5} \tag{5}$$

in which Ω is the reordering score and λ_5 is its weight. The reordering model score in BTG-based translation is calculated using a maximum entropy model:

$$\Omega = P_\theta\left(O|A^l, A^r\right) = \frac{\exp\left\{\sum_i \theta_i h_i\left(O, A^l, A^r\right)\right\}}{\sum_{O'} \exp\left\{\sum_i \theta_i h_i\left(O', A^l, A^r\right)\right\}} \tag{6}$$

where $h_i\left(O, A^l, A^r\right)$ is a binary model feature function, O denotes merging order. Similar to (Xiong et al. 2006), lexical boundary words (leftmost and rightmost) of source and target phrases are employed as features. For example, if

two neighboring translation blocks are $(到_2新_3的_4办_5公_5大楼_6, \textit{to a new office building})$ and $(将_7是_8一个_9挑战_{10}, \textit{will be a challenge})$, the combination order of the target translation will be determined by lexicalized features $(到_2, 大楼_6, 将_7, 挑战_{10}, \textit{to}, \textit{building}, \textit{will}, \textit{challenge})$. θ_i is the weight of the feature function and is tuned with a maximum entropy toolkit.

Given that the decoding process yields the final target translation using n_l lexical rules and n_m merging rules, then the total score of the translation is calculated as:

$$P(e|f) = \prod_{i=1}^{n_l} \Pr(r^{l_i}) \cdot \prod_{i=1}^{n_m} \Pr(r^{m_i}) \cdot \exp(n_l)^{\lambda_6} \cdot \exp(|e|)^{\lambda_7} \cdot P_{LM}^{\lambda_8}(e) \tag{7}$$

The above translation score is usually formulated as a log-linear model, in which $\exp(n_l)$ and $\exp(|e|)$ denote respectively the phrase number penalty and the target length penalty. The first two items denote respectively translation model score and phrase reordering model score, and $P_{LM}(e)$ is the score of the target language model.

4.2 Model adaptation for syntactic rules

From the baseline BTG-based translation, we can see that the reordering model deals with any kind of phrases (syntactic phrases and the non-syntactic phrases). Furthermore, the baseline reordering model predicts the reordering probability of any two phrases with only lexicalized features. However, we know from Sect. 3 that the syntactic reordering rules are all about reordering between syntactic phrases and the reordering is predicted with multiple syntactic features. We believe that the syntactic reordering rules are more accurate compared with the baseline lexicalized reordering model. Thus, it comes to our idea of integrating the syntactic reordering rule: Its syntactic reordering probability will substitute the lexicalized reordering probability if the syntactic reordering rule matches the two neighboring blocks in the block merging process during decoding.

It is worthy to note that the syntactic reordering rules influence only the reordering of syntactic phrases. Thus, in decoding, the reordering of syntactic phrases and the reordering of non-syntactic phrases will depend on different kinds of features. It is natural that syntactic phrase reordering and non-syntactic phrase reordering should not coexist in a single reordering model. Moreover, we believe that the syntactic phrase reordering plays a more important role than the non-syntactic one. As a result, we design two orthogonal reordering models: syntactic reordering model handling reordering between syntactic phrases and non-syntactic reordering model dealing with other cases. The new score of the merging rules will be formulated as follows:

$$\Pr(r^m) = \Omega_N^{\lambda_5 \cdot I_N(A)} \cdot \Omega_S^{\lambda_9 \cdot I_S(A)} \tag{8}$$

where Ω_S and Ω_N denote respectively the syntactic and non-syntactic reordering score. The non-syntactic reordering score Ω_N is calculated with formula (6). The integrated syntactic reordering rules will influence the syntactic reordering score Ω_S. $I_S(A)$ and $I_N(A)$ are indicator functions (1 for true and 0 for false). $I_S(A) = 1$ and

$I_N(A) = 0$ when A is merging two syntactic phrases, and Ω_S is triggered; $I_S(A) = 0$ and $I_N(A) = 1$ otherwise, and Ω_N is triggered in this case.

In SMT, the importance of sub-models (such as the language model, the syntactic reordering model and the non-syntactic reordering model) is determined by its weight. Usually, the model weights are automatically tuned in a development set with an optimization algorithm (minimum error rate training (Och 2003)). We hope that the syntactic reordering model is more important and its weight is bigger than the one of non-syntactic reordering model. However, the weights tuning algorithm cannot guarantee this. Thus, to emphasize the importance of syntactic phrase reordering, we further create a reward feature to enhance syntactic reordering. The final score of merging rules are calculated as follows:

$$\Pr(r^m) = \Omega_N^{\lambda_5 \cdot I_N(A)} \cdot \Omega_S^{\lambda_9 \cdot I_S(A)} \cdot R_S^{\lambda_{10}} \tag{9}$$

in which R_S is a binary feature in order to reward syntactic reordering and it equals 1 if Ω_S is active. All the ten feature weights $\lambda_1 - \lambda_{10}$ in our new model are tuned with the minimum error rate training (MERT) algorithm.

4.3 Algorithm of integrating syntactic rules

After introducing the adapted translation model and the decoding algorithm style to be employed, we turn to the question on how the syntactic reordering rules are applied during decoding.

