문형을 제약 조건으로 하는 CFG기반의 한국어 구문분석

이현영, 황이규, 배우정, 이용석
전북대학교 컴퓨터과학과

Syntax analysis of Korean based on CFG using Sentence Pattern Information as a constraint

Hyeon-Yeong Lee, Yi-Gyu Hwang, Woo-Jeong Bae, Yong-Seok Lee
Dept. of Computer Science, ChonBuk National University

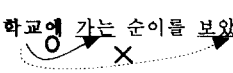
요 약

한국어는 용언이 의미적 제약을 통해 문장을 지배하는 SOV 구조의 언어이다. 또한, 조사나 어미와 같은 기능어의 발달은 물론 관형절을 내포하는 문장이 주류를 이룬다. 따라서 한국어의 구문분석은 부착에 따른 많은 구문 모호성이 발생하게 된다. 본 논문에서는 조건단일화 기반의 CFG 문법을 기술하고 문형을 구문 제약으로 하여 구문모호성을 해결하는 방안을 제시한다. 문형은 한국어의 특성을 용언의 하위범주화에 맞게 재분류한 문장의 구조적 유형을 말한다. 본 논문에서 제안하는 문형은 동사와 형용사를 구분하여 39가지로 설정하였다. 이런 문형 정보를 이용하여 관형형 어미를 갖는 용언이 최대의 정보를 가지도록 함으로써 관형절에서 발생하는 부사 및 체언구 부착의 문제가 해결된다. 또한 문형은 이중주어나 이중 목적어 문장을 처리할 수가 있어 한국어에서 발생하는 많은 구문모호성을 해결할 수 있다.

1. 서론

한국어는 부분 자유 어순을 가지며 용언이 의미적 제약을 통해 문장을 지배하는 언어이다. 또한 한국어는 상황 중심의 언어라 불릴 만큼 의미(semantic)와 화용(discourse)이 중요한 역할을 한다. 따라서 한국어는 서구어와는 달리 정교하게 설계된 문법 규칙과 엄밀한 제약을 갖는 문법 이론보다는 규칙외적인 정보에 의해서 분석이 주도된다[1]. 이러한 관점에서 볼 때 한국어의 구문분석을 위해서는 규칙은 간단히 기술하고 분석 도중에 각 문장소(morpheme)들의 문법적 관계를 검사하면서 분석하는 방법이 타당하다고 볼 수 있다.

구문분석은 문장에서 체언구와 용언사이의 표층적 문법 관계를 밝히는 작업이다. 구문분석 단계에서 발생하는 대부분의 모호성은 용언과 체언의 결합에 따라 나타난다. 즉, "NP+VP"나 "VP+NP"로 결합할 때 발생한다. 예로 다음의 문장에서 '학교에'라는 체언구는 용언 '가다'나 '보다'에 모두 부착이 가능하지만 "N이 N에 V"라는 문형을 갖는 '가다'에 부착되는 것이 타당함을 알 수가 있다.



문형) 가다 : N이 N에 V, N이 N로 V, N이 V
보다 : N이 N을 V

<그림 1> 한국어 문장에서 부착 모호성의 발생 예

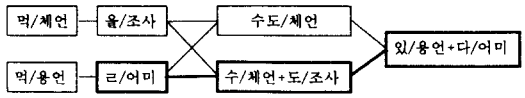
이와 같이 한국어에서는 용언이 요구하는 조사와 명사의 유형이 존재한다. 격틀은 용언의 의미적 자질에 따라 가질 수 있는 심층격을 기술하기 때문에 용언과 체언에 대한 정확한 시멘틱이 필요하다[2]. 따라서 정확한 구문분석은 가능하지만 격틀의 구축은 거의 불가능하다. 반면 문형은 구문 정보에 약간의 의미적 정보를 가한 문장의 구조적 유형을 말한다[3]. 따라서 본 논문에서는 용언의 하위범주화 정보를 한국어의 특성에 맞게 재분류한 문형을 구문 모호성을 해결하기 위한 제약조건으로 사용한다. 또한 문형 정보만으로 구문 모호성을 해결할 수 없는 경우에는 문형에 의미제약을 가한 의미지표를 이용하여 모호성을 해결한다.

