CaptionEye/EK: English-to-Korean Caption Translation System Using the Sentence Pattern

Young-Ae Seo, Yoon-Hyung Roh, Ki-Young Lee, Sang-Kyu Park

Knowledge Processing Research Team, Electronics and Telecommunications Research Institute
161 Kajong-dong, Yusong-gu, Taejon, 305-350, Korea
{yaseo, yhroh, kylee}@etri.re.kr, skpark@computer.etri.re.kr

Abstract

This paper describes CaptionEye/EK, an English-to-Korean caption translation system, which is aiming at translating English broadcasting caption into Korean one. CaptionEye/EK has been designed based on data-driven methodology. This methodology has the characteristics of both shallow bottom-up parsing between protectors and top-down matching by structure-oriented sentence patterns. The shallow bottom-up parsing between protectors is similar to the parsing of noun phrases in rule-based machine translation, and the protectors mean the linguistic part-of-speeches that cause many structural ambiguities in structural analysis. The top-down matching is similar to matching in example-based machine translation, but unlike the bilingual example in EBMT the sentence pattern is the structure-oriented pattern. The sentence patterns are patterns to be built by regarding sentence as translation unit. They consist of the source sentence pattern and the target sentence pattern that corresponds to a source sentence pattern. In order to verify our translation methodology, we made an experiment on 100 sentences that was randomly extracted from CNN news scripts. Each sentence contained average 17.2 words. In the experiment, CaptionEye/EK showed the 61% translation rates with about 28,000 sentence patterns. From the graph on the progress of translation rate, we expect that the more the number of sentence patterns is, the higher translation rate is.

Keywords

English-to-Korean machine translation, caption translation, data-driven methodology, sentence pattern

1. Introduction

With a popularization of the satellite broadcasting and the Internet, we get many opportunities to obtain information about foreign countries. But the language barrier interferes with this acquisition. In order to solve it, the machine translation systems from English to Korean have been developed actively in Korea since 1985. In spite of technological attempts for about 15 years, the translation quality of English-to-Korean machine translation system didn't go up to 40% translation rate (Se-Young, 1999). We have concluded that the major causes would be enumerated as follows (Choi, 1994 ; Hutchins, 1992 ; Sung-Kwon, 1999):

- Many ambiguities are caused by uncertain boundary of right association in parsing.
- Incorrect translation in whole sentence is occurred by considering not global translation pattern such as sentence, but local translation patterns such as phrases or clauses as translation unit of sentence.
- Translation quality comes to a standstill due to conflicts in massive translation rules accumulated every year.

To solve these problems, we have developed new machine translation methodology based on protectors and sentence patterns. According this methodology, CaptionEye/EK, an English-to-Korean caption translation system, has been developed. CaptionEye/EK translates the English news caption of satellite TV broadcasting to Korean one. Using this, The Korean who is not good at English can watch and understand the English news. Figure 1 illustrates the role of automatic caption translation system.

2. System Configuration

Our machine translation methodology belongs to data-driven methodology, which is based on protectors and sentence patterns. This methodology has two main characteristics, that is both shallow bottom-up parsing between protectors and top-down matching by structure-oriented sentence patterns.

The shallow bottom-up parsing between protectors is similar to the parsing of noun phrases in rule-based machine translation. We define protectors as the linguistic part-of-speeches that cause many structural ambiguities in structural analysis. So, parsing of English words that are located between protectors can reduce the structural ambiguity.

The top-down matching is similar to matching in example-based machine translation. But, unlike the bilingual example in EBMT the sentence pattern is the structure-oriented pattern. The sentence patterns are patterns to be built by regarding sentence as translation unit.