The unified format of syntactic reordering rule we adopt is designed as $\langle span(N^L),\ span(N^R),\ P(i)\rangle$, and the merging rules used in decoding always handle two continuous source spans (phrases): if $span(N^L)$ and $span(N^R)$ are successive, then $P(i)$ will serve as the syntactic reordering score Ω_s. However, $span(N^L)$ and $span(N^R)$ will not be consecutive if there is a non-empty $\diamondsuit$ between the two nodes. This brings trouble to the syntactic rule integration. In our current work, a simple strategy is proposed to solve this non-continuous problem.

Transformation strategy: The probabilistic syntactic rule in Fig. 6 is employed as an example to illustrate this detailed strategy. The original rule format is $\langle span(N^L),\ span(N^R),\ P(i)\rangle$ in which $N^L = PP(\text{到}_2\ \text{新}_3\ \text{的}_4\ \text{办公}_5\ \text{大楼}_6)$ and $N^R = VP(\text{是}_8\ \text{一个}_9\ \text{挑战}_{10})$, and thus the real rule is $\langle(2,\ 6),\ (8,\ 10),\ P(i)\rangle$. It is easy to find that these two spans are not continuous. However, it is fortunate to see that if we apply the syntactic reordering rule $\langle(2,\ 6),\ (8,\ 10),\ P(i)\rangle$ and reposition the first node after the last one, the span $(2,\ 10)$ will be $((7,7),(8,\ 10),(2,6)) = ((7,10),(2,6))$. This result is equivalent to the inverted case for spans $(2,\ 6)$ and $(7,\ 10)$. Therefore, the rule $\langle(2,\ 6),\ (8,\ 10),\ P(i)\rangle$ is equivalent to $\langle(2,\ 6),\ (7,\ 10),\ P(i)\rangle$ in which the spans are consecutive. Thus, a discontinuous syntactic reordering rule $\langle(i,\ k),\ (h,\ j),\ P(i)\rangle$ where $i \le k < h \le j$ and $h \neq k+1$, can be simply converted into an equivalent format $\langle(i,\ k),\ (k+1,\ j),\ P(i)\rangle$.[4]

With the transformation strategy, each discontinuous syntactic reordering rule can be converted into an equivalent continuous one which can fit the CYK decoding

[4] In our proposed model, we suppose that the combination of sibling children nodes under a parent node corresponds to a syntactic phrase. Thus, the span $(k + 1, j)$ corresponds to a syntactic phrase.

algorithm. As a result, all the syntactic reordering rules can be applied within the syntactic reordering model during decoding.

5 Experiments and analysis

5.1 Baselines used

The first baseline is the BTG-based phrasal translation system which uses a lexicalized reordering model trained with MaxEnt classifier. It is re-implemented according to Xiong et al. (2006) and it is then further improved and speeded up with cube pruning technique (Chiang 2007; Huang and Chiang 2007). We denote this baseline as MEBTG.[5] We modified the baseline model (NewModel) to incorporate the handcrafted or probabilistic syntactic reordering rules as described in Sects. 4.2 and 4.3.

To show the competitiveness of our approach, we want to compare our usage of handcrafted syntactic rules with the previous usage in Wang et al. (2007), and compare our method of using probabilistic syntactic rules with the previous method in Li et al. (2007). The classical implementation of the previous usage of syntactic rules is to reorder the source sentences of training, development and test data; then train the translation model with the reordered training data, tune the weights of features using development data with source sentence reordered, and finally use a phrase-based system (MEBTG in this paper) to obtain the target translation of the reordered test data. The system using handcrafted rules is named MEBTG+HSR which means MEBTG with handcrafted syntactic rules pre-ordering the source sentences. Likewise, the system using probabilistic rules is called MEBTG+PSR indicating MEBTG with probabilistic syntactic rules pre-ordering the source sentences.[6]

5.2 Corpora and experimental settings

At first, we report the experimental results on medium-scale training data. The experiments conducted on large-scale training data will be discussed in the Sect. 5.5. The medium-scale training set consists of 297K Chinese–English parallel sentences which are filtered from LDC.[7] The development set including 571 Chinese sentences is chosen from the test set of NIST06 and NIST08. The NIST05 test set is used as our test data.

Word-level alignments were obtained using GIZA++ (Och and Ney 2003). The grow-final-diag-and heuristic (Koehn et al. 2007) is employed to refine the alignments before lexical rule extraction. The target 4-gram language model was built with the English part of training data using the SRI Language Modeling

[5] In principle, MEBTG can deal with any kind of reordering. However, the reordering power is limited due to the exclusive use of lexicalized features in MEBTG.

[6] In training, the best reordered source sentence is found to be sufficient. In decoding, following (Li et al. 2007), 10-best reordered test sentences are employed as input.

[7] The catalogs include: LDC2003E14, LDC2005T06, LDC2004T07.

Toolkit (Stolcke 2002). The language model is smoothed with the modified Kneser–Ney algorithm. In order to acquire syntactic rules, we parse the Chinese sentences using the Stanford parser[8] (Klein and Manning 2003) with its default Chinese grammar. We build the maximum entropy model with the MaxEnt Toolkit developed by Zhang (2004) and set the Gaussian prior $g = 1.0$ to avoid overtraining.

