Chinese Syntactic Reordering for Adequate Generation of Korean Verbal Phrases in Chinese-to-Korean SMT

Jin-Ji Li, Jungi Kim, Dong-Il Kim*, and Jong-Hyeok Lee
Department of Computer Science and Engineering,
Electrical and Computer Engineering Division,
Pohang University of Science and Technology (POSTECH),
San 31 Hyoja Dong, Pohang, 790-784, R. of Korea
E-mail: {ljj, yangpa, jhlee}@postech.ac.kr

*Language Engineering Institute,
Department of Computer, Electron and Telecommunication Engineering,
Yanbian University of Science and Technology (YUST),
Yanji, Jilin, 133-000, P.R. of China
E-mail: {dongil}@ybust.edu.cn

Abstract

Chinese and Korean belong to different language families in terms of word-order and morphological typology. Chinese is an SVO and morphologically poor language while Korean is an SOV and morphologically rich one. In Chinese-to-Korean SMT systems, systematic differences between the verbal systems of the two languages make the generation of Korean verbal phrases difficult. To resolve the difficulties, we address two issues in this paper. The first issue is that the verb position is different from the viewpoint of word-order typology. The second is the difficulty of complex morphology generation of Korean verbs from the viewpoint of morphological typology. We propose a Chinese syntactic reordering that is better at generating Korean verbal phrases in Chinese-to-Korean SMT. Specifically, we consider reordering rules targeting Chinese verb phrases (VPs), preposition phrases (PPs), and modality-bearing words that are closely related to Korean verbal phrases. We verify our system with two corpora of different domains. Our proposed approach significantly improves the performance of our system over a baseline phrased-based SMT system. The relative improvements in the two corpora are +9.32% and +5.43%, respectively.

1 Introduction

Recently, there has been a lot of research on encoding syntactic information into statistical machine translation (SMT) systems in various forms and in different stages of translation processes.

During preprocessing source language sentences undergo reordering and morpho-syntactic reconstruction phases to generate more target language-like sentences. Also, fixing erroneous words, generating complex morphology, and reranking translation results in post-processing phases may utilize syntactic information of both source and target languages. A syntax-based SMT system encodes the syntactic information in its translation model of the decoding step.

A number of researchers have proposed syntactic reordering as a preprocessing step (Xia and McCord, 2004; Collins *et al.*, 2005; Wang *et al.*, 2007). In these syntactic reordering approaches, source sentences are first parsed and a series of reordering rules are applied to the parsed trees to reorder the source sentences into target language-like word orders. Such an approach is an effective method for a phrase-based SMT system that employs a relatively simple distortion model in the decoding phase.

This paper concentrates upon reordering source sentences in the preprocessing step of a Chinese-to-Korean phrase-based SMT system using syntactic information. Chinese-to-Korean SMT has more difficulties than the language pairs studied in previous research (French-English, German-English, and Chinese-English). From the viewpoint of language typology, these language pairs are all SVO languages and they have relatively simpler morphological inflections. On the other hand, Korean is an SOV and agglutinative language with relatively free word order and with complex and rich inflections.

For the Chinese-to-Korean SMT, these systematic differences of the two languages make the generation of Korean verbal phrases very difficult. Firstly, the difference in the verb position of the two languages may not be reflected in the simple distortion model of a phrase-based SMT system. Secondly, Morphology generation of

Proceedings of the Fourth Workshop on Statistical Machine Translation , pages 190–196,
Athens, Greece, 30 March – 31 March 2009.

Korean verbs is difficult because of its complexity and the translation direction from a low-inflection language to a high-inflection language.

In the following sections, we describe the characteristics of Korean verbal phrases and their corresponding Chinese verbal phrases, and present a set of hand-written syntactic reordering rules including Chinese verb phrases (VPs), preposition phrases (PPs), and modality-bearing words. In the latter sections, we empirically verify that our reordering rules effectively reposition source words to target language-like order and improve the translation results.