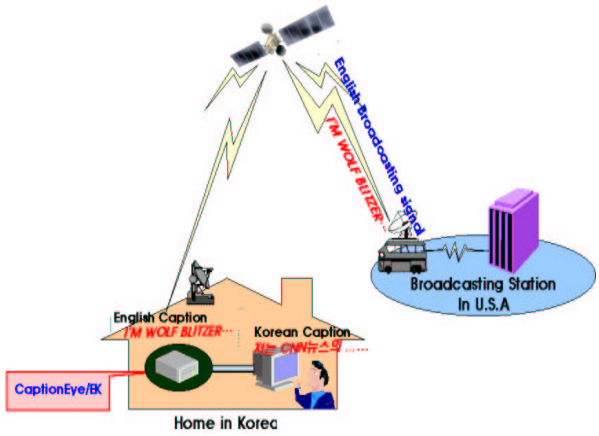

Figure 1: Role of automatic caption translation system

Figure 2 shows the configuration of CaptionEye/EK. CaptionEye/EK has basically the formalism as follows:

This system carries out English sentence analysis by partial parser between protectors. From this analysis, English sentence pattern that is corresponding to input sentence is built. And then, the system tries to match it with the English sentence patterns in DB. After selecting the appropriate English sentence pattern in DB, it generates a Korean sentence from the Korean sentence pattern, which is corresponding to the English sentence pattern.

The following shows the translation process of CaptionEye/EK. The input sentence was extracted from a CNN news script.

[Input Sentence]

The enormous amounts of special interest money that flood our political system have become a cancer in our democracy and the voices of average citizens can barely be heard.

The input sentence is tagged and the fixed pattern such as 'enormous amount of' is recognized.

[Fixed Pattern Recognition]

(det:determiner, prep:preposition, num:number, adj:adjective,punct:punctuation,conj:conjunction)

The(det) enormous_amounts_of(det) special(adj) interest(noun) money(noun) that(conj) flood(verb) our(det) political(adj) system(noun) have(aux) become(verb) a(det) cancer(noun) in(prep) our(det) democracy(noun) and(conj) the(det) voices(noun) of(prep) average(adj) citizens(noun) can(aux) barely(adv) be(aux) heard(verb)

The protectors which are linguistic part-of-speeches consisting of 'verb', 'aux', 'conj', and 'punct' are detected.

[Protector Detection]

det det adj noun noun conj verb det adj noun aux verb det noun prep det noun conj det noun prep adj noun aux adv aux verb

The phrase between protectors is parsed and is reduced to phrase symbols.

[Partial Parsing between Protectors]

(det det adj noun noun ⇒ NP) conj verb (det adj noun ⇒ NP) (aux verb ⇒ verb) (det noun prep det noun ⇒ NP) conj (det noun prep adj noun ⇒ NP) (aux adv aux verb ⇒ verb)

The resulting slot symbol is encoded to the key word in source sentence pattern database. If the key word searching in database is failed, the partial pattern corresponding to simple sentence is recognized and is translated by corresponding simple sentence pattern. Then the resulted pattern in which the partial pattern corresponding to simple sentence is reduced to 's' is translated.

[Partial Sentence Pattern Processor]

(n:NP, C:conj, V:verb, s:SP)

n/CVnVn/CnV
⇒ n/C(Vn)Vn/C(nV)
– Vn → s
– nV → s
⇒ nCsVnCs

For the translation of the reduced sentence pattern, the database key about the resulted sentence pattern is searched in database.

[Source Sentence Pattern Selection]

nCsVnCs
(= NP1 conj1 SP1 verb1 NP2 conj2 SP2)

In many target sentence patterns, the most appropriate target sentence pattern is selected and the sequence of phrases in the translated sentence is determined.

[Target Sentence Pattern Selection]

NP1 conj1 SP1 verb1 NP2 conj2 SP2 ⇒ SP1인 NP1은 NP2를 verb2고 SP2

The sequence of words in the phrases of the translated sentence is determined by slot pattern processing.

[Slot Pattern Processing]

SP1인 NP1(det1 det2 adj1 noun1 noun2 ⇒ det2 adj1 noun1 noun2)은 (det1 noun1 prep det2 noun2 ⇒ det2 noun2에서 noun1)를 verb1고 SP2

The morphological generation is conducted and the whole target sentence is generated.

