Phrase-Pattern-based Korean to English Machine Translation using Two Level Translation Pattern Selection

Kim, Jung-jae and Choi, Key-Sun
Computer Science Division, KAIST
jjkim@nlp.kaist.ac.kr, kschoi@cs.kaist.ac.kr

Chae, Young-Soog
KORTERM, KAIST
yschae@korterm.kaist.ac.kr

Abstract

Pattern-Based Machine Translation is one of the machine translation methods which performs syntactic analysis and structure transfer at the same time using bilingual patterns. PBMT is used to expand the length of patterns up to sentence-length in order to reduce ambiguities in translation, but it brought out the problem of rapidly increased patterns. We propose a model which shortens the length of patterns to phrase-length and reduces ambiguities in translation by using two-level translation pattern selection method. In the first level, the proper translation patterns are selected by using a hybrid method of exact example matching and semantic constraint by thesaurus. In the second level, the most natural translation pattern for the verb phrase is selected among the selected translation pattern categories by using statistic information of the target language. By using this proposed model, we could shorten the length of patterns without raising the ambiguities in translation.

1 Introduction

A Transfer-Based Machine Translation method generally has four steps (Kim, 1994); morphological analysis of source language, syntactic analysis of source language, structure transfer to the target language, and sentence generation of the target language. In structure transfer step, transfer patterns used to be greatly lexicalized in order to raise the accuracy of translation. Pattern-Based Machine Translation (PBMT) performs both syntactic analysis and structure transfer simultaneously using this lexicalized patterns (Takeda, 1996). PBMT can shorten the translation time and raise the accuracy of syntactic analysis by using the lexicalized patterns.

At first, since all patterns were short phrase-length patterns, many syntactic ambiguities occurred in pattern matching. As a result, patterns became longer to sentence-length to reduce ambiguities. But, sentence-length patterns cause pattern sparseness problem, because the same number of sentence-length patterns can cover less sentences than the same number of phrase-length patterns.

To overcome this problem, Watanabe and Takeda (1998) adopted example-based approach. However, example-based approach has some problems. One of the problems is that two different verbs of target language take two semantically similar nouns as objects respectively although a verb of source language takes the nouns as its objects. For example, a Korean verb '*ta-da*' with objects '*bus*(bus)' and '*mal*(horse)' is translated into English verbs 'take' and 'ride' respectively.

Many researches were done to solve them. They used syntactic collocation (Kim *et al.*, 1996; Lee *et al.*, 1999), semantic constraint by thesaurus (Moon *et al.*, 1998), semantic features (Palmer *et al.*, 1999) and statistical information (Brown *et al.*, 1991; Dagan and Itai, 1994). But when syntactic collocation is used, each example for a verb has the same effect to select the proper word sense of the verb. Consequently, it is difficult to obtain the representative examples and to describe senses which have domains of different size. In addition, when only semantic constraint by thesaurus is used, it is difficult to obtain good translated words because of the insufficient thesaurus problem. As a result, we use a hybrid method of both exact example matching technique with syntactic collocation and semantic constraint by thesaurus.

We use phrase-length patterns to solve pattern sparseness problem and propose two level selection method of translation pattern to reduce the ambiguities of pattern matching.

2 Two level Translation Pattern Selection Method

We use only mono-lingual resources to reduce translation ambiguities. It is almost impossible to incorporate semantic knowledge of two languages. The mutual information between two different languages can be described not by one-to-one mapping but by coarse-mapping (Palmer *et al.*, 1999). For example, a Korean verb phrase pattern 'NP+*lul ssu-da*' can be translated into 'wear NP', 'write NP', 'compose NP', 'use NP', or 'spend NP'. If NP has a meaning [head-gear], then the Korean verb phrase pattern can be translated into 'wear NP' or 'put on NP'. If the headword of NP is '*don*(money)', then the verb phrase patterns are changed into 'spend money'. In this way, 'NP+*lul ssu-da*' has five translation pattern categories. It takes two steps to select the most natural English translation pattern as a corresponding pattern of a Korean pattern:

1) to select possible translation pattern categories,

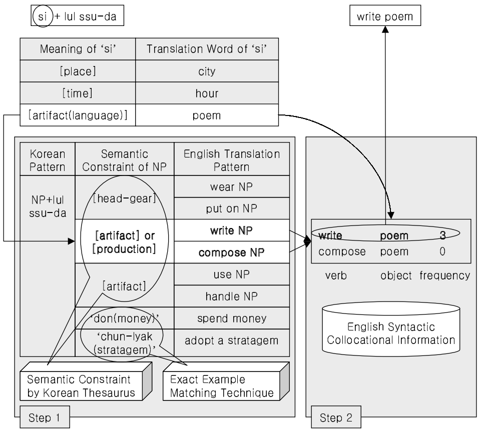
2) to select the most natural English translation pattern among possible translation pattern categories.

The first step is performed in pattern matching, and the second step in pattern transfer.

Among the ambiguities in Korean to English Machine Translation, the ambiguities in Korean pattern matching are reduced by using a hybrid method of exact example matching and semantic constraint by thesaurus, and the ambiguities in Korean to English pattern transfer are reduced by using syntactic collocational information of English (Lee *et al.*, 1999).

2.1 Pattern Matching

There are several English translation patterns of a Korean verb phrase pattern (e.g. 'NP+*lul ssu-da*' can be translated into 'wear NP', 'write NP', 'use NP' and so on.). We focus on selecting the most natural translation pattern among them. In the first step, we divide them into several translation pattern categories with both examples and semantic constraint by Korean thesaurus.


si + lul ssu-da
write poem
Semantic Constraint by Korean Thesaurus
Exact Example Matching Technique
English Syntactic Collocational Information
Step 1
Step 2

Meaning of 'si'	Translation Word of 'si'
[place]	city
[time]	hour
[artifact(language)]	poem

Korean Pattern	Semantic Constraint of NP	English Translation Pattern
NP+lul ssu-da	[head-gear]	wear NP
		put on NP
	[artifact] or [production]	write NP
		compose NP
	[artifact]	use NP
		handle NP
	'don(money)'	spend money
	'chun-lyak (stratagem)'	adopt a stratagem

verb	object	frequency
write	poem	3
compose	poem	0

Figure 1: A example of applying two level translation pattern selection method

Figure 1 shows five translation pattern categories of 'NP+*lul ssu-da*' by the proposed hybrid method. First three categories are semantically constrained by Korean thesaurus. In this case, if the meaning of headword of a noun phrase NP1 is a hyponym or a hypernym of the semantic constraint for NP, then NP1 can be matched to NP. For example, '*si*' can be an object of '*ssu-da*' because the meaning [production(language)] of the meanings of '*si*', is a hyponym of [production], (i.e. '*si-lul ssu-da*' is matched to 'NP[production]+*lul* *ssu-da*' with the possible translation patterns {'write NP', 'compose NP'}). At the same time, '*si*' is translated into 'poem' among all English words for the Korean word '*si*', because only 'poem' has the meaning of [production(language)].

Last two translation pattern categories of 'NP+*lul ssu-da*' are constrained by exact example matching method. For example, if and only if the headword of a noun phrase NP1 is '*chun-lyak*(strategy)', 'NP1+*lul ssu-da*' takes {'adopt a stratagem'} as its translation pattern category. Because the exact example matching method is too rigid, it's better to adopt example-based approach (Watanabe and Takeda, 1998). But, in this paper, we only implemented exact example matching method.

The proposed hybrid method of exact example matching and semantic constraint by thesaurus can reduce the number of possible translation patterns and also the ambiguities of pattern matching.

Ex1) *na-nun si-lul chom muk-ko nan da-um-e ssu-go sip-da.* (I want to write a poem after eating something.)

For example, both Korean verbs '*muk-da*(eat)' and '*ssu-da*(write)' can take '*si-lul*' as objects in example Ex1. '*muk-da*' takes only nouns with meaning [something to eat] as objects, but '*ssu-da*' can take nouns with meaning [production] as an object. As a result, '*si-lul*' is regared as an object of '*ssu-da*', because '*si*' does not have the meaning [something to eat] but has the meaning [production(language)], a hyponymy of [production].