2 Contrastive analysis of Chinese and Korean with a focus on Korean verbal phrase generations

In the Chinese-to-Korean SMT, the basic translation units are morphemes. For Chinese, sentences are segmented into words. As a typical isolating language, each segmented Chinese word is a morpheme. Korean is a highly agglutinative language and an *eojeol* refers to a fully inflected lexical form separated by a space in a sentence. Each *eojeol* in Korean consists of one or more base forms (stem morphemes or content morphemes) and their inflections (function morphemes). Inflections usually include postpositions and verb endings (verb affixes) of verbs and adjectives. These base forms and inflections are grammatical units in Korean, and they are defined as morphemes. As for the translation unit, *eojeol* cause data sparseness problems hence we consider a morpheme as a translation unit for Korean.

As briefly mentioned in the previous section, Chinese and Korean belong to different word-order typologies. The difference of verb position causes the difficulty in generating correct Korean verbal phrases. Also, the complexity of verb affixes in Korean verbs is problematic in SMT systems targeting Korean, especially if the source language is isolated.

In the Dong-A newspaper corpus on which we carry out our experiments in Section 4, Korean function morphemes occupy 41.3% of all Korean morphemes. Verb endings consist of 40.3% of all Korean function words, and the average number of function morphemes inflected by a verb or an adjective is 1.94 while that of other content morphemes is only 0.7.

These statistics indicate that the morphological form of Korean verbal phrases (Korean verbs) [1] are significantly complex. A verbal phrase in Korean consists of a series of verb affixes along with a verb stem. A verb stem cannot be used by itself but should take at least one affix to form a verbal complex. Verb affixes in Korean are ordered in a relative sequence within a verbal complex (Lee, 1991) and express different modality information[2]: tense, aspect, mood, negation, and voice (Figure 1). These five grammatical categories are the major constituents of modal expression in Korean.

K1: **먹(stem)** + 고_ 있(aspect prt.) + 었(aspect prt.) + 었(tense prt.) + 다(*mood* prt.)
E1: had been **eating**

K2. **잡(stem)** + 히(passive prt.) + 었(aspect prt.) + 을_ 수_있(*modality* prt.) + 다(*mood* prt.)
E2: might have been **caught**

Figure 1. Verbal phrases in Korean. Bold-faced content morphemes followed by functional ones with "+" symbols. Prt. is an acronym for particle.

The modality of Korean is expressed intensively by verb affixes. However, Chinese expresses modality using discontinuous morphemes scattered throughout a sentence (Figure 2). Also, the prominence of grammatical categories expressing modality information is different from language to language, and correlations of such categories in a language are also different. The differences between the two languages lead to difficulties in alignment and cause linking obscurities.

C3: 小偷(thief)/可能(might)/被(passive prt.)/警察(police)/**抓(catch)**/了(aspect prt.)/。

K3: 도둑(thief)+은 경찰(police)+에게 **잡(catch)**+히(passive prt.)+었(aspect prt.)+을_수_있(*modality* prt.)+다(*mood* prt.)+./

E3: The thief might have been **caught** by the police.

Figure 2. Underlined morphemes are modality-bearing morphemes in Chinese and Korean sentences. Chinese words are separated by a "/" symbol and Korean *eojeols* by a space.

[1] 'Korean verbal phrase' or 'Korean verbs' in this paper refer to Korean predicates (verbs or adjectives) in a sentence.

[2] Modality system refers to five grammatical categories: tense, aspect, mood (*modality & mood*), negation, and voice. The definition of these categories is described in detail in (Li *et al.*, 2005).

We consider two issues for generating adequate Korean verbal phrases. First is the correct position of verbal phrases, and the second is the generation of verb affixes which convey modality information.

3 Chinese syntactic reordering rules

In this section, we describe a set of manually constructed Chinese syntactic reordering rules.

Chinese sentences are first parsed by Stanford PCFG parser which uses Penn Chinese Treebank as the training corpus (Levy and Manning, 2003). Penn Chinese Treebank adopts 23 tags for phrases (Appendix A). We identified three categories in Chinese that need to be reordered: verb phrases (VPs), preposition phrases (PPs), and modality-bearing words.

3.1 Verb phrases

Korean is a verb-final language, and verb phrase modifiers and complements occur in the pre-verbal positions. However, in Chinese, verb phrase modifiers occur in the pre-verbal or post-verbal positions, and complements mostly occur in post-verbal positions.

