문형과 단문 분할을 이용한 한국어 구문 모호성 해결

이현영[U] 황이규 이용석
전북대학교 컴퓨터과학과 언어정보공학실
{hylee, lethe}@cypher.chonbuk.ac.kr

Resolution of Korean Syntactic Ambiguity using Sentence Patterns Information and Clausal Segmentation

Hyeon-Yeong Lee[U] Yi-Gyu Hwang Yong-Seok Lee
Dept. of Computer Science, Chonbuk University

요 약

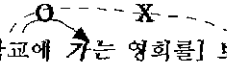
한국어 구문 분석은 체언구 부착이나 부사구 부착의 문제를 가진다. 이런 부착의 문제는 많은 구문 모호성을 만들어 내어 올바른 의미를 가지는 파스 트리의 선택을 어렵게 한다. 한국어에서 이런 부착의 문제는 대부분 한국어 문장이 내포문을 포함하는 복문의 형태로 구성되어 있기 때문이다. 단문에서는 부착의 문제가 발생하지 않지만 복문에서는 체언구나 부사구가 어떤 용언에 부착하느냐에 따라 체언구 부착이나 부사구 부착의 문제가 발생한다. 따라서 용언이 가지는 정보를 이용하여 내포문의 범위를 결정해서 하나의 구문범주의 기능을 가지도록 분할한다. 이를 단문 분할이라 하며 문형이 가지는 필수격들을 최대로 부착하여 이루어진다. 단문분할을 하면 복문의 구조가 단문으로 바뀌므로 이런 부착의 문제가 자연스럽게 해결된다. 본 논문에서는 문형과 단문 분할을 이용하여 많은 구문 모호성을 해결할 수 있음을 제안한다.

1. 서론

전통적인 한국어 구문 분석 결과는 많은 구문 모호성을 가지고 있기 때문에 의미처리나 기계번역 등의 응용분야에 바로 적용하기에는 여러 가지 문제가 있다. 구문 모호성의 원인은 크게 형태소 해석 결과의 과생성과 한국어의 구조적 특성에 의한 체언구 부착이나 부사구 부착 문제 때문이다. 형태소 해석 결과의 과생성에 의한 구문모호성을 해결하기 위한 연구로는 [1,2,3,4,5,6]등의 연구가 있었지만 구문분석의 관점에서 연구한 [6]의 구문 형태소[1]가 가장 타당하다고 생각된다. 따라서 본 논문에서는 구문 형태소를 이용하여 형태론적 모호성을 해결한다. 또한 구문적 특성에 의해 발생되는 구문 모호성은 체언이나 부사와 용언간의 결합에 의해 주로 발생한다. 이것은 다음과 같은 한국어의 특성에 의해 발생되는 모호성으로 본 논문에서는 이 문제를 해결하는 방안을 제안하고자 한다.

1) 여러 형태소가 결합하여 하나의 구문적 단위나 의미적 단위를 가질 수 있는 형태소들

한국어는 한 구절에 중심어인 용언은 고정되어 있고 나머지 어휘 성분은 부분적으로 자유로운 어순 특성을 가지고 있다. 또한 한국어는 단문보다는 내포문을 포함하는 복문의 구성 형태를 가진다. 예로 KAIST-Corpus[2]를 분석한 결과 두 개 이상의 용언으로 구성되는 문장은 90.4%나 차지했다. 따라서 단문에서는 모든 체언구나 부사구를 용언에 부착하면 되지만 내포문을 포함하는 문장에서는 부사구나 체언구를 어떤 용언과 결합하느냐에 따라 많은 구문 모호성이 발생하게 된다.



가) 철수가 [학교에 가는 영희를] 보았다.

문형[3]) 가다 : N이 N에 V
보다 : N이 N을 V

2) KAIST에서 동아일보와 한겨레사설, 소설, 초등학교 교과서와 한글과 컴퓨터의 홍보자료로부터 한국어의 구문 트리를 작성할 목적으로 작성된 태그된 코퍼스

3) 편의상 문장에서 사용되는 문형만 기술함

예를 들면, 문장 가)는 체언구 "학교에"가 용언 "가는" 이나 "보다" 모두에 부착할 수가 있다. 그러나 문형정보 [7]을 이용하여 용언에 부착되는 체언구나 부사구를 제약하면 이런 부착의 문제를 해결할 수가 있다. 가)에서는 문형 "N이 N에 가다"와 관형절의 특성을 이용하여 용언 '가는'에 부착되는 체언구를 제약하면 "학교에"는 용언 "가다"에 부착된다. 따라서 관형절의 범위를 "학교에 가는 영희를"로 제약하여 하나의 단문으로 분할할 수가 있다. 이렇게 분할된 관형절은 목적어 구실을 하는 체언구의 기능을 가짐을 알 수가 있다. 따라서 "보다"의 문형 "N이 N을 보다"에서 "N을"에 해당된다. 이와 같이 내포문에 사용된 용언의 문형 정보를 이용하여 그 용언이 가지는 필수격들을 최대한으로 묶어 하나의 단위로 분할하면 부착의 문제에 의해 발생되는 많은 구문모호성을 제거할 수가 있다.

