

한국어에서 구절의 검출 방법

이 정현\* 최 장선 이 대형 이 주근  
\* 경기대 전산과 인하대 전자공학과

A Method Detecting the Phrase and Clause in Korean

J. H. Lee J. S. Choi D. H. Lee J. K. Lee

\* Dept. of Computer Science Kyonggi Univ. Dept. of Electronics Inha Univ.

ABSTRACT

This paper describes a method which detects the phrase and clause in Korean sentences.

By analysing modificatory relation of modifier and modified word in Korean, two rules are defined.

- (1) The rules of phrase and clause.
- (2) The rules of separation points.

As the result of this method, the conceptual units are easily elicited from the Korean sentence.

1. 서 론

자연언어의 처리를 위한 기계번역 시스템, 질문응답 시스템 등에서는 자연언어를 이해하여 지식 베이스로 구축하는 것이 중요하다. 이런 시스템들은 구문적, 의미적 해석에 의한 parser를 사용하고 있다.[3][4]

본 논문은 한국어의 이해를 위해 구와 절의 규칙을 정의하고, 구절의 분리점 규칙을 정의하였다. 또한 한국어 문장에서 수식어는 반드시 피수식어의 앞에만 존재하는 특징을 가지므로, 이런 규칙들을 이용하여 left-to-right 방향으로 수식의 범위를 결정하는 방법을 제안하였다.

따라서 한국어 문장의 구절로부터 개념단위를 도출하기가 용이해진다.

본 논문에서 형태소 해석과 구문 및 의미 해석은 한일 기계번역 시스템의 방법을 이용하였다.[1][2]

2. 한국어 구절의 규칙

한국어는 문법구조가 강하며, 수식어와 피수식어의 위치 관계가 영어와는 다르게 수식어는 반드시 피수식어의 앞에만 존재하여 left-branch 관계에 의한

수식 관계만이 성립한다. 또한 한국어에서 개념 단위는 용언이 중심이 되어 그의 지배를 받는 체언과 결합하여 형성하는 구조적 결합 관계와 용언이 관형어로서 체언의 의미적 범위를 제한하는 한정적 결합 관계에 의해 이루어진다.[5]

이와 같은 개념 단위를 도출하기 위해서 결합 관계를 해석하기 위한 구와 절의 규칙을 다음과 같이 정의한다.

(1) 구의 규칙

- 1)  $N \leftarrow [NIDetIDINPi] + N$
- 2)  $NPcs \leftarrow N + Pcs$
- 3)  $NPco \leftarrow N + Pco$
- 4)  $NPcc \leftarrow N + Pcc$
- 5)  $NPi \leftarrow N + Pi$
- 6)  $NPs \leftarrow N + Ps$
- 7)  $V \leftarrow [AdiAINPcoINPcc] + [VIVBe]$
- 8)  $VBe \leftarrow V + Be$
- 9)  $VEc \leftarrow V + Ec$
- 10)  $D \leftarrow N + Pcd$
- 11)  $D \leftarrow [VIS] + Ed$
- 12)  $D \leftarrow Ad + Det$
- 13)  $A \leftarrow N + Pca$
- 14)  $A \leftarrow [VIS] + Ea$

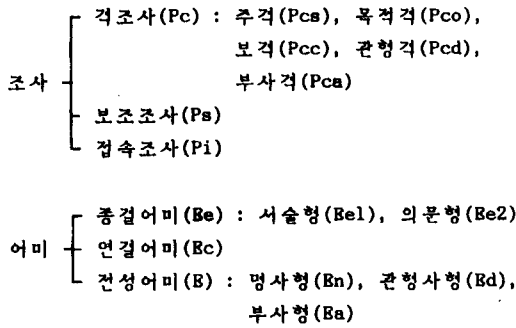
(2) 절의 규칙

- 1)  $S \leftarrow NPcs + [VIVBe|VEc]$
- 2)  $S \leftarrow Nps + [SIV]$

여기서 " $\leftarrow$ ", "[ ]", " | "는 각각 합속, 반복, 선택을 의미하며 기호는 다음과 같다.

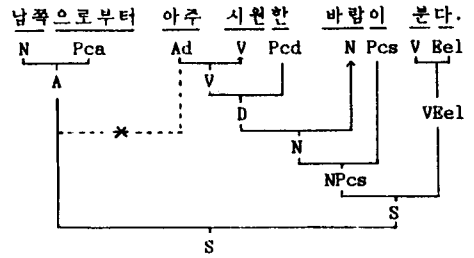
S: 절 Det: 관형사  
N: 체언 Ad: 부사  
V: 용언 B: 어미  
D: 관형어 P: 조사  
A: 부사

특히 본 논문에서는 조사와 어미는 다음과 같이 분류하였다.