본 논문에서는 한국어의 구문분석에서 발생되는 구문모호성의 원인을 파악하고 이를 해결할 수 있는 방안을 제시한다. 구문모호성의 원인은 크게 두 가지로 분류할 수가 있다. 첫째는 구문해석의 이전 단계인 형태소분석 결과에서 발생하는 형태론적 모호성이다. 이것은 [4]가 제안한 구문형태소를 이용하여 해결할 수가 있다. 둘째는 한국어의 구문 특성상 발생하는 모호성이다. 본 논문에서는 이를 해결하기 위하여 한국어의 구문 특성을 살펴보고 모호성을 해결할 수 있는 방안으로 문형을 제약 조건으로 하는 구문분석 방법을 제안한다.

2. 한국어의 특성

2.1 형태론적 특성

한국어는 여러 형태소들이 결합하여 하나의 구문 단위를 이루는 경우가 많다. 이러한 형태소열은 형태론적 모호성과 구문모호성의 원인이 되므로 이를 해결하기 위한 연구[4,5,6]가 많이 진행되었다.



<그림 2>"먹을 수도 있다"의 형태소 분석 결과

<그림 2>는 "먹을 수도 있다"에 대한 형태소분석 결과로 4개의 형태론적 모호성이 발생함을 알 수 있다. 그러나 [4]가 제안한 구문 형태소를 이용하면 "ㄹ 수도 있다"라는 형태소열이 결합하여 '가능'이라는 양상정보로 표현된다. 따라서 '먹다/용언[가능]'이라는 하나의 결과만을 얻을 수가 있다. 이와 같이 구문형태소는 형태론적 모호성을 해결하고 구문분석의 부담을 덜어준다. 따라서 본 논문에서는 구문형태소를 구문분석의 입력 단위로 사용한다.

2.2 구문적 특성

한국어는 부분 자유 어순을 갖는 SOV 형태의 언어로 용언에 따라 다양한 격조사와 체언을 요구한다. 따라서 문장의 구조를 파악하기 위해서는 정형화된 구문적 정보만을 이용할 수는 없다. 그러나 기존에는 체언구와 용언 사이의 표층적 문법 관계로 "주격/목적격/장소격/도구격/기타격"만을 고려하거나 용언 정보인 "자동사/타동사/형용사" 정보만을 이용하여 구문해석을 시도했기 때문에 많은 모호성이 발생하게 되었다. 예를 들어 다음의 문장을 살펴보자.

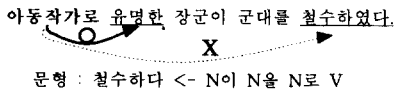
1) 철수가 귀찮게 군다. 1*) 철수가 군다.*
2) 철수가 순이와 싸운다. 2*) 철수가 싸운다.*

1)과 2)에서 '군다'나 '싸우다'는 자동사이므로 주격만을 필수 성분으로 간주할 수 있다. 따라서 위의 4문장은 모두 옳은 문장으로 분석이 된다. 그러나 '군다'라는 용언은 '어떠하게'라는 의미를 가지는 부사를 필요로 하고 '싸우다'는 '~와'라는 체언구를 문장의 필수성분으로 가진다. 따라서 1*)와 2*)는 의미적으로 올바른 문장이 아님을 알 수가 있다.

이와 같이 한국어는 부사나 특별한 격을 수반하는 용언이 많이 존재한다. 이런 경우에 부사나 특별한 격을 보조적인 의미로 파악하면 문장의 올바른 의미를 파악하기 어렵거나 모호성 발생의 원인이 된다. 따라서 이러한 용언들의 구조적 유형을 어떤 틀로 제약할 필요가 있다.

