Doctoral Thesis

Structural Transfer of Morpho-syntactically Divergent Language Pairs for SMT

Jin-Ji Li (이 금희)
Division of Electrical and Computer Engineering
Pohang University of Science and Technology
2011

형태구문적으로 상이한 언어쌍을 위한 통계기계번역에서의 구조변환

Structural Transfer of Morpho-syntactically Divergent Language Pairs for SMT

Structural Transfer of Morpho-syntactically Divergent Language Pairs for SMT

by

Jin-Ji Li
Division of Electrical and Computer Engineering
Pohang University of Science and Technology

A dissertation submitted to the faculty of the Pohang University of Science and Technology in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Electrical and Computer Engineering

Pohang, Korea
12. 22. 2010
Approved by

Jong-Hyeok Lee
Academic Advisor

Structural Transfer of Morpho-syntactically Divergent Language Pairs for SMT

Jin-Ji Li

The undersigned have examined this dissertation and hereby certify that it is worthy of acceptance for a doctoral degree from POSTECH

12/22/2010

Committee Chair	Jong-Hyeok Lee	(Seal)
Member	Geunbae Lee	(Seal)
Member	Chee-Ha Kim	(Seal)
Member	Seung-won Hwang	(Seal)
Member	Hyopil Shin	(Seal)

DECE
20053418

이 금희. Jin-Ji Li. Structural Transfer of Morpho-syntactically Divergent Language Pairs for SMT. 형태구문적으로 상이한 언어쌍을 위한 통계기계번역에서의 구조변환. Division of Electrical and Computer Engineering. 2011. 86P. Advisor: Jong-Hyeok Lee. Text in English.

Abstract

In this Thesis, we present a framework that first resolves structural differences as syntax-aided preprocessing then resolves lexical differences using a phrase-based SMT for morpho-syntactically divergent language pairs such as Chinese-Korean and English-Japanese. This framework strengthens the structure transfer of phrase-based SMT whose capacity for lexical transfer is widely proved. We contrastively analyze the morpho-syntactic differences of source and target languages from the viewpoint of word order and morphological typology. What kind of linguistically-motivated features are utilized and how to effectively incorporate them into the phrase-based SMT is our main concern.

First, we consider the totally different word orders between Chinese and Korean. A Chinese syntactic reordering approach with an emphasis on predicate-predicate patterns is proposed for the phrase- and hierarchical phrase-based SMT. We examine the predicate-predicate patterns relating to long-distance reordering, and investigate which specific constructions contribute to better translation through syntactic reordering. Useful linguistic knowledge is explored to detect the constructions which need to be reordered.

Then, we consider the translation direction of both language pairs from a morphologically-poor language (isolating language) to a morphologically-rich one (agglutinative language) which is more difficult than translating in the opposite direction. Our proposed method handles the structural differences to generate complex morphology in the target languages. Among various kinds of structural differences, we focus on the thematic divergences of syntactic roles such as *subject* and *object* between source and target languages. *Subject* and *object* relations in Chinese and English are implicitly expressed by the word order, while in Korean and Japanese they are explicitly expressed by case markers. Furthermore, these syntactic roles are frequently transferred into other syntactic roles when translating. Our proposed approach fills the morpho-syntactic gaps with the transferred syntactic roles to help the generation of adequate case markers in the target languages. The process of resolving thematic divergences is realized as a structured prediction model.

Finally, we combine the two aforementioned approaches in a cascaded model to perform a more generalized structural transfer because they are complementary to each other. The input sentences are syntactically reordered, then the thematic divergences of *subject* and *object* relations of the reordered sentences are resolved, and vice versa.

Contents

List of Tables

List of Figures

I. Introduction

Resolving lexical and structural ambiguities both within language (monolingual ambiguity) and between languages (bilingual ambiguity) are the major problems of all machine translation (MT) systems [21]. In most MT systems, resolution of monolingual ambiguity is usually conducted in the analysis phase of the source language. Bilingual ambiguity refers to the translational (transfer) ambiguity caused by lexical and structural differences between languages. Therefore, effective lexical and structural transfer is directly related to the performance of MT system.

Statistical machine translation (SMT) systems without exception develop various technologies tackling transfer ambiguities. The capacity for lexical transfer is widely proved in today's SMT systems, but the structural transfer is still poor. Augmenting the SMT systems with the power of resolving structural differences has become the mainstream in current research e.g. syntax-aided phrase-based SMT and syntax-based SMT.

Syntax-aided phrase-based SMT refers to independent subcomponents such as pre- or post-processing approaches that augment the phrase-based system on the structural transfer. Syntax-aided phrased-based SMT is a loosely-coupled method, while syntax-based SMT is a tightly-coupled one that directly embeds syntax in the translation model.

The main objective of this Thesis is to propose structural transfer methods as preprocessing of phrase-based SMT for morpho-syntactically divergent language pairs such as Chinese-Korean and English-Japanese. Phrase-based SMT is still the state-of-the-art system, however it lacks the capability to generate more grammatical translations even most of the content words are correctly translated. Since lexical transfer is one of the strong points of phrase-based model, we compensate the structural transfer as preprocessing by reflecting the systematic differences between source and target languages. We explore useful linguistic knowledge through contrastive analysis and present how to effectively encode it to resolve the structural differences. In this Thesis, we use Chinese-Korean language pair as a main target of study while English-Japanese is used as verification purposes since the two language pairs have similar contrastive linguistic phenomena.

We also propose annotation guidelines of Chinese-Korean word alignment in order to achieve an objective, correct, and consistent word-aligned parallel corpus. This topic is not deeply involved in the main objective of this Thesis, however we include it because word alignment itself is also an important task in the SMT system.

1.1 Contributions of this Thesis

This Thesis presents the following scientific contributions.

- Contrastively analyze morpho-syntactic differences of Chinese-Korean language pair from the viewpoint of language typology
- Propose annotation guidelines of Chinese-Korean word alignment by contrastively analyzing the complex morpho-syntactic encoding of Korean verbs
- Propose Chinese syntactic reordering as preprocessing with an emphasis on predicate-predicate patterns by exploiting useful linguistic knowledge to resolve long-distance reordering problem
- Resolve thematic divergences of *subject* and *object* relations as preprocessing to generate correct case markers of target languages realized as a structured prediction model

- Present more generalized structural transfer by combining approaches of syntactic reordering and resolving thematic divergences

1.2 Outline of this Thesis

The summary of each Chapter is given as follows.

Chapter III: Annotation guidelines for Chinese-Korean word alignment

Word-aligned parallel corpus serves as an important source of knowledge for SMT. However, to achieve an objective, correct, and consistent word-aligned parallel corpus is a difficult task especially for linguistically-distant language pairs such as Chinese and Korean. In this Chapter, we propose annotation guidelines for Chinese-Korean word alignment by analyzing the complex morpho-syntactic encoding system to resolve uncertain cases where correct counterparts of the languages are difficult to find.

We annotate a gold standard set of word alignment of Chinese-Korean parallel corpus according to the proposed guidelines and use it to extract correction rules automatically. Word alignment results are refined by the correction rules. This process decreases the alignment error rate and finally improves the overall Korean-to-Chinese SMT performance.

Chapter III is slightly extrinsic to the main research topic, however we include it in the Thesis because word alignment is also an important component of SMT system, and we deal with a language pair from totally different language families.

Chapter IV: Syntactic reordering for SMT

From the viewpoint of word order typology, Chinese is a SVO language with rigid word order while Korean belongs to a SOV language with flexible word order.

The difference of verb position causes the difficulty in generating correct verbal phrases of target languages. A verbal phrase of Korean consists of a series of verb affixes along with a verb stem. Verb affixes are ordered in a relative sequence within a verbal complex and express different modality information. However, Chinese expresses modality using discontinuous morphemes scattered throughout a sentence.

Chapter IV describes a Chinese syntactic reordering as a preprocessing approach for SMT to disambiguate the word order differences between two languages. For the general syntactic reordering, we first move modality-bearing words near their verbal heads to organize a verb group and move it to the end of the sentences since Korean is a verb-final language. We further focus on a specific structure, namely a predicate-predicate pattern which leads to a long-distance reordering problem when translating from Chinese into Korean. We explore useful linguistic knowledge for the purpose of effective long-distance reordering in Chinese dependency trees. The proposed method indicates the solidness in the hierarchical phrase-based SMT as well.

Chapter V: Resolving thematic divergences of *subject* and *object* for SMT

Chinese is a typical isolating language, while Korean is a highly agglutinative language from the viewpoint of morphological typology. Our translation direction is from a morphologically-poor language to a morphologically-rich one, which is more difficult than the opposite direction [25]. 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 the SMT system.

In this Chapter, we propose a framework that handles the structural differences to generate the complex morphology in the target language. Among various kinds of structural differences, we focus on the thematic divergences of syntactic roles such as *subject* and *object* between source and target languages. The reasons are as follows. First, *subject* and *object* relations in Chinese are implicitly expressed by the word order, while in Korean they are explicitly expressed by case markers. Furthermore, these syntactic roles are frequently transferred into other syntactic roles when translating.

Our proposed approach fills the morpho-syntactic gaps with the transferred syntactic roles to help generation of adequate case markers in the target languages. The process of resolving thematic divergences is realized as a structured prediction model.

Chapter VI: Structural transfer as preprocessing for SMT

Both Chapter IV and V describe how to resolve structural differences of source and target languages by tackling different points such as word order differences and thematic divergences. In this Chapter, we combine the two methods to perform a more generalized structure transfer as preprocessing.

Chapter VII: Conclusion and future work

Conclusion of this Thesis and future work will be given in this Chapter.

II. Background

2.1 Related work

We will review the development of SMT systems focusing on how to resolve the structural differences. Syntax-aided phrase-based SMT and syntax-based SMT are included. The fundamental knowledge of SMT is well written in the textbook [26], in this Thesis, we only summary the issues about structural transfer.

2.1.1 Syntax-aided phrase-based SMT

Phrase-based SMT uses ‘phrases’ as basic translation units. These phrases are any contiguous sequences of words which are not necessarily linguistically motivated. The input sentences are first segmented into a sequences of phrases and translated one-to-one into target phrases which are possibly reordered.

Phrase-based SMT is motivated by the frequent one-to-many mappings (vice versa) between source and target words. As a result, phrases as atomic translation units help resolve lexical ambiguities better than words as atomic units. Good lexical choice is one of the strong points of the phrase-based SMT.

The reordering is simply explained as ‘distortion’ that is modeled by a distance-based reordering cost which discourages the reordering in general. For linguistically distant language pairs which have totally different word orders, such a simple distortion model is not sufficient enough to rearrange the translated phrases into the correct word orders of target languages. The translation output usually suffers from the problem of overall grammaticality, although most of the words are translated correctly and the local reordering seems quite reasonable.

Syntax-aided phrase-based SMT means to introduce linguistic knowledge into the phrase-based SMT system to strengthen the structural transfer. It is efficient to encode the linguistically motivated features in the pre- or post-processing of phrase-based SMT because they are independent subcomponents which do not introduce any other complexity to the decoder.

Syntax-aided preprocessing method

1. Morpho-syntactic reconstruction

The purpose of mopho-syntactic reconstruction is two-fold. One is to decrease the morpho-syntactic differences between source and target languages, and the other is to access syntactic information at a word level.

For morpho-syntactically divergent language pairs, usually the granularity of lexical units and the representation methods of syntax are totally different. Some researchers insert pseudo words into source sentences such as target language unique functional words [70] or syntactic relations [20, 31] to fill the morpho-syntactic gaps between two languages.

[19, 5] adopt supertags or microtags to enrich the word with syntactic information. Each word is supertagged by Lexicalized Tree-Adjoining Grammar (LTAG) or Combinary Categorial Grammar (CCG) supertag sets or enriched with microtags i.e. per-word projections of chunk levels.

2. Syntactic reordering

Long-distance reordering is one of the key weaknesses of phrase-based SMT. Syntactic reordering method means to restructure the source sentences into more target-like word order by the guidance of syntactic information. The input sentences are first syntactically analyzed and reordered by reordering rules as preprocessing of SMT. Reordering rules can be hand-crafted rules [66, 9, 64, 69, 31] or automatically generated ones [74, 34, 14, 18, 62].

Syntactic reordering method effectively compensates the low long-distance reordering power of phrase-based SMT without introducing complexity to the decoding process. We can still turn on the distortion models to further capture local reordering which are not captured in the preprocessing.

Most of the syntactic reordering methods deterministically reorder input sentences, therefore once there is faulty reordering, we cannot recover the mistakes. To resolve this problem, word lattice instead of one reordering as input is proposed [74, 34, 14], in which the preference is encoded as the path probability in the lattice.

Syntax-aided post-processing method

1. N-best re-ranking of translation outputs

N-best candidate translations can be generated by a baseline SMT system. From these candidate translations, the best translation will be chosen by re-ranking method. Usually each candidate translation is represented by a set of features. Each candidate is enriched with additional linguistic features to choose the more grammatically correct output [50, 55, 59, 57].

2. Reordering output sentences

[22] parses the translation results that follow the source language word order into non-projective dependency trees, then reorder dependency trees to obtain fluent target sentences.

A reordering method guided by shallow semantic parsers is proposed by [65]. The output sentences of traditional phrase-based SMT are first semantically parsed and reordered to maximize the match of semantic predicates and arguments between source and target sentences. This method is designed under the assumption that most semantic roles preserve consistently when translation.

3. Generating complex morphology

A number of researchers have tried to develop independent components to handle complex morphology generation. Often the relevant information of complex morphology is scattered widely over the input sentence or implicitly expressed. [59], [60] and [47] suggested post-processing 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 which is implemented as a structured prediction model.

