Verb Pattern Based Korean-Chinese Machine Translation System

Kim, Changhyun
NLP Team
Human Information Processing Dept.
ETRI,Korea
chkim@etri.re.kr

Kim, Young Kil
NLP Team
Human Information Processing Dept.
ETRI,Korea
kimyk@etri.re.kr

Hong, Munpyo
NLP Team
Human Information Processing Dept.
ETRI,Korea
Hmp63108@etri.re.kr

Seo, Young Ae
NLP Team
Human Information Processing Dept.
ETRI,Korea
yaseo@etri.re.kr

Yang,Sung Il
NLP Team
Human Information Processing Dept.
ETRI,Korea
siyang@etri.re.kr

Sung-Kwon Choi
NLP Team
Human Information Processing Dept.
ETRI,Korea
choisk@etri.re.kr

Abstract

This paper describes our ongoing Korean-Chinese machine translation system, which is based on verb patterns. A verb pattern consists of a source language pattern part for analysis and a target language pattern part for generation. Knowledge description on lexical level makes it easy to achieve accurate analyses and natural, correct generation. These features are very important and effective in machine translation between languages with quite different linguistic structures including Korean and Chinese. We performed a preliminary evaluation of our current system and reported the result in the paper.

1 Introduction

Machine translation requires correct analysis of source languages, appropriate generation into target languages and a large amount of knowledge such as rules, statistics or patterns. Especially persistent and consistent knowledge acquisition and management, monotonic improvement of performance according to knowledge accumulation are the keys to machine translation.

Rule based methods suffer from knowledge acquisition and consistent management. Statistical methods show no connections between the previous statistical knowledge and the new statistical knowledge and have difficulty in reflecting linguistic phenomena and peculiarities directly into knowledge. Patterns have several formats such as sentence-based patterns(Kaji Hiroyuki(1992)), phrase-based patterns(Watanabe Hideo(1993)) and collocation-based patterns(Smadja(1996), Kevin McTait(1999)). Sentence-based patterns uses a whole sentence as a pattern and transfer the input sentence in one time. It suffers mainly from data sparseness. Phrase-based patterns can be used for both syntactic analysis and transfer. The transfer is done phrase by phrase. Collocation-based patterns are used for lexical transfer, that is, the transfer unit is a word.

ETRI performed a verb-pattern based Korean-English machine translation from 1999 to 2001 and experienced strong points on the side of knowledge acquisition, consistent management of linguistic peculiarities between two languages and monotonic increase in system performance according to the number of patterns(Kim, Y.K. et al.(2001); Seo,Y.A. et al(2000)). A verb pattern consists of a source language pattern for analysis and a target language pattern for generation. Knowledge description on lexical level makes it easy to achieve accurate analyses and natural, correct generations. So, accurate

analysis directly leads to correct generation, which is very effective in machine translation between languages with quite different linguistic structures. With respect to the reusability of knowledge, verb patterns for Korean-English machine translation can be reused after just modifying the English pattern part into Chinese, thus saving the cost of knowledge construction.
In section 2 we show the system overview and a simulation example translating a Korean sentence into Chinese. Verb patterns are explained in more detail in section 3 and a parser using verb patterns are described in section 4. A generation module is explained in section 5 and an evaluation is made in section 6. In section 7 we conclude our system with some remarks.

2 System Overview

Our verb-pattern based Korean-Chinese machine translation system consists of a Korean morphological analyzer, a verb-pattern based parser and a generation module consisting of a verb phrase linker and a word generator. Figure 1 is the system overview.
Figure 2 is a simple simulation example translating a Korean input sentence into a Chinese sentence. The morphological analyzer first readjusts words into appropriate morphological units, performs morphological analysis and finally ranks the results using statistical information. The parser first readjusts the results of the morphological analyzer into syntactic units and performs predicate-argument-adjunct analysis and predicate-predicate structure analysis. The generation module determines the Chinese translation for connectives and arranges the order of each connective clause. And it finally generates Chinese words.