또한 한국어에는 용언이 두 개 이상 나오는 문장이 대부분을 차지한다. 그 중에서도 관형절을 포함하는 문장에서 체언구나 부사의 부착문제는 많은 구문모호성의 원인이 된다. 관형절에서 용언의 뒤에 오는 체언구는 대부분 그 용언의 필수격 구실을 한다. 따라서 <그림 1>의 문장은 "순이가 학교에 가다", "철수가 그 순이를 보다"라는 의미를 가지도록 분석이 되어야 한다. 문형을 이용하면 "순이가 학교에 가다"는 "N이 N에 V"라는 문형을 만족하므로 용언 '가다'와 '철수가'가 결합하는 것을 막을 수 있다. 이와 같이 문형 정보를 이용하여 용언과 부착할 수 있는 체언구나 부사를 최대로 가지도록 제약하면 많은 구문모호성을 쉽게 해결할 수가 있다.

그러나 문형정보만으로는 구문 모호성을 해결하기 어려운 경우도 있다. 예를 들어 <그림 3>에서 문형 정보만을 이용하면 '아동작가로'는 '유명하다' 뿐만 아니라 '철수하다'와도 결합할 수 있다.



문형 : 철수하다 <- N이 N을 N로 V
[N로 : 구체적 장소]
유명하다 <- N이 N로 V
[N로 : 직업-신분/구체물]

<그림 3> 문형과 의미지표에 의한 모호성 해결

이러한 문제를 해결하기 위해서는 문형에 나오는 체언구를 제약하면 된다. 즉, '유명하다'에서 부사격 조사 'N(으)로'를 가지는 체언은 반드시 '직업-신분'을 나타내도록 하고 '철수하다'는 '구체적 장소'를 나타내도록 제약을 가하는 것이다. 그러면 '아동작가로'라는 체언은 '유명하다'와 결합한다. 이와 같이 체언구를 제약하는 정보를 의미지표라고 한다. 따라서 문형정보만으로 해결이 불가능한 구문모호성은 의미지표를 사용하여 해결한다.

3. 문형을 제약조건으로 하는 CFG

3.1 문형과 의미지표

한국어는 용언 중심의 문장 구조를 가진다. 이는 체언보다는 용언에 의해 문장의 구조가 결정된다는 의미이다. 이런 이유로 대부분의 한국어 분석에서는 용언을 이용하여 구문분석을 수행하고 있다. 대표적인 예로는 격틀과 문형이 있는데 격틀은 용언에 대한 정확한 시멘틱을 요구하기 때문에 정확한 구문분석은 가능하지만 구축이 불가능한 실정이다. 반면에 문형은 순수한 구문적 정보만을 이용하며 약간의 의미적 제약을 가할 수 있기 때문에 본 논문에서는 문형을 이용하여 구문분석을 수행한다. 문형을 설정하기 위해 대량의 코퍼스로부터 문장을 추출하고 용언의 특성에 따라 유형을 분류하였다. 또한, 연세 한국어 사전[7] 정보를 활용하여 동사는 31개, 형용사는 8개의 문형을 설정하였다. 본 연구에서 설정된 문형의 일부는 다음과 같다.

V1) N(이/는/은/가)+V	A1) N(이)+A
V2) N(이)+N(에/에게)+V	A2) N(이)+N(에)+A
:	:
V30) N[이]+N[와]+N[에]+V	A7) N1(이)+N(로)+N2(이)+A
V31) N[이]+N[에게]+N[에서]+V	A8) N1(이)+N(와)+N2(이)+A

한국어는 같은 의미적 계층을 가지는 용언이라도 개개의 용언에 따라 문형이 다를 수가 있다. 예를 들어 감각동사인 "맡다, 시청하다, 보다"의 경우 "N이 N을 V"라는 문형을 가지지만 목적어는 다음과 같이 의미적 제약이 따른다.

말다 : 주체가 냄새를 맡다 -- **[추상물, 냄새]**
시청하다 : 주체가 TV를 시청하다 -- **[구체물]**
보다 : 주체가 구체물을 보다 -- **[구체물]**

이러한 '의미지표를 가장 일반적으로 표현하는 것이 공기정보를 이용한 것으로 볼 수 있다. 그러나 공기 정보를 이용할 경우, 코퍼스로부터 추출한 공기 정보는 자료부족 문제를 야기할 수 있다. 이는 용언과 체언이나 부사에 대한 부분적인 공기관계만을 추출할 수 있음을 의미한다. 우리가 분류한 문형은 이러한 자료 부족 문제를 어느 정도 해소할 수 있으며 명사와 용언 및 부사와 용언에 대한 의미 표지는 [7]의 분류를 참조하여 사전을 구축하였다.

