Transferring Syntactic Relations of Subject-Verb-Object Pattern 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

Since most Korean postpositions signal grammatical functions such as syntactic relations, generation of incorrect Korean postpositions results in producing ungrammatical outputs in machine translations targeting Korean. Chinese and Korean belong to morpho-syntactically divergent language pairs, and usually Korean postpositions do not have their counterparts in Chinese. In this paper, we propose a preprocessing method for a statistical MT system that generates more adequate Korean postpositions. We transfer syntactic relations of subject-verb-object patterns in Chinese sentences and enrich them with transferred syntactic relations in order to reduce the morpho-syntactic differences. The effectiveness of our proposed method is measured with lexical units of various granularities. Human evaluation also suggest improvements over previous methods, which are consistent with the result of the automatic evaluation.

1 Introduction

Translating from a morphologically poor language to a morphologically rich one is more difficult than the opposite case (Koehn, 2005). If the source language is a morphologically poor language, surface words only cannot provide sufficient linguistic clues to generate the complex morphology needed for the morphologically rich target language in a Statistical Machine Translation (SMT) system. Chinese and Korean are a morpho-syntactically divergent language pair and to generate adequate Korean postpositions is a challenging task in Chinese-to-Korean SMT. Wrong postposition generation leads to ungrammatical output sentences because most postpositions indicate grammatical relations in Korean.

This paper describes a method for transferring the syntactic relations of subject-verb-object (SVO) patterns, and enriching the Chinese sentences by inserting the corresponding transferred relations as pseudo words. The SVO pattern refers to a predicate with immediate children that have a *subject* or an *object* relation in a dependency tree.

Specifically, we adopt grammatical relations that are produced by Stanford Chinese typed dependency parser (Levy and Manning, 2003; Chang et al., 2009). The previous work provides the following 7 grammatical relations that are related to *subject* and *object* relation: *nsubj*, *xsubj*, *nsubjpass*, *top*, *dobj*, *range*, and *attr*.[1] In this paper, the SVO pattern is a general term which represents constructions that consist of any number of above 7 grammatical relations with a corresponding head predicate.

SVO patterns frequently occur in Chinese dependency trees and cause incorrect postposition generations when they are translated into Korean. Our proposed method has the following characteristics. First, since Korean postpositions indicate grammatical functions such as syntactic relations, transferring the syntactic relations is identical to resolving the structural transfer ambiguities when translating. Second, by inserting the transferred syntactic relations as pseudo words, the Chinese sentences be-

[1] *nsubj*: nominal subject; *xsubj*: controlling subject; *nsubjpass*: nominal passive subject; *top*: topic; *dobj*: direct object; *range*:dative object that is a quantifier phrase; *attr*: attributive (complement of a copular verb).

come more morpho-syntactically similar to Korean sentences.

We convert this transfer task into a structured prediction one, for which we train and tune using the bilingual corpus provided for the SMT system. Though we use language-specific knowledge in our experiment, the framework of supplementing the source language with morpho-syntactic knowledge from the target language is applicable to other language pairs that suffer from the same issue.

We analyze and contrast the morpho-syntactic differences between Chinese and Korean in Section 2. Related work is given in Section 3. Section 4 describes our proposed method which has three independent components. The experimental results and discussion are given in Section 5.

2 Morpho-syntax of Chinese and Korean

Chinese is a typical isolating language and has few functional markers that signal the grammatical functions such as syntactic relations. In Chinese, these grammatical functions are generally expressed by means of word order and prepositions (Li and Thompson, 1989). Syntactic relations such as *subject* and *object* are expressed by word order only, and *adverb* mostly by prepositions. On the other hand, Korean is a highly agglutinative language with rich functional morphemes such as postpositions and verbal endings. Korean postpositions include case markers, auxiliary particles, and conjunctive particles. Most of the case markers are utilized to signal the grammatical relations of the complement Noun Phrase(NP) and its corresponding predicate. In our training corpus, there are 290 unique postpositions. Among them, 79 are case markers. Japanese, which belongs to the same language family as Korean, has only 18 case markers(Toutanova and Suzuki, 2007). As Korean postpositions are quite diverse and indicate the syntactic relations in a sentence, correct postposition generation directly leads to producing grammatical sentences in SMT systems.