Factored translation model

Factored translation models integrate additional linguistic features at the word level called a factor. It is an extension of the phrase-based SMT model. The linguistic features may be morphological, syntactic, or semantic which are generally considered in the syntax-aided pre- or post-processing method. Factored models can tightly combine linguistic features into the decoding phase, while expanding the search space at the same time.

CCG supertags are adopted in [2] to achieve more grammatical outputs. [1] proposed a method that extracts information from the syntax of source sentences to enrich the morphologically poor language to reduce the grammatical agreement errors on the generated sentences. Also, [54] employed factored models to factorize syntactic/semantic relations and suffixes to help generate inflections and case markers.

2.1.2 Syntax-based SMT

Syntax-based SMT systems directly integrate syntax into the translation model to translate input sentences by the guidance of syntactic structures. Either Phrase Structure Grammar (PSG) or Dependency Grammar (DG) is used. Considering in which side the syntax is used, usually the translation model is categorized into 1) tree-to-string 2) string-to-tree and 3) tree-to-tree model.

[39, 38, 68, 73, 72] present tree-to-string translation models which use syntax in the source side. The parsed source tree is transformed into target string through applying the tree-to-string translation rules. String-to-tree model makes use of target side syntax only and the translation process is like an parsing problem [17, 16, 41, 11, 56, 61]. Tree-to-tree translation models use syntactic representation both in source and target language [12, 53, 43, 73, 45, 40, 46].

In order to increase the expressive power of translation rules, the concept of tree-sequence is proposed [38, 73, 56, 72] instead of tree. Tree-sequence loosens the translation rule that the tree must be syntactified by some syntax tree fragment, therefore it describes both the syntactic and non-syntactic phrase to leverage on the strengths of both phrase- and syntax-based method.

Since linguistically syntax-based SMT uses parsed trees, it is not free from the error propagation problem of parsing errors. The concept of packed forest is introduced into syntax-based SMT by [45, 40, 46, 61], which encodes n-best parsing results to offer more alternatives.

Syntax-based SMT is a more tightly coupled method compared with the syntax-aided phrase-based method to resolve the structural differences between two languages. However, modeling the syntax in the translation model is a hard task and it also significantly increase the decoding complexity.

2.2 Contrastive analysis of Chinese and Korean

This section provides the contrastive analysis and presents the systematic differences between two languages from the viewpoint of language typologies. According to the contrastive analysis, we further explore useful linguistic knowledge to resolve the structural differences in the following Chapters.

2.2.1 Morphological typology

Chinese is a typical isolating language while Korean is a highly agglutinative one. The morphological form of Korean is much more complex than that of Chinese.

As an isolating language, it is generally true that in Chinese, each word consists of only one morpheme and cannot be further analyzed into component parts. In other words, Chinese has a very little morphological complexity [33]. Grammatical functions are expressed by means of word order and some independent morphemes. Usually, an translation unit in Chinese is a segmented word from a morphological analyzer.

For each *eojeol* in Korean, it consists of one or more base forms (stem morphemes or content morphemes) and their inflections (function morphemes) that have a very productive inflectional system. An *eojeol* refers to a fully inflected lexical form separated by a space in a sentence. Inflections usually include postpositions and verbal endings (verbal affixes) on verbs and adjectives (predicate). After morphological analysis has been performed, the basic unit in a given sentence is a Korean morpheme. We generally consider it as a translation unit.

The translation direction is from a morphologically-poor to -rich language, therefore complex morphology generation is one of major issues in the Chinese-Korean SMT. Chinese surface words only cannot provide sufficient clues for correct generation of complex morphology such as Korean postpositions and verbal affixes. Korean postpositions indicate grammatical relations and verbal affixes convey modality information of a sentence, therefore correct generation of Korean functional morphemes directly leads to producing grammatical outputs.

Enriching Chinese sentences with morpho-syntactic information corresponding to Korean complex morphology should be considered in the Chinese-Korean SMT.

2.2.2 Word order typology

Chinese is an SVO language with a rigid word order, while Korean is an SOV language with a flexible word order. The difference of verb position causes the difficulty in generating correct Korean verbal phrases. Also, the complexity of verb affixes in Korean verbs (2.2.1) is problematic in SMT systems targeting Korean, especially if the source language is isolated.

The modality of Korean is expressed intensively by verb affixes. However, Chinese expresses modality using discontinuous morphemes scattered throughout a sentence. 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.

We should 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.

III. Annotation Guidelines for Chinese-Korean Word Alignment

3.1 Introduction

Word alignment is defined as an object indicating the correspondence between words in a parallel text[3], and usually it serves as an important source of knowledge for SMT. However, the notion of "correspondence" between words is subjective [51]. For a language pair such as Chinese and Korean that belong to entirely different language families in terms of typology and genealogy, finding the correspondences is quite unclear in word alignment. Especially problematic is the difference in morpho-syntactic encoding of the two languages.

To achieve more objective, correct, and consistent evaluation results of word alignment, reasonable annotation guidelines are desired to resolve uncertain cases where correct counterparts of the languages are difficult to find. Relatively little research has been carried out on this issue, especially from the perspective of contrastive analysis of morpho-syntactic encodings.

There are several annotation guidelines for other languages such as Blinker project [42], ARCADE project [63], PLUG project [44], guidelines for Chinese-English word alignment of LDC, and [29]. However, these guidelines enumerate specific annotation rules classified by lexical categories such as Part of Speech (POS) of the source languages; these schemes cannot systematically describe linguistic phenomena occurring in morpho-syntactically distant language pairs such as Chinese and Korean.

This Chapter proposes annotation guidelines for Chinese-Korean word alignment. We analyze differences of morpho-syntactic encoding systems of Chinese and Korean. Korean is a typical agglutinative language, whose morphological form of verbs is much more complex than that of Chinese one. Most linking obscurities in word alignment between Chinese and Korean are caused by this difference.

We analyze the complex morpho-syntactic encoding system of Korean verbs to investigate the grammatical categories which the system conveys. Hence, the corresponding elements in Chinese are relatively easy to study. The perspective from grammatical categories provides a more comprehensible and general view for the alignment process. It improves the linking consistency in language pairs with highly different linguistic phenomena.

The scope of our guidelines is limited to the alignment between Chinese and Korean, but the instruction methods exemplified in this Chapter are also applicable in developing systematic alignment guidelines for other languages having such different linguistic phenomena.

3.2 Some issues in annotation guidelines

There are general alignment issues valid in most language pairs. We list general alignment instructions that are also reasonable in Chinese-Korean word alignment.

Two major rules for the word alignment are summarized well by [63].

1. Mark as many words as necessary on both the target and source side.
2. Mark as few words as possible on both the target and source side.

In general, parallel texts are translated non-literally. Hence, using only word-to-word links is not sufficient enough to contain all the information conveyed in the given sentence pair. Therefore, sometimes it is better to use a group of words as alignment units but making it as small as possible according to the above rules.

We allow S(ure) and P(ossible) links in our annotation guidelines [51]. S and

P link mean unambiguous and ambiguous link respectively. There exist many ambiguities of manual alignment because of non-literal translation and systematic differences between the language pair. We think various alignments are acceptable for annotating P links and no need to reach an agreement on this.

Unlinked cases are also needed when the corresponding parts are ‘not translated’ in the target language. To judge the ‘not translated’ case, we adopt the judgment rule defined by Blinker project as follows: “when you can answer ‘Yes’ to the following question: If the seemingly extraneous words were simply deleted from their verse, would the two verses become more similar in meaning?”

Unlike above rules that generally apply to all languages, there exist some language pair-specific issues. Following section describes such characteristics of Chinese-Korean language pair and proposes guidelines utilizing the contrastive analysis of the given language pair with an emphasis on verbal phrases in Korean.

3.3 Verbal phrase in Korean

The complex form of Korean verbs (verbal phrases)[1] frequently causes annotation ambiguities in Chinese-Korean word alignment. A verbal phrase in Korean consists of a series of verbal 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. Verbal affixes in Korean are ordered in a relative sequence within a verbal complex [30] and express various modality information[2] viz. tense, aspect, mood, negation, and voice as Figure 3.1 shows. These five grammatical categories are the major constituents of modal expression in Korean. Table 3.1 shows the modality encodings of Korean verbal phrases.

[1] ‘Korean verb’ or ‘verbal phrase’ in this Chapter refers to Korean predicates in a sentence.

[2] Modality system refers to five grammatical categories such as tense, aspect, mood (*modality* & *mood*), negation, and voice. The definition of these categories is described in [37] in detail.

Table 3.1: Relative orderings of verbal affixes in Korean.

Order	Type
1	Verb Stem
2	Causative & Passive
3	Honorific
4	Aspect Tense *Modality*
5	Negation
6	*Modality* - Evidential
7	*Mood* - Illocutionary Force

1. 먹(stem)고_있(aspect)었(aspect)었(tense)다(*mood*)
(had been eating)

2. 잡(stem)히(passive)었(aspect)겠(modality)다(*mood*)
(may have been captured)

Figure 3.1: Verbal phrases in Korean.

The prominence and correlations of modality system is different from language to language, and such difference increases the annotation ambiguity.

The modality of Korean is expressed intensively by verbal affixes of complex inflectional forms. However, as a typical isolating language, Chinese expresses modality using discontinuous morphemes around lexical verbs. Modal expression in Korean is much more various than that of Chinese. Many-to-few assignment of modality expression causes linking obscurity in word alignment.

Languages, in general, do not give equal prominence to modality information. Chinese is an aspect- and topic-prominent language. As an aspect-prominent language, Chinese does not have grammatical markings for expressing tense. Unlike Korean, Chinese does not have a specific grammatical form in the voice system, which is natural for a typical topic-prominent language.

There exist some correlations among the grammatical categories, and such categories tend to share affixes for conveying modality information. For example, tense and aspect are interconnected as they both are involved with the 'temporal structure' of an event. In Korean, '었(eoss)' can be used as temporal and aspectual marker as Figure 3.1 shows.

Chinese has different ways of expressing modality: modal information is scattered throughout a sentence. We locate such modal expressing elements to provide correct candidate words.

3.4 Annotation guidelines

In this Section, we dedicate much space to explaining how Korean verbal phrases are linked to corresponding Chinese words because they are where most linking obscurities occur.

Since Korean is a verb-final language, identification of verbal phrases is much easier than Chinese. For efficiency, consistency, and accuracy, we propose an annotation principle that first judge Korean verbal phrases, then match the correspondent words in Chinese. The correspondences in Chinese are mainly composed of features used to display Chinese modality information.

Because of linguistic differences and liberal translations in parallel corpora, there exist phrasal correspondences and different link types: S-link, P-link, and not-translated. Explicit and unambiguous correspondences are S-linked and implicit correspondences are P-linked. As mentioned before, annotators may have disagreements on P-links.

3.4.1 Guidelines based on Korean verbal system

We propose special guidelines based on Korean verbal system as follows. To find the correspondences of Korean verbal phrases, we need to clarify the method for expressing modality information in Chinese. Such clarification can help link the verbal affixes in Korean because verbal affixes in Korean also convey modal expression.

We will give an explanation based on five grammatical categories such as tense, aspect, mood, negation, and voice, which, in Chinese, compose most of the modal expression.

Tense

Chinese does not have a grammatical category of tense, because the concept of tense is indicated by content words such as temporal adverbs, times[3], and auxiliary verbs. It is also inferred using aspect markers and attribute of Main Predicate (MP) by inspecting whether it is a motion verb of instantaneity or not. Table 3.2 shows the features which provide temporal information in Chinese.

Time & MP-motion verb of instantaneity Times and motion verbs of instantaneity indicate tense information and also their counterparts in Korean can be

[3] Time is a category of Part of Speech in Chinese, which shows the temporal information.

Table 3.2: Tense markers in Chinese.

Tense Marker	Example

Time	明天(MingTian)
	去年(QuNian),下星期(XiaXingQi)

MP_motion verb of instantaneity	送(Song), 告诉(GaoSu)

Temporal adverb	将(Jiang),将要(JiangYao),已经(YiJing)
	总是(ZongShi)

Auxiliary verb	会(Hui), 要(Yao)
Aspectual particle	了(Le),着(Zhe),过(Guo)

found readily. In Ex 2, the attribute of main verb ‘送(Song)’ indicates past tense. In this case, we can simply link the counterparts together.

Eojeols are separated by a space. For each *eojeol*, bold-faced content morphemes followed by functional ones with ‘+’ sign. Corresponding morphemes in each language are italicized and main predicates are underlined. Italicized morphemes in each language have high chances to be linked each other.

Ex 1.

[cn] 明天*(tomorrow)*/我(I)/去*(go)*/北京(Beijing)/。

[kr] **나(I)**+는 ***내일(tomorrow)*** **북경(Beijing)**+에 ***가(go)****+ㄹ 것+이+다+.*

[en] I will go to Beijing tomorrow.

Ex 2.

[cn] 老王(Mr. Wang)/送*(give)*/我(me)/一(one)/本(Classifier) /书(book)/。

[kr] **왕(Wang)**+**씨(Mr.)**+는 **나(me)**+에게 **책(book)**+을 ***선물(give)****+하+였+다+.*

[en] Mr. Wang gave me a book as a present.

Temopral adverb Some temporal adverbs such as '将(Jiang)' and '已经(YiJing)' in Chinese almost behave as function words since they only provide tense information. The actual translation is usually omitted in target sentence. Ex 3 shows this phenomenon. In this case, '已经(YiJing)' is glued to main verb '回家(HuiJia)' and linked to the verbal phrase in Korean.

Ex 3.
[cn] 他(he)/已经*(already)*/回家*(go home)*/了*(Particle)* /。
[kr] **그(he)**+는 ***집(home)***+에 ***가(go)****+았+다+.*
[en] He went home.

Ex 4.
[cn] 我(I)/将*(soon)* /去*(go)*/北京(Beijing)/。
[kr] **나(I)**+는 **북경(Beijing)**+에 ***가(go)****+ㄹ 것+이+다+.*
[en] I will go to Beijing.