관형절, 명사절, 부사절 등과 같은 내포문을 포함하는 문장에서 내포문에 포함된 용언의 문형정보에 의해 필수격을 최대한 부착하고 보조격들은 가까운 거리에 부착하여 단문으로 분할하는 것을 본 논문에서는 **단문 분할**이라고 정의한다. 또한 이러한 단문 분할을 통해 얻어진 결과에서 관형절과 명사절은 체언구의 기능을 가지며 부사절은 부사의 기능을 가지도록 한다. 이렇게 함으로서 복합문의 구성 형식이 단문의 형식으로 변환되는 원리를 이용하여 문장 내의 부착 문제를 단문내의 부착 문제로 축소한다. 예로 가)문장에서는 관형절 "학교에 가는 영희를"을 체언구인 목적어로 인식하면 "철수가 영희를 보았다"라는 단문 형식으로 변환되는 것이다. 이런 원리를 이용하여 구문적 특성에 의해 발생하는 구문 모호성을 해결한다.

2. 관련연구

지금까지 개발된 한국어 구문분석 시스템은 대부분 생성 가능한 모든 파스트리를 출력하므로 의미분석이나 기계번역과 같은 응용 분야에 적용하기 위해서는 파스트리를 제약하여 구문모호성을 해결할 필요가 있다. 구문 모호성을 해결하기 위한 연구로는 [3,4,5,6,8,9] 등이 있다. [3]과 [5]는 복합동사구나 본용언에 뒤따르는 보조용언을 하나의 구문단위로 처리함으로서 구문모호성을 축소하는 방안을 제안했다. 이는 한국어의 통사적 특성을 고려한 것으로 구문 분석의 부담을 덜고 구문 모호성을 사전에 방지할 수 있는 장점이 있다. 그러나 이들은 [6]이 제안한 구문 형태소의 일부분이며 본 논문에서는 구문형태소를 사용하여 구문 모호성을 해결한다.

[4]는 중심어 주도의 파싱 기법으로 한국어에서 중심어인 용언을 이용하여 개념단위나 최장 묶음으로 분리하여 구문 분석을 수행하였다. 말뭉치를 분석하여 미리 구축된 사전 정보를 이용하여 용언을 중심으로 최대한의 묶음 단위로 나누어서 한국어를 분석한다. 따라서 모호성의 발생 빈도가 감소되고 구문 분석의 효율을 기대할 수 있다. 그러나 많은 예문과 사전 정보가 필요하며 영어 문장을 생성하기 위한 구문 단위로 묶이기 때문에 "큰 나무 아래에서"의 경우 "큰 나무"로 묶이는 것이 아니라 "나무 아래에서"로 묶여 의미적으로 타당하지 못하다는 단점을 가진다.

[8]과 [9]는 동사 패턴과 명사의 의미표지를 이용하여 한국어 문장을 용언을 중심으로 하는 개념 단위로 분리했다. 즉, 내포문을 포함하는 문장에서 용언의 패턴과 명사의 의미표지를 이용하여 내포문을 단문으로 분할하는 원리로서 [10]과 [11]의 방법과 유사하다. 그러나 이들 논문에서는 복합 명사나 관형구의 처리가 미흡하고 복합동사의 처리가 이루어지지 않았다. [11]는 정보검색을 위해서 단문을 분할했고 [10]은 한국어 문장을 같은 의미를 가지는 여러 개의 단문으로 분할했다. 또한 다음과 같은 점에서 본 논문과 차이가 있다. 본 논문에서는 분할된 내포문이 문장 내에서 하나의 구문적 기능을 가진다는 것이다. 예를 들면 나)문장에서 명사절 "영희가 효녀임을"은 문장 전체에서 목적어의 기능을 한다. 또한 다)문장에서는 부사절 "소리도 없이"가 부사의 기능을 한다. 이와 같이 분할된 내포문이 하나의 구문적 기능을 하며 의미분석이나 기계번역을 위해서는 이러한 정보가 분석 결과에 반영되어야 한다는 점이다.

나) 철수는 [영희가 효녀임을] 안다. - 목적어
다) 철수가 [소리도 없이] 갔다. - 부사

또한 [8, 9, 10, 11]은 체언구를 중심으로 용언의 패턴 정보를 분류했기 때문에 부사가 필수적으로 필요한 경우에는 부사구 부착의 문제가 발생할 수 있다. 라)는 부사구 부착 문제가 발생한 예로 부사 "예쁘게"는 용언 "생긴"과 "보았다"에 부착이 가능하다. 그러나 본 논문에서는 부사구도 고려한 문형 정보를 사용하여 이런 부사구

부착 문제를 해결한다.