The basic translation units in Chinese-to-Korean SMT are usually morphemes. In Chinese, the sentences are segmented into words, and each segmented word is a morpheme. In Korean, *eojeol* (similar to bunsetsu in Japanese) is a fully inflected lexical form separated by a space in a sentence. Each *eojeol* consists of one or more base forms (content morphemes) and inflections (functional morphemes, postpositions or verbal endings). *Eojeol* easily causes data sparseness problems and we have to consider a morpheme as a translation unit for Korean. In our corpus, each *eojeol* contains 2.2 morphemes on average.

3 Related work

Recently, a number of researchers have studied complex morphology generation in SMT systems where the translation direction is from a morphologically-poor language to a morphologically-rich one.

Avramidis and Koehn (2008) proposed a method that extracts information from the syntax of source sentences to enrich the morphologically poor language using the framework of factored SMT. Also, Ramanathan et al. (2009) adopted factored models to factorize syntactic/semantic relations and suffixes to help generate inflections and case markers. Factored models can tightly combine linguistic features into the decoding phase, while expanding the search space at the same time.

Some researchers have tried to develop independent components to handle complex morphology generation. This kind of research has the advantage that it does not introduce any other complexity to the SMT decoder. Toutanova and Suzuki (2007), Toutanova et al. (2008) and Minkov et al. (2007) suggested postprocessing models that predict inflected word forms utilizing morpho-syntactic information from both source and target sentences. The inflection prediction model chooses the correct inflections of given target language stems.

Hong et al. (2009) proposed bridging morpho-syntactic gaps as a preprocessing to an English-to-Korean SMT system. They utilized a set of syntactic relations from source sentences and directly inserted them as pseudo words to generate intermediate sentences. The main aim of their work was to decrease the null alignments of Korean functional morphemes, and as a result to generate appropriate functional words. However, this method only considers the syntax of source sentences, and therefore it cannot sufficiently reflect the structural differences between the source and target sentences.

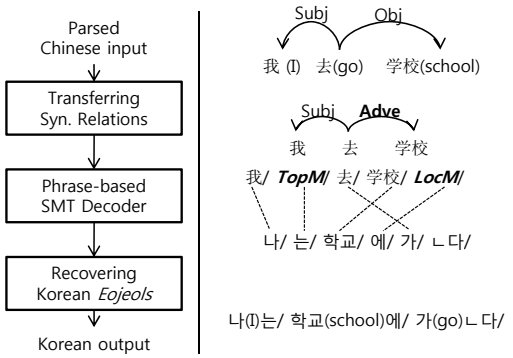


Figure 1: System architecture of the proposed method. *TopM* is a pseudo word representing a topic marker, and *LocM* a locative case marker.

4 Proposed method

In this paper, we propose a preprocessing method that not only transfers the syntactic relations, but also enriches the source sentences by inserting the transferred relations as pseudo words. The transfer phase is realized with a structured prediction model using automatically annotated training data. It is similar to the transfer phase of a traditional transfer-based machine translation but without the lexical transfer. We leave the lexical transfer to the SMT decoder which is one of the greatest strengths of a phrase-based SMT system. Finally, the output of the Korean morphemes is recovered as Korean *eojeols*. Figure 1 shows the system architecture and we will explain each module in detail in the following sections.

4.1 Transferring syntactic relations of SVO patterns

SVO patterns are basic and frequently occurring patterns (3 patterns per sentence in our training corpus) in Chinese sentences which retain structural transfer ambiguities when they are translated into Korean. SVO patterns can be transferred into various structures such as subject-adverb-verb, adverb-subject-verb, and adverb-object-verb. Words with *subject* and *object* relations to a predicate are strong candidates of complements. However, without explicit functional markers, it is difficult to correctly translate the patterns. We convert this transfer problem into a structured prediction one, and train a prediction model using a word-aligned bilingual corpus.