Auxiliary verb & Aspectual particle Auxiliary verbs and aspectual particles are completely translated into verbal affixes in Korean. These two markers also convey the modal and aspectual information.

Ex 5.
[cn] 明天*(tomorrow)*/会*(will)*/下*(fall)*/雨(rain)/。
[kr] ***내일(tomorrow)*** **비(rain)**+가 ***오(fall)****+ㄹ 것+이+다+.*
[en] It will rain tomorrow.

Ex 6.
[cn] 我(I)/去*(go)*/过*(Particle)*/北京(Beijing)/。
[kr] **나(I)**+는 **북경(Beijing)**+에 ***가(go)****+았+었+다+.*
[en] I have been to Beijing.

Aspect

Chinese is recognized as an aspect prominent language with a complete set of markers to express aspectual distinctions. Conforming to the aspect system classified by [67], we see that there are several types of aspect markers as Table 3.3 shows.

Table 3.3: Aspect markers in Chinese.

Aspect Marker	Example
Aspectual Particle	了(Le),着(Zhe),过(Guo)
Adverbs	在(Zai),正在(ZhengZai),正(Zheng),曾经(CengJing),曾(Ceng)

Temporal adverb	笑一笑(Xiao),看看(Kan), 讨论讨论(TaoLun)
	过过瘾(GuoYin), 看了一看(Kan)

Reduplication	会(Hui), 要(Yao)
RVC	(跳)下去(XiaQu), (交)上来(ShangLai) (携)起(手)来(QiLai), (写)清楚(Qingchu)

RVC is an acronym of Resultative Verb Complement like open in ‘push the door open’ [67].

Aspectual particle & Adverb As mentioned before, aspectual particle indicates temporal information as well as aspectual one. In Korean, tense and aspect also share verbal affixes to express temporal structures such as tense and aspect.

Ex 7.
[cn] 他(he)/在*(now)*/写*(do)*/作业(homework)/。
[kr] **그(he)+는 숙제(homework)+를** *하(do)+고 있+다+.*
[en] He is doing homework.

Ex 8.

[cn] 我(I)/曾*(already)*/去*(go)*/过*(Particle)*/北京(Beijing)/。

[kr] **나(I)**+는 **북경(Beijing)**+에 ***가(go) 보***+ㄴ *적+이 있+다+.*

[en] I have been to Beijing.

Reduplication Verb reduplication is an idiosyncratic linguistic form in Chinese. Some verbs can be reduplicated to convey delimitative aspect in a sentence. There are several formats for verb copying such as VV, V了(Le)V, V一(Yi)V and V了(Le)一(Yi)V.

Ex 9.

[cn] 给(prep.)/我(me)/看/看*(see)*/报纸(news)/吧*(Particle)*/。

[kr] **저(me)**+에게 **신문(news)** 좀 ***보(see)***+여 *주+세+요+.*

[en] Let me see the news, please.

Ex 10.

[cn] 我(I)/看/了*(Particle)*/看*(read)*/报纸(news)/。

[kr] **나(I)**+는 **신문(news)**+을 ***보(read)***+*았+다+.*

[en] I glanced at the news.

RVC RVCs not only convey the aspectual values, but also retain their original lexical meanings. Therefore, they can be translated into auxiliary predicates, as well as independent lexical verbs in Korean. In the latter case, we link RVCs to the correspondent verbs in Korean such as Ex 12. The RVC '清楚(QingChu)' is translated into an adverb "똑바로(ddok-ba-ro)" in Korean.

Ex 11.

[cn] 大家(everybody)/把(Prep.)/作业(homework)/交*(submit)*/上来*(RVC)*/。

[kr] **모두(everybody) 숙제(homework)**+를 ***내(submit)*** *주+세+요+.*

[en] Everybody, submit your homework.

Ex 12.

[cn] 写*(write)*/清楚*(clearly)*/你(your)/的(Particle)/名字(name)/。

[kr] **당신(your)**+의 **이름(name)**+을 **똑바로(clearly)** ***적(write)****+어 주+세+요+.*

[en] Please write down your name clearly.

Mood Mood refers to a general linguistic term: a grammatical category signaling the expression of the speaker's attitude towards a proposition. It includes the concepts of both *'mood'* and *'modality'*.

Usually the category of mood is defined as a morphological verbal category which indicates the modal value of a sentence. It is usually expressed by inflection in most languages. In a broader category, it covers so-called sentence-moods. However, as an isolating language, mood system of Chinese is not expressed by verbal inflection.

'Modality' is expressed by various means of modal encoding ranging from lexical to highly grammaticalized ones. In particular, as an isolating language, Chinese mainly uses modal auxiliaries to express the *modalities*. The correlation between future tense and *modality* makes it possible that future events also can be expressed temporally or modally. In fact, auxiliary verbs for future tense are developed from the modal auxiliaries in Chinese.

Table 3.4: Mood markers in Chinese.

Mood Marker	Example
Auxiliary verb	应该(YingGai),能(Neng),可以(KeYi),必须(BiXu), 得(Dei)
Sentence-final particle	呢(Ne),呀(Ya),吗(Ma),了(Le)

Auxiliary verbs In some cases, the auxiliary verbs can translate into adverbs in Korean as well as indicate modal information. Such auxiliary verb should have links to both of the counterparts as in Ex 14.

Ex 13.
[cn] 你(you)/应该*(should)*/先(first)/做*(do)*/作业(homework)/。
[kr] **너(you)**+는 **먼저(first) 숙제(homework)**+를 ***하(do)****+어야* ***하****+ㄴ+다+.*
[en] You should do your homework first.

Ex 14.
[cn] 你(you)/必须*(ought to)*/先(first)/做*(do)*/作业*(homework)*/。
[kr] **너(you)**+는 ***반드시(ought to)*** **먼저(first) 숙제(homework)**+를 ***하(do)****+어야* ***하****+ㄴ+다+.*
[en] You ought to do your homework first.

Sentence-final particle Sentence-final particle shows the information of sentence-type mood. In Korean, it is expressed by inflection of verbal affixes with respect to honorific information.

Ex 15.
[cn] 您(you)/明天*(tomorrow)*/去*(go)*/北京*(Beijing)*/吗*(Particle)*/？
[kr] **당신(you)**+은 ***내일(tomorrow)*** **북경(Beijing)**+에 ***가(go)****+시+ㅂ니까+?*
[en] Are you going to Beijing tomorrow?

Negation

The negation systems in Chinese and Korean are very similar. In general, there are standard negation, double negation, and imperative/propositive negation. There are four negative forms commonly use in Chinese: '不(Bu)', '别(Bie)', '没(Mei)', and '没有(MeiYou)'. The most general and neutral form of negation is '不(Bu)'.

There are also special negative formats in Chinese. One is a negative particle '不(Bu)' before an RVC and the other is '不了(BuLiao)/不得(BuDe)' after main predicates to show negative view. Besides these two formats, some negative particles such as '未能(WeiNeng)' and '别(Bie)' also show other modality information like aspect and mood.

Table 3.5: Negation markers in Chinese.

Negation Marker	Example

Negative particle	不(Bu), 没(有)(Mei(You)), 别(Bie), 未能(WeiNeng)
	从未(CongWei), 甭(Beng)

MP_bu_RVC	(吃)不(Bu)(下去), (看)不(Bu)(过去)
MP_buliao/bude	(开)不了(BuLiao), (听)不了(BuLiao), (吃)不得(BuDe)

Negative particle Although the main usage of negative particles is to show negative view in a sentence, it negates different modal situations in Chinese. For example, '没(有)(Mei(You))' negates completion of an event and '别(Bie)' is a negative imperative.

Ex. 16.

[cn] 他(he)/不*(not)*/在*(Particle)*/学习*(study)*/。

[kr] **그(he)**+는 ***공부(study)****+하+지* ***않(not)****+고 있+다+.*

[en] He is not studying now.

Ex 17.

[cn] 我(I)/没有*(not)*/吃*(eat)*/饭(meal)。

[kr] **나(I)**+는 **밥(meal)**+을 ***먹(eat)****+지* ***않(not)****+았+다+.*

[en] I did not eat a meal.

Ex 18.

[cn] 别*(not)*/让*(Particle)*/她(she)/出去*(go out)*/。

[kr] **그녀(she)**+가 ***나가(go out)+게 하+지 말+라+.***

[en] Do not let her go out.

MP_bu_RVC & MP_buliao(bude) These two formats not only indicate negation information, but also give *modality* information.

Ex 19.

[cn] 我(I)/实在(really)/是(be)/吃*(eat)*/不*(not)*/下去*(RVC)*/。

[kr] **나(I)**+는 **정말(really) 더(more) 이상(over)** ***먹(eat)+을 수 없+다+.***

[en] I can not eat anymore.

Ex 20.

[cn] 我(I)/听*(hear)*/不了*(can not)*/音乐(music)/。

[kr] **나(I)**+는 **음악(music)**+을 ***듣(hear)+을 수 없+다+.***

[en] I can not hear the music.

Voice

Generally there are two kinds of voice construction in Chinese: with, or without voice markers.

The typical passive marker is '被(Bei)', however the non-adversity usage of passive sentence makes it possible to express passive voice without any markers. Usually topic–comment structure in Chinese can function as a passive sentence as in Ex 23.

A variety of notional causative forms are adopted in Chinese to express the causative voice. Some RVC constructions in Chinese convey the causative meaning as in Ex 24 and Ex 25. The typical causative markers are '使(Shi)', '让(Rang)' and '叫(Jiao)'.

Generally, sentences without any passive/causative markers are more productive than sentences with markers in Chinese. Although they help to convey voice information using special constructions such as '被字(BeiZi)' phrases or '使字(ShiZi)' phrases, these voice markers do not directly provide voice information of the lexical verbs.

Table 3.6: Voice markers in Chinese.

Voice Marker	Example

Passive/Causative particle	被(Bei), 让(Rang), 使(Shi), 叫(Jiao)
	令(Ling), 给(Gei)

Topic-comment construction	那本书(Topic)已经出版了(Comment).
RVC_causative	(写)好(Hao), (搞)清楚(QingChu)
MP_causative	放沉(FangChen), 加强(JiaQiang), 弄醒(NongXing)

Passive/Causative marker

Ex 21.

[cn] 他(he)/被*(Particle)*/老师(teacher)/骂*(scold)*/了*(Particle)*/。

[kr] **그(he)**+는 **선생님(teacher)**+께 ***야단(scold)***+*맞+았+다+.*

[en] He was scolded by the teacher.

Ex 22.

[cn] 这(this)/件(Classifier)/事(thing)/使*(Particle)*/我(me)/非常(very)/ 高兴*(happy)*/。

[kr] **이(this) 일(thing)**+은 **나(me)**+로 **하여금 매우(very)** *기쁘+게 하+였+다+.*

[en] This thing makes me very happy.

Topic-Comment construction

Ex 23.

[cn] 那*(that)*/本*(Classifier)*/书*(book)*/已经*(already)*/出版*(publish)*/了*(Particle)*/。

[kr] **그(that) 책(book)**+은 **이미(already)** ***출판(publish)***+되+었+다+.

[en] That book has already been published.

RVC_causative

Ex 24.

[cn] 信(letter)/写*(write)*/好*(good)*/了*(Particle)*/。

[kr] **편지(letter)**+를 **다(all)** ***쓰(write)***+었+다+.

[en] The letter was written.

Ex 25.

[cn] 把(Prep.)/问题(problem)/搞*(make)*/清楚*(clear)*/。

[kr] **문제(problem)**+를 ***명확(clear)***+하+게 하+라+.

[en] Make the problem clear.

MP_causative

Ex 26.

[cn] 噪音(noise)/弄醒*(wake)*/了*(Particle)*/我(me)/。

[kr] **소음(noise)**+은 **나(me)**+를 ***깨(wake)***+게 **하**+였+다+.

[en] The noise made me wake up.

Other cases

There are two special constructions and both of them are not appropriate to be classified into the five grammatical categories we have discussed.

Table 3.7: Exceptional cases in Chinese.

	Example
Nominalization	看书的(De)
Separated verb	理(Li)(了)发(Fa), 上(Shang)(了)风(Feng)

Nominalization

Ex 27.
[cn] 这(this)/篇/(Classifier.)论文(paper)/是*(be)*/我们(our)/发表*(publish)*/过*(Particle)*/的*(Particle)*/ 。
[kr] **이(this) 논문(paper)**+은 **우리(our)**+가 ***발표(publish)****+하+였+던 것+이+다+.*
[en] This was published by us.

Ex 28.
[cn] 看*(read)*/书(book)/的*(Particle)*/是*(be)*/我(I)/的(Particle)/朋友(friend)/。
[kr] **책(book)**+을 ***보(read)****+는* **이(person)**+는 **나(my)**+의 ***친구(friend)****+이+다+.*
[en] The person who is reading a book is my friend.

Separated verb Chinese has a kind of verbs whose internal construction is a verb-object compound. The first constituent, like a verb in a sentence, can take aspect markers. Also, it can be separated by a measure phrase, modifiers of object constituents and so on.

Ex 29.
[cn] 他(he)/昨天*(yesterday)*/理*(cut)*/了*(Particle)*/发*(hair)*/。
[kr] **그(he)**+는 **어제(yesterday)** ***이발(cut hair)****+하+였+다+.*
[en] He had his hair cut yesterday.

Ex 30.

[cn] 他(he)/昨天*(yesterday)*/理*(cut)*/了*(Particle)*/一(one)/次(Classifier)/发*(hair)*/。

[kr] **그(he)+는 어제(yesterday)** ***이발(cut hair)+하+였+다+.***

[en] He had his hair cut yesterday.