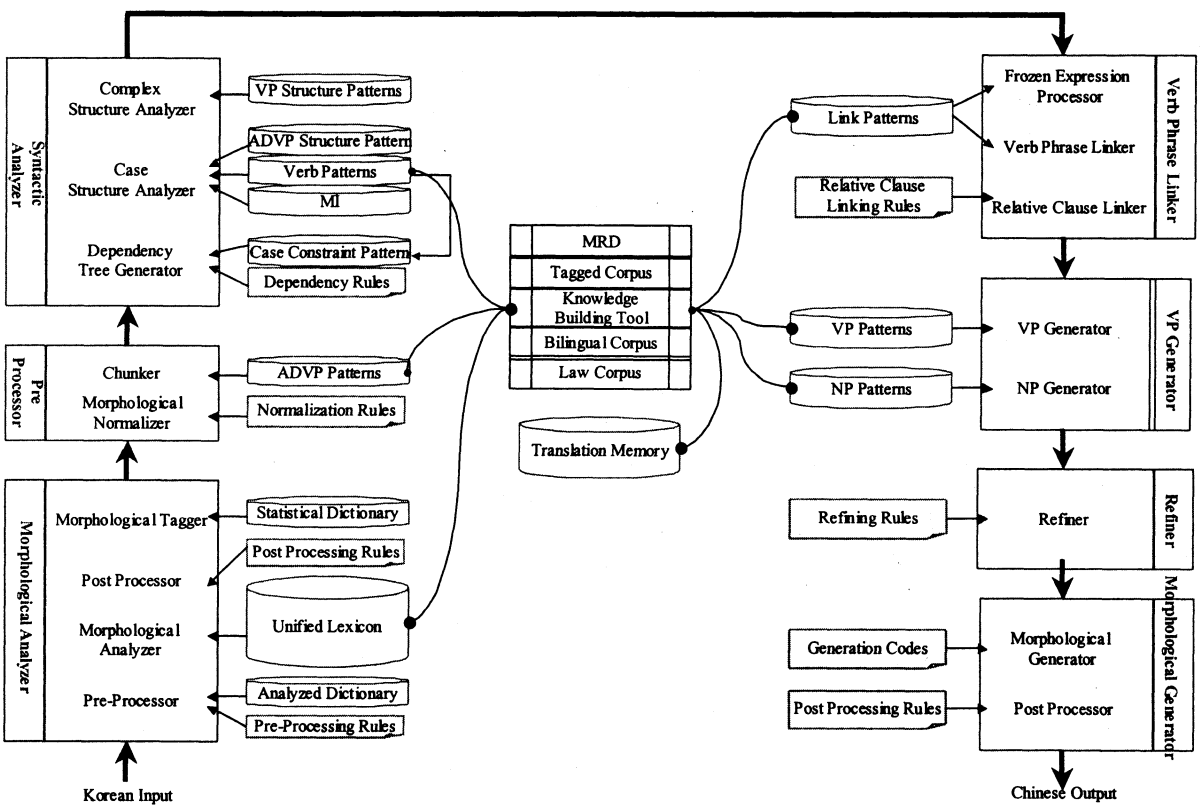

Figure 1 : Korean-Chinese Machine Translation System

3 Verb Pattern

Phrase-based patterns can be used for both syntactic analysis and transfer(Watanabe Hideo(1993)). The term 'verb pattern' we are using is to be understood as a kind of subcategorization frame of a predicate.