Further structural transfer such as syntactic reordering could be conducted; however we only transfer the syntactic relations to investigate the effectiveness of the proposed method more precisely.

Table 1: 7 representative Korean postposition categories in our structured prediction model.

Korean syntactic relation	Corresponding Postposition Category
Subject	Nominative case marker (Topic marker)
Object	Accusative case marker
Adverb	Dative case marker Locative case marker Instrumental case marker Quotative case marker Collaborative case marker

4.1.1 Task description

Given an SVO pattern, the transfer module predicts a value for each syntactic relation from a set of representative postpositions in Korean. Since Korean postpositions indicate grammatical relations, this process is identical to resolving the ambiguity of the SVO pattern when translating into Korean.

As mentioned earlier, Korean postpositions have great diversity. However, linguists usually consider the case markers listed in Table 1 and genitive case markers in Korean sentence generation. In our task, we exclude the genitive case marker because Chinese *subject* and *object* cannot be transferred into the genitive relation of a verb. We also include topic markers with the *subject* relation because Chinese is a topic-prominent language. These postpositions cover over 80% of overall usage of Korean postpositions in our corpus. Finally, we include a 'null' category in our prediction model. The 'null' category indicates that the words with a *subject* or *object* relation in Chinese are translated into Korean content words without any postposition, or with other postpositions not listed in Table 1, or with verbal endings.

Because *subject* and *object* relations are mutually constrained when transferred into Korean, we build a structured prediction model using conditional random fields (CRF) for this task rather than to transfer each syntactic relation independently. Instances in each SVO pattern are predicted as a sequential la-

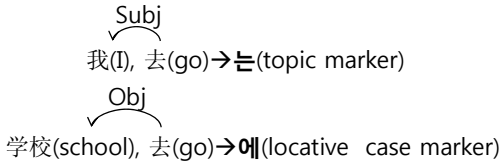
beling. CrfSgd toolkit[2] is used to construct our classifier.

4.1.2 Training data construction

The training data for the prediction model is automatically constructed using a word-aligned, POS tagged, and dependency parsed Chinese-Korean bilingual corpus. We extracted all SVO patterns from every Chinese parse tree. For each word with a *subject* or *object* relation, a Korean postposition category is assigned via the word-alignment information. If the counterpart Korean word is a content morpheme, then we find the Korean *eojeol* that contains it, and the corresponding postposition that the *eojeol* contains.

We assign each postposition to one of the postposition categories, using the dependency relation of *eojeol* that the postposition is part of. *subject* and *object* relations are mapped to nominative and accusative case markers. For *adverb* relations, if the postposition matches one of the five adverbial case markers, then we assign the category to the corresponding subject-verb or object-verb instance in Chinese. Considering the assignment precision, we use the intersection of the bidirectional word alignments by GIZA++ (Och, 2000).

For example, the sentence in Figure 1 will be annotated as follows. The first instance shows that the *subject* remains as *subject* relation in Korean; while there is a transfer in the second instance from *object* to *adverb* relation.



Some words with a *subject* or *object* relation cannot be mapped to a Korean postposition because there is no word alignment link. In this case, we also tag it with the 'null' category. Finally, we extract 295,589 SVO patterns as training data, and 1,609 patterns as test data.

4.1.3 Feature engineering

Table 2 shows the detailed features used in the prediction model. Besides the lexical, POS, and syntactic information, we adopted semantic features as well. The semantic classes are obtained from a Chinese thesaurus (*CiLin*) (Mei et al., 1984). *CiLin* is a conceptual hierarchy with 5 levels. Because of the data sparseness problem, we use up to level-2 tags. When a Chinese word maps to several semantic classes, we choose the most frequently used one as its semantic class.

Table 2: Feature description for CRF classifier.