All the models are optimized and tested using the case-sensitive BLEU-4 with shortest reference length penalty. The statistical significance test is performed using the pairwise re-sampling approach (Koehn 2004).

5.3 Experimental results

Before giving the experimental results, some notations of our new systems have to be introduced first. The system INcorporating the Handcrafted Syntactic Rules into our new model is named IN-HSR-NewModel. Likewise, IN-PSR-NewModel is used to denote the system incorporating the Probabilistic Syntactic Rules into the proposed new model.

In Table 3, we present the experimental results. Like Wang et al. (2007) and Li et al. (2007), we find that pre-ordering the source sentences with either handcrafted rules or with probabilistic rules can both obtain a significant improvement ($p < 0.05$) over the baseline MEBTG by absolute 0.58 and 0.60 BLEU percent points respectively. Since these two approaches may cause many pre-ordering errors, the gains are not very promising. However, after using our new approach, the system integrating the handcrafted rules into the new model IN-HSR-NewModel achieves a significantly larger improvement ($p < 0.01$) of up to 1.02 BLEU percent points over MEBTG, and also significantly outperforms the system pre-ordering with the handcrafted rules ($p < 0.05$). Furthermore, the system incorporating the probabilistic rules IN-PSR-NewModel performs even better. It outperforms both MEBTG and MEBTG+PRP significantly by 1.35 and 0.75 BLEU percent points with $p < 0.01$. The significant improvements achieved by the systems IN-HSR-NewModel and IN-PSR-NewModel indicate that our approach of using syntactic reordering rules within syntactic reordering model to help phrase reordering in the decoding stage is more effective than the previous approach for pre-ordering source sentences.

5.4 Analysis

In this section, we have a detailed analysis on the translation results.

5.4.1 Why do MEBTG+HSR and MEBTG+PSR perform similarly?

It is interesting that pre-ordering with the handcrafted rules has a similar performance to pre-ordering using the probabilistic rules. We find that because of the abundance of Chinese parsing errors, the accuracy of the handcrafted rules is not high; only 62.1 % of the rules are reported as correct in Wang et al. (2007). This results in many pre-ordering errors. Although the system pre-ordering with

[8] The precision of this parser in Chinese was reported to be 78.8 in F1-value (Levy and Manning 2003).

Table 3 Translation results on development set and test set

System	Dev (BLEU %)	Test (BLEU %)
MEBTG	25.67	32.96
MEBTG+HSR	26.35	33.54*
MEBTG+PSR	26.52	33.56*
IN-HSR-NewModel	26.71	**33.98**$^{**,+}$
IN-PSR-NewModel	27.13	**34.31**$^{**,\#\#}$

* or ** Significantly better than baseline MEBTG ($p < 0.05$ or $p < 0.01$ respectively). $^{+}$ Significantly better than MEBTG+HSR ($p < 0.05$). $^{\#\#}$ Significantly better than MEBTG+PSR ($p < 0.01$)

Src:	$\big((\text{迁移})_{NP}\ \underline{(\text{到}_{P}\ \text{新的办公大楼})_{PP}}\ \text{将}_{ADVP}\ \underline{(\text{是一个挑战})_{VP}}\big)_{VP}$
Ref:	relocation to a new office building will be a challenge
MEBTG:	relocation to new office building will be a challenge
MEBTG+HSR:	migration will be a challenge to the new office building
IN-HSR-NewModel:	relocation to a new office building will be a challenge

Fig. 7 An example that the handcrafted rule is wrong because the NP and PP are parsed with error, and since the pre-ordering system MEBTG+HSR reorders first the source sentence resulting in a wrongly reordered source sentence, it leads to a wrong translation which is even worse than the baseline MEBTG. However our approach IN-HSR-NewModel gets a correct translation

Src:	$\big(\underline{\text{首家}\,\big(\text{获准}\,(\text{在中国})_{PP}\ \text{经营人民币业务的}_{DEC}\big)_{CP}}\ \underline{(\text{比利时银行})_{NP}}\big)_{NP}$
Ref:	the first Belgian bank authorized to operate renminbi business in China
MEBTG:	the first authorized to operate rmb business Belgian bank in China
MEBTG+SRP:	the first authorized to operate rmb Belgian bank in China
IN-PSR-NewModel:	the first Belgian bank authorized to operate rmb business in China

Fig. 8 An example that the probabilistic rules miss the reordering instance that the CP should be repositioned after its sibling NP, and the reordering system MEBTG+PSR causes a wrong translation just as the baseline MEBTG does; however, our approach IN-PSR-NewModel obtains the correct one

probabilistic rules does not produce as many errors as MEBTG+HSR since many probabilistic rules are not applied if the reordering probability is smaller than 0.5, it may miss some correct reordering instances which should be applied. Thus, the two systems have similar translation quality. Two translation examples are illustrated in Figs. 7 and 8 to show the situations which handcrafted rules and probabilistic rules may encounter.