3.2 조건단일화를 이용한 Context Free Grammars

우리는 구문 분석을 위한 기본 틀로 문형을 제약조건으로 사용할 수 있는 조건단일화 기반의 CFG를 이용하였다. 이는 구구조 규칙의 간결함과 구구조 의존적 언어의 특성을 조건 단일화 제약을 통해 문장을 분석하는 것이다. <표 1>은 구구조 규칙 "SV -> NP SV"가 적용되기 위해 필요한 제약들을 문형을 기술한 예를 보여준다.

우리는 PATR II를 이용하여 위와 같은 CFG 기반의 문법을 작성하고 이를 구문 분석을 위한 LR 파싱 테이블과 조건 제약을 위한 함수로 번역하여 LR 파서 기반의 구문 분석기를 이용하고 있다.

```
(<SV> -> (<NP> <SV>)   ;;;;; Context Free Grammar Rule
 ((x0 = x2)
   (*or*
      (((x1 jform) =c jcs)
       (*or*
          (((x0 subj) = *defined*)
           ((x0 sp-info) =c (*or* v6 v10))
           (*or*
              (((x0 subj jform) =c jxc)
               ((x0 comp) = x1))
                    :
              (((x0 subj) = *undefined*)
               ((x0 subj) = x1))) )))
      (((x1 jform) =c jco)
       (*or*
          (((x0 sp-info) =c v2)
           ((x0 dest) = *undefined*)
           ((x0 dest) = x1))
          (((x0 sp-info) =c (*or* v26 v27))
           ((x0 about) = *undefined*)
           ((x0 about) = x1))
```

<표 1>문형을 제약조건으로 한 조건단일화 기반의 CFG

4. 문형을 이용한 구문 모호성 해결

관형절을 내포하는 문장이 주류를 이루는 한국어 문장은 의미를 보다 명확히 하기 위해서 부사나 보조적 의미를 가지는 체언구가 많이 사용된다. 따라서 부사나 체언구 부착의 문제로 많은 구문모호성이 발생한다. 예를 들어 예문 3)은 공동격 조사 '와'와 부사구 부착에 관련된 모호성이 발생한다.

3) 철수가 **영수와** *학교에서* 싸운 순이를 보았다.
가) 철수가 *[영수와 학교에서 싸운 순이를]* 보았다.
나) 철수가 영수와 *[학교에서 싸운 순이를]* 보았다.
다) 철수가 영수와 학교에서 *[싸운 순이를]* 보았다.

문형) 싸우다 : N이 N와 V
보다 : N이 N을 V

본 논문에서는 문형을 이용하여 이러한 부착의 문제를 해결한다. 먼저 관형절에 있는 용언의 뒤에 따르는 체언(순이)은 대부분 그 용언의 필수격 구실을 한다. 또한 관형절에 있는 용언에는 문형정보를 최대로 만족하도록 부착을 한다. 따라서 용언 '싸우다'는 "N이 N와 V"문형이므로 자연스럽게 "영수와 학교에서 싸운 순이를"이 묶이게 된다. 이와 같이 본 논문에서 제안한 방법으로 구문분석을 수행하면 공동격 조사 '와'나 부사구 부착에 따른 모호성을 해결할 수가 있다.

또한 문형을 이용하면 용언이 필수적으로 요구하는 부사나 필수격의 처리가 가능하다. 예문 1)에서는 '군다'가 가지는 문형 "N이 ADV[-게] V"를 이용하고 예문 2)에서는 '싸우다'가 가지는 문형 "N이 N와 V"에 의해 비문을 가려낼 수가 있다. 이와 같이 구문적으로는 바르지만 의미적으로 비문인 구조를 쉽게 처리할 수가 있다.

아울러, 문형을 이용하면 이중주어나 이중 목적어 문장을 처리할 수가 있다. 예문 4)는 이중 주어 문장이고 예문 5)는 이중목적어 문장이다. 자동사나 타동사의 정보만을 이용하면 구문분석에 실패하지만 문형을 이용하면 구문 분석할 수가 있다.