2.2 Pattern Transfer

After the selection of possible translation pattern categories in pattern matching, the most natural translation pattern among those patterns is selected in pattern transfer. To acquire the most natural English sentence, we use English syntactic collocational information, especially for subject-verb relation and verb-object relation. We regard the English pattern of the most frequent syntactically related pair as the most natural translation pattern.

As explained above, 'write NP', 'compose NP' are selected as the translation patterns for 'NP+*lul ssuda*', and 'poem' is selected as the translated word for '*si*' in pattern matching. Therefore, 'write poem' and 'compose poem' are the possible translation for '*si-lul ssu-da*'. English verb 'write, compose' and English noun 'poem' appear as verb-object relation in the corpus four times and zero times respectively. Therefore, 'write poem' is selected as the most natural translation for '*si-lul ssu-da*'.

Pattern transfer is needed especially when a verb of source language can translated into several words of target language according to objects, although the verb takes semantically similar nouns as objects. For example, NP in 'NP+*lul ta-go ka-da*' must have the meaning [something to ride], and its translation patterns can be 'ride NP' or 'go by NP'. There are many Korean nouns with the meaning [something to ride] like 'horse', 'car', 'bus', 'train' and so on. When the headword of NP is 'horse' or 'car', 'NP+*lul ta-go ka-da*' is usually translated into 'ride NP', but when the headword of NP is 'bus' or 'train', 'NP+*lul ta-go ka-da*' is usually translated as 'go by NP'. But the Korean thesaurus, which we used, does not bring out the differences. This problem shows that Korean and English have very different semantic hierarchies of thesaurus construction (Palmer *et al.*, 1999).

2.3 Pattern Scoring

We used Generalized LR parsing (GLR) algorithm (Tomita, 1991) for pattern matching. To prune out nodes made by GLR algorithm during pattern matching, we score each node and remove nodes which have lower score than those of the top ten nodes with the same range and the same syntactic category. To score each node, several methods can be used; the frequency of each pattern in the corpus (Sorniertlamvanich, 1998) and the Korean syntactic collocational information (Yoon, 1998). Since we use very lexicalized patterns, the frequency of each pattern in the corpus is not available. Also we didn't describe syntactic information of the

Korean pattern, the Korean syntactic collocational information is not useful yet. Therefore, we used following preferences for scoring patterns.

- to prefer more lexicalized patterns
- when semantically constrained by thesaurus, to prefer patterns whose meaning of the argument is closer to the constraints for the argument
- when constrained by exact example matching, to allow patterns whose headword of the argument is included in the examples
- to give penalty to arguments which have no constraint

A pattern scoring method according to these preferences is shown at Figure 3. In the case of Figure 2, *score*(VPS) is the sum of $(P(c)+\alpha\times score(c))$ of all child node c of VPS. If c is lexical, then $score(c)=\eta$, otherwise, $score(c)$ is calculated recursively. For example, *score*('*ga*')=*score*('*lul*')=*score*('*ssuda*')=η, and *score*(NP1) and *score*(NP2) are calculated recursively. $P(c)$ is determined by the constraints for c. If c is lexical, $P(c)=\beta$. If β is higher, lexicalized patterns are more preferred. If c is semantically constrained by thesaurus with SC, $P(c)$ is determined by the distance between SC and the real semantic code of c in the thesaurus. If they are closer, $P(c)$ is higher, i.e. second scoring prefence is applied. Function *distance*(p, SC) means the distance between the semantic code of p and SC. For example, if p has semantic code [111120(human role)] and SC is [111000(human)], *distance*(111120,111000)=2. If c is constrained by exact example matching with EC and the headword of c is included in EC, high score is given to $P(c)$. If c has no constraint, then $P(c)$ is given as a penalty. Let α=1.0, β=2.0, γ=3.0, δ=3.0, $\theta 1$=3.0, ρ=1.0, η=0.0, ϕ=0.8 and NTTLENG=6[1].