Feature	Description
LEX_c	Surface form of a word with *subject/object* relation
LEX_h	Surface form of a head verb
POS_c	Part of Speech of a word with *subject/object* relation
POS_h	Part of Speech of a head verb
SEM_c	Semantic class of a word with *subject/object* relation
SEM_h	Semantic class of a head verb
SYN	Grammatical relation of Chinese dependency

Table 3: Feature template of combination and context features for CRF classifier.

Combination feature for current position i	
$LEX_c/LEX_h/SYN$	$POS_c/LEX_h/SYN$
$LEX_c/POS_h/SYN$	$POS_c/POS_h/SYN$
$LEX_c/SEM_h/SYN$	$POS_c/SEM_h/SYN$
$POS_c/SEM_c/LEX_h/SYN$	$LEX_c/LEX_h/POS_h/SYN$
$POS_c/SEM_c/POS_h/SYN$	$LEX_c/POS_h/POS_h/SYN$
$POS_c/SEM_c/SEM_h/SYN$	$LEX_c/SEM_h/POS_h/SYN$
Context feature for current position i	
$SYN_{i-1}, SYN_i, SYN_{i+1}$	

Using the above features, we further construct a feature template for the classifier (Table 3). For combination features, we use features of subject/verb and object/verb pairs, and for the context feature, we refer to the syntactic relations of neighboring instances with a window size of 3. With the proposed feature template, 144,935 features are extracted.

4.1.4 Result and discussion

We tested the performance of the prediction model using the test corpus (500 sentences). The distribution of the 8 Korean postposition categories is shown in Table 4.

[2]http://leon.bottou.org/projects/sgd, Parameter setting is as follows. c=1.2; f=3; r=40.

Table 4: The distribution of 8 Korean postposition categories automatically annotated using word alignment result (intersection).

Syn. Rel.	Corresponding Postposition Category	Freq.	Ratio (%)
Subject	1.Nominative case marker (Topic marker)	403	18.8
Object	2.Accusative case marker	286	13.3
Adverb	3.Dative case marker	69	3.2
	4.Locative case marker	18	0.8
	5.Instrumental case marker	47	2.2
	6.Quotative case marker	15	0.7
	7.Collaborative case marker	1	0.0
	8. 'null'	1,305	60.9
	Total	**2,144**	**100.0**

Table 5: The confusion matrix for Korean postposition prediction. P: columns show the distribution of prediction results. C: rows show the real distribution (correct answers). #1∼ #8 indicates the corresponding postposition categories shown in Table 4.

P	C							
	#1	#2	#3	#4	#5	#6	#7	#8
#1	**160**	17	5	2	1	1	0	113
#2	25	**87**	2	1	1	0	0	72
#3	0	0	**8**	0	0	0	0	3
#4	0	1	0	**4**	0	0	0	2
#5	1	1	1	2	**7**	0	0	3
#6	0	0	0	0	0	**0**	0	0
#7	0	0	0	0	0	0	**0**	0
#8	217	180	53	9	38	14	1	**1,112**
Total	403	286	69	18	47	15	1	**1,305**
Acc.(%)	39.7	30.4	11.6	22.2	14.9	0.0	0.0	85.2

The accuracy of our proposed method is 64.3%. Considering that the 'null' category occupies 60.9% (Table 4), the prediction accuracy only slightly improved (3.4%) from the *'null'* as the default category classifier. Intuitively, 64.3% is a quite low accuracy that may lead to much noise in translation. However, the small gain in accuracy improve the overall SMT performance significantly (Section 5.2). We address how SMT performance improves by analyzing the confusion matrix of prediction results in Table 5.

1. Since the majority category is 'null' in our training corpus, the classification system has a tendency to predict the 'null' category. 512 instances (23.9%) are misclassified as the 'null' category. However, it is better to predict as 'null' than other incorrect categories because the 'null' category includes uncertain instances, in which case no pseudo word is generated.

