Chinese Syntactic Reordering through Contrastive Analysis of Predicate-predicate Patterns in Chinese-to-Korean SMT

Jin-Ji Li, Jungi Kim and Jong-Hyeok Lee
Division of Electrical and Computer Engineering
Pohang University of Science and Technology, Pohang, Republic of Korea
{ljj,yangpa,jhlee}@postech.ac.kr

Abstract
We propose a Chinese dependency tree reordering method for Chinese-to-Korean SMT systems through analyzing systematic differences between the Chinese and Korean languages. Translating predicate-predicate patterns in Chinese into Korean raises various issues such as long-distance reordering. This paper concentrates on syntactic reordering of predicate-predicate patterns in Chinese dependency trees through contrastively analyzing construction types in Chinese and their corresponding translations in Korean. We explore useful linguistic knowledge that assists effective syntactic reordering of Chinese dependency trees; we design two experiments with different kinds of linguistic knowledge combined with the phrase- and hierarchical phrase-based SMT systems, and assess the effectiveness of our proposed methods. The experiments achieved significant improvements by resolving the long-distance reordering problem.

1 Introduction

In the SMT community, word reordering has been treated as one of the most important tasks for resolving word-order differences when translating from a source language into a target one. Although many effective reordering methods have been proposed, long-distance reordering is still considered difficult. State-of-the-art SMT systems such as the phrase- and the hierarchical phrase-based SMTs, also are not free from this problem.

In this paper, we consider a specific structure, namely a predicate-predicate pattern which leads to a long-distance reordering problem when translating from Chinese into Korean. We define a predicate-predicate pattern as a pattern that consists of a pair of predicates in a dependency parse tree (D-tree) where a head predicate has another predicate as an immediate child. In Chinese, these patterns assume several different structures with little or no morphological differences, since Chinese is a morphologically poor language. We define long-distance reordering as the relocation of one predicate across another in predicate-predicate patterns in the D-tree. Without any linguistic clues from the surface forms, it is difficult to compile reordering rules for the predicate-predicate patterns. In this paper, we explore various linguistic knowledge for the purpose of effective long-distance reordering of Chinese D-trees.

As a preprocessing to a phrase-based SMT, a number of researchers have proposed syntactic reordering approaches to phrase structure parse trees (PS-trees) (Xia and McCord, 2004; Collins et al., 2005; Wang et al., 2007; Li et al., 2009) and D-trees (Chang et al., 2009; Xu et al., 2009; Hong et al., 2009). Previous work on deterministic syntactic reordering in a phrase-based SMT has been effective for language pairs that belong to different word-order typologies such as Chinese and Korean. This kind of reordering approach is very flexible to combine with various decoding models without adding computational complexity to the decoding phase.

Syntactic reordering methods for PS-trees and D-trees have their advantages and disadvantages due to the differences in their constituent and dependency structures. PS-trees contain hierarchy and precedence information of syntactic units (words or
phrases), and D-trees directly encode syntactic or semantic relations between words. Recent studies have shown that more flexible and high coverage reordering can be achieved with D-trees (Xu et al., 2009). Xu et al. (2009) described a set of manually constructed precedence rules using the dependency relations and showed great efficiencies in SMT systems targeting 5 subject-object-verb (SOV) languages including the Korean language.

In our Chinese-to-Korean SMT, we adopt the principles of compiling reordering rules in a D-tree as Xu et al. (2009) proposed. Since Chinese is a morphologically poor language with insufficient linguistic clues, more careful concern is required when compiling syntactic reordering rules, especially for predicate-predicate patterns.

In Section 2, we describe why it is difficult to re-order predicate-predicate patterns with comparison to other languages such as English and Korean. Section 3 analyzes the various structures of predicate-predicate patterns and deduces the structures that need reordering when translating. General reordering rules and specific reordering rules on predicate-predicate patterns will be given in Section 4. Finally, Section 5 shows the experimental results and discussion.

2 Reordering predicate-predicate patterns is difficult.

Translating predicate-predicate patterns from one language to another is problematic, especially from a morphologically poor language to a rich one. Chinese is a typical isolating language, and predicate-predicate patterns in Chinese may represent several structures (Table 1) with less (almost no) morphological differences than other languages such as English and Korean. Therefore, identifying the correct structure of a predicate-predicate pattern is a challenging task for Chinese dependency parsing. The patterns also frequently appear in Chinese sentences; on average, there are 1.78 predicate-predicate patterns per sentence in our training corpus.