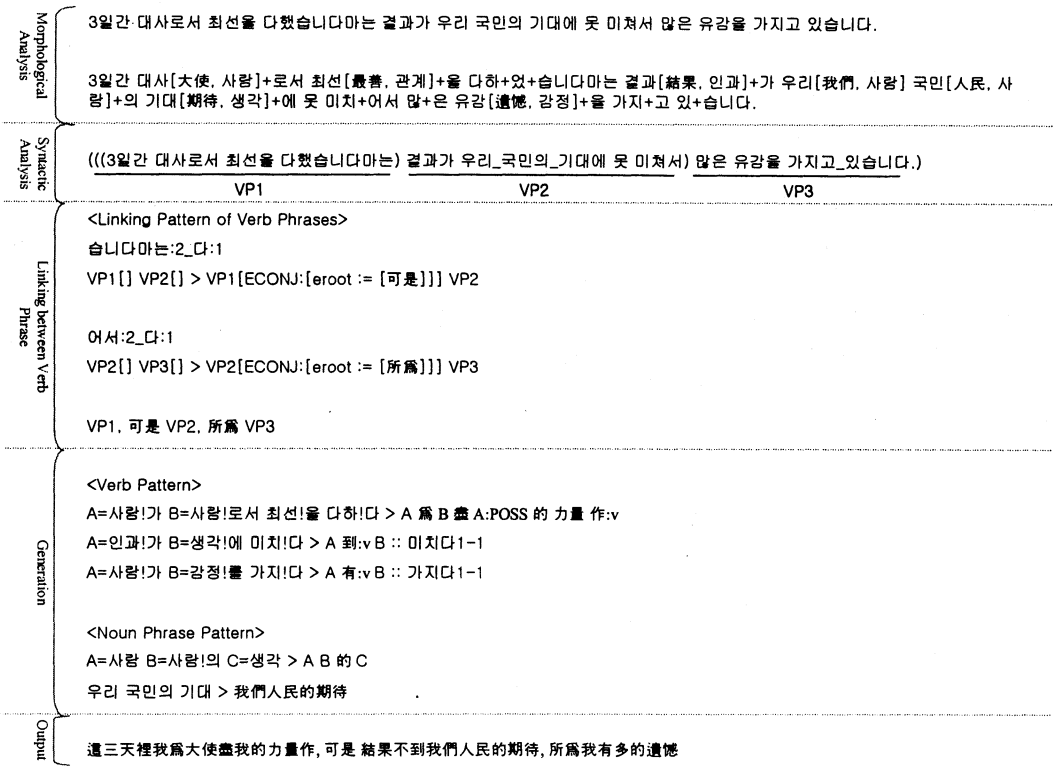

Figure 2 : A Simulation for Korean-Chinese Machine Translation

The main difference between a verb pattern and a subcategorization frame lies in the fact that a verb pattern is always linked to the target language word (predicate of the target language). A verb pattern is employed not only in the analysis but also in the transfer phase. That's why a source language verb pattern is linked to a target language verb pattern. Because we consider target words during analysis, a verb pattern in our approach is slightly different from a subcategorization frame in the traditional sense. In the theoretical linguistics a subcategorization frame always contains arguments of the predicate. An adjunct of a predicate or a modifier of an argument is usually not included in the subcategorization frame. However, we think that for the purpose of machine translation, these words must be taken into account. In reality adjuncts of a verb or modifiers of an argument can seriously affect the selection of target words, as can be seen in the following example.

Korean : 이 달도 다 갔다
English : This month is up

Verb Pattern: A=시간!이 다:b 가!다:PAST > A be:v up :: 가다 1-2

In this example, the Korean verb '가다 (to pass by)' is more appropriate and natural to be translated into 'be up', if modified by the adverb '다 (completely)'. This kind of conflational divergence can be handled in pattern-based approaches and our verb patterns annotate the adverb with a marker (b) and link the adverb and the verb to a conflated English expression. Idiomatic usages of a verb can also be treated easily within verb patterns. A frozen argument in an idiomatic expression is just to be equated with a variable. For example, a Korean idiomatic expression "호감이 가다 (= be favorably disposed toward)" can be described as the following:

A=사람!가 B=사람!에게 호감!가 가!다 > A be:v favorably disposed toward B:OBJ :: 가다 1-3

The noun '호감 (=a favorable impression)' is not overtly expressed in the target language expression. If an expression in the source language side is not marked, it will not be considered any more in the further phase of the translation. Postpositions such as '에게', '가', are normalized Korean postpositions which correspond to syntactic case markers.