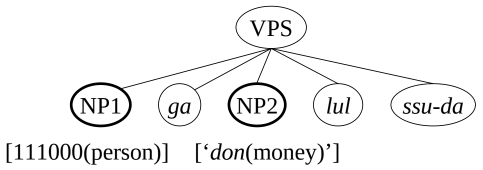


Figure 2: A example of pattern

For example, if *score*(NP1)=4.0 and *score*(NP2)=7.0 and NP1 is semantically constrained by [111000(human)] and NP2 is constrained by examples {'*don*(money)'}, and the real semantic code of NP1 is [111120(human role)] and the headword of NP2 is '*don*', then *score*(VPS) is $2.0\times 3+3.0\times(0.8+0.2\times\frac{6-2}{6})+3.0+4.0+7.0 = 22.8$.

$$score(p) = \sum_{\forall c} (P(c) + \alpha \times score(c))$$

- c : p's child node
- N : the number of p's child nodes
- if c is lexical, $score(c) = \eta$

$$P(c) = \begin{cases} \text{1. if } c \text{ is lexical, } \beta \\ \text{2. if } c \text{ is semantically constrained by thesaurus with SC,} \\ \quad \gamma \times sem(c, \mathrm{SC}) \\ \text{3. if } c \text{ is constrained by exact example method with ES,} \\ \quad \text{1) if } headword(c) \in \mathrm{ES},\ \delta \\ \quad \text{2) if } headword(c) \notin \mathrm{ES},\ -\theta 1 \\ \text{4. if } c \text{ has no constraint, } -\rho \end{cases}$$

$$sem(p, \mathrm{SC}) = \phi + (1-\phi) \times \frac{\mathrm{NTTLENG} - distance(p, \mathrm{SC})}{\mathrm{NTTLENG}}$$

Figure 3: A pattern scoring method

After all the nodes are scored recursively with the pattern scoring method, the root node of the best score is regarded as the root of the best Korean pattern tree. After the best Korean pattern tree is selected, this tree is transferred to English pattern tree in pattern transfer and the English pattern tree is transferred to the English phrase structure automatically. Then, according to the English phrase structure, the final English sentence is generated.

3 Experiment

3.1 Experimental Environment

We translated 100 sentences in letters of trade field. The thesaurus has six levels and includes about 1,800 words[2]. And we manu-

[1] NTTLENG is the depth of thesaurus.

[2] We made the Korean thesaurus by translating NTT thesaurus for nouns and verbs. The full depth of

ally made 486 translation patterns for test sentences. These also include much general patterns such as N+N, Subj+Verb, and Object+Verb. The English syntactic collocational information (about 54,000 different subject-verb pairs and about 75,000 different verb-object pairs included) was obtained from Penn Treebank[3].

3.2 Pattern Length Comparison Experiment

Experiment I (ExI) means the system of Seo *et al.* (1998) and Experiment II (ExII) means our system.

	ExI	ExII
Average length of sentence	21.7	22.3
The number of patterns	364	415
Average length of pattern	6	4
The number of patterns × Average length of pattern	2184	1660
The number of errors	125	115

Table 1: The result of pattern length comparison experiment

Table 1[4] shows the result of comparison between two systems. The table shows that proposed model reduce the length of pattern to $\frac{2}{3}$ than Experiment I and the number of ambiguities of the proposed model is less than Experiment I. Therefore, we concluded that our proposed model is effective to reduce the length of pattern and translation ambiguities. The number of patterns of the proposed model is more than the number of patterns of Experiment I. But as the number of patterns × average length of pattern of the proposed model is less than that of Experiment I, it is expected that as the size of corpus is bigger, the number of patterns of the proposed model will be fewer.

3.3 Error Analysis

The errors of our system are roughly divided into the errors in pattern matching and the errors in pattern transfer and sentence generation. The errors appeared in pattern matching are almost the errors of govern and governor relation.

Relation	The number of errors
Subject:Verb	9
Adverb(phrase):Verb	7
Adjective(phrase):Verb	6
Noun(phrase):Noun	2
Total	26

Table 2: The error analysis result of pattern matching

Table 2 shows the error analysis result of pattern matching. To solve these problems, it is needed to use the Korean syntactic analysis (Yoon, 1998).

We counted the errors in pattern transfer and sentence generation subjectively. Thus, the number of errors may not be exact, but the error ratio of error types seems be meaningful.