3.5 Experiment

3.5.1 Corpus profile

We automatically collected and constructed a sentence-aligned parallel corpus from the DongA newspaper[4]. Strictly speaking, it is a non-literally translated Korean-to-Chinese corpus. The corpus consists of 101,226 sentence pairs and we randomly selected 50 sentence pairs as test data. The corpus profile is shown in Table 3.8.

Table 3.8: Statistics for test corpus.

	Chinese	Korean
# of sentences	50	50
# of words	1,323	1,502
# of singletons	741	645
Avg. length	26.5	30.4

3.5.2 Experimental result

Our aim is to examine the effectiveness of proposed guidelines. Usually it is measured by agreements between annotators with the same test corpus. We adopt the Kappa statistic [4] to measure the agreements between annotators. Although our method presents guidelines regarding verbal systems, the experiment is performed

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

to evaluate the effectiveness of the whole annotation guidelines for Chinese-Korean word alignment. The experimental scenario is as follows:

1. Kappa value between two skilled annotators (A1 and A2) who are very familiar with the annotation guidelines;
2. Kappa values between each skilled annotator and a beginner (B) who was never involved in corpus annotation;
3. Kappa values between each skilled annotator and the beginner acquainted (B_acquainted) with the annotation guidelines;

Table 3.9 shows Kappa values according to our proposed experimental scenario.

Table 3.9: Kappa values between annotators.

	Kappa Value
A1 vs. A2	0.892
A1 vs. B	0.799
A2 vs. B	0.805
A1 vs. B_acquainted	0.858
A2 vs. B_acquainted	0.848

The Kappa values between a beginner who is not familiar with the annotation process of Chinese-Korean and each skilled annotator is relatively low; some literature adopts assessment scales with Kappa values between 0.67 and 0.8 as only allowing tentative conclusion [28]. After acquainted with proposed guidelines, the Kappa values between the beginner and skilled annotators improves by about .05, in the range of definite conclusion of the assessment scale defined by [28]. We deduce that a novice annotator is able to achieve high agreements with skilled annotators with our suggested annotation guidelines.

The improvement ratio of A1 vs. B and A1 vs. B_acquainted is greater than A2 vs. B and A2 vs. B_acquainted. B acquired the annotation guidelines through

Question-and-Answering period with skilled annotator A1. We speculate that B could be influenced by the annotation style of annotator A1. This is fairly possible because many cases, especially regarding P links, are open to different interpretations according to the linguistic intuitions of annotators.

3.6 Conclusion

We presented annotation guidelines for Chinese-Korean word alignments through contrastive analysis of morpho-syntactic encodings. We discuss the differences in verbal systems that cause most linking obscurities in Chinese-Korean annotation process. Systematic comparison of verbal systems is conducted by analyzing morpho-syntactic encodings. Such approach from the viewpoint of grammatical category allowed us to define consistent and systematic instructions for linguistically distant languages such as Chinese and Korean. The proposed approach is also applicable to other language pairs with different morpho-syntactic encodings.

To validate the reliability of proposed guidelines, we adopted Kappa statistic. We achieved high Kappa value of 0.892 between two skilled annotators. 0.858 and 0.844 are also achieved between each skilled annotator and a beginner. Therefore, we believe the proposed guidelines produce consistent annotation results.

We annotated a gold standard set of word alignment of Korean-Chinese parallel corpus according to the annotation guidelines proposed in this Chapter and used it to extract a correction rule automatically [23]. The word alignment result of a phrase-based SMT system is refined by the correction rule. This approach led to 17.6% relative decrease in alignment error rate compared to the baseline system, and it also improved the SMT performance. This result shows that the annotation guidelines are efficient to construct a gold standard set of word alignment.

IV. Syntactic Reordering for SMT

4.1 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.

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. 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) [66, 9, 64, 35] and D-trees [6, 69, 20]. 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 [69]. [69] 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.

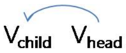
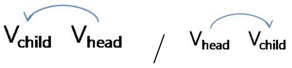
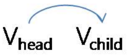
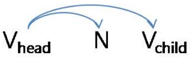
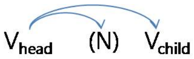
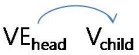
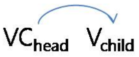
In our Chinese-to-Korean SMT, we adopt the principles of compiling reordering rules in a D-tree as [69] proposed. Since Chinese is a morphologically poor language with insufficient linguistic clues, more careful concern is required when compiling syntactic reordering rules, especially for predicate-predicate patterns.

4.2 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 4.1) with less (almost no) morphological differences than other languages such as English and Korean.[1] 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.

[1] For convenience and consistency, we describe the POS of predicate in the Penn Chinese Treebank style [71]. V_{head} is either VV or VA. VV: common verb; VA: predicative adjective; VE: existential verb; VC: copula verb.

Table 4.1: Types and structures of predicate-predicate patterns.

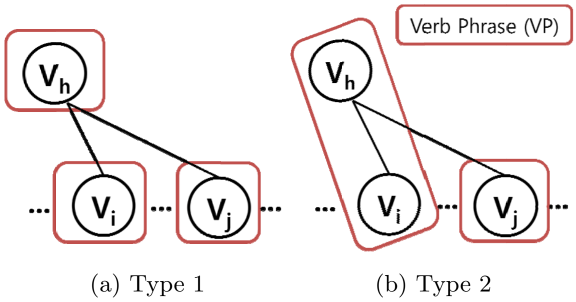
Type	Structure
Clausal subject construction	
Complex/Compound construction	 V_{child} V_{head} / V_{head} V_{child}
Serial verb construction (SVC)	
Pivot construction	
Clausal complement construction	
Existential construction	
Emphasis construction	

In Figure 4.1, '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 '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.

Ch1.	张三(ZhangSan)/ **相信(believe)**/ 李四(LiSi)/ **有(have)**/ 才能(talent)/ 。
En1.	ZhangSan **believes** *that(complementizer)* LiSi **has** the talent .
Kr1.	ZhangSan은 LiSi가 재능(talent)이 **있(have)**다고*(complementizer)* **믿(believe)** 는다.
Ch2.	张三(ZhangSan)/ **买(buy)**/ 菜(vegetable)/ **做(cook)**/ 饭(meal)/ 。
En2.	ZhangSan **bought** vegetables *and(conjunction)* **cooked** a meal .
Kr2.	ZhangSan은 채소(vegetable)를 **사(buy)**서*(conjunctive verb ending)* 밥(meal)을 **지(cook)**었다.

Figure 4.1: Examples of predicate-predicate patterns extracted from the Penn Chinese Treebank 4.0.



(a) Type 1

(b) Type 2

Figure 4.2: Two types of a head predicate according to the predicate's VP formation with the first predicate from its immediate right children.

4.3 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 Chapter.

A head predicate V_h may have several predicates as immediate right children such as V_i and V_j (Figure 4.2). 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 4.1 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 4.1 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

Ch3.	他(he)	不能(cannot)	去(go)	北京(Beijing)	了(Aspect particle)
Syn.	subj	mmod	Root	dobj	asp
	L_Children		Predicate	*R_Children*	

Re.	他(he)	北京(Beijing)	不能(cannot)	去(go)	了(Aspect particle)
	L_Other	*L_FromRight*	*L_Modal*	Predicate	*R_Modal*

Figure 4.3: An example of reordered Chinese sentence after applying the general reordering method. Ch3.: a Chinese sentence; Syn.: dependency structure; Re.: reordered Chinese sentence;

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.

4.4 Chinese syntactic reordering on predicate-predicate patterns

4.4.1 General syntactic reordering

We and [69] proposed to compile a set of general Chinese syntactic reordering rules under the same principles. 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.

We move modality-bearing words near their verbal heads. 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 preverbal positions in Korean sentences. [69] use verb precedence rules to organize a verb group and move it to the end of the sentence. 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 4.3. 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 4.4.2. 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

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*

```
for node N in L_Children do
  if dep.relation of N $\in$ {mmod, neg, pass} then
    L_Modal $\Leftarrow$ L_Modal + {N}
  else
    L_Other $\Leftarrow$ L_Other + {N}
  end if
end for
for node N in R_Children do
  if dep.relation of N $\in$ {comod, asp, neg, rcomp} then
    R_Modal $\Leftarrow$ R_Modal + {N}
  else if dep.relation of N $\in$ {ccomp, punct} then
    R_Remnant $\Leftarrow$ R_Remnant + {N}
  else
    L_FromRight $\Leftarrow$ L_FromRight + {N}
  end if
end for
```

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]

Table 4.2: Features for the SVM binary classifier.

Feature	Description
Lexical	Surface form of V_h
	V_h is a pivot construction verb with the lexical meaning of command
	V_h is a verb that can take a clausal complement
Syntactic	V_h has a direct object
	V_i has a "," or ":" punctuation as left sibling
	V_i has a nominal subject

Table 4.3: Accuracy of the SVM classifiers for predicate-predicate pattern reordering.

Feature	Accuracy (%)
All features	93.3
w/o lexical feature	77.1
w/o syntactic features	90.7

4.4.2 Syntactic reordering of predicate-predicate patterns

In this Section, we present the syntactic reordering methods for predicate-predicate patterns. As pointed out in Section 4.3, a very specific type of predicate-predicate pattern needs our attention for long-distance reordering: a head predicate of Type 2 corresponding to a clausal complement and some pivot constructions. These patterns need be discovered.

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

[2] Following 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.

[3] http://svmlight.joachims.org, version 6.02.

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 4.2. 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 4.3). 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.

[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 INFL.) that acts as a sentential object in the sentences.

4.5 Experiment

4.5.1 Experimental setting

Our baseline system is the state-of-the-art phrase- and hierarchical phrase-based SMT system built in Moses [8, 27] with 5-gram SRI language modeling [58] tuned with Minimum Error Rate Training (MERT) [49]. We adopt NIST [13] and BLEU[5] [52] as our evaluation metrics. A significance test is also conducted using a paired bootstrap resampling method[6] [24].

We use the Stanford Chinese typed dependency parser [32, 7] 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 analyzer[7].

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.

4.5.2 Corpus profile

We automatically collected and manually aligned a parallel corpus from the Dong-A news.[8] 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. The corpus profile is displayed in Table 4.4.

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

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

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

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

Table 4.4: Corpus profile of Dong-A news.

		Chinese	Korean
Training	# words	2,673,422	3,120,466
	Sen. length	27.09	31.62
Development	# words	14,452	16,738
	Sen. length	28.90	33.48
Test	# words	14,619	16,925
	Sen. length	29.24	33.85

Table 4.5: BLEU scores under different experimental settings. † mark shows significant improvement over the general syntactic reordering method with a confidence level over 95%, and ‡ with a confidence level over 99%.

Method	Lexicalized reordering		Hiero	
	NIST	BLEU (gain)	NIST	BLEU (gain)
Baseline	5.8428	22.19	6.1020	23.77
General syntactic reordering	6.0288	23.84	6.1207	24.59
Method 1: PKU dictionary	6.1348	24.26 (+0.42)†	6.1622	25.05 (+0.46)
Method 2: SVM Classifier	6.1242	24.73 (+0.89)‡	6.2258	25.28 (+0.69)†

4.5.3 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 4.5). 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 4.4) 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 4.3. 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.

4.6 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-

Ch4:	三星/ 电子 (Samsung Electronics)/ 11/ 日/ **表示 (announce)**/ , 对/ “ 诺贝尔 博物馆(Nobel Museum) ”/ 提供(supply)/ 薄膜/ 晶体/ 液晶/ 显示器/ (TFT-LCD)/ 之后/ , 将/ 对/ 4 月(April)/ 完工(complete)/ 的/ 该(this)/ 馆(museum)/ 继续(continuously)/ **提供(supply)**/ DVDP/ 设备(equipment)/ 。/
	phrase-based SMT with lexicalized reordering
B:	삼성전자 는 11 일 노벨상 박물관 를 제공 하 는 필름 晶体 액정표시장치 (TFT - LCD) 에 잇 어 4 월 완공 되 ㄴ 이 과 DVDP 계속 하 였 다 .
G:	삼성전자 는 11 일 노벨 박물관 초 박막 트랜지스터 액정표시장치 (TFT - LCD) 晶体 을 제공 하 ㄴ 뒤 4 월 에 완공 되 ㄴ 이 과 DVDP 장비 를 **제공(supply)** 하 고 있 다 .
P:	삼성전자 는 11 일 노벨 박물관 晶体 초 박막 트랜지스터 액정표시장치 (TFT - LCD) 이 제공 하 ㄴ 뒤 4 월 에 완공 되 ㄴ 이 성균관 DVDP 장비 를 **제공(supply)** 하 였 다 고 **밝히(announce)** 었 다 .
	hierarchical phrase-based SMT
B:	삼성전자 는 11 일 노벨 박물관 에 대하 ㄴ 필름 晶体 액정표시장치 (TFT - LCD) 에 잇 어 4 월 완공 되 ㄴ 이 제공 과 계속 DVDP 장비 를 **제공(supply)** 하 고 있 다 .
G:	삼성전자 는 11 일 노벨 박물관 을 필름 晶体 액정표시장치 (TFT - LCD) 을 제공 하 ㄴ 뒤 4 월 에 완공 되 ㄴ 이 과 DVDP 장비 를 **제공(supply)** 하 기 로 하 였 다 .
P:	삼성전자 는 11 일 노벨 박물관 을 필름 晶体 액정표시장치 (TFT - LCD) 을 제공 하 ㄴ 뒤 4 월 에 완공 되 ㄴ 이 성균관 DVDP 장비 를 **제공(supply)** 하 였 다 고 **밝히 (announce)** 었 다 .
Ref:	삼성전자 (Samsung Electronics) 는 스웨덴(Sweden) 노벨상(Nobel Prize) 재단(Foundation) 이 올해(this year) 4 월(April) 완공(complete) 을 목표(purpose) 로 짓(build) 고 있는 노벨 박물관 (Nobel Museum) 에 초 박막 트랜지스터 액정표시장치 (TFTLCD) 모니터(Monitor) 에 잇 어 DVDP 도 **공급(supply)** 하 게 되 었 다 고 11 일 **밝히(announce)** 었 다 .