In Figure 1, ‘相信(believe)’ is a head predicate that dominates predicate ‘有(have)’ in Ch1. and ‘买(buy)’ dominates ‘做(cook)’ in Ch2, but there is no contextual evidence to suggest these relations. In English, the complementizer ‘that’ and the conjunction ‘and’ signals clausal complement and coordinate constructions. In Korean, such structures are indicated by conjunctive verb-endings ‘고(ko)’ and

<table>
<thead>
<tr>
<th>Type</th>
<th>Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clausal subject construction</td>
<td>![ Diagram ]</td>
</tr>
<tr>
<td>Complex/Compound construction</td>
<td>![ Diagram ]</td>
</tr>
<tr>
<td>Serial verb construction (SVC)</td>
<td>![ Diagram ]</td>
</tr>
<tr>
<td>Pivot construction</td>
<td>![ Diagram ]</td>
</tr>
<tr>
<td>Clausal complement construction</td>
<td>![ Diagram ]</td>
</tr>
<tr>
<td>Existential construction</td>
<td>![ Diagram ]</td>
</tr>
<tr>
<td>Emphasis construction</td>
<td>![ Diagram ]</td>
</tr>
</tbody>
</table>

Table 1: Types and structures of predicate-predicate patterns.

For convenience and consistency, we describe the POS of predicate in the Penn Chinese Treebank style (Xue et al., 2000). $V_{\text{head}}$ is either VV or VA: VV: common verb; VA: predicative adjective; VE: existential verb; VC: copula verb.
‘\text{서}(seo)’. In English and Korean, identifying particular structures of predicate-predicate patterns is relatively easy.

The predicate-predicate pattern in Ch1 needs reordering while the pattern in Ch2 does not, considering the order of predicates in the corresponding Korean sentences. In addition to identifying the structures of predicate-predicate patterns, further analysis on the characteristics of translation from the constructions will help us infer more effective syntactic reordering rules.

3 Predicate-predicate patterns of Chinese

In this section, we analyze the types of predicate-predicate patterns to deduce the structures that need to be reordered when translating.

A predicate-predicate pattern where the immediate child is located to the left of the head predicate forms either a clausal subject construction or a complex construction. These constructions seldom require long distance reordering in Chinese-to-Korean MT, so we do not treat the issue in this paper.

A head predicate $V_h$ may have several predicates as immediate right children such as $V_i$ and $V_j$ (Figure 2). A head predicate is categorized into two types; if the predicate forms a base verb phrase (VP) without its child predicates, the predicate falls into Type 1, and if the predicate constructs a base VP with the first predicate from its right children, the predicate is categorized as Type 2.

Construction types listed in Table 1 fall into one of the following types; a compound construction and SVC correspond to Type 1, and pivot, clausal complement, existential, and emphasis constructions correspond to Type 2. The constructions which correspond to Type 2 are strong candidates for reordering. If reordered, we relocate the head predicate $V_h$ right after the first right-child predicate $V_i$, since the head predicate only dominates the first right-child predicate. ‘相信(believe)’ in Figure 1 belongs to Type 2, and ‘买(buy)’ Type 1.

The Type 1 constructions do not require reordering of its predicates. SVC in Chinese is where two or more predicates are juxtaposed sharing a subject. Dominantly, the predicates in the SVC and compound construction are translated into Korean in sequential order.

For the four constructions of Type 2, the head predicate $V_h$ takes the child predicate $V_i$ as its sentential argument. However, in some cases $V_i$ remains in its original position when translating.

In the pivot construction ($V_{head} + N + V_{child}$), N functions as the object of $V_{head}$ as well as the subject of $V_{child}$. This kind of head verbs is described as the object control verb in the Penn Chinese Treebank. We do not reorder pivot constructions when the lexical meaning of the head verb is command; this construction is usually translated non-literarily and improving translation results of such constructions requires more than syntactic reordering. Only object control verbs with other lexical meanings are reordered.

The head predicate in the clausal complement construction ($V_{head} + N + V_{child}$) takes up to two objects. If N exists, it functions as a direct object. $V_{child}$ functions as a sentential object. The objects have a strong tendency to be translated at the pre-verbal position in Korean sentences. The Penn Chinese Treebank describes verbs belonging to this construction as psychological verbs, subject control verbs, and other verbs such as ‘告诉(tell) and 通知(announce)’.

In existential and emphasis constructions, VE and VC have functional roles as well as lexical meanings. To translate these structures adequately, a linguistic process more complex than syntactic reordering is necessary.