[Output Sentence]

우리의 정치적인 시스템을 범람시킨 방대한 양의 특별한 금리는 우리의 민주주의에 암이 되었고 평균 시민의 목소리가 거의 들릴 수 없습니다.

And then, we describe the black boxes in the Figure 2 and the detail translation processing of CaptionEye/EK.

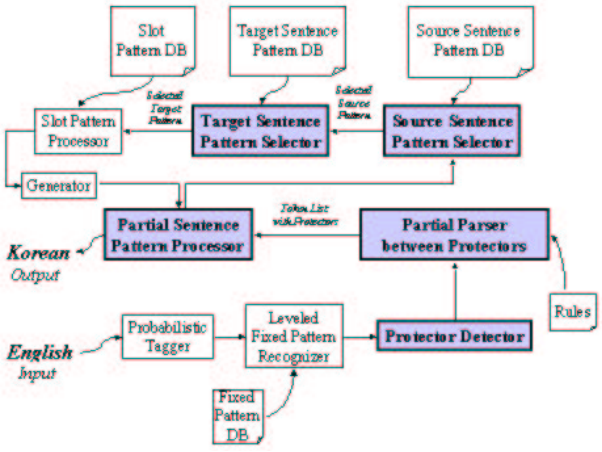

Figure 2: System configuration of CaptionEye/EK

3. Protector Detector

In rule-based English-to-Korean machine translation, one of major causes of translation failure was structural ambiguity caused by the boundary of right association in parsing. The protectors mean the linguistic part-of-speeches that cause many structural ambiguities in structural analysis. We found that parsing of English words that are located between protectors can reduce the structural ambiguity. Protector detector is a module to find protectors in English sentence. They consist of verb, conjunction such as 'and', 'that', and 'if' and punctuation mark such as 'comma' in English.

4. Partial Parser

Partial parser transforms the list of morphological part-of-speeches between protectors into phrases. The purpose of this parser is to reduce structural ambiguities in sentences and to understand the grammatical function of constituents in sentences.

Partial parser has the format of augmented context-free grammar consisting of rule description, condition checking, and action execution.

The following shows an example of rules of partial parser where the phrase 'the voices of average citizens' in the example sentence presented in Chapter 2 is reduced to 'NP PP'.

```
det noun prep adj noun ⇒ NP PP
{AND(lpro:epos==[verb],rpro:epos==[illegal]);}
{ rhs[1]:start := lhs[1];
  rhs[1]:end := lhs[2];
  rhs[2]:start := lhs[3];
  rhs[2]:end := lhs[5];
  rhs[1]:etype := lhs[2]:etype;
  rhs[2]:etype := lhs[5]:etype;}
```

The above pattern describes that the list of morphological part-of-speeches 'det noun prep adj noun' is reduced to NP PP when the left protector is verb and the right one is end of sentence.

5. Partial Sentence Pattern Processor

In pattern-based English-to-Korean machine translation, with the increase of the length, the coverage of sentence patterns decreases remarkably. So, partial sentence pattern that corresponds to sub-clause is recognized and translated first. Then the partial sentence pattern is reduced to one symbol 's' and the reduced whole sentence pattern is translated. So, the sentence pattern size is reduced and the coverage of sentence patterns increases. The next sections explain the detail processing.

5.1. Simple Sentence Recognition

For the recognition of simple sentence, all starting point of sub-clause is recognized. The starting points are recognized by the fixed rule and the patterns in DB. The starting points of simple sentence are the points which don't have that-clause and relational clause for next starting points. Next starting point generally determines the ending point of simple sentence. In case of relational clause, it demands more complicate processes. One is the match of the starting points and the main verb for the recognition of ending points. The other is sentence restoration for its translation. For example, to translate the relative clause in the sentence 'The enormous amounts of special interest money that flood our political system have become a cancer in our democracy', the ending point of the clause starting with conjunction 'that' can be obtained by matching the first verb 'flood' and the conjunction 'that', and by matching the second verb 'become' and the beginning point of sentence. Then the partial sentence 'flood our political system' is extracted as simple sentence and restored to the sentence 'the money flood our political system' for the search of the simple sentence pattern in DB.