5.4.2 Why does IN-PSR-NewModel outperform IN-HSR-NewModel?

The gap between the two versions of the system that uses syntactic rules for pre-ordering sentences (the version with handcrafted rules and the version with probabilistic ones) is only 0.02 BLEU. Why is the gap between the two

corresponding versions of the system that applies syntactic reordering rules during decoding so much greater (0.33 BLEU more for the system with probabilistic rules than for the one with handcrafted rules)? We know that in the latter version which applies syntactic reordering rules during decoding, the two systems are almost the same except that they incorporate different syntactic rules: probabilistic rules versus handcrafted ones. Instead of using them directly to reorder the source sentence, the systems IN-HSR-NewModel and IN-PSR-NewModel apply these syntactic reordering rules to help phrase reordering in the decoding stage with the same algorithm. Therefore, we believe the difference might lie in the number of rules they have employed. We find that an average of only 4.18 handcrafted rules are acquired from each test sentence while 17.08 probabilistic rules[9] on average are obtained. During decoding, we believe that the more syntactic information is applied, the better the phrase reordering will be. As a result, the system IN-PSR-NewModel can outperform the system IN-HSR-NewModel.

5.4.3 The effect of new features?

As described in Sects. 4.2 and 4.3, our system introduces three new features: (1) the syntactic and non-syntactic phrase reordering models are designed to replace the baseline lexicalized reordering model [formula (8)]; (2) syntactic rules are incorporated into syntactic reordering model in the decoding step and (3) a binary rewarding feature is used to enhance the syntactic reordering [formula (9)]. Thus, it is interesting to investigate the effectiveness of each new feature. The IN-PSR-NewModel is employed to conduct this experiment. Table 4 shows the results. We can see that just distinguishing syntactic phrase reordering from non-syntactic one (SynNon) gives a significant improvement over the baseline MEBTG ($p < 0.05$). This corroborates our conjecture that syntactic reordering and non-syntactic reordering play different roles and should not be considered within the same reordering model. On this basis, we integrate the probabilistic rules (SynNon+PSR) and the result is promising with 0.69 BLEU percent points of improvement. It indicates that the syntactic rules can help phrase reordering in decoding to a large extent. Finally, we add a rewarding feature to encourage syntactic phrase reordering. The result shows that this feature can also improve the translation quality. It should be noted that the central contribution of this paper is the combination of these three new features for syntactic reordering and their integration into decoding. The experimental results show that this approach yields a significant performance improvement.

5.4.4 Are syntactic rules better than lexicalized ones?

The key idea in our paper is employing syntactic rules to replace lexicalized ones if they match. One may argue if the syntactic rules are indeed more reliable than

[9] It should be noted that the handcrafted rules are extracted only on three kinds of tree nodes (VP, NP, LCP) while the probabilistic rules can be extracted on any tree node with two children beside on the tree node of VP and NP. Therefore, the probabilistic rules are much more than handcrafted rules. For pre-ordering methods, MEBTG+HSR averagely used 4.18 handcrafted rules whereas MEBTG+PSR averagely used 6.26 probabilistic rules (with probability more than 0.5) per test sentence before decoding.

Table 4 The effect of new features

Features	BLEU (%)
MEBTG	32.96
SynNon	33.47*
SynNon+PSR	34.16**,@@
SynNon+PSR+Reward (IN-PSR-NewModel)	34.31**,@@

"SynNon" means syntactic and non-syntactic reordering model distinguished; "PSR" denotes probabilistic rules integrated. * or ** Significantly better than baseline MEBTG ($p < 0.05$ or $p < 0.01$). @@ Significantly better than "SynNon" ($p < 0.01$)

lexicalized ones, and the experimental results have empirically proven that they are. According to our analysis on probabilistic rules, we find that the syntactic rules are better than lexicalized ones if the parse tree is correctly parsed. For example, the probability $P(i)$ in probabilistic rule $\langle CP, NP, P(i)\rangle$ in Fig. 8 is 0.9796 recommending strong reordering (correct case) while the probability predicted by the lexicalized reordering model with boundary words as features is 0.6687. And we also find that when the tree is parsed with error, most syntactic rules in low quality can still be remedied during decoding with the help of other sub-models such as the translation model and the language model. For example, the probability $P(i)$ of the probabilistic rule $\langle PP, VP, P(i)\rangle$ in Fig. 7 is 0.6826, which is slightly bigger than 0.6094 from the lexicalized reordering model. Thus the syntactic rule has a slightly bigger tendency for wrong reordering; however this incorrect rule is remedied in our approach and a similar translation to that of MEBTG (using lexicalized phrase reordering) is obtained as Fig. 7 shows. Based on the analysis above, we can say that the syntactic rules are a better choice for phrase reordering compared with a lexicalized reordering model.