We move the verb phrase modifiers and complements located before the verbal heads to the post-verbal position as demonstrated in the following examples. A verbal head consists of a verb (including verb compound) and an aspect sequence (Xue and Xia, 2000). Therefore, aspect markers such as "了 (perfective prt.)", "着 (durative prt.)", "过(experiential prt.)" positioned immediately after a verb should remain in the relatively same position with the preceding verb. The third one in the example reordering rules shows this case. Mid-sentence punctuations are also considered when constructing the reordering rules.

Examples of reordering rules of VPs[3]:

$VV_0\ NP_1 \rightarrow NP_1\ VV_0$
$VV_0\ IP_1 \rightarrow IP_1\ VV_0$
$VV_0\ AS_1\ NP_2 \rightarrow NP_2\ VV_0\ AS_1$
$VV_0\ PU_1\ IP_2 \rightarrow IP_2\ PU_1\ VV_0$

Original parse tree:
VP
  PP (P 按)
    NP (NN 需要)
  PP (P 对)
    NP (PN 它们)
  VP (VV 进行)
    NP (NN 配置)

Reordered parse tree:
VP
  PP (P 按)
    NP (NN 需要)
  PP (P 对)
    NP (PN 它们)
  NP (NN 配置)
    VP (VV 进行)

3.2 Preposition phrases

Chinese prepositions originate from verbs, and they preserve the characteristics of verbs. Chinese prepositions are translated into Korean verbs, other content words, or particles. We only consider the Chinese prepositions that translate into verbs and other content words. We swap the prepositions with their objects as demonstrated in the following examples.

Examples of reordering rules of PPs:

Case 1: translate into Korean verbs
$P(按)_0\ NP_1 \rightarrow NP_1\ P(按)_0$
$P(通过)_0\ IP_1 \rightarrow IP_1\ P(通过)_0$
$P(除了)_0\ LCP_1 \rightarrow LCP_1\ P(除了)_0$

Case 2: translate into other content words
$P(由于)_0\ IP_1 \rightarrow IP_1\ P(由于)_0$
$P(因为)_0\ NP_1 \rightarrow NP_1\ P(因为)_0$

Original parse tree:
VP
  PP (P 按)
    NP (NN 需要)
  PP (P 对)
    NP (PN 它们)
  VP (VV 进行)
    NP (NN 配置)

Reordered parse tree:
VP
  NP (NN 需要)
    PP (P 按)
  PP (P 对)
    NP (PN 它们)
  VP (VV 进行)
    NP (NN 配置)

[3] VV: common verb; AS: aspect marker; P: preposition; PU: punctuation; PN: pronoun;

3.3 Modality-bearing words

Verb affixes in Korean verbal phrases indicate modality information such as tense, aspect, mood, negation, and voice. The corresponding modality information is implicitly or explicitly expressed in Chinese. It is important to figure out what features are used to represent modality information. Li *et al*. (2008) describes in detail the features in Chinese that express modality information. However, since only lexical features can be reordered, we consider explicit modality features only.

Modality-bearing words are scattered over an entire sentence. We move them near their verbal heads because their correspondences in Korean sentences are always placed right after their verbs.

When constructing reordering rules, we consider temporal adverbs, auxiliary verbs, negation particles, and aspect particles only. The following example sentences show the results of a few of our reordering rules for modality-bearing words.

Examples of reordering rules of modality-bearing words:

Original parse tree:

VP
  ADVP (AD 将) ← Temporal adverb
    PP (P 在)
      LCP
        NP (NN 法律) (NN 许可) (NN 范围)
        (LC 内)
    VP (VV 受到)
      NP (NN 起诉)

Reordered parse tree:

VP
  PP (P 在)
    LCP
      NP (NN 法律) (NN 许可) (NN 范围)
      (LC 内)
  ADVP (AD 将)
  VP (VV 受到)
    NP (NN 起诉)

Original parse tree:

VP (VV 要) ← Auxiliary verb
  VP
    PP (P 从)
      LCP
        NP (NN 文件) (NN 组)
        (LC 中)
    VP (VV 排除)