Let's see verb patterns for Korean-Chinese machine translation. A Korean verb pattern is linked to its corresponding Chinese verb pattern by the symbol '>'. The arguments in the left-hand side of a verb pattern are basically represented with semantic features such as '시간(time)', '공간(location)', '교통(transportation)', etc. We are currently using about 200 hierarchical semantic features to cover the semantic information of nouns, the arguments of a verb. The right-hand side of '>' is the corresponding target word expression. In the current stage of the development we have about 40,000 Korean-Chinese verb patterns and we expect about 60,000 at the end of the year. We present some examples of Korean-Chinese verb patterns in the below.

Adjectival Verbs

뚱뚱하다 1 : A=사람!가 뚱뚱하!다 > A 胖:v :: 뚱뚱하다 1-1

Copular

이다 1 : A 가 B 이 이!다 > A 是:v B

이다 2 : A=사람!가 약:b B=단위!이 이!다 > A 有:v B :: 이다 1-1

Other Intransitive Verbs

날다 7 : A=동물!가 B=환경!에서 날!다 > A 在 B 飛:v

울다 6 : A=사람!가 울!다 > A 哭:v :: 울다 1-1

Transitive Verbs

보다 74 : A=사람!가 B=생산품!를 보!다 > A 看:v B :: 보다 1-1

이상화하다 1 : A=사람!가 B=시간!를 이상화하!다 > A 把 B 理想化:v

알다 1 : A=사람!가 알!다 > A 知道:v :: 알다 1-1

도와주다 3 : A=사람!가 B=사람!를 도와주!다 > A 幇:v B 的忙

졸업하다 1 : A=사람!가 B=조직!를 졸업하!다 > A B 畢業:v :: 졸업하다 1-1

주다 1 : A=사람!가 B=사람!에게 C=사물!를 주!다 > A 把 C 給 B:: 주다 1-1

Verbs taking NPs with Adverbial Postpositions

살다 1 : A=사람!가 B=장소!에서 살!다 > A 住 在 B:: 살다 1-1

Each entry denotes different meaning or different translation. For example, in copular, 이다 1 and 이다 2 have different translation results. In the above verb pattern examples we can see that many divergence problems in translation can be solved easily by verb patterns. For example, a structural divergence problem between Korean and Chinese can be properly solved using the verb pattern of '도와주다 (help)' as in the below.

나는 그 여자를 도와주었다 ⇔ 我幇皮女的忙

The direct object of '도와주다'(help), '그 여자 (the woman)' is combined with a post-nominal particle '的' and used as a kind of modifier of the grammatical object '忙'. This kind of divergence is directly addressed in the verb pattern by combining the direct object with '的' as in the below.

도와주다 3 : A=사람!가 B=사람!를 도와주!다 > A 幇:v B 的忙

4 Parser

The parser consists of pre-processing the results of morphological analysis into syntactic units, grasping the predicate-argument-adjunct structure for each predicate using verb patterns, finding the head of each unresolved adverb phrase and finally linking predicates using predicate-predicate structure patterns. In this section, we are going to explain about analyzing the predicate-argument-adjunct structures and predicate-predicate structures. Verb patterns represent predicate-argument-adjunct structures and also provide the information for resolving the syntactic cases of auxiliary particles and particle ellipses. Currently predicate-predicate structure patterns use statistics of verb endings to represent structural preferences between predicates in multi-predicate sentences.

4.1 Predicate-Argument-Adjunct Structure Analysis

A dependency structure is used in predicate-argument-adjunct analysis. As described previously, verb patterns describe not only arguments but also adjuncts. In predicate-argument-adjunct analysis, verb patterns are used in two steps, that is, in binary pattern matching and in full pattern matching. In binary pattern matching, the information of the form <a noun meaning, a postposition, a verb meaning> are extracted from verb patterns to filter out the unnecessary dependency relations. In full pattern matching, each verb on a dependency tree is compared with verb patterns and evaluated according to the matched proportion. In case of a pronominal clause, the modifee of the clause can fill in an argument of the predicate. So the pronominal clause including the modifee is compared together with verb patterns. The higher the proportion is, the higher the score of the evaluation is.
Auxiliary predicates and suffixes can change the argument of a verb, which demands another verb patterns different from the original and causes difficulties in manual construction and management of verb patterns. So, we use rules to deal with auxiliary predicates and suffixes instead of constructing another verb patterns. This is possible because there exist regularities, for example, in transforming syntactic cases to and from active/passive, active/causative forms as the following.