Figure 4.4: Translated results of Baseline (B), General reordering method(G), and Proposed method (P) with phrase- and hierarchical phrase-based SMT systems.

predicate patterns through contrastive analysis of the source and the target languages. We examined the predicate-predicate patterns 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.

V. Resolving Thematic Divergences of *Subject* and *Object* Relations for SMT

5.1 Introduction

In this Chapter, we propose a preprocessing approach for the phrase-based SMT for Chinese-Korean and English-Japanese language pairs. Both language pairs are morpho-syntacitcally divergent and the translation direction is from a morphologically-poor language to a morphologically-rich one. Our proposed framework handles the structural differences to generate the complex morphology in the target languages.

Among various kinds of structural differences, we focus on the thematic divergences of syntactic roles such as *subject* and *object* between source and target languages. The reasons are as follows. First, *subject* and *object* relations in Chinese and English are implicitly expressed by the word order, while in Korean and Japanese they are explicitly expressed by case markers. Furthermore, these syntactic roles are frequently transferred into other syntactic roles when translating. Our proposed approach fills the morpho-syntactic gaps with the transferred syntactic roles to help the generation of adequate case markers in the target languages. The process of resolving thematic divergences is realized as a structured prediction model. The training corpus can easily be constructed automatically with word alignment and syntactic information using any given bilingual corpora prepared for SMT systems.

5.2 Morpho-syntax of Chinese-Korean and English-Japanese

First, we contrastively analyze the language pairs of Chinese-Korean and English-Japanese.

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 [33]. Syntactic relations such as *subject* and *object* are expressed by word order only, and *adverbial* 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. 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-Korean SMT are usually morphemes. In Chinese, the sentences are segmented into words, and each segmented word is a morpheme. In Korean, an *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 cause data sparseness problems and we have to consider a morpheme as a translation unit for Korean. Under the same reason, in Japanese, the translation unit is a morpheme.

English is also an isolating language and *subject* and *object* are expressed by word order. Japanese belongs to the same language family as Korean, while the system of functional morphemes is much simpler than that of Korean. For example, it has only 18 case markers [59]. We report the statistics of functional morphemes in our training corpus for Korean and Japanese in Section 5.5.1. From Table 5.4 and Table 5.5, we see that Korean has a more complex morphology system than

Japanese.

5.3 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.

[1] 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, [54] 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. [59], [60] and [47] suggested post-processing 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. This method achieved significant improvements when combined with the syntax-based SMT, but not with the phrase-based SMT.

[20] 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.

[70] have also proposed a similar principle. The framework of their system is a two-stage translation. In the first stage, it structurally reconstructs the source language to generate an intermediate language, and in the second stage, the SMT system performs the lexical transfer such as word sense disambiguation and lexical mapping. The intermediate language maintains most of the source words, and also reflects target-unique structure by inserting functional words of the target language. However, the first stage requires very expensive linguistic processing consisting of language understanding and language generation. The main difference between their system and ours is that our system transforms the problem of structural transfer into a structured prediction model, which is relatively easy to construct.

5.4 Proposed method

In both language pairs of Chinese-Korean and English-Japanese, Subject-Verb-Object (SVO) patterns retain structural transfer ambiguities such as thematic divergences when translating. In this Chapter, we propose a preprocessing method that transfers the syntactic roles of SVO patterns, and as a result, the transferred syntactic roles promote the generation of correct case markers in the target languages. The transfer phase is realized as a structured prediction model. This process is similar to the structural 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.

The prediction module is relatively light to construct using an easily and automatically constructed training corpus. How to build the prediction module and what kind of linguistic features are employed are two important issues in our work. Figure 5.1 shows the system architecture and we explain each module in detail in the following sections.

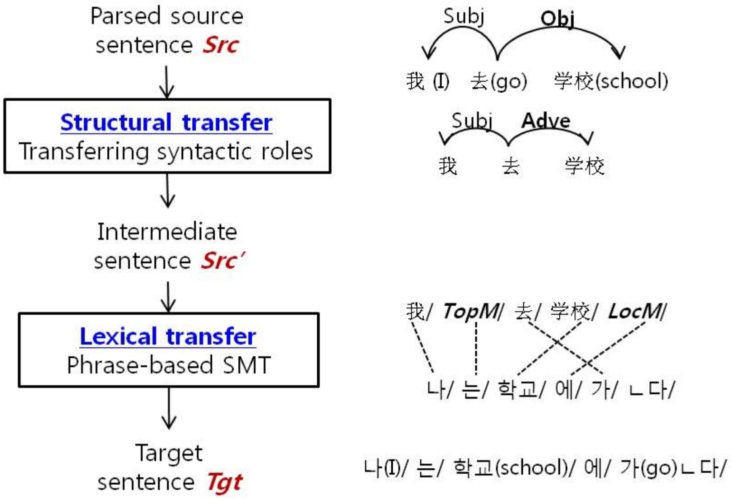

Parsed source sentence *Src*
Structural transfer
Transferring syntactic roles
Intermediate sentence *Src′*
Lexical transfer
Phrase-based SMT
Target sentence *Tgt*
Subj
Obj
我 (I) 去(go) 学校(school)
Subj
Adve
我 去 学校
我/ *TopM*/ 去/ 学校/ *LocM*/
나/ 는/ 학교/ 에/ 가/ ㄴ다/
나(I)/ 는/ 학교(school)/ 에/ 가(go)ㄴ다/

Figure 5.1: System architecture of the proposed method. *TopM* is a pseudo word representing a topic marker, and *LocM* a locative case marker.

5.4.1 Structural transfer as preprocessing: transferring syntactic roles of SVO patterns

To transfer syntactic roles of SVO patterns, we identify grammatical relations in source languages. Specifically, we adopt grammatical relations that are produced by the Stanford Chinese [32, 7] and English [10] typed dependency parsers. The previous work provides 7 grammatical roles that are related to *subject* and *object* in Chinese[1] and English[2].

In this Chapter, the SVO pattern is a general term that represents a construction that consist of any number of the above 7 grammatical relations with a corresponding head predicate.

[1] Chinese typed dependency. *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).

[2] English typed dependency. *nsubj*: nominal subject; *nsubjpass*: passive nominal subject; *csubj*: clausal subject; *csubjpaa*: passive clausal subject; *dobj*: direct object; *iobj*:indirect object; *cop*: copula.

Table 5.1: Statistics of thematic divergences for Chinese *subject* in the Dong-A news corpus.

Chinese Syntactic Relation	Korean Syntactic Relation	Frequency
Chinese *subject*	Korean *subject*	54,080
	Korean *object*	9,759
	Korean *adverbial*	6,092
	Other cases	101,403

Why SVO patterns?

Words with *subject* and *object* relations are strong candidates of arguments, and predicate-argument structures are considered skeletons of sentences. Therefore, well-translated SVO patterns gives a strong impression when judging whether or not the whole sentence is correctly translated.

SVO patterns retain thematic divergences when translating. For example, Chinese SVO patterns can be transferred into various structures such as subject-object-verb, adverb-subject-verb, and adverb-object-verb (Table 5.3)[3]. The statistics are calculated using the word alignment and the syntactic information of source and target languages. Chinese *subjects* are frequently translated into other syntactic relations such as *object* and *adverb*.

In Chinese and English, *subject* and *object* relations are expressed implicitly by word order, while in Korean and Japanese, there exist explicit case markers to show the related syntactic roles. Because we assume the isomorphism of the original SVO patterns and the transferred ones, structural transfer of SVO patterns is equal to the prediction of corresponding case markers in the target languages.

In summary, given an SVO pattern, the transfer module predicts a value for each

[3] Other cases indicate the following three cases: 1. Chinese *subject* is translated into other syntactic relations besides *subject*, *object*, and *adverb*. 2. Or it is null aligned to the corresponding target word. 3. Or it is not translated at all.

syntactic relation from a set of corresponding case marker categories in Korean and Japanese. Since they indicate syntactic roles, this process is identical to resolving the thematic divergences of the SVO pattern when translating into the target languages.

Structured prediction model

As the structured prediction model using preprocessing is an independent subcomponent, various models can be plugged in.

In our previous work, we only extracted the SVO patterns from each sentence and performed the structural transfer [36]. Because *subject* and *object* relations are mutually constrained when transferred into target languages, 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 labeling.

Now, we adopt a phrase-based SMT as our prediction model. The SMT system could be considered a more generalized version when compared to the previous work. The previous model performed the structural transfer without the context information of the sentences, while SMT system refers to the context features of given source sentences.

The input data (***Src***) of the prediction model is syntactically-enriched source sentences with syntactic roles. After translated by the SMT system, the output data (***Src'***) will be equipped with the transferred syntactic roles as the following example. As a result, the phrase-based SMT functions as structured prediction model.

Src: 明天(tomorrow)/ 我(I)/ Subj/ 去(go)/ 学校(school)/ Obj/ 。/
Src': 明天(tomorrow)/ 我(I)/ TopM/ 去(go)/ 学校(school)/ LocM/ ./

The overall SMT performance of our proposed method is much better than that

Table 5.2: 7 representative Korean case marker categories in our structured prediction model.

Korean grammatical functions	Corresponding case marker category
Subject	Nominative case marker (Topic marker)
Object	Accusative case marker
Adverbial	Dative case marker
	Locative case marker
	Instrumental case marker
	Quotative case marker
	Collaborative case marker

Table 5.3: 6 Japanese case markers in our structured prediction model.

Japanese grammatical functions	Corresponding case markers
Subject; Object	が(ga)
Object; Path	を(wo)
Genitive; Subject	の(no)
Dative object; Location	に(ni), には(niwa)
Topic	は(wa)

of the previous one.

Intermediate language generation

Training data for the prediction model is automatically constructed using a word-aligned and dependency-parsed Chinese-Korean/English-Japanese bilingual corpora. In other words, we generate the gold standard data of intermediate sentences (***Src'***). More specifically, for each word with a *subject* or *object* relation in the source sentences, a case marker of the target language is assigned via the word-alignment information.

This process should consider the following two aspects: first, which one of the case markers in the target languages could be observed through the structural transfer? Second, how many case markers should we consider as the output of the structured prediction? If there are too many case markers to predict, it will decrease the prediction accuracy and the overall translation performance. We show such a phenomenon through experiments in Section 5.5.2.

As mentioned earlier, Korean postpositions have great diversity. However, linguists usually consider the case markers listed in Table 5.2 and genitive case marker in Korean sentence generation. For 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. Case markers that belong to these 7 categories cover over 80% of overall usage of Korean postpositions in our corpus.

For the English-Japanese language pair, we refer to the case marker list which was considered in [59]. From the 18 case markers, we only choose 6 case markers listed in Table 5.3 which could be transferred from English *subject* and *object*.

For each word with a *subject* or *object* relation in the source sentence, if the counterpart word is a content morpheme, then we find the Korean *eojeol* or Japanese *bunsetsu* that contains it, and the corresponding postposition that it contains. When the corresponding postpositions do not belong to one of the case markers listed in Table 5.2 or Table 5.3. then we set this case to 'null'.

For Korean, we assign each postposition to one of the case marker categories using the dependency relation of *eojeol* that the postposition is part of. Korean *subject* and *object* relations are mapped to nominative and accusative case markers. For *adverb* relations, we only consider postpositions that match one of the five adverbial case markers in Table 5.2. For Japanese, the process is relatively easy. We only consider the case markers listed in Table 5.3.

To maximize the assignment precision, we use the intersection of the bidirectional word alignments by GIZA++ [48].

5.4.2 Lexical transfer using a phrase-based SMT

After performing structural transfer as a structured prediction, the lexical transfer is conducted using another phrase-based SMT system. As the structural transfer described in our work is a limited version, we still turn on the distortion in the phrase-based SMT, and further let the system deal with the structural differences.

We construct a phrase-based SMT system with the modified Chinese-Korean/English-Japanese bilingual corpora. The Chinese training corpus is converted into an intermediate language enriched with 7 Korean case marker categories, and the English training corpus is enriched with 6 Japanese case markers. 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 baseline is 22.13 BLEU using the original Chinese-Korean bilingual corpus. The oracle system performance is 24.55 BLEU and this will be the upper bound of our proposed method. This suggests that there is much room for improvement using the proposed method. For the English-Japanese language pair, the oracle BLEU is 33.12, while the baseline is 28.50.

5.5 Experiment

Our baseline system is Moses, a state-of-the-art phrase-based SMT system [27], with 5-gram SRI language modeling [58] tuned with Minimum Error Rate Training

Table 5.4: Corpus profile of Dong-A news.

Training	Chinese	Korean	
(98,671 sent.)		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	Chinese	Korean	
(500 sent.)		Content	Function
# words	14,452	9,863	6,875
# singletons	4,012	4,166	162
Sen. length	28.90	19.73	13.75
Test	Chinese	Korean	
(500 sent.)		Content	Function
# words	14,619	9,997	6,928
# singletons	4,009	4,229	154
Sen. length	29.24	19.99	13.86

(MERT) [49]. We adopted NIST [13] and BLEU [52] as our evaluation metrics. [4] Also, a significance test was conducted using a paired bootstrap resampling method [24]. [5] All systems conducted the lexicalized reordering.

Source sentences (***Src***) in the test corpus were first parsed, and transferred into intermediate sentences (***Src'***) as preprocessing. Then they were translated by the SMT system as described in Section 5.4.2.

5.5.1 Corpus profile

We use the same Chinese-Korean parallel corpus in Chapter IV named the Dong-A news. The training corpus has 98,671 sentence pairs, and the development and test

[4] ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v12.pl

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

Table 5.5: Corpus profile of NTCIR-8 patent translation.