5.2. Simple Sentence Reduction and Whole Sentence Translation

Once the simple sentences are recognized, they are translated and reduced to sentence phrase symbol 's'. Then, the reduced whole sentence pattern is searched in DB and translated. The translation process of the simple sentence is the same as that of reduced whole sentence described in the following chapters.

6. Source Sentence Pattern Selector

The source sentence pattern means analysis results of partial parsing between protectors about English partial or whole input sentences. It is made up of protectors and slots such as noun phrases and prepositional phrases.

6.1. Source Sentence Pattern DB

The key word in source sentence pattern database is given as encoded composition form of each protector and slot to reduce data base memory and to increase readability at the same time. Its content is described as form 'S-keyword' where S means source sentence pattern. Table 1 shows that the list of encoded form of protectors and slots.

Part-Of-Speech	Code	
CONJ(CONJunction)	C	Protectors
VERB(VERB)	V	
PUNCT(PUNCTuation)	T	
ADP(ADverbialPhrase)	v	Slots
AP(Adjective Phrase)	a	
DETP(DETerminer Phrase)	d	
NP(Noun Phrase)	n	
PP(Prepositional Phrase)	p	
IPREP(IsolatedPREPpositon)	i	
SP(SentencialPhrase)	s	

Table 1: Encoded name of protector and slot

The following is an example of source sentence pattern database about 'I love you':

```
Example: I love you
Key word: nVn
Content: S-nVn
```

6.2. Source Sentence Pattern Selection

Several matched source sentence patterns can be founded in parsing between protectors in case we can not know whether a prepositional phrase is a noun-modifier or it is a

verb-modifier. Because such ambiguities depend on generally context or meaning of sentence, we don't reduce structural ambiguities in module of source sentence pattern selection. Instead of that, several candidate source sentence patterns matched in source sentence pattern selector are passed over to target sentence pattern selector without ambiguity resolution.

7. Target Sentence Pattern Selector

The target sentence pattern means a target sentence pattern that corresponds to a source sentence pattern. Target sentence pattern selector is a module to select a correct target sentence pattern among target sentence patterns that corresponds to a source sentence pattern. For target sentence pattern selector, we have two types of target sentence pattern databases. One is source sentence pattern database with grammatical constraint feature and the other is bilingual sentence pattern database where Korean word order is reflected.

7.1. Target Sentence Pattern DB

7.1.1. Source Sentence Pattern DB with Grammatical Constraint Feature

There is a grammatical constraint feature in source sentence pattern DB of target sentence pattern DB. It plays a role to classify source sentence pattern to match target sentence pattern with the Korean word order. Such constraint feature is only given to protectors and consists of eform(English morphological form information, for example, comparative and superlative for adjective) and etype(English syntactic type information, for example, t1 for transitive verb). For example, in the sentence 'I love you', we can see the following source sentence pattern with constraint feature:

Example: I love you
Key word: S-nVn
Content: { NP verb:[t1,vb] NP2} T-nnV3
{ NP verb:[t1,vg] NP2} T-nVn1
{ NP verb:[t1,vn] NP2} T-nVn2

The above example shows that target sentence pattern should be T-nnV3 in case that S-nVn(key-word of source sentence pattern) has etype:t1(transitive verb) and eform:vb (declarative sentence form) in verb, T-nVn1 in case that S-nVn has etype:t1 and eform:vg ('ing' form) in verb, and T-nVn2 in case that S-nVn has etype:t1 and eform:vn (past participle form) in verb.