5.4.5 Some evidence about why integrating rules in decoding is better than pre-ordering

We argued before that incorporating syntactic reordering rules for decoding is better than pre-ordering source sentence because many pre-ordering errors can be avoided in our approach. Nevertheless, we did not give any experimental statistics to corroborate our speculation. For a better demonstration, we randomly choose 50 sentences from the test set and manually analyze the usage of handcrafted rules in the pre-ordering system (MEBTG+HSR) and in the proposed integrated system (IN-HSR-NewModel). Table 5 reports the detailed statistics. As shown in the table, there are 233 handcrafted rules extracted from the 50 test sentences in which only 127 rules are correct and others are incorrect due to the incorrect parse trees. Thus, for the pre-ordering approach, 106 handcrafted rules are used incorrectly to reorder the source sentences prior to decoding. It is very interesting that more than half of the errors are remedied during decoding. It is because that we not only reorder the test source sentences, but also reorder the source part of parallel training sentences to keep consistence. Therefore, some incorrect handcrafted rules may appear both in test sentence and training data. In this case, the incorrectly reordered source phrase

Table 5 Usage of handcrafted rules in pre-ordering system and integrated system for randomly selected 50 test sentences

Systems	# of reordering rules	# of correct rules	# of wrong rules before decoding	# of wrong rules after decoding
MEBTG+HSR	233	127	106	51
IN-HSR-NewModel	233	127	N/A	35

would correspond to the correct target translation since there is a lexical rule matching the wrongly reordered test source phrase in the translation phrase table. For instance, there exists a handcrafted rule *NP*(*DNP*(!*PN*)◇*NP*) → *NP*(◇*NP DNP*(!*PN*)) where DNP is (近期 的(*recent*)) and the child NP is (调查(*survey*)) in test set. Just as shown in Fig. 9, this handcrafted rule incorrectly reorders the original phrase into "(调查(*survey*)) (近期 的(*recent*))". However, the same handcrafted rule is also applied in the training data, and correspondingly a lexical rule[10] "调查 近期 的 ||| *recent survey*" is extracted from the training data. Thus, even though the handcrafted rule incorrectly reorders the test sentence, it can still obtain the correct translation. For the system integrating handcrafted rules during decoding, we cannot say how many rules are incorrectly used before decoding since they are not applied prior to decoding. After analyzing the target translation, we find that only 35 out of 233 rules are mistakenly applied in our proposed system IN-HSR-NewModel. That is to say that more original incorrect syntactic reordering rules are remedied during decoding in the new system, and the error rate of phrase reordering is reduced by 31.4 % ((51 − 35)/51) compared to pre-ordering system MEBTG+HSR. On the whole, the system integrating handcrafted rules in decoding causes much less reordering errors than pre-reordering system does. Therefore, our approach leads to better translation quality.

5.4.6 Comparing and Combining with Soft Syntactic Boundary Constraints

Besides using syntax to improve phrase reordering, Zollmann and Venugopal (2006), Marton and Resnik (2008), and Cherry (2008) focused on utilizing source-side syntax to model translation boundaries. Modeling the translation boundary addresses the issue that which source-side span partition is preferred to obtain the best translation.

As discussed in Sect. 2, Marton and Resnik (2008) designed various features for each specific constituent label to reward the translation whose source span covers the constituent and penalize the translation otherwise. For example, the feature NP = means that if the translation span exactly covers an NP in the source-side parse tree, a feature value will be added to the hypothesis score. Accordingly, NP+ means that if the translation span crosses a source-side NP, a feature value will be subtracted from the hypothesis score. In their work, they defined XP = {NP, VP, CP, IP, PP, ADVP, QP, LCP, DNP} and found that the feature XP+ performs best in

[10] A lexical rule is a translation equivalent in the form of "source language phrase ||| target language phrase" in the phrase table and can be viewed as A → (x, y).

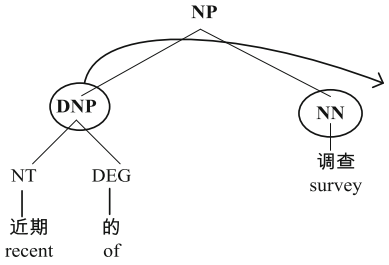

Fig. 9 An incorrect handcrafted syntactic rule appear both in the test and training data

Table 6 Translation results on development set and test set, SSC denotes Soft Syntactic Constraint (XP+feature)

System	Dev (BLEU %)	Test (BLEU %)
MEBTG	25.67	32.96
MEBTG+SSC	26.62	33.89**
IN-PSR-NewModel+SSC	27.53	**34.76**$^{**,++}$

** Significantly better than baseline MEBTG ($p < 0.01$). $^{++}$ Significantly better than MEBTG+SSC ($p < 0.01$)

Chinese-to-English translation. They addressed the translation boundary and we focus on phrase reordering, they are complementary in theory. Therefore, we compare and combine the idea of (Marton and Resnik 2008)'s XP+ feature with our idea in this section. We conduct the experiment with the same setting as the above experiments.

Table 6 gives the detailed results. Like (Marton and Resnik 2008), the XP+ feature (MEBTG+SSC) outperforms the baseline MEBTG significantly with an improvement of 0.93 absolute BLEU in the test set. However, combining XP+ feature and our idea of incorporating probabilistic syntactic reordering rules into decoding (IN-PSR-NewModel+SSC) achieves even larger improvements. It obtains a significant improvement of 1.80 BLEU over the baseline and performs significantly better than XP+ feature with a gain of 0.87 BLEU points. The experimental results show that the soft syntactic boundary constraints and our integration of syntactic reordering rules are complementary with each other.

5.5 Experiments on large-scale data

If we need to judge if a translation system is robust and sufficiently good, which factors should we consider? We believe one of the most important factors lies in whether or not the system is also effective on large-scale data set.