Reordered parse tree:

VP
  PP (P 从)
    LCP
      NP (NN 文件) (NN 组)
      (LC 中)
  VP (VV 要)
  VP (VV 排除)

Original parse tree:

VP
  ADVP (AD 不) ← Negation particle
  VP (VV 应该) ← Auxiliary verb
  VP
    PP (P 以)
      NP (NN 管理员) (NN 身份)
    VP (VV 运行)

Reordered parse tree:

VP
  PP (P 以)
    NP (NN 管理员) (NN 身份)
  ADVP (AD 不)
  VP (VV 应该)
  VP (VV 运行)

Generally speaking, Chinese does not have grammatical forms for voice. Although, voice is also a grammatical category expressing modality information, we have left it out of the current phase of our experiment since voice detection is another research issue and reordering rules for voice are unavoidably complicated.

4 Experiment

Our baseline system is a popular phrase-based SMT system, Moses (Koehn *et al.*, 2007), with 5-gram SRILM language model (Stolcke, 2002), tuned with *Minimum Error Training* (Och, 2003). We adopt NIST (NIST, 2002) and BLEU (Papineni *et al*., 2001) as our evaluation metrics.

Chinese sentences in training and test corpora are first parsed and are applied a series of syntactic reordering rules. To evaluate the contribution of the three categories of syntactic reordering rules, we perform the experiments applying each category independently. Experiments of various combinations are also carried out.

4.1 Corpus profile

We automatically collected and constructed a sentence-aligned parallel corpus from the online

Dong-A newspaper[4]. Strictly speaking, it is a non-literally translated Korean-to-Chinese corpus. The other corpus is provided by MSRA (Microsoft Research Asia). It is a Chinese-Korean-English trilingual corpus of technical manuals and a literally translated corpus.

Chinese sentences are segmented by Stanford Chinese word segmenter (Tseng *et al*., 2005), and parsed by Stanford Chinese parser (Levy and Manning, 2003). Korean sentences are segmented into morphemes by an in-house morphological analyzer.

The detailed corpus profiles are displayed in Table 1 and 2. The Dong-A newspaper corpus is much longer than the MSRA technical manual corpus. In Korean, we report the length of content and function words.

	Training (99,226 sentences)		
	Chinese	Korean	
		Content	Function
# of words	2,692,474	1,859,105	1,277,756
# of singletons	78,326	67,070	514
avg. sen. length	27.13	18.74	12.88
	Development (500 sentences)		
	Chinese	Korean	
		Content	Function
# of words	14,485	9,863	6,875
# of singletons	4,029	4,166	163
avg. sen. length	28.97	19.73	13.75
	Test (500 sentences)		
	Chinese	Korean	
		Content	Function
# of words	14,657	10,049	6,980
# of singletons	4,027	4,217	164
avg. sen. length	29.31	20.10	13.96

Table 1. Corpus profile of Dong-A newspaper.

	Training (29,754 sentences)		
	Chinese	Korean	
		Content	Function
# of words	425,023	316,289	207,909
# of singletons	5,746	4,689	197
avg. sen. length	14.29	10.63	6.99
	Development (500 sentences)		
	Chinese	Korean	
		Content	Function
# of words	6,380	4,853	3,214
# of singletons	1,174	975	93
avg. sen. length	12.76	9.71	6.43
	Test (500 sentences)		
	Chinese	Korean	
		Content	Function
# of words	7,451	5,336	3,548
# of singletons	1,182	964	99
avg. sen. length	14.90	10.67	7.10

Table 2. Corpus profile of MSRA technical manual.

[4] http://www.donga.com/news/ (Korean) and http://chinese.donga.com/gb/index.html (Chinese)

4.2 Result and discussion

The experimental results are displayed in Table 3 and 4. Besides assessing the effectiveness of each reordering category, we test various combinations of the three categories.