In conclusion, predicate-predicate patterns form clausal complement constructions and some of the pivot constructions require a long-distance reordering of verbs when translated into Korean.
4 Chinese syntactic reordering on predicate-predicate patterns

4.1 General syntactic reordering

We compile a set of general Chinese syntactic reordering rules under the same principles proposed by Li et al. (2009) and Xu et al. (2009). The two work describe syntactic reordering of PS-trees and D-trees respectively, and translation directions are both from SVO to SOV languages. Despite the differences in parse structures, their main principles of syntactic reordering are similar.

Li et al. (2009) move modality-bearing words near their verbal heads. They argue that Chinese expresses the modality information using discontinuous morphemes scattered throughout a sentence; while the modality of Korean is expressed intensively by verb endings. Since Korean is a verb-final language, all the other elements should take the pre-verbal positions in Korean sentences. Although they did not use the term ‘modality-bearing word’, the elements which they grouped are closely related to ‘modality-bearing words’ such as phrasal verb particle, auxiliary verb, passive auxiliary verb, and negation.

We also apply reordering rules to prepositions in Chinese, which originate from verbs and preserve the characteristics of verbs. Objects of the prepositions are positioned as a right child, and it will move to the left side for reordering.

We will describe the principle of the general syntactic reordering rules as follows in which the head word is a predicate. Here is an example shown in Figure 3. Every predicate in a Chinese D-tree consists of left children (L_Children) and right children (R_Children). From the left children, the modality-bearing words (L_Modal) are relocated near the predicate, and the other elements (L_Other) remain on the left side of the predicate.

For the right children, the process is slightly different. Modality-bearing words (R_Modal) are relocated near the predicate, as L_Modal. However, as Korean is a verb-final language, most right children will be moved to the left side of the predicate (L_FromRight). A right child belonging to R_Remnant always forms predicate-predicate patterns with the head predicate if the right child is a predicate. The movement of child nodes in this case will be carefully controlled using the reordering rules of predicate-predicate patterns which we proposed in Section 4.2. In other words, in general reordering rules, all of the child nodes of predicate-predicate patterns remain in R_Remnant due to the lack of linguistic knowledge.

After applying the reordering rules, such as in Algorithm 1, each predicate will have newly constructed children such as, L_Other, L_FromRight, L_Modal, R_Modal, and R_Remnant, in which the children reserve the relative orders of the original sentence.²

4.2 Syntactic reordering of predicate-predicate patterns

In this section, we present the syntactic reordering methods for predicate-predicate patterns. As pointed out in Section 3, a very specific type of predicate-predicate pattern needs our attention for long-distance reordering: a head predicate of Type

<table>
<thead>
<tr>
<th>Ch3.</th>
<th>他 (he)</th>
<th>不能 (cannot)</th>
<th>去 (go)</th>
<th>北京 (Beijing)</th>
<th>了 (Aspect particle)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syn.</td>
<td>subj</td>
<td>mmod</td>
<td>Root</td>
<td>dobj</td>
<td>asp</td>
</tr>
<tr>
<td>L_Children</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Re.</td>
<td>他 (he)</td>
<td>北京 (Beijing)</td>
<td>不能 (cannot)</td>
<td>去 (go)</td>
<td>了 (Aspect particle)</td>
</tr>
<tr>
<td>L_Other</td>
<td></td>
<td>L_FromRight</td>
<td></td>
<td>L_Modal</td>
<td>R_Remnant</td>
</tr>
</tbody>
</table>

Figure 3: An example of reordered Chinese sentence after applying the general reordering method. Ch3.: a Chinese sentence; Syn.: dependency structure; Re.: reordered Chinese sentence;

²Following are a set of dependency relations defined in Stanford Chinese typed dependency parser. mmod: modal verb modifier; neg: negative modifier; pass: passive marker; asp: aspect marker; rcomp: resultative complement; comod: coordinated verb compound modifier; ccomp: clausal complement; punct: punctuation.
Algorithm 1 General syntactic reordering rules