2. The last column shows the 193 (113+72+3+2+3+0+0=193) instances of the 'null' category misclassified as other postposition categories. Among 113 instances predicting as nominative case markers, the grammatical relations in 96 instances are *subject* in Chinese. Since most of *subject* and *object* retain the same grammatical relation when translating from Chinese to Korean, this kind of prediction error cannot be considered as a fatal error. For 72 instances which are predicated as accusative case markers, there are 65 instances of the grammatical relation in Chinese as *object*.

Moreover, since the training data is automatically constructed using the word alignment result, it contains incorrect instances which influence the prediction results.

4.2 Phrase-based SMT system

We construct a phrase-based SMT system with the modified Chinese-Korean bilingual corpus. The Chinese training corpus is converted into an intermediate language enriched with 7 Korean postposition categories using the same algorithm mentioned in Section 4.1.2. For the 'null' category, we do not insert any pseudo word.

In order to evaluate the oracle performance of the SMT system, we also transform the test corpus (which needs to be translated for the evaluation of the SMT system) using the word alignment information. In other words, we assume the word alignment information is provided for the test corpus.

The oracle system performance is 24.56 in morpheme-BLEU and this will be the upper bound of our proposed method. The baseline is 22.19 morpheme-BLEU using the original Chinese-Korean bilingual corpus. This suggests that there is much room for improvement using the proposed method.

4.3 Recovering Korean *eojeols*

The output of the SMT system is Korean morphemes. In most SMT systems, only morpheme-BLEU is reported where the target language is

a morphologically rich language such as Korean. However, *eojeol* is the basic lexical unit in Korean and contains functional markers; *eojeol*-BLEU provides more meaningful evaluation than morpheme-BLEU.

In order to recover *eojeol*, we first omit all the spaces in the Korean output, and then re-segment it into *eojeols*. The segmentation problem can be resolved by a CRF model as a sequence labeling problem. We adopt BIO[3] tags for this segmentation problem, and utilize up to character trigram features with a window size 5. Korean sentences in the training corpus for the SMT system are used to model detecting Korean *eojeol* boundaries. The segmentation accuracy is 97.7% in the test corpus.

5 Experiment

Our baseline system is Moses, a state-of-the-art phrase-based SMT system (Koehn et al., 2007), with 5-gram SRI language modeling (Stolcke, 2002) tuned with Minimum Error Rate Training (MERT) (Och, 2003). We adopted NIST (Doddington, 2002) and BLEU (Papineni et al., 2001) as our evaluation metrics.[4] Also, a significance test was conducted using a paired bootstrap resampling method (Koehn, 2004).[5]

Chinese sentences in the test corpus were first parsed, and the syntactic relations of SVO patterns were transferred as preprocessing. The enriched Chinese sentences with transferred syntactic relations were translated by the SMT system as described in Section 4.2. Finally, the output of the Korean morphemes was recovered as Korean *eojeols*.

5.1 Corpus profile

We automatically collected and manually aligned a parallel corpus from the Dong-A newspaper.[6] Strictly speaking, 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. For Korean, we reported the length of content and function words separately (Table 6).

Table 6: Corpus profile of Dong-A newspaper.

Training (98,671 sent.)	Chinese	Korean	
		Content	Function
# words	2,673,422	1,848,798	1,271,668
# singletons	78,243	66,872	510
Sen. length	27.09	18.74	12.89
Development (500 sent.)	Chinese	Korean	
		Content	Function
# words	14,452	9,863	6,875
# singletons	4,012	4,166	162
Sen. length	28.90	19.73	13.75
Test (500 sent.)	Chinese	Korean	
		Content	Function
# words	14,619	9,997	6,928
# singletons	4,009	4,229	154
Sen. length	29.24	19.99	13.86

We used the Stanford Chinese typed dependency parser to parse the Chinese sentences. The Korean sentences were segmented into morphemes and dependency-parsed using an in-house morphological analyzer and an in-house dependency parser.[7]

5.2 Result and discussion

The proposed method shows significant improvements compared to the baseline phrase-based Chinese-Korean SMT system using *eojeol*-BLEU (Table 7). There are about 2.2 Korean morphemes in each *eojeol*; hence even bigram *eojeol*-BLEU is meaningful for performance evaluations. We tested the proposed method with 4-gram morpheme-BLEU and bigram *eojeol*-BLEU.