For pre-ordering systems, the whole source sentences of training set need to be parsed for consistency with reordered test source sentences. When it comes to large-scale data, the parsing time may be beyond tolerance. It could cost a few weeks or even several months. For instance, suppose we have a training set consisting of 4 million sentences and the average parsing time of one sentence is 2 s (an optimistic estimate). Then we will find that parsing all the source sentences of

Table 7 The statistics of the experimental data

	Size
Training data	3.8 M bilingual
Language model	3.8 M bilingual + 10 M Reuters
Syntactic reordering model	1.2 M
Development set	3,276
Test set	4,007

training data will take us about 93 days if not using parallel computing. It is obvious that pre-ordering approaches are not very suitable for large-scale data.

Our approach integrating handcrafted syntactic rules in decoding does not need to parse the source sentences of training data. As a result, this usage of handcrafted rules has no difficulty to be applied to large-scale data set. For the approach incorporating probabilistic syntactic rules in decoding, we are required to train a reordering model using multiple syntactic features (as discussed in Sect. 3.2). For this situation, we also do not need to parse all the source sentences of training data but to choose only a small part for parsing. Because we can see from the last section that the system incorporating probabilistic rules performs better than the system integrating handcrafted rules, we test our proposed system incorporating probabilistic rules in large-scale data set to show its effectiveness and efficiency.

The experimental background is the 2009 Chinese Workshop of Machine Translation (CWMT2009).[11] And all the corpora are from this workshop. The statistics are illustrated in Table 7. The training data contains about 3.8 million bilingual Chinese-English sentences. The large 5-gram language model is trained using the target part of bilingual data and 10 million Reuters English news. We train the syntactic reordering model with 1.2 million bilingual data (the source part needs to be parsed). The development set includes 3,276 sentences and the test set has 4,007 sentences. It should be noted that all the preprocessing is the same as what we used in Sect. 5.3.

To have a better comparison, we have also conducted the experiment using the widely used open source translation toolkit Moses (Koehn et al. 2007). Table 8 reports the final results which are measured by case-sensitive BLEU-SBP[12] (Chiang et al. 2008). From this table, we see that, like Zhang and Li (2009), the system MEBTG significantly outperforms Moses by 0.62 BLEU-SBP percent points in development set and 0.66 BLEU-SBP percent points in test set. It is because MEBTG employs a generalized MaxEnt-based lexicalized reordering model using boundary words as features, but Moses uses a lexicalized reordering model which lacks generalization ability. Thanks to the probabilistic syntactic rules and their skillful integration algorithm, our proposed system IN-PSR-NewModel obtains a significant improvement over MEBTG ($p < 0.01$). The gains are 0.77 BLEU-SBP percent points in the development set and 0.92 BLEU-SBP percent points in the test set. It is worthy to be noted that the improvements are quite promising because they

[11] http://www.icip.org.cn/cwmt2009.

[12] SBP stands for Strictly Brevity Penalty. Since the CWMT2009 workshop scores all the results with BLEU-SBP, we tune and test our system with BLEU-SBP.

Table 8 The experimental results on large-scale data set, the decoding time is the average decoding time on the development set and the test set in seconds per sentence

Systems	Dev (BLEU-SBP %)	Test (BLEU-SBP %)	Decoding time
Moses	26.52	22.51	2.761
MEBTG	27.14	23.17**	4.187
IN-PSR-NewModel	27.91	24.09**,++	4.245

** Denotes statistically better than Moses, and ++ means statistically better than MEBTG (the significance test was conducted by the organization of CWMT2009 and we didn't do this by ourselves as they didn't release the references)

are achieved over a baseline incorporating a competitive 5-gram language model. It is widely acknowledged that it can be very difficult to outperform high-order n-gram models in large-scale experiments (Galley and Manning 2009). Furthermore, we can see from the table that, compared with baseline MEBTG, our proposed system IN-PSR-NewModel performs the translation with nearly the same speed.

According to the experiments, we can conclude that our proposed approach of using syntactic reordering rules in the decoder of a phrase-based system can not only significantly improve the translation quality, but also show effectiveness and efficiency in large-scale experiments.

6 Conclusion and future work

In this paper, we have presented a framework for effectively incorporating source-side syntactic reordering rules into the phrase-based SMT. We designed a unified format to represent both the handcrafted and probabilistic syntactic reordering rules. To facilitate the integration of the syntactic reordering rules in decoding, we distinguished the syntactic phrase reordering model from the non-syntactic phrase reordering model. The syntactic phrase reordering model was finely designed so that it can accommodate the syntactic reordering rules. Furthermore, we have created a binary feature to reward the syntactic reordering in order to attach more importance to syntactic phrase reordering. For a test sentence to be translated, we first acquire the syntactic reordering rules from the source parse trees. Instead of using them to reorder the source sentences arbitrarily, we incorporate these rules to guide phrase reordering within the syntactic phrase reordering model in the decoding stage. The experiments have shown that our approach of using syntactic reordering rules significantly outperforms the previous approaches whether for handcrafted rules or for probabilistic rules. Moreover, we have found that our proposed approach also shows effectiveness and efficiency in large-scale experiments.