Training	English	Japanese	
(1,997,549 sent.)		Content	Function
# words	48,731,806	41,448,462	11,534,611
# singletons	185,329	123,379	115
Sen. length	24.56	20.75	5.77
Development	English	Japanese	
(1,000 sent.)		Content	Function
# words	34,115	29,232	8,116
# singletons	3,456	3,112	58
Sen. length	34.12	29.23	8.15
Test	English	Japanese	
(1,119 sent.)		Content	Function
# words	36,378	31,135	8,239
# singletons	3,800	3,360	55
Sen. length	32.51	27.82	7.36

corpora each have 500 sentence pairs. For Korean, we reported the length of content and function words separately (Table 5.4).

For the English-Japanese language pair, we used the NTCIR-8 patent translation corpus [15]. After cleaning the sentences with the sentence length of 40, the training corpus size becomes 1,997,547. We present the details in Table 5.5.

We used the Stanford Chinese and English typed dependency parser to parse the Chinese and English sentences. The Korean sentences were segmented into morphemes and dependency-parsed using an in-house morphological analyzer and an in-house dependency parser.[6] ChaSen[7] and CobaCha[8] were employed to segment

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

[7] http://chasen.naist.jp/hiki/ChaSen/

[8] http://www.chasen.org/ taku/software/cabocha/

Table 5.6: Experimental results of the Chinese-Korean translation w/ Dong-A news corpus. The BLEU performance with ‡ mark shows significant improvement over the baseline system with a confidence level over 99%.

Method	NIST	BLEU(gain)	Oracle BLEU
Baseline	5.8282	22.13	–
Hong et al.(2009)	5.9638	22.56(+0.43)	–
Proposed	6.0169	23.09(+0.96)‡	24.55

Table 5.7: Experimental results of the English-Japanese translation w/ NTCIR-8 patent translation corpus.The BLEU performances with ‡ mark show significant improvements over the baseline system with a confidence level over 99%.

Method	NIST	BLEU(gain)	Oracle BLEU
Baseline	7.1058	28.50	–
Hong et al.(2009)	7.5091	30.94(+2.44)‡	–
Proposed	7.6011	31.50(+3.00)‡	33.12

and parse Japanese sentences.

5.5.2 Experimental results

Automatic evaluation

Our proposed method shows significant improvement with a confidence level of 99% compared to the baseline phrase-based SMT system both in the Chinese-Korean and English-Japanese translations (Table 5.6 and Table 5.7). NIST values are also consistent with BLEU scores and show improvements.

[20]'s method is also a preprocessing method that enriches the source language with the syntactic relations of source sentences as pseudo words. For comparison purposes, we implemented and carried out their proposed method by inserting the Chinese or English syntactic relations as pseudo words. Their method did not show

Table 5.8: Human evaluation result of the Chinese-Korean translation comparing [20]'s method 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	**9**	4	1
P	2	**33**	5
E	11	19	**16**

significant improvements in the Chinese-Korean Dong-A news corpus.

In the English-Japanese translation, our proposed approach shows significant improvement over [20]'s method with a confidence level of 99%.

Human evaluation of Chinese-Korean translation

Since the BLEU metric does not always correlate with the human evaluation, we randomly selected 100 sentences on which to perform the human evaluations. The comparison target to our proposed method is [20]'s method, because our proposed method does not show significant improvement over their method in the Chinese-Korean language pair.

We adopted the human evaluation measure proposed by [60]. Two annotators compared the translation quality in terms of adequacy and fluency (Table 5.8). The reference translation was given to annotators, but without the source sentence.

The diagonal values in Table 5.8 show the agreement between two annotators for the Chinese-Korean language pair. We further measured the agreement between the annotators using the Kappa statistic. The Kappa value is only 0.348 when considering Hong's (H), Proposed (P), and Equal quality (E) categories. However, excluding the uncertain evaluation result E, the Kappa value is 0.667. This value falls within the scope of a substantial agreement.

Chinese-Korean translation using various linguistic features

We also conducted experiments to observe the trade-off between the structured prediction model and the translation model in Chinese-Korean translation using various linguistic features (Table 5.9).

In our prediction model (Section 5.4.1), source languages are enriched with the syntactic roles such as *subject* and *object* (***Src***). Now we further enrich them with the semantic relations. In other words, we consider the predicate-argument structures of the given sentences instead of SVO patterns. The source sentences are first semantically parsed [9] and for the arguments of the predicate we inserted semantic relations as pseudo words such as Arg$n$ (n=0, 1, 2, 3, 4, or 5).

As the intermediate language (***Src'***) is enriched with the case markers of target languages, we try another set of case markers besides the 7 representative case marker categories described in Section 5.4.1. We select the 30 most frequently used case markers in the Korean corpus. This set of case markers covers over 95% of the overall usage of Korean postpositions, which is more of the data than covered by that of 7 representative case marker categories.

When the top 30 case markers are used, the oracle translation BLEU score is better than only using the 7 representative case marker categories. However, at the same time the performance of the prediction model becomes lower because of the increase in the number of case marker categories that need to be predicted. As a result, the overall translation is worse than the one using the 7 case marker categories and is not a significant improvement.

Similar BLEU scores are achieved by enriching source sentences with syntactic or semantic relations. The prediction BLEU of semantic relations and the overall translation performance are slightly lower than those of syntactic relations.

[9] http://hlt030.cse.ust.hk/research/c-assert/

Table 5.9: Chinese-Korean translation using various linguistic features.

System Description	NIST	BLEU(gain)	Prediction BLEU	Oracle Translation BLEU
Baseline	5.8282	22.13	–	–
Src: syntactic relations ***Src'***: 7 representative case marker categories	6.0169	23.09(+0.96)‡	89.36	24.55
Src: syntactic relations ***Src'***: top 30 case markers	5.9588	22.58(+0.45)	83.30	25.66
Src: semantic relations ***Src'***: 7 representative case marker categories	5.9825	23.00(+0.87)‡	89.05	24.55

Table 5.10: English-Japanese translation using various corpus size.

Size (sentence pairs)	Baseline		Proposed Method		Oracle	
	NIST	BLEU	NIST	BLEU(gain)	NIST	BLEU
100K	6.3398	23.72	6.6873	25.83(+2.11)	6.7732	27.15
500K	6.9754	27.60	7.4395	30.27(+2.67)	7.5478	31.74
1M	6.9857	27.90	7.4675	30.86(+2.96)	7.6119	32.55
All	7.1058	28.50	7.6011	31.50(+3.00)	7.7333	33.12

English-Japanese translation using various corpus size

Various corpus size is used to verify that our proposed work is robust enough regardless of varing the corpus size (Table 5.10). The improvement gain is gradually grown as the corpus size becomes larger.

5.6 Conclusion

In this Chapter, we focus on the thematic divergences of syntactic roles such as *subject* and *object* between source and target languages. We resolved the thematic divergence problem as a structured prediction model with an automatically constructed training corpus. Our method also contributes to the generation of grammatical outputs as the syntactic roles of Korean and Japanese languages are expressed by the case markers.

Our proposed method showed significant improvements both in Chinese-Korean and English-Japanese language pairs. For comparison purposes, we implemented a previous work and compared the translations through automatic and human evaluations, and we showed that our method performs better than the previous method. Although we use some language-specific linguistic knowledge, the overall procedure can be easily applied to other language pairs which suffer from similar linguistic issues.

From the extended experiments, we find that the trade-off between the prediction and the translation accuracy needs further investigation which we leave as our future work. Also, we would like to cover a more general structural transfer in addition to the thematic divergences.

VI. Structural Transfer as Preprocessing for SMT

6.1 Structural transfer as preprocessing

In Chapter IV and V, we introduce syntactic reordering and resolving thematic divergences as preprocessing for SMT, respectively. Both approaches are related to disambiguating structural differences of source and target languages and they are complementary to each other. Therefore, in this Chapter we combine the two approaches to perform a more generalized structural transfer.

We simply cascade the two approaches as follows. The input sentences are syntactically reordered, then the thematic divergences of *subject* and *object* relations of the reordered sentences are resolved by the approach proposed in Chapter V. Or we first resolve the thematic divergences then perform syntactic reordering.

Since in Chapter IV, only Chinese syntactic reordering is addressed, we briefly describe the English syntactic reordering in this Section. English syntactic reordering rules are compiled using the same principle suggested for general syntactic reordering of Chinese (Algorithm 2 [1]).

Every predicate in a English 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

[1] Followings are a set of dependency relations defined in Stanford English typed dependency parser. aux: auxiliary; auxpass: passive auxiliary; neg: negation modifier; cop: copula; prt: phrasal verb particle; advcl: adverbial clause modifier; conj: conjunction; cc: coordination; punct: punctuation;

Algorithm 2 English syntactic reordering rules of predicate

Input: *L_Children*, *R_Children* of a Predicate P

Output: *L_Advcl*, *L_Other*, *L_FromRight*, *L_Modal*, *R_Modal*, *R_Remnant*

```
for node N in L_Children do
    if dep.relation of N $\in$ {aux, auxpass, neg, cop} then
        $L\_Modal \Leftarrow L\_Modal + \{N\}$
    else
        $L\_Other \Leftarrow L\_Other + \{N\}$
    end if
end for
for node N in R_Children do
    if dep.relation of N $\in$ {prt} then
        $R\_Modal \Leftarrow R\_Modal + \{N\}$
    else if dep.relation of N $\in$ {advcl} then
        $L\_Advcl \Leftarrow L\_Advcl + \{N\}$
    else if dep.relation of N $\in$ {conj, cc, punct} then
        $R\_Remnant \Leftarrow R\_Remnant + \{N\}$
    else
        $L\_FromRight \Leftarrow L\_FromRight + \{N\}$
    end if
end for
```

left side of the predicate.

For the right children, the modality-bearing words (*R_Modal*) are relocated near the predicate, as *L_Modal*. However, most right children are moved to the left side of the predicate (*L_FromRight*) since Japanese is a verb-final language. A right child belonging to *R_Remnant* forms coordination or when the right child is a punctuation.

For the adverbial clause modifier, we move it to the leftmost position (*L_Advcl*) of the given predicate considering it is usually translated in the beginning of Japanese

Table 6.1: Evaluation results of Chinese-Korean language pair.

Method	Morpheme (4-gram)		*Eojeol* (4-gram)	
	NIST	BLEU	NIST	BLEU
Baseline	5.8282	22.13	3.3010	6.97
1. Syntactic reordering	6.1135	24.63	3.6878	9.63
2. Resolving thematic divergences	6.0296	23.09	3.4999	7.47
First 1, then 2	**6.1859**	**24.66**	3.7221	9.40
First 2, then 1	6.1381	24.59	**3.7681**	**9.66**

Table 6.2: Evaluation results of English-Japanese language pair.

Method	Morpheme (4-gram)		*Bunsetsu* (4-gram)	
	NIST	BLEU	NIST	BLEU
Baseline	7.1058	28.50	2.2684	1.64
1. Syntactic reordering	7.4763	31.09	2.4876	2.83
2. Resolving thematic divergences	**7.6011**	31.50	2.3667	2.12
First 1, then 2	7.4610	31.69	**2.5684**	**2.98**
First 2, then 1	7.4547	**31.89**	2.5489	2.74

sentences.

6.2 Experimental results and discussion

The corpus and experimental setting is identical to used in Chapter V. We combine both approaches of syntactic reordering and resolving thematic divergences to perform structural transfer, and achieve better performances than applying each approach independently (Table 6.1 and Table 6.2).

We suggest to measure the effectiveness of our proposed method with lexical units of various granularity. The output of the Chinese-Korean and English-Japanese SMT system is morphemes. In most SMT systems, only morpheme-based result is reported where the target language is a morphologically rich language such as Korean and Japanese. However, *eojeol* and *bunsetsu* are the basic lexical units in Korean and Japanese respectively which contains functional markers to indicate grammatical roles; *eojeol*/*bunsetsu*-based metric provides more meaningful evaluation result than morpheme-based. We test our proposed method with 4-gram morpheme-BLEU/NIST and 4-gram *eojeol*/*bunsetsu*-BLEU/NIST.

In order to recover *eojeol*, we first omit all the spaces in the Korean output, and then re-segment it into *eojeol*. The segmentation problem can be resolved by a CRF model as a sequence labeling problem. We adopt BIO[2] 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. The similar process is also performed to recover Japanese *bunsetsu* and the accuracy is 99.6%.

The *bunsetsu*-BLEU/NIST values are much lower than the *eojeol*-BLEU/NIST although the morpheme-based evaluation values in the English-Japanese SMT are higher than in the Chinese-Korean SMT. The reason lies in that each *bunsetsu* contains 3.2 morphemes, while *eojeol* only includes 2.2 morphemes in average.

In Chinese-Korean language pair, either of the combined approaches show significant improvement in morpheme-NIST with a confidence level over 99%, while insignificant in morpheme-BLEU. In English-Japanese language pair, the combined approaches show improvement in morpheme-BLEU, while not in morpheme-NIST. Morpheme-based evaluation shows inconsistent evaluation result in BLEU and NIST, while *eojeol*/*bunsetsu*-based evaluation leads to consistent improvement in both SMT systems. As *eojeol*/*bunsetsu*-based evaluation measure is more meaningful

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

for translating into morphologically rich target languages, the evaluation results reveal the effectiveness of our proposed methods.

BLEU and NIST metrics are both precision- and n-gram-based, while they highlight different points when evaluation. BLEU favors fluency by incorporating matches of high n-grams. NIST emphasizes lexical choice over word order and does not consider structural information. Korean and Japanese both belong to relatively free word order language family, and the grammatical roles are indicated by the functional markers, therefore *eojeol*/*bunsetsu*-NIST metric is more reliable than BLEU. Our experimental results also support this conclusion.