Hong et al. (2009)'s method is also a preprocessing method which enriches the source language with the syntactic relations as pseudo words. For comparison purposes, we carried out their proposed method by inserting the Chinese syntactic relations as pseudo words. This method did not show significant improvements when using both 4-gram morpheme-BLEU and bigram *eojeol*-BLEU.

As *eojeol* includes both the content and functional morphemes, *eojeol*-BLEU is more suitable for testing whether the output sentence is grammatically

[3] B: current morpheme is the start of an *eojeol*; I: current morpheme is a middle of an *eojeol* ; O: an *eojeol* with single morphemes;

[4] ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v11b.pl

[5] http://www.nlp.mibel.cs.tsukuba.ac.jp/bleu_kit/

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

[7] http://kle.postech.ac.kr:8000/demos/KOMA_KTAG/koma_and_tagger.html and http://kle.postech.ac.kr:8000/demos/KoPA/parser.html

Table 7: Performance of the proposed method. The BLEU performance with † mark show significant improvements over the baseline system with the confidence level over 95%. All the systems conduct the lexicalized reordering.

Method	Morpheme (4-gram)		*Eojeol* (2-gram)	
	NIST	BLEU	NIST	BLEU
Baseline	5.8428	22.19	3.3000	17.63
Hong et al. (2009)'s method	5.9772	22.61	3.3703	18.08
Proposed method	**6.0133**	**22.67**	**3.3992**	**18.36**†

correct or not. Bigram *eojeol*-BLEU showed significant improvements compared with the baseline system.

Although the accuracy of transferring the syntactic relations of SVO patterns is not as high as we expected, and thereby the SMT system suffers from error propagation, the overall performance of the proposed method improved over baseline with statistical significance. A gold standard bilingual corpus would be more helpful to construct an effective transfer module.

5.3 Human evaluation

Since BLEU metric does not always correlate to the human evaluation, we selected 100 sentences on which to perform the human evaluation. The comparison target to our proposed method is Hong et al. (2009)'s method. We adopted the human evaluation measure proposed by Toutanova et al. (2008). Two annotators compared the translation quality in terms of adequacy and fluency (Table 8). The reference translation is given to annotators, but without the source sentence.

The diagonal values in Table 8 show the agreement between two annotators. We further measured the agreement between the annotators using the Kappa statistic. The Kappa value is only 0.320 when considering Hong's (H), Proposed (P), and Equal quality (E) categories. However, excluding the uncertain evaluation result E, the Kappa value is 0.732. This value falls within the scope of a substantial agreement.

Although the morpheme-BLEU of Hong's and our proposed method are similar, *eojeol*-BLEU and human evaluation result suggest that the proposed method is better than theirs.

Korean is a relatively free word-order language, and the postpositions enable such free movements of *eojeols* because they indicate the grammatical relations in a sentence. If correct postpositions are produced, humans will feel the sentence is well translated. Ramanathan et al. (2009) also point out a similar observation from their "experience of large-scale English-Hindi MT, ... [they are] convinced that fluency and fidelity in the Hindi output get an order of magnitude facelift if accurate case marker and suffixes are produced."

Table 8: Human evaluation result comparing Hong et al. (2009) vs. proposed method. H: Hong's method is better; P: proposed method is better; E: equal quality.