From the experimental results, we also know that just distinguishing syntactic reorderings from non-syntactic ones can improve the translation quality significantly, and at the same time, facilitate the integration of the syntactic reordering rules. The question arises whether it is also true for translation systems in different decoding styles such as Moses and the hierarchical phrase-based system Hiero (Chiang 2007). We leave this to our future work. Furthermore, we plan to

investigate the syntactic reordering rules acquired from dependency structures (Xu et al. 2009) and to design methods for integrating these rules in decoding in order to better guide phrase reordering.

References

Andreas, J., Habash, N., & Rambow, O. (2011). Fuzzy syntactic reordering for phrase-based statistical machine translation. In *Proceedings of the 6th workshop on statistical machine translation*, Edinburgh, Scotland, UK, July 30th–31th, 2011.

Badr, I., Zbib, R., & Glass, J. (2009). Syntactic phrase reordering for English-to-Arabic statistical machine translation. In *Proceedings of the 12th conference of the European chapter of the association for computational linguistics* (pp. 86–93). Athens, Greece, March 30th–April 3rd, 2009.

Brown, P. F., Cocke, J., Della, S. A., Pietra, V. J., Pietra, D., Jelinek, F., et al. (1990). A statistical approach to machine translation. *Computational Linguistics, 16*(2), 79–85.

Brown, P. F., Della, S. A., Pietra, V. J., Pietra, D., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics, 19*(2), 263–311.

Cherry, C. (2008). Cohesive phrase-based decoding for statistical machine translation. In *Proceedings of the 46th annual meeting of the association for computational linguistics: Human language technology* (pp. 72–80). Columbus, Ohio, USA, June 15th–20th, 2008.

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

Chiang, D., Marton, Y., & Resnik, P. (2008). Online large-margin training of syntactic and structural translation features. In *Proceedings of the 2008 conference on empirical methods in natural language processing* (pp. 224–233). Waikiki, Honolulu, USA, October 25th–27th, 2008.

Collins, M., Koehn, P., & Kučerová, I. (2005). Clause restructuring for statistical machine translation. In *Proceedings of the 43rd annual meeting on association for computational linguistics* (pp. 531–540). Michigan, USA, June 26th–30th, 2005.

Costa-jussà, M. R., Crego, J. M., Lambert, P., Khalilov, M., Fonollosa, J. A. R., Marino, J. B., et al. (2007). Ngram-based statistical machine translation enhanced with multiple weighted reordering hypotheses. In *Proceedings of the second workshop on statistical machine translation* (pp. 167–170). Prague, Czech Republic, June 27th–30th, 2007.

Crego, J. M., & Yvon, F. (2010). Improving reordering with linguistically informed bilingual n-grams. In *Proceedings of the 23rd international conference on computational linguistics* (pp. 197–205). Beijing, China, August 23rd–27th, 2010.

Du, J. & Way, A. (2010). The impact of source-side syntactic reordering on hierarchical phrase-based SMT. In *Proceedings of the 14th annual conference of the European association for machine translation* (pp. 82–89). Saint-Raphaël, France, May 27th–28th, 2010.

Elming, J. (2008). Syntactic reordering integrated with phrase-based SMT. In *Proceedings of the 22nd international conference on computational linguistics* (pp. 209–216). Manchester, UK, August 18th–22nd, 2008.

Galley, M., & Manning, C. D. (2009). Quadratic-time dependency parsing for machine translation. In *Proceedings of the joint conference of the 47th annual meeting of the association for computational linguistics and the 4th international joint conference on natural language processing* (pp. 773–781). Singapore, August 2nd–7th 2009.

Genzel, D. (2010). Automatically learning source-side reordering rules for large scale machine translation. In *Proceedings of the 23rd international conference on computational linguistics* (pp. 376–384). Beijing, China, August 23rd–27th, 2010.

Habash, N. (2007). Syntactic preprocessing for statistical machine translation. In *Proceedings of the 11th machine translation summit* (pp. 215–222). Copenhagen, Denmark, September 10th–14th, 2007.

Huang, L. & Chiang, D. (2007). Forest rescoring: Faster decoding with integrated language models. In *Proceedings of the 45th annual meeting of the association of computational linguistics* (pp. 144–151). Prague, Czech Republic, June 27th–30th, 2007.

Klein, D., & Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st annual meeting on association for computational linguistics* (pp. 423–430). Sapporo, Japan, July 7th–12th, 2003.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 conference on empirical methods in natural language processing* (pp. 388–395). Barcelona, Spain, July 25th–26th, 2004.

Koehn, P., Hoang, H., Birch, A., Federico, M., Bertoldi, N., Cowan, B., et al. (2007) Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting on association for computational linguistics on interactive poster and demonstration sessions* (pp. 177–180). Prague, Czech Republic, June 27th–30th, 2007.

Koehn, P., Och, F. J., & 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 language* (pp. 48–54). Edmonton, Canada, May 27th–June 1st, 2003.

Lee, Y.-S., Zhao, B., & Luo, X. (2010). Constituent reordering and syntax models for English-to-Japanese statistical machine translation. In *Proceedings of the 23rd international conference on computational linguistics* (pp. 626–634). Beijing, China, August 23rd–27th, 2010.