passive → active	causative → active
① for transitive verbs ② subject → object ③ adverb → subject ④ subject, adverb(에 서) → adverb(로), object(를)	① for verb/adjective ② subject → adverb ③ object,adverb(에 게) → object,subject ④ adverb(에 게) → subject ⑤ object → subject

Auxiliary particles and particle ellipsis can be interpreted into several syntactic cases and cause difficulties in syntactic case resolution. The current parser restricts the interpretation only to objective and subjective cases and applies the same method to both phenomena. The modifee of a pronominal clause also falls in the case of particle ellipsis. The parser applies the subjective particle '가' and the objective particle '를' in turn to auxiliary particles and particle ellipsis and compares them with verb patterns. If the meanings between nouns are the same and the case is still empty then the case resolution is succeeded.

4.2 Predicate-Predicate Structure Analysis

Predicate-predicate structure analysis determines the structure between predicate phrases. Below shows an example.

Korean Sentence : 그-는 자신-이 범인-이 아니-라고 밝히-고 잠적했다.
He disappeared after declaring that he is not the criminal

After predicate-argument-adjunct analysis :

Cand1. (그는 ((자신이 범인이 아니라고) 밝히고) 잠적했다)
Cand2. (그는 (자신이 범인이 아니라고) (밝히고) 잠적했다)

Figure 3 shows the candidate structure between predicate phrases.

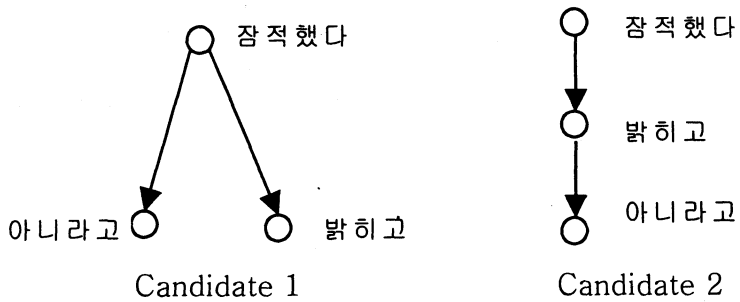

Figure 3 : Predicate Structure Candidates

To determine the structure of predicates, the relation between them needs to be explained, thus inevitably requiring semantic information of any kind to quite a large amount, which is very difficult and time consuming work. So, at present we are considering only statistical information between connective endings of predicates. In Korean, postpositions and predicate endings are well developed and are possible to represent the meanings of sentences(Nam,K.S. et al(19930). Connective predicate endings connect two predicates with specific relations such as cause, reason, expectation, or condition. So, we also assume that connective verb endings can play the role of weak semantic information. Basically a predicate-predicate structure pattern describes the dependency structure of predicates by their lexical connective endings with frequency information now. But the semantic relations denoted by connective verb endings are too simple and we are going to consider the semantic classification of predicates and the relations between predicates based on the classification in the future. Any number of predicates in a predicate-predicate pattern is possible from 2 to n, but we are currently considering 2 and 3 predicate patterns from the practical point of view. For each dependency tree, preference value of each predicate-predicate structure candidate is computed as follows.