6.3 Conclusion

In this Chapter, we combine the structural transfer methods suggested in Chapter IV and V as a cascaded model. Through the experiment we proved the effectiveness of our proposed method using lexical units of various granularity such as 4-gram morpheme-based, and 4-gram *eojeol*/*bunsetsu*-based BLEU/NIST.

The combination method is relatively simple and loose, therefore developing other incorporation methods will be our future work.

VII. Conclusion and Future Work

7.1 Conclusion

As a main research topic of this Thesis, we presented a framework that first resolve the structural differences as syntax-aided preprocessing then resolve the lexical differences using a phrase-based SMT. In other words, we strengthen the structural transfer of the phrase-based SMT whose capacity for lexical transfer is widely proved. We explored useful linguistic knowledge and encoded it into SMT systems to effectively disambiguate the structural differences of Chinese-Korean and English-Japanese language pairs. Besides this, we also presented annotation guidelines for Chinese-Korean word alignment by contrastively analyzing the complex morpho-syntactic encoding systems in Chapter III.

First, Chinese syntactic reordering approach with an emphasis on predicate-predicate patterns was proposed for the phrase- and hierarchical phrase-based SMT. We examined the predicate-predicate patterns 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.

Then, we focused on the thematic divergences of syntactic roles such as *subject* and *object* between source and target languages. We resolved the thematic divergence problem as a structured prediction model with an automatically constructed training corpus. This method contributed to the generation of grammatical outputs since the syntactic roles of Korean and Japanese languages are expressed by the case markers. Our proposed method showed significant improvements both in Chinese-Korean and English-Japanese language pairs. Although we use some language-specific linguistic knowledge, the overall procedure can be easily applied to other language pairs which suffer from similar linguistic issues.

Finally, we combined the two approaches to perform a more generalized structural transfer between two languages that are complementary to each other. The input sentences were syntactically reordered, then the thematic divergences of *subject* and *object* relations of the reordered sentences were resolved. Or we first resolve the thematic divergences, then perform the syntactic reordering.

In conclusion, we tackled the low capability of structural transfer in today's phrase-based SMT systems, while not compromising the advantages of phrase-based SMT such as lexical transfer and local reordering. Various linguistically motivated features were exploited and we showed how to effectively encode them into the phrase-based SMT.

7.2 Future work

The proposed structural transfer was performed as preprocessing of phrase-based SMT in a loosely-coupled way. Sometimes although the structural transfer is conducted well, the phrase-based model cannot fully reflect the correctly enriched information. We will study a more tight incorporation method to prevent loosing the offered information from the structural transfer.

Besides the structural differences proposed in this Thesis, there are still many differences between the two language pairs. For example, one of the differences is the morpho-syntactic encoding system of modality information as mentioned in Chapter III. To generate correct verbal endings of Korean and Japanese is a difficult task because the corresponding information is scattered widely over the sentence or implicitly expressed in the source languages. The nearest future task will be mining useful linguistic features to support correct generation of Korean/Japanese verbal endings which indicate modality information.

Summary in Korean
요 약 문

본 논문에서는 중한 또는 영일과 같이 형태구문적으로 상이한 언어쌍에 대한 통계기계번역에서, 언어학적 정보를 이용한 전처리를 통하여 구조변환을 하고 다음 구기반 통계기계번역 시스템을 이용하여 어휘변환을 진행하는 방법을 제안한다. 이 방법은 널리 알려진 구기반 통계기계번역(phrase-based SMT)의 어휘변환 능력을 훼손시키지 않으면서 약점인 구조변환 능력을 강화할 수 있는 방법이다. 본 논문은 중한/영일 언어쌍의 형태구문적 상이성을 어순 및 형태소 유형론적 측면에서 세밀히 대조분석(contrastive analysis)하고, 이런 상이성을 해소하기 위하여 관련 언어학적 정보를 개발하고, 또 구기반 통계기계번역에서 어떻게 효과적으로 활용하는 지를 중점적으로 보여준다.

먼저 어순 유형론적으로 볼때 중국어와 한국어는 서로 매우 다른 어순을 가지고 있는데, 특히 중국어 문장에서 용언-용언 패턴을 중심으로 효과적인 구문적 재배치(syntactic reordering) 방법을 제안하고, 구기반과 계층적 구기반 통계기계번역 시스템에서 그 유효성을 검증한다. 용언-용언 패턴은 한국어로 번역될 때 먼거리 재배치(long-distance reordering) 문제를 야기시키는데, 우선 구체적으로 어떤 구조가 재배치를 통하여 번역향상에 도움이 되는지를 판별하고, 유용한 언어학적 지식을 개발하여 이 패턴의 재배치 문제를 해결한다.

다음으로 중한/영일 모두 형태적으로 빈약한(morphologically-poor) 언어(고립어)에서 형태적으로 풍부한(morphologically-rich) 언어(교착어)로의 번역방향을 가지는데 이는 그 반대인 경우에 비해 번역하기가 훨씬 어려운 특징을 가지고 있다. 목표언어의 복잡한 형태소를 효과적으로 생성하는 방법으로서 특히 두 언어쌍의 번역에서 주어와 목적어의 주제불일치(thematic divergence) 문제를 해결한다. 그 이유는 중영에서는 주어와 목적어 구문관계(syntactic relation)를 어순을 통하여

암묵적으로 표현을 하지만 한일에서는 격조사를 통하여 명시적으로 표현을 할 뿐만아니라, 목표언어로 번역 될 때 구문관계의 불일치 현상 즉 주어와 목적어가 빈번하게 다른 구문관계로 번역이 되기 때문이다. 전처리를 통하여 주어와 목적어의 주제불일치 문제를 구조적 예측 모델(structured prediction model)로 전환하여 풀고, 미리 예측된 주어와 목적어의 구문관계 정보를 원시언어에 삽입하는 방법을 통하여 형태구문적 차이를 줄이고자 한다.

마지막으로 위에서 제안한 두 가지 방법을 순차적으로 결합하여 보다 일반적인 구조변환 방법을 제시한다. 먼저 원시언어의 구문적 재배치를 하고 주어와 목적어 구문관계의 주제불일치 문제를 해결하거나, 또는 먼저 주제불일치 문제를 해결하고 다음으로 구문적 재배치를 실행하는 두 가지 방법을 제안한다.

References

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

[2] Alexandra Birch, Miles Osborne, and Philipp Koehn. Ccg supertags in factored statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 9–16, Morristown, NJ, USA, 2007. Association for Computational Linguistics.

[3] Peter F. Brown, Vincent J.Della Pietra, Stephen A. Della Pietra, and Robert. L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19:263–311, 1993.

[4] Jean Carletta. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22:249–254, 1996.

[5] Federico Marcello Pighin Daniele Cettolo, Mauro and Bertoldi Nicola. Shallow-syntax phrase-based translation: Joint versus factored string-to-chunk models. In *8th Conference of the Association for Machine Translation in the Americas (AMTA 2008)*, 2008.

[6] Pi-Chuan Chang, Daniel Jurafsky, and Christopher D. Manning. Disambiguating "DE" for Chinese-English machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 215–223, Athens, Greece, March 2009. Association for Computational Linguistics.

[7] Pi-Chuan Chang, Huihsin Tseng, Dan Jurafsky, and Christopher D. Manning. 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 2009. Association for Computational Linguistics.

[8] David Chiang. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 263–270, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[9] Michael Collins, Philipp Koehn, and Ivona Kucerova. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 531–540, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[10] Maire-Catherine de Marneffe and Christopher D. Manning. Stanford typed dependencies manual. Technical report, Stanford University, 2008.

[11] Steve DeNeefe, Kevin Knight, Wei Wang, and Daniel Marcu. What can syntax-based MT learn from phrase-based MT? In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 755–763, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[12] Yuan Ding and Martha Palmer. Machine translation using probabilistic synchronous dependency insertion grammars. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 541–548, Morristown, NJ, USA, 2005. Association for Computational Linguistics.

[13] George Doddington. 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, 2002. Morgan Kaufmann Publishers Inc.

[14] Jakob Elming. Syntactic reordering integrated with phrase-based SMT. In *Proceedings of the ACL-08: HLT Second Workshop on Syntax and Structure in Statistical Translation (SSST-2)*, pages 46–54, Columbus, Ohio, June 2008. Association for Computational Linguistics.

[15] Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, Takehito Utsuro, Terumasa Ehara, Hiroshi Echizen-ya, and Sayori Shimohata. Overview of the patent translation task at the ntcir-8 workshop. In *Proc. of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access*, June 2010.

[16] Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. Scalable inference and training of context-rich syntactic translation models. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 961–968, Sydney, Australia, July 2006. Association for Computational Linguistics.

[17] Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. What's in a translation rule? In Daniel Marcu Susan Dumais and Salim Roukos, editors, *HLT-NAACL 2004: Main Proceedings*, pages 273–280, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics.

[18] Dmitriy Genzel. Automatically learning source-side reordering rules for large scale machine translation. In *Proceedings of the 23rd International Conference*

on Computational Linguistics (Coling 2010), pages 376–384, Beijing, China, August 2010. Coling 2010 Organizing Committee.

[19] Hany Hassan, Khalil Sima'an, and Andy Way. Supertagged phrase-based statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 288–295, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[20] Gumwon Hong, Seung-Wook Lee, and Hae-Chang Rim. 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 2009. Association for Computational Linguistics.

[21] W. John Hutchins. Machine translation: history and general principles. In R. E. Asher, editor, *The encyclopedia of languages and linguistics*, volume 5, pages 2322–2332. Oxford: Pergamon Press, 1994.

[22] Jungi Kim Hwidong Na, Jin-Ji Li and Jong-Hyeok Lee. Improving fluency by reordering target constituents using mst parser in English-to-Japanese phrase-based smt. In *MT Summit XII: proceedings of the twelfth Machine Translation Summit*, pages 276–283, Ottawa, Ontario, Canada, August 2009. International Association for Machine Translation hosted by the Association for Machine Translation in the Americas.

[23] Changhu Jin. Improving korean-to-chinese phrase-based statistical machine translation using enhanced word alignment. Master's thesis, POSTECH, 2010.

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

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

[26] Philipp Koehn. *Statistical Machine Translation.* Cambridge university press, 2010.

[27] 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. 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 2007. Association for Computational Linguistics.

[28] Klaus Krippendorff. *Content Analysis: an Introduction to its Methodology.* Sage Publications, 1980.

[29] Patrik Lambert, Adrià de Gispert, Rafael E. Banchs, and José B. Mariño. Guidelines for word alignment evaluation and manual alignment. *Language Resources and Evaluation*, 39(4):267–285, 2005.

[30] Hyo Sang Lee. *Tense, aspect, and modality: A discourse-pragmatic analysis of verbal affixes in Korean from a typological perspective.* PhD thesis, University of California, Los Angeles, 1991.

[31] Young-Suk Lee, Bing Zhao, and Xiaoqian Luo. Constituent reordering and syntax models for english-to-japanese statistical machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics*

(Coling 2010), pages 626–634, Beijing, China, August 2010. Coling 2010 Organizing Committee.

[32] Roger Levy and Christopher D. Manning. 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 2003. Association for Computational Linguistics.

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

[34] Chi-Ho Li, Minghui Li, Dongdong Zhang, Mu Li, Ming Zhou, and Yi Guan. A probabilistic approach to syntax-based reordering for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 720–727, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[35] Jin-Ji Li, Jungi Kim, Dong-Il Kim, and Jong-Hyeok Lee. Chinese syntactic reordering for adequate generation of Korean verbal phrases in Chinese-to-Korean SMT. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 190–196, Athens, Greece, March 2009. Association for Computational Linguistics.

[36] Jin-Ji Li, Jungi Kim, and Jong-Hyeok Lee. Transferring syntactic relations of Subject-Verb-Object pattern in Chinese-to-Korean SMT. In *Proceedings of the Association for Machine Translation in the Americas 2010*, October 2010.

[37] Jin-Ji Li, Ji-Eun Roh, Dong-Il Kim, and Jong-Hyeok Lee. Contrastive analysis and feature selection for korean modal expression in chinese-korean machine translation system. *Int. J. Comput. Proc. Oriental Lang.*, 18(3):227–242, 2005.

[38] Yang Liu, Yun Huang, Qun Liu, and Shouxun Lin. Forest-to-string statistical translation rules. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 704–711, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[39] Yang Liu, Qun Liu, and Shouxun Lin. Tree-to-string alignment template 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*, ACL-44, pages 609–616, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

[40] Yang Liu, Yajuan Lü, and Qun Liu. Improving tree-to-tree translation with packed forests. 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: Volume 2 - Volume 2*, ACL-IJCNLP '09, pages 558–566, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[41] Daniel Marcu, Wei Wang, Abdessamad Echihabi, and Kevin Knight. Spmt: statistical machine translation with syntactified target language phrases. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 44–52, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

[42] Dan Melamed. Manual annotation of translational equivalence: The blinker project. Technical report, University of Pennsylvania, 1998.

[43] Arul Menezes and Chris Quirk. Using dependency order templates to improve generality in translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 1–8, Morristown, NJ, USA, 2007. Association for Computational Linguistics.

[44] Magnus Merkel. Annotation style guide for the plug link annotater. Technical report, Linköping University, 1999.

[45] Haitao Mi and Liang Huang. Forest-based translation rule extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 206–214, Morristown, NJ, USA, 2008. Association for Computational Linguistics.

[46] Haitao Mi and Qun Liu. Constituency to dependency translation with forests. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1433–1442, Morristown, NJ, USA, 2010. Association for Computational Linguistics.

[47] Einat Minkov, Kristina Toutanova, and Hisami Suzuki. 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 2007. Association for Computational Linguistics.

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

[49] Franz Josef Och. 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 2003. Association for Computational Linguistics.