Error type	The number of errors
Restoration error	64
Analysis error	26
Dictionary construction error	22
Translated word selection error	5
Sentence generation error	3
Total	112

Table 3: The error analysis result of pattern transfer and sentence generation

Table 3 shows the error analysis result of pattern transfer and sentence generation. Restoration error has occurred when the necessary information in English sentence was not restored, e.g. article, tense, number.

NTT thesaurus is 10. But we use only 6 levels which seem to be useful in Korean

[3] http://www.cis.upenn.edu/treebank/home.html

[4] The length of pattern was calculated by the number of morphems and the length of sentence was calculated by the number of words.

4 Conclusion

We proposed two-level translation pattern selection model to reduce the length of pattern and reduce the ambiguities occurred due to short patterns. In the first step, the ambiguities in the pattern matching is reduced and several translation pattern categories are selected by the hybrid method of exact example matching and semantic constraint by thesaurus. In the second step, the most natural translation pattern is selected by the English syntactic collocational information.
In the future, it is needed to use the syntactic collocational information of Korean to reduce ambiguities in pattern matching step. Also it is needed to expand the syntactic collocational information of English, e.g. adjective-noun relation and verb-adverb relation. And research on restoring the necessary information in English sentence has to be done.

Acknowledgement

This research is supported by the Korea Science and Engineering Foundation(KOSEF) through the "Multilingual Informaton Retrieval" project at the Advanced Information Technology Research Center(AITrc), and also supported by Korea Terminology Research Center for Language and Knowledge Engineering, under the project "On Development of Deep-Level Processing and Quality Management Technology for Very Large Korean Information Base", a project of plan STEP2000 by the fund of MInistry of Science and Technology, during the period from 1997 through 2000.

References

Brown, Peter F., Stephen A. Della, and Vincent J. Della Pertra. 1991. Word Sense Disambiguation using Statistical Methods. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*, pages 264–270.

Dagan, Ido and Alon Itai. 1994. Word Sense Disambiguation using a Second Language Monolingual Corpus. In *Computational Linguistics*, 20(4):563–596.

Kim, Nari and Yung Taek Kim. 1996. Ambiguity Resolution of Korean Sentence Analysis and Korean-English Transfer Based on Korean Verb Patterns. In *Journal of KISS*, 23(7):766–775.(*in Korean*).

Kim, Yung Taek. 1994. *Natural Language Processing.* Kyo-Hak-Sa.

Lee, Hyun Ah, Jong C. Park, and Gil Chang Kim. 1999. Lexical Selection with a Target Language Monolingual Corpus and an MRD. In *Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 150–160.

Moon, Kyong-Hi, Jong-Hyeok Lee, Jung-In Kim, and Gijoo Yang. 1998. Resolution of Word Sense Ambiguities using Collocation Patterns in Japanese-to-Korean MT System. *Journal of KISS*, 25(8).(*in Korean*).

Palmer, Martha, Dania Egedi, Chunghye Han, Fei Xia, and Joseph Rosenzweig. 1999. Constraining Lexical Selection Across Languages using Tree Adjoining Grammars. *TAG+3 Workshop Proceedings, CSLI volume.*

Seo, Kwang-joon *et al.* 1998. A Study on Example-Based Korean to English Machine Translation System Development. *KAIST.*(*in Korean*)

Sorniertlamvanich, Virach. 1998. Probabilistic Language Modeling for Generalized LR Parsing. *Department of Computer Science Tokyo Institute of Technology, Technical Report.*

Takeda, Koichi. 1996. Pattern-based Machine Translation. *COLING-96*, pages 1155–1158.

Tomita, Masaru. 1991. *Generalized LR Parsing.* Kluwer Academic Publishers.

Watanabe, Hideo and Koichi Takeda. 1998. A Pattern-based Machine Translation System Extended by Example-based Processing. *COLING-98*, pages 1369–1373.

Yoon, Juntae. 1998. *Syntactic Analysis for Korean Sentences Using Lexical Association Based on Co-occurrence Relation.* Ph.D. thesis, Yonsei University, Dept. of Computer Science.(*in Korean*)