1. Extract all possible predicate-predicate structure candidates consisting of 2 and 3 predicates from a dependency tree
2. Check whether extracted predicate-predicate structure candidates match any of predicate-predicate structure patterns and if matched, compute preference values based on frequency information.
3. Find the highest value pattern set which covers the whole dependency tree

5 Generator

5.1 Verb phrase Linker

Due to the resemblance, between Korean Chinese, of the order of predicates in complex sentences, Chinese sentences can be generated in many cases by just combing the translated Chinese verb phrases through appropriate conjunction words and that's the work of the verb phrase linker. The verb phrase linker uses verb phrase link patterns to link verb phrases. Below is the basic structure of a verb phrase link pattern :

VP1[어서] VP2[다] > VP1[ECONJ:[eroot := [所爲]]] VP2

"VP1[어서] VP2[다]" means two verb phrases are linked with a connective ending '어서'. 'ECONJ' is the Chinese conjunction corresponding to '어서'. Here we can see that the order of VP1 VP2 is the same in both Korean and Chinese. The verb phrase linker traverses a dependency tree to detect dependency relations between verb phrases and produces generation information using verb phrase link patterns.

5.2 Verb phrase generator

The verb phrase generator translates Korean verb phrases into Chinese. It consists of verb phrase generation, adverb phrase generation and noun phrase generation. The words covered by verb patterns in parsing have their Chinese translations already, but the uncovered ones don't. In verb phrase generation, the uncovered ones are dealt with using co-occurrence patterns. The co-occurrence patterns have a quadruple format, "(noun meaning or lexical word, functional word, verb, frequency of co-occurrences)". For example, "(장소*[place]*, 에*[to]*, 가*[go]*, 12)" shows that an article, "에" and the noun meaning "장소" are appeared with a verb, "가" 12 times in verb patterns. The uncovered nouns determine their meanings as the one having the most frequent co-occurrence entries. The uncovered adverb phrases use adverb translation patterns. Noun phrase generation uses rules and patterns both. Noun phrase generation in Chinese is much simpler than in English. For example, a particle '의(of), used for adnominal noun phrase, is translated into "的" only, in comparison with various Korean-English translations of it.

6 Evaluation

To decide whether the verb-pattern based approach is suitable for Korean-Chinese translation, the evaluation of our Korean-Chinese MT-system was conducted in the first stage of the development. The test suite is composed of 100 sentences extracted from primary school textbooks. The average length of the test sentence is approximately 10.5 words. A Chinese mother-tongue has given scores to the sentences according to the following criteria:

Score	Criteria
4	the meaning of the sentence is preserved
3	the meaning of the sentence is partially preserved (the predicate of the sentence is correctly translated, so that the skeletal meaning of the sentence is preserved, however some arguments or adjuncts of the sentence are not correctly translated)
2	at least one phrase is correctly translated
1	at least one word is correctly translated
0	no output

The translation rate calculated in this way was 44.5%. The errors can be classified roughly into unknown word problem, knowledge insufficiency and inaccurate analysis. Most unknown words are proper nouns, which are not dealt with appropriately in our current prototype system. Below is an example.

Korean Sentence : 두레가 콜록거리는 모습을 보고, 식구들이 모두 웃었습니다.
Seeing Dure cough, all the family laughed at it.
Generated Output : 看 ( ) 咳咳咔咔地咳嗽 的 模样，家人 全部 微笑。

Currently our system doesn't have any module to deal with unknown proper nouns. So, the generated output omits the translation of '두레'(represented by ( )). To handle unknown proper nouns reasonably, we need to first classify unknown proper nouns into such as people's name or location name or organization name etc. Contrary to Koran, each Chinese character has meanings and different Chinese characters are used for people, location and organization. After classifying, appropriate Chinese characters have to be generated according to the similarity of pronunciation between each Korean syllable and Chinese characters and the domain of the word. Most of the errors arose from the lack of verb patterns.

Korean Sentence : *집으로 오는* 길에 누구를 만났습니까?
Whom did you meet on the way home ?
Generated Output : 遇见 谁 *在 来 到 家*的路？
After Correction : *回家*的路

The corresponding translation of '집으로 오다' is '回家' and requires the below verb pattern :

A=사람!가 집!로 오!다 > A 回家

Without the pattern, each word is considered separately for generation and the default Chinese word 来 is selected for the translation of the verb '오다'. There occurred several errors in the arrangement of Chinese words. Usually adverb phrases come before verb in Chinese but our system doesn't do that right currently.