Annotator 2	Annotator 1		
	H	P	E
H	**10**	2	4
P	3	**34**	14
E	7	12	**14**

In the first example of translation results (Ch1, Table 9), 负责人(party) is the *subject* of 派往(dispatch), and 日本(Japan) is the *object*. In this SVO pattern, the *object* relation should be transferred into *adverb* (locative case marker). Hong's method produced the *object* case marker '을', while our proposed method correctly generated the locative case marker '에'. The *subject* 负责人(party) should be transferred into *object* relation. Both Hong's and the proposed method do not correctly generate the corresponding case markers. However, Hong's method generated two case markers '는' and '에', which are grouped into one *eojeol*. Since '는 에' is an inexistent case marker, it makes the annotator immediately judge that the translation is ungrammatical.

In the second example (Ch2), 访问团(visitors) and 平壤(Pyongyang) are the *subject* and *object* of 返回(return). Hong's method generated two case markers and they are grouped as '이 에서' which is an ungrammatical complex case marker. Although

Table 9: Translated results of Hong's method (H) and Proposed method (P).

Ch1.	Nasdaq/ 公司(company)/ 打算(scheduled)/ 于(in)/ 20/ 日(day)/ 将(will)/ 有关(related)/ 负责人(party)/ 直接(directly)/ 派往(dispatch)/ 日本(Japan)/ , 正式(officially)/ 发表(announce)/ 撤出(withdraw)/ 的(DE)/ 方针(policy)/ 。
H.	나스닥(Nasdaq)이 20일 관계자(party)**는에** 일본(Japan)**을** 철수(withdraw)/하ㄴ다는 방침(policy)**을** 공식(officially) 발표(announce)하였다.
P.	나스닥(Nasdaq)은 20일 관계자(party)들(pl.)**은** 일본(Japan)**에** 파견(dispatch)하어 철수(withdraw) 방침(policy)**을** 공식(officially) 발표(announce)하였다.
Ref.	나스닥(Nasdaq)은 20일 관계자(party)**를** 직접(directly) 일본(Japan)**에** 보내(send) 철수(withdraw) 방침(policy)**을** 공식(officially) 발표(announce)할 예정(scheduled).
Ch2.	第三/ 次(3rd)/ 离散/ 家属(separated families)/ 访问团(visitors)/ 在/ 经过(passing)/ 三/ 天/ 两/ 夜/(3 days) 短暂(short)/ 的/ 相逢(reunion)/ 之后(after)/ , 在/ 28/ 日/ 各自(respectively)/ 返回(return)/ 汉城(Seoul)/ 和(and)/ 平壤(Pyongyang)/ 。
H.	3차(3rd) 이산가족(separated families) 방문단(visitors)**이에서** 2박3일간(3 days)을 거치(passing)/ㄴ 뒤(after) 상봉 짧(short)은 각자(each)의 28일 서울(Seoul)과 평양(Pyongyang) 귀환(return)하였다.
P.	3차(3rd) 이산가족(separated families) 방문단(visitors)**은** 2박3일간(3 days)의 짧(short)은 상봉(reunion)을 하ㄴ 뒤(after) 28일 각각(respectively) 서울(Seoul)과 평양(Pyongyang) 귀환(return)하기로 하였다.
Ref.	제3차 이산가족(separated families) 교환 방문단(visitors)**이** 2박 3일간(3 days)의 아쉬운(short) 만남(reunion)을 뒤로 한 채 28일 서울(Seoul)과 평양(Pyongyang)**으로** 각각(respectively) 귀환(return)했다.

our proposed method produced '은', which is different from '이' in the reference sentence, '은' is a topic marker and '이' is a subjective case marker in Korean and both of them fall in the category 1 in Table 4. Annotators easily judge that '은' is also a correct generation while the automatic evaluation cannot. In this example, the line morpheme-BLEU of Hong's method is 36.23, which is much higher than that of the proposed method (29.20). However, annotators were in favor of the translation quality of the proposed method.

Both Hong's and the proposed method do not produce the correct case marker '으로' for 平壤(Pyongyang), which is an *object* but should be transferred into *adverb* in Korean. Although our proposed method correctly transferred the syntactic relation and inserted it as a pseudo word, the phrase-based system did not successfully generate the corresponding case marker. This phenomenon may result from the loosely coupled transferred syntactic relations to the translation model. How to effectively conquer this phenomenon will be our future work.

