Semi-Automatic Construction of Korean-Chinese Verb Patterns Based on Translation Equivalency

Munpyo Hong
Dept. of Speech/Language Technology Research, ETRI
Korea
Hmp63108@etri.re.kr

Young-Kil Kim
Dept. of Speech/Language Technology Research, ETRI
Korea
kimyk@etri.re.kr

Sang-Kyu Park
Dept. of Speech/Language Technology Research, ETRI
Korea
parksk@etri.re.kr

Young-Jik Lee
Dept. of Speech/Language Technology Research, ETRI
Korea
ylee@etri.re.kr

Abstract

This paper addresses a new method of constructing Korean-Chinese verb patterns from existing patterns. A verb pattern is a subcategorization frame of a predicate extended by translation information. Korean-Chinese verb patterns are invaluable linguistic resources that are not only used for Korean-Chinese transfer but also for Korean parsing. Usually a verb pattern has been either hand-coded by expert lexicographers or extracted automatically from bilingual corpus. In the first case, the dependence on the linguistic intuition of lexicographers may lead to the incompleteness and the inconsistency of a dictionary. In the second case, extracted patterns can be domain-dependent. In this paper, we present a method to construct Korean-Chinese verb patterns semi-automatically from existing Korean-Chinese verb patterns that are manually written by lexicographers.

1 Introduction

PBMT (Pattern-based Machine Translation) approach has been adopted by many MT researchers, mainly due to the portability, customizability and the scalability of the approach. cf. Hong et al. (2003a), Takeda (1996), Watanabe & Takeda (1998). However, major drawback of the approach is that it is often very costly and time-consuming to construct a large amount of data enough to assure the performance of the PBMT system. From this reason many studies from PBMT research circles have been focused on the data acquisition issue. Most of the data acquisition studies were about automatic acquisition of lexical resources from bilingual corpus.

Since 2001, ETRI has developed a Korean-Chinese MT system, TELLUS K-C, under the auspices of the MIC (Ministry of Information and Communication) of Korean government. We have adopted verb pattern based approach for Korean-Chinese MT. The verb patterns play the most crucial role not only in the transfer but also in the source language analysis. In the beginning phase of the development, most of the verb patterns were constructed manually by experienced Korean-Chinese lexicographers with some help of editing tools and electronic dictionaries. In the setup stage of a system, the electronic dictionary is very useful for building a verb pattern DB. It provides with a comprehensive list of entries along with some basic examples to be added to the DB. In most cases, however, the examples in the dictionary with which the lexicographers write a verb pattern are basic usages of the verb in question, and other various usages of the verb are often neglected. Bilingual corpus can be useful

resources to extract verb patterns. However, as for language pairs like Korean-Chinese for which there are not so much bilingual corpus available in electronic form, the approach does not seem to be suitable. Another serious problem with the bilingual corpus-based approach is that the patterns extracted from the corpus can be domain-dependent.

The verb pattern generation based on translation equivalency is another good alternative to data acquisition from bilingual corpus. The idea was originally introduced by Fujita & Bond (2002) for Japanese to English MT.

In this paper, we present a method to construct Korean-Chinese verb patterns from existing Korean-Chinese verb patterns that are manually written by lexicographers. The clue for the semi-automatic generation is provided by the idea that verbs of similar meanings often share the argument structure as already shown in Levin (1993). The synonymy among Korean verbs can be indirectly inferred from the fact that they have the same Chinese translation.

We have already applied the approach to TELLUS K-C and increased the number of verb patterns from about 110,000 to 350,000. Though 350,000 patterns still contain many erroneous patterns, the evaluations in section 5 will show that the accuracy of the semi-automatically generated patterns is noteworthy and the pattern matching ratio improves significantly with 350,000 pattern DB.

2 Related Works

When constructing verb pattern dictionary, too much dependence on the linguistic intuition of lexicographers can lead to the inconsistency and the incompleteness of the pattern dictionary. Similar problems are encountered when working with a paper dictionary due to the insufficient examples. Hong et al (2002) introduced the concept of causative/passive linking to Korean word dictionary. The active form 'mekta (to eat)' is linked to its causative/passive forms 'mekita (to let eat)', and 'mekhita (to be eaten)', respectively. The linking information of this sort helps lexicographers not to forget to construct verb patterns for causative/passive verbs when they write a verb pattern for active verbs. The semi-automatic generation of verb patterns using translation equivalency was tried in Hong et al (2002). However, as only the voice information was used as a filter, the over-generation problem is serious.