Korean Sentence : 산에는 먼저 온 사람들이 나무를 심고 있었습니다.
On the mountain people who had come before us was planting trees.
Generated Output : 先 来 的 人 栽种 树 *在 山*。
After Correction : *在 山上*先 来 的 人 栽种 树。

In generation we usually maintain the relative order of two verb phrases in a Korean sentence as in the verb phrase linker. But, it causes problems when verb phrases are in special relations.

Korean Sentence :나는 동생이 신을 신도록 도와 줍니다.
I helped my younger brother put on his shoes.
Generated Output : 我 弟弟 *穿* 鞋 *帮助*。
After Correction : 我*帮助*弟弟*穿*鞋。

In the noun phrase '마루 청소', '청소(cleaning)' has a meaning of action and '마루' is the theme of the action. The generation has to consider such cases and make them as predicate-noun structure in Chinese.

Korean Sentence :나는 오빠와 함께 *마루 청소*를 합니다.
I do the floor cleaning with my elder brother.
Generated Output : 我 与 哥哥 *地板扫除* 。
After Correction : 我与哥哥*扫除地板*。

As can be seen from the above error list, the major problems concerning the knowledge part are the lack of verb patterns and generation information. Generation can be improved in the next year and also as the size of verb patterns grows larger, the lack of verb patterns is expected to be overcome.

7 Conclusion

Our Korean-Chinese machine translation system basically uses pattern-based knowledge, which shows strong points especially on consistent management of linguistic peculiarities between language pairs, monotonic increase in system performance and reusability of knowledge. Now we are under development of the prototype system and still have many works to do especially in the increase of patterns, consideration of linguistic characteristic between Korean and Chinese such as word order.

References

Choi.Y.S. and Lee,J.H. and Choi,K.S. 1999. Research on Automatic Case Frame Construction and Evaluation, *Conference on Korean Information Processing.*

Kaji Hiroyuki and Yuuko Kida and Yasutsugu Morimoto, 1992, Learning Translation Templates From Bilingual Text, in Proceeding of the 15th International Conference on Computationla Linguistics, Nantes, France, pp.678-678.

Kevin McTait and Arturo Trujillo. 1999. A Language-Neutral Sparse Data Algorithm for Extracting Translation Patterns, in Proceedings of the 8th International Conference on heoretical and Methodological Issues in Machine Translation, England, pp.98-108.

Kim,T.W. 1999. Automatic Construction of Japanese-Korean Bilingual Dictionary using Alignment Techniques. Doctoral Thesis. Computer Science Department. KAIST.

Kim, Y.K, Seo, Y.A., Choi, S.K., Park, S.K. 2001. Estimation of Feasibility of Sentence Pattern-Based Method for Korean to English Translation System, *Int'l Conference on Computer Processing of Oriental Languages.*

Lee,H.B. and Kang,I.S.and Lee,J.H. 1998. Determination of Unknown Syntactic Relation in Korean using Concept patterns and Statistical Information, *Conference on Korean Information Processing*, pp.261-266

Nam, K.S. and Ko, Y.G. 1993. The Standard Theory of Korean Grammar, Top publication.

R.Jain and R.M.K.Sinha and A.Jain. 1995. A Pattern-Directed Hybrid Approach to Machine Translation Through Examples, *SNLP'95 : 2nd Symp. on Natural Lang. Processing*, Bangkok, Thailand, August 2-4, pp 324-335.

Seo, Y.A. and Kim, Y.K. and Seo, K.J.and Choi, S.K. 2000. Korean to English Machine Translation System based on Verb phrase:CaptionEye/KE, *Proceedings of the 14th KIPS Fall Conference*, 2000

Smadja, Frank and McKeown, K. and Hatzivassiloglou, V. 1996. Translating Collocations for Bilingual Lexicons:A Statistical Approach, in Computational Linguistics, Vol.21(4), pp.1-38

Watanabe Hideo. 1993. A method for extracting translation patterns from translation patterns. In Proceeding of the 5th international conference on theoretical and methodological issues in machine translation. Kyoto. Japan. pp.292-301.