Method	**NIST**	**BLEU**
Baseline	5.7801	20.49
Reorder.VP	5.8402	22.12 (+7.96%)
Reorder.PP	5.7773	20.10 (-1.90%)
Reorder.Modality	5.7682	20.93 (+2.15%)
Reorder.VP+PP	5.8176	21.96 (+7.17%)
Reorder.VP+Modality	5.9198	22.24 (+8.54%)
Reorder.All	*5.9361*	*22.40 (+9.32%)*

Table 3. Experimental results on the Dong-A newspaper corpus.

Method	**NIST**	**BLEU**
Baseline	7.2596	44.03
Reorder.VP	7.2238	44.57 (+1.23%)
Reorder.PP	7.2793	44.22 (+0.43%)
Reorder.Modality	7.3110	44.25 (+0.50%)
Reorder.VP+PP	7.3401	45.28 (+2.84%)
Reorder.VP+Modality	*7.4246*	*46.42 (+5.43%)*
Reorder.All	7.3849	46.33 (+5.22%)

Table 4. Experimental results on the MSRA technical manual corpus.

From the experimental result of the Dong-A newspaper corpus, we find that the most effective category is the reordering rules of VPs. When the VP reordering rules are combined with the modality ones, the performance is even better. The gain of BLEU is not significant, but the gain of NIST is significant from 5.8402 to 5.9198. The PP reordering rules do not contribute to the performance when they are singly applied. However, when combined with the other two categories, they contribute to the performance. The best performance is achieved when all three categories' reordering rules are applied and the relative improvement is +9.32% over the baseline system.

In the MSRA corpus, the performance of various combinations of the three categories is better than those of the individual categories. The PP category shows improvement when it is combined with the VP category. The combination of VP and modality category improves the performance by +5.43% over the baseline.

These results agree with our expectations: resolving the word order and modality expression differences of verbal phrases between Chinese and Korean is an effective approach.

4.3 Error Analysis

We adopt an error analysis method proposed by Vilar *et al*. (2006). They presented a framework for classifying error types of SMT systems. (Appendix B.)

Since our approach focuses on verbal phrase differences between Chinese and Korean, we carry out the error analysis only on the verbal heads. Three types of errors are considered: word order, missing words, and incorrect words. We further classify the incorrect words category into two sub-categories: wrong lexical choice/extra word, and incorrect form of modality information. 50 sentences are selected from each test corpus on which to perform the error analysis. For each corpus, we choose the best system: Reorder.All for the Dong-A corpus and Reorder.VP+modality for the MSRA corpus.

The most frequent error type is wrong word order in both corpora. When a verb without any modality information appears in a wrong position, we only count it as a wrong word order but not as a wrong modality. Therefore, the number of wrong modalities is not as frequent as it should be.

Table 5 and 6 indicate that our proposed method helps improve the SMT system to reduce the number of error types related to verbal phrases.

Error type	Frequency	
	Baseline	Reorder.All
wrong word order	34	7
missing content word	18	5
wrong lexical choice/ extra word	6	1
wrong modality	10	6

Table 5. Error analysis of the Dong-A newspaper corpus.

Error type	Frequency	
	Baseline	Reorder. VP+Modality
wrong word order	19	11
missing content word	4	2
wrong lexical choice/ extra word	8	3
wrong modality	11	6

Table 6. Error analysis of the MSRA technical manual corpus.

5 Conclusion and future work

In this paper, we proposed a Chinese syntactic reordering more suitable to adequately generate Korean verbal phrases in Chinese-to-Korean SMT. Specifically, we considered reordering rules targeting Chinese VPs, PPs, and modality-bearing words that are closely related to Korean verbal phrases.

Through a contrastive analysis between the two languages, we first showed the difficulty of generating Korean verbal phrases when translating from a morphologically poor language, Chinese. Then, we proposed a set of syntactic reordering rules to reorder Chinese sentences into a more Korean like word order.

We conducted several experiments to assess the contributions of our method. The reordering of VPs is the most effective, and improves the performance even more when combined with the reordering rules of modality-bearing words. Applied to the Dong-A newspaper corpus and the MSRA technical manual corpus, our proposed approach improved the baseline systems by 9.32% and 5.43%, respectively. We also performed error analysis with a focus on verbal phrases. Our approach effectively decreased the size of all errors.