Fujita & Bond (2002) and Bond & Fujita (2003) introduced the new method of constructing a new valency entry from existing entries for Japanese-English MT. Their method creates valency patterns for words in the word dictionary whose English translations can be found in the valency dictionary. The created valency patterns are paraphrased using monolingual corpus. The human translators check the grammaticality of the paraphrases.

Yang et al. (2002) used passive/causative alternation relation for semi-automatic verb pattern generation. Similar works have been done for Japanese by Baldwin & Tanaka (2000) and Baldwin & Bond (2002) .

3 Verb Pattern in TELLUS K-C

The term 'verb pattern' is understood as a kind of subcategorization frame of a predicate. However, a verb pattern in our approach is slightly different from a subcategorization frame in the traditional linguistics. The main difference between the verb pattern and the subcategorization frame is that a verb pattern is always linked to the target language word (the predicate of the target language). Therefore, a verb pattern is employed not only in the analysis but also in the transfer phase so that the accurate analysis can directly lead to the natural and correct generation. In the theoretical linguistics, a subcategorization frame always contains arguments of a predicate. An adjunct of a predicate or a modifier of an argument is usually not included in it. However, in some cases, these words must be taken into account for the proper translation. In translations adjuncts of a verb or modifiers of an argument can seriously affect the selection of target words. (1) exemplifies verb patterns of "cata (to sleep)":

(1)
cata1 : A=WEATHER!ka ca!ta[1] > A 停:v
[param(A)ka cata: *The wind has died down*]

[1] The slot for nominal arguments is separated by a symbol "!" from case markers like "ka", "lul", "eykey", and etc. The verb is also separated by the symbol into the root and the ending.

cata2 : A=HUMAN!ka ca!ta > A 睡觉:v
[ai(A)ka cata: *A baby is sleeping*]
cata 3 : A=WATCH! ka ca!ta > A 停:v
[sikye(A)ka cata: *A watch has run down*]
cata 4 : A=PHENOMENA!ka ca!ta > A 平静:v
[phokpwungwu(A)ka cata: *The storm has abated*]

On the left hand of ">" Korean subcategorization frame is represented. The argument position is filled with a variable (A, B, or C) equated with a semantic feature (WEATHER, HUMAN, WATCH, PHENOMENA). Currently we employ about 410 semantic features for nominal semantic classifications. The Korean parts of verb patterns are employed for syntactic parsing.

On the right hand of ">" Chinese translation is given with a marker ":v". To every pattern is attached an example sentence for better comprehensibility of the pattern. This part serves for the transfer and the generation of Chinese sentence.

4 Pattern Construction based on Chinese Translation

In this chapter, we elaborate on the method of semi-automatic construction of Korean-Chinese verb patterns. Our method is similar to that of Fujita & Bond (2002) and inspired by it as well, i.e. it makes most use of the existing resources.

The existing resources are in this case verb patterns that have already been built manually. As every Korean verb pattern is provided with the corresponding Chinese translation, Korean verb patterns can be re-sorted to Chinese translations. The basic assumption of this approach is that the verbs with similar meanings tend to have similar case frames, as is pointed out in Levin (1993). As an indication to the similarity of meaning among Korean verbs, Chinese translation can be employed. If two verbs share Chinese translation, they are likely to have similar meanings. The patterns that have translation equivalents are seed patterns for automatic pattern generation.

Our semi-automatic verb pattern generation method consists of the following four steps:

Step1: Re-sort the existing Korean-Chinese verb patterns according to Chinese verbs

Example:

Chinese Verb 1: 给 (to give)

tulita	A=HUMAN!ka B=CAR!lul tuli!ta
cwuta	A=HUMAN!ka B=HUMAN!eykey C=VEGETABLE!lul cwu!ta
swuyehata	A=HUMAN!ka B=MONEY!lul swuyeha!ta

Chinese Verb 2: 停止 (to stop)

kumantwuta	A=HUMAN!ka B=CONSTRUCTION!lul kumantwu!ta
kwantwuta	A=ORGANIZATION!ka B=VIOLATION!lul kumantwu!ta

When the re-sorting is done, we have sets of synonymous Korean verbs which share Chinese translations, such as {tulita, cwuta, swuyehata} and {kumantwuta, kwantwuta }.