**Input:** \(L_{Children}, R_{Children}\) of a Predicate \(P\)

**Output:** \(L_{Other}, L_{FromRight}, L_{Modal}, R_{Modal}, R_{Remnant}\)

\[
\begin{align*}
\text{for node } N \text{ in } L_{Children} & \text{ do} \\
& \quad \text{if dep. relation of } N \in \{\text{mmod, neg, pass}\} \text{ then} \\
& \quad \quad L_{Modal} \leftarrow L_{Modal} + \{N\} \\
& \quad \text{else} \\
& \quad \quad L_{Other} \leftarrow L_{Other} + \{N\} \\
& \text{end if} \\
\text{end for} \\
\text{for node } N \text{ in } R_{Children} & \text{ do} \\
& \quad \text{if dep. relation of } N \in \{\text{comod, asp, neg, rcomp}\} \text{ then} \\
& \quad \quad R_{Modal} \leftarrow R_{Modal} + \{N\} \\
& \quad \text{else if dep. relation of } N \in \{\text{ccomp, punct}\} \text{ then} \\
& \quad \quad R_{Remnant} \leftarrow R_{Remnant} + \{N\} \\
& \quad \text{else} \\
& \quad \quad L_{FromRight} \leftarrow L_{FromRight} + \{N\} \\
& \text{end if} \\
\text{end for}
\]

Table 2: Features for the SVM binary classifier.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lexical</strong></td>
<td>Surface form of (V_h)</td>
</tr>
<tr>
<td>(V_h) is a pivot construction verb with the lexical meaning of command</td>
<td></td>
</tr>
<tr>
<td>(V_h) is a verb that can take a clausal complement</td>
<td></td>
</tr>
<tr>
<td><strong>Syntactic</strong></td>
<td>(V_i) has a direct object</td>
</tr>
<tr>
<td>(V_i) has a “,” or “:” punctuation as left sibling</td>
<td></td>
</tr>
<tr>
<td>(V_i) has a nominal subject</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Accuracy of the SVM classifiers for predicate-predicate pattern reordering.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All features</td>
<td>93.3</td>
</tr>
<tr>
<td>w/o lexical feature</td>
<td>77.1</td>
</tr>
<tr>
<td>w/o syntactic features</td>
<td>90.7</td>
</tr>
</tbody>
</table>

We adopt a binary classifier using SVMlight\(^3\) for robust classification. The task is simplified as determining whether a reordering is necessary or not, given a head predicate \((V_h)\) of \(VV\) or \(VA\) and its first right-child predicate \((V_i)\) in \(R_{Remnant}\).

The features used by the classifier are described in Table 2. Two kinds of features are used: lexical and syntactic information from D-trees.

We collect positive instances from the Penn Chinese Treebank 4.0. Positive instances are either 1) a head predicate \((V_h)\) (\(VV\) or \(VA\)) with its sentential argument \((V_i)\) corresponding to IP-OBJ\(^4\), or 2) a head predicate \((V_h)\) (\(VV\) or \(VA\)) without a lexical meaning of command, and has a sentential argument \((V_i)\) corresponding to IP.

PKU dictionary is a dictionary of “the Grammatical Knowledge-base of Contemporary Chinese” from Peking University which contains about 80,000 entries. It was developed for the purpose of Chinese language processing with various information including morphology, syntax and semantics. We refer to the PKU dictionary to collect the pivot construction verbs with the lexical meaning of command. It also provides a list of verbs that can take a clausal complement.

From 18,487 valid instances extracted from the Penn Chinese Treebank, the number of positive instances is 5,544. The accuracy of the SVM classifier is measured using 10-fold cross validation (Table 3). It reveals that the lexical information of head predicate is the most important feature.

For comparison purposes, we estimate the classification accuracy of heuristic rules which only uses the PKU dictionary information. If \((V_h)\) has the property of taking a clausal complement, reordering is performed. Its performance is 87.2\%, 6.1\% lower than the SVM classifier.

\(^3\)http://svmlight.joachims.org, version 6.02.

\(^4\)The Penn Chinese Treebank is annotated with the functional tags of phrase such as IP-OBJ. IP-OBJ is an IP (simple clause headed by INF.,) that acts as a sentential object in the sentences.
5 Experiment

5.1 Experimental setting

Our baseline system is the state-of-the-art phrase- and hierarchical phrase-based SMT system built in Moses (Chiang, 2005; Koehn et al., 2007) with 5-gram SRI language modeling (Stolcke, 2002) tuned with Minimum Error Rate Training (MERT) (Och, 2003). We adopt NIST (Doddington, 2002) and BLEU (Papineni et al., 2001) as our evaluation metrics. A significance test is also conducted using a paired bootstrap resampling method⁵ (Koehn, 2004).

We use the Stanford Chinese typed dependency parser (Levy and Manning, 2003; Chang et al., 2009) to parse Chinese sentences. Chinese sentences in training and test corpora are first parsed into dependency trees and are applied to a series of syntactic reordering rules recursively from the root to the bottom. Korean sentences are segmented into morphemes using an in-house morphological analyzer⁶.