Levy, R., & Manning, C. D. (2003). Is it harder to parse Chinese, or the Chinese Treebank? In *Proceedings of the 41st annual meeting of the association of computational linguistics* (pp. 439–446).

Li, C.-H., Zhang, D., Li, M., Zhou, M., Li, M., & Guan, Y. (2007). A probabilistic approach to syntax-based reordering for statistical machine translation. In *Proceedingd of the 45th annual meeting of the association of computational linguistics* (pp. 720–727). Prague, Czech Republic, June 27th–30th, 2007.

Marton, Y., & Resnik, P. (2008). Soft syntactic constraints for hierarchical phrased-based translation. In *Proceedings of the 46th annual meeting of the association for computational linguistics: human language technology* (pp. 1003–1011), Columbus, Ohio, USA, June 15th–20th, 2008.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st annual meeting on association for computational linguistics* (pp. 160–167). Sapporo, Japan, July 7th–12th, 2003.

Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics, 29*(1), 19–51.

Och, F. J., & Ney, H. (2004). The alignment template approach to statistical machine translation. *Computational Linguistics, 30*(4), 417–449.

Stolcke, A. (2002). SRILM-an extensible language modeling toolkit. In *Proceedings 7th International conference on spoken language processing* (pp. 901–904). Denver, Colorado, USA, September 16th–20th, 2002.

Tillmann, C., & Zhang, T. (2005). A localized prediction model for statistical machine translation. In *Proceedings of the 43rd annual meeting on association for computational linguistics* (pp. 557–564). Michigan, USA, June 26th–30th, 2005.

Visweswariah, K., Navratil, J., Sorensen, J., Chenthamarakshan, V., & Kambhatla, N. (2010). Syntax-based reordering with automatically derived rules for improved statistical machine translation. In *Proceedings of the 23rd international conference on computational linguistics* (pp. 1119–1127) Beijing, China, August 23rd–27th, 2010.

Wang, C., Collins, M., & Koehn, P. (2007). Chinese syntactic reordering for statistical machine translation. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 737–745). Prague, Czech Republic, June 27th–30th, 2007.

Wu, D. (1997). Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics, 23*(3), 377–403.

Wu, X., Sudoh, K., Duh, K., Tsukada, H., & Nagata, M. (2011). Extracting pre-ordering rules from predicate-argument structures. In *Proceedings of the 5th international joint conference on natural language processing* (pp. 29–37). Chiang Mai, Thailand, November 8th–13th, 2011.

Xiang, B., Ge, N., & Ittycheriah, A. (2011). Improving reordering for statistical machine translation with smoothed priors and syntactic features. In *Proceedings of the fifth workshop on syntax, semantics and structure in statistical translation* (pp. 61–69). Portland, Oregon, USA, June 19th–24th, 2011.

Xiong, D., Liu, Q., & Lin, S. (2006). Maximum entropy based phrase reordering model for statistical machine translation. In *Proceedings of the 21st international conference on computational linguistics and the 44th annual meeting of the association for computational linguistics* (pp. 521–528). Sydney, Australia, July 17th–21st, 2006.

Xiong, D., Zhang, M., Aw, A., & Li, H. (2008). Linguistically annotated BTG for statistical machine translation. In *Proceedings of the 22nd international conference on computational linguistics* (pp. 1009–1016). Manchester, UK, August 18th–22nd, 2008.

Xiong, D., Zhang, M., & Li, H. (2011). Enhancing language models in statistical machine translation with backward N-grams and mutual information triggers. In *Proceedings of the 49th annual meeting of the association for computational linguistics* (pp. 1288–1297). Portland, Oregon, USA, June 19th–24th, 2011.

Xu, P., Kang, J., Ringgaard, M., & Och, F. (2009). Using a dependency parser to improve SMT for subject-object-verb languages. In *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics* (pp. 245–253). Boulder Colorado, May 31th–June 5th, 2009.

Xue, N., Xia, F., Chiou, F.-D., & Palmer, M. (2005). The Penn Chinese Treebank: Phrase structure annotation of a large corpus. *Natural Language Engineering, 11*(02), 207–238.

Younger, D. H. (1967). Recognition and parsing of context-free languages in time n^3. *Information and Control, 10*(2), 189–208.

Zens, R., Ney, H., Watanabe, T., & Sumita, E. (2004). Reordering constraints for phrase-based statistical machine translation. In *Proceedings of the 20th international conference on computational linguistics* (pp. 205–262). Geneva, Switzerland, August 23rd–27th, 2004.

Zhang, L. (2004). Maximum entropy modeling toolkit for Python and C++. Available at http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html.

Zhang, M., & Li, H. (2009). Tree kernel-based SVM with structured syntactic knowledge for BTG-based phrase reordering. In *Proceedings of the 2009 conference on empirical methods in natural language processing* (pp. 698–707). Singapore, August 6th–7th, 2009.

Zhang, D., Li, M., Li, C.-H., & Zhou, M. (2007). Phrase reordering model integrating syntactic knowledge for SMT. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 533–540) Prague, Czech Republic, June 27th–30th, 2007.

Zollmann, A., & Venugopal, A. (2006). Syntax augmented machine translation via chart parsing. In *Proceedings of NAACL 2006—Workshop on statistical machine translation.* New York. June 4–9.