Step2: Pair verbs with the same Chinese translation

Example:

Chinese Verb 1: 给 (to give)

Pair1:

tulita	A=HUMAN!ka B=CAR!lul tuli!ta
cwuta	A=HUMAN!ka B=HUMAN!eykey C=VEGETABLE!lul cwu!ta

Pair2:

tulita	A=HUMAN!ka B=CAR!lul tuli!ta
swuyehata	A=HUMAN!ka B=MONEY!lul swuyeha!ta

Pair3:

cwuta	A=HUMAN!ka B=HUMAN!eykey C=VEGETABLE!lul cwu!ta
swuyehata	A=HUMAN!ka B=MONEY!lul

	swuyeha!ta

Step3: Exchange the verbs, if the following three conditions are met:

- The two Korean verbs of the pair have the same voice information
- Neither of the two verbs is idiomatic expressions
- The Chinese translation is not 加以, 进行, 做, 作

Example:

tulita	A=HUMAN!ka B=HUMAN!eykey C=VEGETABLE!lul tuli!ta
tulita	A=HUMAN!ka B=MONEY!lul tuli!ta
cwuta	A=HUMAN!ka B=CAR!lul cwu!ta
cwuta	A=HUMAN!ka B=MONEY!lul cwu!ta
swuyehata	A=HUMAN!ka B=CAR!lul swuyeha!ta
swuyehata	A=HUMAN!ka B=HUMAN!eykey C=VEGETABLE!lul swuyeha!ta

Step4: If the newly-generated pattern already exists in the verb pattern dictionary, it is discarded.

The three conditions to be met in the third step are the filters to prevent the over-generation of patterns. The following examples shows why the first condition, i.e., "the voice of the verbs in question must agree", must be met.

(2) 漂 (to float)

ttuta : A=PLANT!ka B=PLACE!ey ttu!ta > A 漂:v 在 B 上 [namwutip(A)i mwulwi(B)ey ttuta: *A leaf is floating on the water*]

ttiwuta : A=HUMAN!ka B=PLACE!ey C=PLANT!lul ttiwu!ta > A 使 C 漂:v 在 B 上 [ai(A)ka mwulwi(B)ey namwutip(C)ul ttiwuta: A baby floated a leaf on the water]

(3) 滥用 (to use)

sayongtoyta : A=HUMAN!eyuyhay B=MEDICINE!ka sayongtoy!ta > B 被 A 滥用:v [hankwuksalamtul(A)eyuyhay yak(B)i hambwulo sayongtoyta: *The drug is misused by Koreans*]

sayonghata : A=HUMAN!ka B=MEDICINE!lul sayongha!ta > A 滥用:v B [hankwuksalamtul (A)un yak(B)ul hambwulo sayonghanta: *Koreans are misusing the drug*]

As we re-sort the existing patterns according to the Chinese verbs which are marked with ":v", the verbs of different voice may be gathered together. However, as the above examples show, the voice (active vs. causative in (2), passive vs. active in (3)) affects the argument structure of verbs. We conclude that generating patterns without considering the voice information can lead to the over-generation of patterns. The voice information of verbs can be obtained from the linking information between the verb pattern dictionary and the word dictionary. We will not look into the details of the linking relation between the verb pattern dictionary and the word dictionary of TELLUS K-C system in this paper. cf. Hong et al. (2002)

The second condition relates to the lexical patterns of Korean. Lexical patterns are used for collocational expressions. As the nature of collocation implies, a predicate that shows a strict co-occurrence relation with a certain nominal argument cannot be arbitrarily combined with any other nouns.

The third condition deals with the support verb construction of Chinese. The four verbs, 加以, 进行, 做, 作, belong to the major verbs in Chinese that form support verb construction with predicative nouns. In support verb construction, the argument structure of the sentence is not determined by a verb but by a predicative noun. Because of this, the same Chinese translation cannot be the indication of similar meaning of Korean verbs, as followed:

(4) 作:v (to make)
ttallangkelita (to ring): A=BELL!ka
ttallangkeli!ta > A 作:v 厩当声
[pangwul(A)i ttallangkelita: *A bell is ringing*]

ssawuta1 (to fight) : A=HUMAN!ka B=PROPERTY!wa ssawu!ta > A 为 B 作:v 斗争 [kunye(A)ka mwulka(B)wa ssawunta: *She is struggling with high price*]

wuntonghata (to exercise) : A=HUMAN!ka B=PLACE!eyse wuntongha!ta > A 在 B 作:v 运动 [ku(A)ka chewyukkwan(B)eyse wuntonghanta: *He is exercising in the gymnasium*]

Although the Korean verbs "ttallangkelita (to ring)", "ssawuta (to fight)", "wuntonghata (to exercise)" share the Chinese verb "作", the argument structure of each Chinese translation is determined by the predicative nouns that are syntactically objects of the verbs.