7.1.2. Bilingual Sentence Pattern DB with Korean Word Order

Bilingual sentence pattern DB is constituted by English sentence pattern and Korean sentence pattern. English sentence pattern has grammatical constraint feature such as 'eform' as well as additional linguistic information for translation. Korean sentence pattern is equivalent to English sentence pattern and has Korean word order and Korean linguistic information as feature. Bilingual sentence pattern of 'I love you' is as follows:

Example: I love you
Key Word: T-nnV3
Content: { NP1 VERB1!:[etype == [t1]] NP2 } → { NP1:[kcase := [topic]] NP2:[kcase := [obj]] VERB1! }

The above pattern describes that English word order NP1(subject) VERB1 NP2(object) is changed into NP1(subject) NP2(object) VERB1 when VERB1 is a type of transitive verb and both subject and object are demonstrative pronoun.

7.2. Target Sentence Pattern Selection

It is possible that one English sentence pattern matches to several Korean sentence patterns. In this case, we select the appropriate sentence pattern by the heuristic rules such that the sentence pattern that meets the more constraint features and has less phrasal slots is selected.

8. Experiment

In order to verify our translation methodology, we made an experiment on 100 sentences, which was extracted randomly from CNN news scripts of March 2001. Each sentence contained average 17.2 words. The Table 2 describes the characteristics of data that is used in the experiments.

Number of sentence	100
Average word number of sentences	17.2
Type of sentences	Random
Number of sentence patterns	28,397

Table 2: Characteristics of experimental data

Figure 3 shows the relation between the number of sentence patterns and the translation rate. From this, we can see that translation rate grows with the number of sentence patterns. It is the major characteristics of data-driven machine translation approach. CaptionEye/EK shows the 61% translation rates with about 28,000 translation patterns.

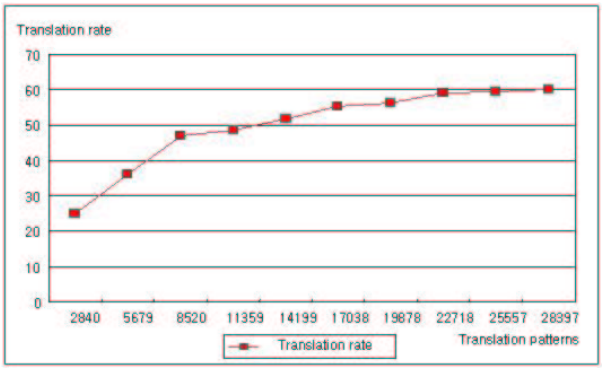

Figure 3: Relation between the number of sentence patterns and the translation rate

Figure 4: Running shot of the caption translation system

Figure 4 shows the running shot of the caption translation system. In this figure, we can see the English caption and its translation result, Korean caption. Caption translation is processed in almost real time.

9. Conclusion

In this paper, we have described new machine translation methodology based on protectors and sentence patterns that is realized in the EKMT system for translating broadcasting caption. The protectors mean the linguistic part-of-speeches that cause such structural ambiguities in structural analysis. The sentence patterns are pattern to be built by regarding sentence as translation unit. They consist of source sentence pattern and target sentence pattern that corresponds to a source sentence pattern.

New machine translation methodology based on protectors and sentence patterns is solving considerably various problems that existing English-to-Korean machine translation systems, that is, generation of many ambiguities caused by the uncertain boundary of right association in parsing, occurrence of incorrect translation in whole sentence by considering local translation patterns as translation unit of sentence, and conflicts in massive translation rules accumulated every year. But this methodology still has the unsolved problems such as the coordination scope, and we are trying to solve these problems.

CaptionEye/EK is still under growing, scaling up the dictionaries, and accumulating the massive sentence pattern to achieve the better translation quality.

References

Choi K.S., Lee S.M., Kim H.G., and Kim D.B. (1994) An English-to-Korean Machine Translator: MATES/EK. COLING94, 129-133.

Hutchins W.J. and Somers H.L. (1992) An Introduction to Machine Translation. Academic Press.

Se-Young Park, Gil-Rok Oh, “Machine Translation in Korea”, MT-Summit VII, 100-104, 1999

Sung-Kwon Choi, Taewan Kim, Sang-Hwa Yuh, Han-min Jung, Chul-Min Sim, Sangkyu Park(1999) "English-to-Korean Web Translator: "FromTo/Web-EK"", MTSUMMIT99, Singapore.