We designed two experiments with different types of knowledge: the first is to assess the effectiveness of the heuristic classifier with verb lists from the PKU dictionary, and the second with the SVM classifier that shows the highest performances in the classification.

5.2 Corpus profile

We used the same corpus introduced in (Li et al., 2009), namely Dong-A newspaper corpus. It is a non-literally translated Korean-to-Chinese corpus. The training corpus has 98,671 sentence pairs, and the development and test corpora each have 500 sentence pairs. The original training corpus size is 99,226 sentence pairs. However, we only use 98,671 pairs because of the parsing errors of Stanford Chinese typed dependency parser. The corpus profile is displayed in Table 5.

5.3 Result and discussion

The experimental results show that the proposed methods improve the baseline of phrase- and hierarchical phrase-based Chinese-to-Korean SMT effectively (Table 4). All the performances using the hierarchical phrase-based SMT (Hiero) is much better than the phrase-based SMT with lexicalized reordering. Our proposed method using the SVM classifier indicates significant improvements, and the gain is smaller in the Hiero than in the phrase-based SMT. Since the domains of the training corpora for the SVM classifier and the SMT system are vastly different, we consider that the SVM classifier is very robust even in an out-of-domain text.

Hiero has stronger reordering power than the phrase-based SMT with lexicalized reordering, it still cannot overcome the long-distance reordering problem. The translated results (Figure 4) show the effectiveness of our proposed method for resolving the long-distance reordering problem. In the given Chinese sentence, ‘表示(announce)’ and ‘提供(supply)’ consist a predicate-predicate pattern where ‘表示(announce)’ dominates ‘提供(supply)’. In other words, ‘表示(announce)’ belongs to Type 2 described in Section 3. The baseline of phrase-based system even cannot translate both of the predicates into Korean. The baseline of hierarchical phrase-based system only translated ‘提供(supply)’. Both of the general reordering methods translated the predicate ‘提供(supply)’ however not the main predicate. Our proposed method translated both predicate correctly. Though ‘提供(supply)’ is translated as ‘공급(supply)’ in the reference sentence, in Korean ‘제공(supply)’ and ‘공급(supply)’ are synonyms and they mean the same thing.

6 Conclusion

We have presented an effective Chinese syntactic reordering method for the phrase- and hierarchical phrase-based Chinese-to-Korean SMT with an emphasis on predicate-predicate patterns through contrastive analysis of the source and the target languages. We examined the predicate-predicate pat-

<table>
<thead>
<tr>
<th>Table 5: Corpus profile of Dong-A newspaper.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Training</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Sen. length</td>
</tr>
<tr>
<td>Development</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Sen. length</td>
</tr>
<tr>
<td>Test</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Sen. length</td>
</tr>
</tbody>
</table>

⁵http://www.nlp.mibel.cs.tsukuba.ac.jp/bleu_kit/
⁶http://kle.postech.ac.kr:8000/demos/KOMA_KTAG/koma_and_tagger.html
Table 4: BLEU scores under different experimental settings. † mark shows significant improvement over the general syntactic reordering method with the confidence level over 95%, and †† with the confidence level over 99%.

<table>
<thead>
<tr>
<th>Method</th>
<th>Lexicalized reordering</th>
<th>Hiero</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NIST</td>
<td>BLEU (gain)</td>
</tr>
<tr>
<td>Baseline</td>
<td>5.8428</td>
<td>22.19</td>
</tr>
<tr>
<td>General syntactic reordering</td>
<td>6.0288</td>
<td>23.84</td>
</tr>
<tr>
<td>Method 1: PKU dictionary</td>
<td>6.1348</td>
<td>24.26 (+0.42)</td>
</tr>
<tr>
<td>Method 2: SVM Classifier</td>
<td>6.1242</td>
<td>24.73 (+0.89)††</td>
</tr>
</tbody>
</table>

Figure 4: Translated results of Baseline (B), General reordering method(G), and Proposed method (P) with phrase- and hierarchical phrase-based SMT systems.
terns relating to long-distance reordering, and inspected which specific constructions contribute to better translation through syntactic reordering. Useful linguistic knowledge is explored to detect the constructions which need to be reordered. Different experimental settings with different kinds of knowledge were proposed and tested for effectiveness.

Acknowledgments

This work is supported in part by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (MEST) (2009-0075211), in part by the BK 21 project in 2010, and in part by the POSTECH Information Research Laboratories (PIRL) project.

References