4) 철수가 돈이 모자랐다. **문형) N1이 N2이 V**

5) 어머니가 철수를 아침을 굶겼다. **문형) N1이 N2을 N3을 V**

그러나 한국어는 <그림 2>에서 본 것처럼 구문구조만으로는 해결할 수 없는 모호성이 존재한다. 또한 체언구의 역할은 조사에 의해서 대부분 결정되지만 보조사인 경우에는 어떤 역할을 하는지 결정하기가 어렵다. 6)에서 '철수'와 '밥'의 의미지표는 '사람'과 '음식'이므로 '먹다'의 문형에 의해 '철수'가 주어로 '밥'이 목적어로 결정할 수가 있다. 또한, 7)과 같이 공동격 조사 '와/과'도 어떻게 묶이느냐에 따라 모호성이 발생할 수가 있다. 의미지표를 이용하여 '빵'과 결합하는 것은 '철수'가 아니라 '우유'임을 알 수 있다.

6) 철수는 밥은 먹는다. [N이 N을 V]
의미지표 : [철수:사람], [밥:음식]

7) **빵과** *철수가* 먹은 **우유**
의미지표 : [철수:사람], [빵:음식], [우유:음식]

KAIST corpus와 초등학교 사회 교과서에서 추출한 232문장으로 실험한 결과 많은 구문모호성이 감소되었다. 따라서 문형과 의미지표는 구문모호성을 해결하기 위한 제약정보로 활용할 수 있다.

실험 문장	문형의 사용 안함	문형의 사용
Kaist Corpus (132)	147.3	14.9
사회교과서(100)	162.5	15.4
평균 (232)	154.9	15.14

<표 2> 문형을 사용할 때와 안할 때의 구문모호성의 수

5. 결론

부분 자유 어순을 가지고 기능어가 발달된 한국어를 구문분석하기 위해서 문형을 제약조건으로 하는 조건단일화 기반의 CFG를 사용하였다. 문형은 관형절을 포함하는 문장에서 부사나 체언구의 부착의 문제를 해결하고 필수격을 필요로 하는 용언이나 이중 주어, 이중목적어의 문제를 해결하는 데 좋은 제약이 됨을 보였다. 또한 문형 정보만으로 해결되지 않는 구문모호성은 문형의 필수격에 의미를 제약하여 해결하였다.

본 논문에서는 문형을 이용한 조건 단일화 기반의 CFG를 사용하여 어순이 자유로운 한국어 문장에서 발생하는 많은 구문 모호성을 해결할 수 있음을 보였다. 이는 일본어와 같이 문법을 기술하기 어려운 언어라도 문형만 파악된다면 효율적인 구문분석이 가능함을 의미한다. 향후 연구과제로는 지금까지 개발된 문형규칙이 한국어의 여러 현상을 처리할 수 있도록 개선하는 것이다. 또한 의미지표를 좀 더 세분화하여 문형에 대한 제약으로 활용하기 위한 연구가 필요하다

참고문헌

[1] S. W. Yang, G. O. Lee, Y. S. Lee, "An Analysis Technique for Korean Sentences Using the Conditional Unification," Proceedings of the International Conference on Computer Processing of Oriental Languages, pp.257-262, 1994.

[2] 김승곤, "한국어의 격 이론", 국어 토씨 연구, 서광학술자료사,pp.343-409, 1994.

[3] 강은국, 조선어 문형 연구, 박이정출판사, 1996.

[4] Y. G. Hwang, H. Y. Lee, Y. S. Lee, "Resolution Strategy of Morphological Ambiguity for Korean Parsing," Proceedings of the International Conference on Computer Processing of Oriental Languages, pp.53-58, 1999.

[5] 강승식, 음절정보와 복수어 단위 정보를 이용한 한국어 형태소 분석, 서울대 박사학위 논문, 1993.

[6] 김창제, 정천영, 김영훈, 서영훈, "부분적인 어절 결합을 이용한 효율적인 한국어 구문 분석기", 제22회 정보과학회 가을 학술발표논문집, pp. 597-600, 1995.

[7] 연세대학교, 연세 한국어 사전, 두산 동아 출판사, 1998.