성립하지 않는 구조적 분리점으로서 위 문장의 주어부와 서술부가 구분된다.

(예 2)



위의 (예 2)에서 A \* Ad 는 부사어가 부사를 수식하지 않는 경우로서 구조적 분리점이 된다.

3. 구절의 분리

한국어의 일반 문장으로부터 개념 단위를 도출하기 위해서는 수식어는 반드시 피수식어의 앞에만 존재하므로 앞에서 제시한 구절의 규칙에 의해 left-to-right 방향으로 수식어와 피수식어의 수식 관계를 조사하고 수식 관계가 성립되지 않는 문장의 분리점을 검출함으로써 가능해진다.

문자로부터 구와 절을 구분해내기 위하여 구문적 분리점과 의미적 분리점으로 구분한다.

B. 의미적 분리점

구문적으로는 수식 관계가 성립하지만 강제적으로 분리점을 결정하여야 하는데 이것을 의미적 분리점이라 한다.

의미적 분리점의 규칙은 다음과 같다.

[규칙 2]

$$[NPcs \mid NPco \mid NPcc \mid NPs] \# V$$

여기서 #는 의미적 분리점이다.

A. 구문적 분리점

문장에서 연속되는 단어사이에 수식어와 피수식어로서 수식관계가 성립하지 않는 것을 구조적 분리점이라 한다. 이 구조적 분리점의 규칙은 다음과 같다.

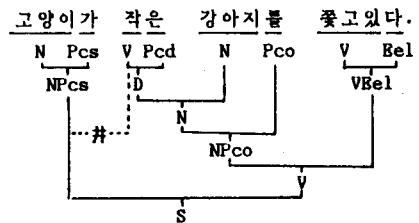
[규칙 1]

$$[NPcs \mid NPco \mid NPcc \mid NPs \mid A] * [N \mid Ad \mid Det]$$

여기서 "[]"는 반복, "\*"는 선택을 의미하며, #는 구문적 분리점이다.

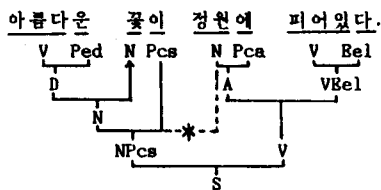
구문적 분리점과 구와 절의 규칙은 배타적 관계(Exclusive relation)이다.

(예 3)



NPcs + V 는 수식 관계가 성립하지만 올바른 구문 해석 tree를 얻기 위하여 강제적으로 분리점을 설정함으로써 주어의 탈락을 막을 수 있다.

(예 1)



위의 (예 1)에서 NPcs \* N 은 수식 관계가

4. 실험 결과

1. 어러분은 \* 이제 \* 모교의 \* 살아있는 역사의  
일부가 되었을 뿐 아니라 \* 민족의 미래를 이끌어갈  
주체로서의 역량을 발휘해야할 \* 중대한 결심의  
순간을 # 맞아했다.

2. 어러분은 \* 이제부터 전개되는 미지의 세계를

참고 문헌

향하여 냉철한 지성과 용기있는 도전으로 \*

조국번영과 민족통일을 이기기위한 희생과 화합의 \*

정당한 길을 \* 걸어가야 할것이다.

(중앙일보 기사 1987년 2월 25일)

5. 결 본

한국어는 수식어가 반드시 피수식어의 앞에만 존재하는 특징을 가지고 있다.

이러한 특징을 이용하여 구와 절의 규칙을 정의하고, 구분적 분리점과 의미적 분리점의 규칙을 제안하였다.

따라서 left-to-right 방향으로 수식의 범위를 결정할 수 있으며 한국어의 문장으로부터 개념단위의 도출이 용이해지므로 지식베이스의 구축이 쉬워질 것으로 기대된다.

[1] 이 주근, 이 수현, 박 찬근, "한일 기계번역 시스템 (I)," 전자공학회 학술 논문집, vol.8, No.1, pp.202-204, 1985.

[2] Joo Keun Lee, Kwang Rok Han, "Korean-Japanese Machine Translation System(III)," Proceeding of 1985 SISSE, pp.134-138, Nov. 1985.

[3] Margaret King, Parsing Natural Language, Academic Press, 1983.

[4] K. S. Jones and Y. Wilks, Automatic Natural Language Parsing, Ellis Horwood Ltd., 1985.

[5] 최 창렬, 국어 의미구조 연구, 한신문화사, 1980.