[50] Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, and Dragomir Radev. A smorgasbord of features for statistical machine translation. In Daniel Marcu Susan Dumais and Salim Roukos,

editors, *HLT-NAACL 2004: Main Proceedings*, pages 161–168, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics.

[51] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

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

[53] Chris Quirk, Arul Menezes, and Colin Cherry. Dependency treelet translation: syntactically informed phrasal smt. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 271–279, Morristown, NJ, USA, 2005. Association for Computational Linguistics.

[54] Ananthakrishnan Ramanathan, Hansraj Choudhary, Avishek Ghosh, and Pushpak Bhattacharyya. 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 2009. Association for Computational Linguistics.

[55] Hermann Ney Saˇsa Hasan, Oliver Bender. Reranking translation hypotheses using structural properties. In *In Proceedings of the EACL'06 Workshop on Learning Structured Information in Natural Language Applications*, pages 41–48, Trento, Italy, April 2006.

[56] Libin Shen, Jinxi Xu, and Ralph Weischedel. A new string-to-dependency machine translation algorithm with a target dependency language model. In

Proceedings of ACL-08: HLT, pages 577–585, Columbus, Ohio, June 2008. Association for Computational Linguistics.

[57] Lucia Specia, Baskaran Sankaran, and Maria Das Graças Volpe Nunes. n-best reranking for the efficient integration of word sense disambiguation and statistical machine translation. In *Proceedings of the 9th international conference on Computational linguistics and intelligent text processing*, CICLing'08, pages 399–410, Berlin, Heidelberg, 2008. Springer-Verlag.

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

[59] Kristina Toutanova and Hisami Suzuki. 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 Conference*, pages 49–56, Rochester, New York, April 2007. Association for Computational Linguistics.

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

[61] Zhaopeng Tu, Yang Liu, Young-Sook Hwang, Qun Liu, and Shouxun Lin. Dependency forest for statistical machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1092–1100, Beijing, China, August 2010. Coling 2010 Organizing Committee.

[62] Karthik Visweswariah, Jiri Navratil, Jeffrey Sorensen, Vijil Chenthamarakshan, and Nandakishore Kambhatla. Syntax based reordering with automatically

derived rules for improved statistical machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1119–1127, Beijing, China, August 2010. Coling 2010 Organizing Committee.

[63] Jean Vronis and Philippe Langlais. Evaluation of parallel text alignment systems - the arcade project, 2000.

[64] Chao Wang, Michael Collins, and Philipp Koehn. 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 (EMNLP-CoNLL)*, pages 737–745, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[65] Dekai Wu and Pascale Fung. Semantic roles for smt: a hybrid two-pass model. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, NAACL '09, pages 13–16, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[66] Fei Xia and Michael McCord. Improving a statistical mt system with automatically learned rewrite patterns. In *Proceedings of Coling 2004*, pages 508–514, Geneva, Switzerland, Aug 23–Aug 27 2004. COLING.

[67] Richard Xiao. *A corpus-based study of aspect in Mandarin Chinsese.* PhD thesis, University of Lancaster, 2002.

[68] Deyi Xiong, Qun Liu, and Shouxun Lin. A dependency treelet string correspondence model for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 40–47, Morristown, NJ, USA, 2007. Association for Computational Linguistics.

[69] Peng Xu, Jaeho Kang, Michael Ringgaard, and Franz Och. 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*, pages 245–253, Boulder, Colorado, June 2009. Association for Computational Linguistics.

[70] Yushi Xu and Stephanie Seneff. Two-stage translation: A combined linguistic and statistical machine translation framework. In *Proceedings of the Association for Machine Translation in the Americas 2008*, pages 222–231, Waikiki, Hawaii, USA, Oct 21–Oct 25 2008. AMTA.

[71] Nianwen Xue, Fei Xia, Shizhe Huang, and Anthony Kroch Kroch. The bracketing guidelines for the penn chinese treebank (3.0). Technical report, IRCS Technical Report, University of Pennsylvania, 2000.

[72] Hui Zhang, Min Zhang, Haizhou Li, Aiti Aw, and Chew Lim Tan. Forest-based tree sequence to string translation model. 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: Volume 1 - Volume 1*, ACL-IJCNLP '09, pages 172–180, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[73] Min Zhang, Hongfei Jiang, AiTi Aw, Haizhou Li, Chew Lim Tan, and Sheng Li. A tree sequence alignment-based tree-to-tree translation model. In *ACL*, pages 559–567, 2008.

[74] Yuqi Zhang, Richard Zens, and Hermann Ney. Chunk-level reordering of source language sentences with automatically learned rules for statistical machine translation. In *Proceedings of the NAACL-HLT 2007/AMTA Workshop on*

Syntax and Structure in Statistical Translation, SSST '07, pages 1–8, Morristown, NJ, USA, 2007. Association for Computational Linguistics.

Curriculum Vitae

Education

1995. 9. – 1999. 7. Computer Science and Engineering, Jilin University of Technology, P. R. China (B.S.)

2001. 3. – 2003. 2. Computer Science and Engineering, Pohang University of Science and Technology, R. Korea (M.S.)

2005. 3. – 2011. 2. Computer Science and Engineering, Pohang University of Science and Technology, R. Korea (Ph.D.)

Experience

1999. 9. – 2001. 2. Instructor, Yanbian University of Science and Technology, P. R. China

2003. 3. – 2005. 2. Researcher, POSTECH Information Research Laboratories, Pohang University of Science and Technology

Publication

International Journal

1. **Jin-Ji Li**, Jungi Kim, Jong-Hyeok Lee, "Resolving Thematic Divergences of Subject and Object Relations for SMT," Int'l J. of Computer Processing of Oriental Languages, 2010 (submitted)

2. **Jin-Ji Li**, Ji-Eun Roh, Dong-Il Kim, Jong-Hyeok Lee, "Contrastive Analysis and Feature Selection for Korean Modal Expression in Chinese-Korean Machine Translation System," Int'l J. of Computer Processing of Oriental Languages, 2005, Vol. 18(3), pp. 227-242 (Sep. 2005)

3. Dong-il Kim, **Jin-Ji Li**, Jong-Hyeok Lee, "Matching Abnormality in Hybrid Machine Translation," Int'l J. of Computer Processing of Oriental Languages, 2004, Vol. 17(4), pp. 253-272 (Dec. 2004)

4. Dong-il Kim, Zheng Cui, **Jin-Ji Li**, Jong-Hyeok Lee, "Resolving Structural Transfer Ambiguity in Chinese-to-Korean Machine Translation," Int'l J. of Computer Processing of Oriental Languages, 2003, Vol. 16(2), pp. 119-142 (Jun. 2003)

Domestic Journal

1. 김한경, 나휘동, **이금희**, 이종혁, "문장구조 유사도와 단어 유사도를 이용한 클러스터링 기반의 통계기계번역", 정보과학회논문지 : 소프트웨어 및 응용, 2010, 제37권 제4호, pp. 297 304 (2010.4)

2. 김한경, 나휘동, **이금희**, 이종혁, "일영 통계기계번역에서 의존문법 문장 구조와 품사 정보를 사용한 클러스터링 기법", 정보과학회논문지 : 컴퓨팅의 실제, 2009, 제15권 제12호, pp. 993 997 (2009.12)

3. 백설매, **이금희**, 김동일, 이종혁, "확장청크와 세분화된 문장부호에 기반한 중국어 최장명사구 식별", 정보과학회논문지 : 소프트웨어 및 응용, 2009, 제36

권 제4호, pp. 320 328 (2009.4)

International Conference

1. **Jin-Ji Li**, Jungi Kim, and Jong-Hyeok Lee. "Transferring Syntactic Relations of Subject-Verb-Object Pattern in Chinese-to-Korean SMT", AMTA, Denver/Colorado: 237-246. (Oct. 31 - Nov. 4, 2010)
2. **Jin-Ji Li**, Jungi Kim, and Jong-Hyeok Lee. "Chinese Syntactic Reordering through Contrastive Analysis of Predicate-predicate Patterns in Chinese-to-Korean SMT", AMTA, Denver/Colorado:277-285. (Oct. 31 - Nov. 4, 2010)
3. Hwidong Na, **Jin-Ji Li**, Yeha Lee and Jong-Hyeok Lee, "A Synchronous Context Free Grammar using Dependency Sequence for Syntax-base Statistical Machine Translation", AMTA student workshop, Denver/Colorado: 341-349 (Oct. 31 - Nov. 4, 2010)
4. Jungi Kim, **Jin-Ji Li**, Jong-Hyeok Lee, "Evaluating Multilanguage-Comparability of Subjectivity Analysis Systems", ACL, Uppsala/Sweden: 595-603, (Jul. 11-16, 2010)
5. Hwidong Na, **Jin-Ji Li**, Jungi Kim and Jong-Hyeok Lee, "Improving Fluency by Reordering Target Constituents using MST Parser in English-to-Japanese phrase-based SMT", MT Summit XII, Ottawa/Canada: 276-283 (Aug. 26-30, 2009)
6. Jungi Kim, **Jin-Ji Li**, Jong-Hyeok Lee, "Discovering the Discriminative Views: Measuring Term Weights for Sentiment Analysis", ACL-IJCNLP, Singapore: 253-261 (Aug. 2-7, 2009)
7. **Jin-Ji Li**, Jungi Kim, Dong-Il Kim, Jong-Hyeok Lee, "Chinese Syntactic Reordering for Adequate Generation of Korean Verbal Phrases in Chinese-to-Korean SMT", EACL - 4th WS on SMT, Athens/Greece: 595-603, (Mar. 30-31, 2009)

8. **Jin-Ji Li**, Hwi-Dong Na, Hankyong Kim, Chang-Hu Jin, Jong-Hyeok Lee, “The POSTECH Statistical Machine Translation Systems for NTCIR-7 Patent Translation Task,” NTCIR-7, Tokyo/Japan: 445-449 (Dec. 16-19, 2008)

9. **Jin-Ji Li**, Dong-Il Kim, Jong-Hyeok Lee, “Annotation Guideline for Chinese-Korean Word Alignment” CD Proceeding of LREC 2008 (6th Int’l Conf. on Language Resources and Evaluation), Marrakech/Morocco: - (May. 28-30. 2008)

10. **Jin-Ji Li**, Dong-il Kim, Jong-Hyeok Lee, “Annotation Guidelines for Chinese-Koran Word Alignment through Contrastive Analysis of Morpho-syntactic Encodings,” SSMT’2007 (The 3rd Symposium on Statistical Machine Translation), Harbin/China: 39-45 (Aug.12-13, 2007)

11. Dong-il Kim, **Jin-Ji Li**, Jong-Hyeok Lee, “Ambiguity Resolution of Complement Transfer in Chinese-to-Korean Machine Translation,” ICMIP’07 (Int’l Conf. on Multilingual Information Processing), Yanbian/China: 29-42 (Jul. 3-5, 2007)

12. Xue-Mei Bai, **Jin-Ji Li**, Dong-Il Kim, Jong-Hyeok Lee, “Identification of Maximal-Length Noun Phrases Based on Expanded Chunks and Classified Punctuations in Chinese,” Lecture Notes in Artificial Intelligence, Vol. 4285 (2006): Computer Processing of Oriental Languages, Beyond the Orient: The Research Challenges Ahead – ICCPOL 2006 (Singapore, Dec. 17-19, 2006) (Eds.: Yuji Matsumoto, Richard Sproat, et al.), Springer, pp. 268-276 (Dec. 2006) (IF=0.302, LNAI, SCI-E)

13. **Jin-Ji Li**, Dong-Il Kim, Jong-Hyeok Lee, “Contrastive Analysis and Feature Selection for Korean Predicate Generation in Chinese-Korean Machine Translation,” ICCLC2004 (Int’l Conf. on Chinese Language Computing), San Francisco/USA: 507-513 (Sep. 8-10, 2004)

Domestic Conference

1. 김한경, 나휘동, **이금희**, 이종혁, “문장구조 유사도와 단어 유사도를 이용한 클러스터링 기반의 통계기계번역”, 한글 및 한국어 정보처리 학술대회: 44-49 (2009년 10월 9-10일)
2. 김장호, **이금희**, 나휘동, 김동일, 이종혁, “통계적 수정규칙을 이용한 한국어-중국어 단어정렬 개선방법”, 제21회 한글 및 한국어 정보처리 학술대회: 23-30 (2009년 10월 9-10일)
3. 김한경, 나휘동,**이금희**,이종혁, “일영 통계기계번역에서 의존문법 문장 구조와 품사 정보를 사용한 클러스터링 기법”, 한국정보과학회 2009 한국컴퓨터종합학술대회 논문집 제36권 제1호(A): 88-89 (2009년 7월 1-3일)
4. 김장호, **이금희**, 나휘동, 이종혁, “한국어 형태소 유형에 따른 한국어-중국어 단어정렬 결과분석”, 한국정보과학회 2009 한국컴퓨터종합학술대회 논문집 제36권 제1호(C): 325-330 (2009년 7월 1-3일)
5. 나휘동, **이금희**, 이종혁, “일영 통계기계번역에서 먼 거리 의존관계를 이용한 일본어 어순 조정,” 2008 한국컴퓨터종합학술대회 논문집 35(1-A): 76-77 (2008년 6월 30 - 7월 2일)
6. 백설매, 김미훈, **이금희**, 정유진, 이종혁, “문장부호 정보와 확장된 청크에 기반한 중국어 최장명사구 식별 ,” 제17회 한글 및 한국어정보처리 학술대회: 언어처리의 현황과 전망 17(1): 112-119 (2005년 10월 21-22일)
7. 백설매, **이금희**, 김미훈, 정유진, 이종혁, “문장부호를 사용한 효과적인 중국어 최장명사구 식별기법,” 한국컴퓨터종합학술대회 2005 논문집 32(1-B): 454-456 (2005년 7월 6-8일)