There remain several issues as possible future work. We only considered the explicit modality features and relocated them near the verbal heads. In the future, we may improve our system by extracting implicit modality features.

In addition to generating verbal phrases, there is the more general issue of generating complex morphology in SMT systems targeting Korean, such as generating Korean case markers. There are several previous studies on this topic (Minkov *et al*., 2007; Toutanova *et al.*, 2008). This issue will also be the focus of our future work in both the phrase- and syntax-based SMT frameworks.

Acknowledgments

This work was supported in part by MKE & IITA through the IT Leading R&D Support Project and also in part by the BK 21 Project in 2009.

References

Charles N. Li, and Sandra A. Thompson 1996. *Mandarin Chinese: A functional reference grammar*, University of California Press, USA.

David Vilar, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. 2006. *Error Analysis of Statistical*

Machine Translation Output. In Proceedings of LREC.

Einat Minkov, Kristina Toutanova, and Hisami Suzuki. 2007. *Generating Complex Morphology for Machine Translation*. In Proceedings of ACL.

Fei Xia and Michael McCord. 2004. *Improving a statistical MT system with automatically learned rewrite patterns.* In Proceedings of COLING.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky and Christopher Manning. 2005. *A Conditional Random Field Word Segmenter.* In Fourth SIGHAN Workshop on Chinese Language Processing.

HyoSang Lee 1991. *Tense, aspect, and modality: A discourse-pragmatic analysis of verbal affixes in Korean from a typological perspective*, PhD thesis, Univ. of California, Los Angeles.

Jin-Ji Li, Ji-Eun Roh, Dong-Il Kimand Jong-Hyeok Lee. 2005. *Contrastive Analysis and Feature Selection for Korean Modal Expression in Chinese-Korean Machine Translation System*. International Journal of Computer Processing of Oriental Languages, 18(3), 227--242.

Jin-Ji Li, Dong-Il Kim and Jong-Hyeok Lee. 2008. *Annotation Guidelines for Chinese-Korean Word Alignment*. In Proceedings of LREC.

Kristina Toutanova, Hisami Suzuki, and Achim Puopp. 2008. *Applying Morphology Generation Models to Machine Translation*. In Proceedings of ACL.

Nianwen Xue, and Fei Xia. 2000. *The bracketing guidelines for the Penn Chinese Treebank (3.0)*. IRCS technical report, University of Pennsylvania.

Nianwen Xue, Fei Xia, Fu-Dong Chiou, and Martha Palmer. 2005. *The Penn Chinese Treebank: Phrase structure annotation of a large corpus*. Natural Language Engineering, 11(2):207–238.

Michael Collins, Philipp Koehn, and Ivona Kuˇcerov´a. 2005. *Clause restructuring for statistical machine translation.* In Proceedings of ACL, pages 531–540.

NIST. 2002. *Automatic evaluation of machine translation quality using n-gram co-occurrence statistics*.

Och, F. J. 2003. *Minimum error rate training in statistical machine translation*. In Proceedings of ACL.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris-Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constrantin, and Evan Herbst. 2007. *Moses: Open source toolkit for statistical machine translation*. In Proceedings of ACL, Demonstration Session.

Roger Levy and Christopher D. Manning. 2003. *Is it harder to parse Chinese, or the Chinese Treebank?* In Proceedings of ACL.

Stolcke, A. 2002. *SRILM - an extensible language modeling toolkit*. In Proceedings of ICSLP, 2:901-904.

Appendix A. Tag for phrases in Penn Chinese Treebank.

ADJP	adjective phrase
ADVP	adverbial phrase headed by AD (adverb)
CLP	classifier phrase
CP	clause headed by C (complementizer)
DNP	phrase formed by "XP+DEG"
DP	determiner phrase
DVP	phrase formed by "XP+DEV"
FRAG	fragment
IP	simple clause headed by I (INFL)
LCP	phrase formed by "XP+LC"
LST	list marker
NP	noun phrase
PP	preposition phrase
PRN	parenthetical
QP	quantifier phrase
UCP	unidentical coordination phrase
VP	verb phrase

Appendix B. Classification of translation errors proposed by Vilar *et al*. (2006).