5 Evaluation

The 114,581 verb patterns we have constructed for 3 years were used as seed patterns for semi automatic generation of patterns. After the steps 1 and 2 of the generation process were finished, the sets of possible synonymous verbs were constructed. To filter out the wrong synonym sets, the whole sets were examined by two lexicographers. It took a week for two lexicographers to complete this process. The wrong synonym sets were produced mainly due to the homonymy of Chinese verbs.

From the original 114,581 patterns, we generated 235,975 patterns. We performed two evaluations with the generated patterns. In the first evaluation, we were interested in finding out how many correct patterns were generated. The second evaluation dealt with the improvement of the pattern matching ratio due to the increased number of patterns.

Evaluation 1

In the first evaluation we randomly selected 3,086 patterns that were generated from 30 Chinese verbs. The expert Korean-Chinese lexicographers examined the generated patterns. Among the 3,086 patterns, 2,180 were correct. The accuracy of the semi-automatic generation was 70.65%. Although the evaluation set was relatively small in size, the accuracy rate seemed to be quite promising, considering there still remain other filtering factors that can be taken into account additionally.

Chinese Verbs	30
Unique generated patterns	3,086
Correct patterns	2,180
Erroneous patterns	906
Accuracy	70.65%

Table 1: Accuracy Evaluation

The majority of the erroneous patterns can be classified into the following two error types:

- The verbs share similar meanings and selectional restrictions on the arguments. However, they differ in selecting the case markers for argument positions (the most prominent error).

  Ex) ~**eykey** masseta/ ~**wa** taykyelhata (to face somebody)

- The verbs share similar meanings, but the selectional restrictions are different.

  Ex) **PAPER**!lul kyopwuhata (to deliver) / **MONEY**!lul nappwuhata (to pay)

Evaluation 2

In the second evaluation, our interest was to find out how much improvement of pattern matching ratio can be achieved with the increased number of patterns in comparison to the original pattern DB. For the evaluation, 300 sentences were randomly extracted from various Korean newspapers. The test sentences were about politics, economics, science and sports. In the 300 sentences there were 663 predicates.

With the original verb pattern DB, i.e. with 114,581 patterns, the perfect pattern matching ratio was 59.21%, whereas the perfect matching ratio rose to 64.40% with the generated pattern DB.

	114,581 Verb patterns	**350,556 Verb patterns**

Num. Of Sentences	300	
Num. of. Predicates	663	
Perfect Matching	392	427
No Matching	73	66
Perfect Matching Ratio	59.21 %	64.40 %

Table 2: Pattern Matching Ratio Evaluation

6 Conclusion

Korean-Chinese verb patterns are invaluable linguistic resources that cannot only be used for Korean-Chinese transfer but also for Korean analysis. In the set-up stage of the development, a paper dictionary can be used for exhaustive listing of entry words and the basic usages of the words. However, as the verb patterns made from the examples of a dictionary are often in-sufficient, a PBMT system suffers from the coverage problem of the verb pattern dictionary. Considering there are not so many Korean-Chinese bilingual corpus available in electronic form till now, we believe the translation-based approach, i.e. Chinese-based pattern generation approach provides us with a good alternative.

The focus of our future research will be given on the pre-filtering options to prevent over-generation more effectively. Another issue will be about post-filtering technique using monolingual corpus with minimized human intervention.

References

T. Baldwin and F. Bond. 2002. Alternation-based Lexicon Reconstruction, *TMI 2002*

T. Baldwin and H. Tanaka. 2000. Verb Alternations and Japanese – How, What and Where? *PACLIC2000*

F. Bond and S. Fujita. 2003. Evaluation of a Method of Creating New Valency Entries, *MT-Summit 2002*

S. Fujita and F. Bond. 2002. A Method of Adding New Entries to a Valency Dictionary by Exploiting Existing Lexical Resources, *TMI2002*

M. Hong, Y. Kim, C. Ryu, S. Choi and S. Park. 2002. Extension and Management of Verb Phrase Patterns based on Lexicon Reconstruction and Target Word Information, *The 14th Hangul and Korean Language Processing* (in Korean)

M. Hong, K. Lee, Y. Roh, S. Choi and S. Park. 2003. Sentence-Pattern based MT revisited, *ICCPOL 2003*

B. Levin. 1993. English verb classes and alternation , The University of Chicago Press

K. Takeda. 1996. Pattern-based Machine Translation, *COLING 1996*

H. Watanabe and K. Takeda. 1998. A Pattern-based Machine Translation System Extended by Example-based Processing, *ACL 1998*

S. Yang, M. Hong, Y. Kim, C. Kim, Y. Seo and S. Choi. 2002. An Application of Verb-Phrase Patterns to Causative/Passive Clause, *IASTED 2002*