Human evaluation is more sensible to ungrammaticality than the automatic one. Since Korean postpositions represent the grammatical roles, *eojeol*-BLEU is more similar to the human evaluation and is a more appropriate measure than the morpheme-BLEU. Our experimental results sufficiently support our argument in this regard.

6 Conclusion and future work

We have presented a novel method which is effective in generating adequate Korean postpositions. Most Korean postpositions indicate grammatical relations; however they do not have the counterparts in Chinese. We tried to fill in the morpho-syntactic differences between Chinese and Korean, by transferring the syntactic relations of SVO patterns, and using the transferred syntactic results, we further enriched the Chinese sentences. Our proposed method showed significant improvements measured with bigram *eojeol*- BLEU. For comparison purposes, we implemented the previous work and compared the translations through automatic and human evaluations, and we showed that our method is better than the previous approach.

The mechanism of transferring syntactic relations in our framework is similar to that of the traditional transfer phase in transfer-based MT approaches. Therefore it can be combined with a rule-based transfer system. Also, our proposed method trains its prediction model on the bilingual corpus

for an SMT system. Therefore it can be easily applied to other language pairs which suffer from similar linguistic issues.

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

Eleftherios Avramidis and Philipp Koehn. 2008. Enriching morphologically poor languages for statistical machine translation. In *Proceedings of ACL-08: HLT*, pages 763–770, Columbus, Ohio, June. Association for Computational Linguistics.

Pi-Chuan Chang, Huihsin Tseng, Dan Jurafsky, and Christopher D. Manning. 2009. Discriminative reordering with Chinese grammatical relations features. In *Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation (SSST-3) at NAACL HLT 2009*, pages 51–59, Boulder, Colorado, June. Association for Computational Linguistics.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Gumwon Hong, Seung-Wook Lee, and Hae-Chang Rim. 2009. Bridging morpho-syntactic gap between source and target sentences for english-korean statistical machine translation. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 233–236, Suntec, Singapore, August. Association for Computational Linguistics.

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 Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain, July. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit X, the tenth machine translation summit*, pages 79–86, Phuket, Thailand.

Roger Levy and Christopher D. Manning. 2003. Is it harder to parse chinese, or the chinese treebank? In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 439–446, Sapporo, Japan, July. Association for Computational Linguistics.

Charles N. Li and Sandra A. Thompson. 1989. *Mandarin Chinese: A functional reference grammar*. University of California Press, USA.

Jia-ju Mei, Yi-Ming Zheng, Yun-Qi Gao, and Hung-Xiang Yin. 1984. *TongYiCi CiLin*. Shanghai: the Commercial Press.

Einat Minkov, Kristina Toutanova, and Hisami Suzuki. 2007. Generating complex morphology for machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 128–135, Prague, Czech Republic, June. Association for Computational Linguistics.

Franz Josef Och. 2000. Giza++: Training of statistical translation models. http://www.fjoch.com/GIZA++.html.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan, July. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. Technical report, Research report RC22176, IBM.

Ananthakrishnan Ramanathan, Hansraj Choudhary, Avishek Ghosh, and Pushpak Bhattacharyya. 2009. Case markers and morphology: Addressing the crux of the fluency problem in english-hindi smt. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 800–808, Suntec, Singapore, August. Association for Computational Linguistics.

Andreas Stolcke. 2002. Srilm—an extensible language modeling toolkit. In *In Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002*, pages 901–904.

Kristina Toutanova and Hisami Suzuki. 2007. Generating case markers in machine translation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Con-*

ference, pages 49–56, Rochester, New York, April. Association for Computational Linguistics.

Kristina Toutanova, Hisami Suzuki, and Achim Ruopp. 2008. Applying morphology generation models to machine translation. In *Proceedings of ACL-08: HLT*, pages 514–522, Columbus, Ohio, June. Association for Computational Linguistics.