라) 철수가 [예쁘게 생긴 영희를] 보았다.
'생기다'의 문형 : N이 ADV-게 생기다.

3. 한국어의 특성

3.1 문형

한국어는 부분 자유 어순을 가지며 용언에 따라 다양한 격조사와 체언을 요구한다. 따라서 문장의 구조를 파악하기 위해서 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 발생한다[12]. 예를 들어 다음의 문장을 살펴보자.

마) 철수가 귀찮게 군다.　　마*) 철수가 군다.*
바) 철수가 영희와 싸운다.　　바*) 철수가 싸운다.*

마)와 바)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 마*)와 바*)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. 대표적인 예로는 격틀과 문형이 있는데 격틀은 용언에 대한 정확한 의미지식을 요구하기 때문에 정확한 구문분석은 가능하지만 구축이 어려운 실정이다. 반면에 문형은 순수한 구문적 정보만을 이용하며 약간의 의미적 제약을 가할 수 있기 때문에 본 논문에서는 문형을 이용하여 구문분석을 수행한다.

일반적으로 동사는 사물의 움직임을 나타내고 형용사는 사물의 상태를 나타내기 때문에 동사와 형용사를 구분하여 문형을 설정하였다. KAIST-Corpus와 연세 한국어 사전[13]을 활용하여 동사를 31가지 문형으로 하고 형용사는 8가지 문형으로 설정하였으며 [표 1]은 그 일부이다. 또한 13,175개의 용언에 대해서 문형으로 분류하였고 하나의 용언은 평균 1.33개의 문형을 가졌다. [표 2]는 상위빈도를 가지는 문형의 일부이다.

V1) N(이/는/은/가)+V	A1) N(이)+A
V2) N(이)+N(에/에게)+V	A2) N(이)+N(에)+A
V3) N(이)+N(와/과) + V	A3) N(이)+N(와)+A
:	:
V29) N(이)+S(기/움)+V	A6) N(이)+N(로)+A
V30) N(이)+S(기로)+V	A7) N1(이)+N(로)+N2(이)+A
V31) N(이)+S(기로)+N(에게/와)+V	A8) N1(이)+N(와)+N2(이)+A

[표 1] 문형의 일부

동사문형	빈도	형용사문형	빈도
V1	41.29%	A1	18.28%
V11	34.49%	A5	2.19%
V2	13.96%	A2	1.9%
V12	5.98%		

[표 2] 상위 빈도의 문형정보

3.2 단문분할을 위한 한국어의 구조적 특성

중문은 단문과 단문이 대등적 연결어미나 종속적 연결어미 등에 의해 연결된 문으로 내포문의 범위는 연결어미를 가지는 용언을 만날 때까지로 하면 된다. 그러나 주절과 종속절로 구성되는 복문은 종속절의 범위를 어떻게 결정하느냐에 따라 체언구나 부사구 부착의 문제가 발생된다. 종속절은 주절과의 관계와 기능에 따라 명사절, 관형절, 부사절로 대별된다[14]. 16,195 문장으로 구성된 KAIST-Corpus를 분석해 본 결과 중문과 복문의 구성 비율[4]이 [표 3]과 같았다.

문장 수	중문	관형절	명사절	부사절	단문 수
16,195	12,334	10,661	1,912	2,013	1,551
100 %	76.2 %	65.9 %	11.8 %	12.4 %	9.6 %

[표 3] 문장의 구성 비율

3.2.1 명사절

3.2.1.1 인용 명사절

내포문에 인용격 조사 '-고/-라고'가 붙어 인용명사절을 이룬다. 내포문은 완전한 문장의 형식을 갖고 평서문,

4) 한 문장에서 종속절이 여러 개인 경우에는 각각 계산함
예) 밥을 먹은 철수를 보기가 쉽다 -- 관형절, 명사절

의문문, 명령문, 청유문을 모두 포함한다. 내포문의 범위를 정할 때 명령문과 청유문은 문장의 주어가 생략되는 특징을 이용한다. 또한, 인용 명사절은 다음의 예문에서 보는 바와 같이 체언구의 기능보다는 부사어에 속하지만 용언에 따라서 인용명사절을 이룰 수도 있고 아닐 수도 있다. 따라서 본 논문에서는 새로운 문형 정보(N이 S고 V)[5]로 분류하여 처리한다.

- 우리는 철수가 밥을 먹었다고 믿는다.
- 우리는 철수가 밥을 먹었느냐고 물었다.
- 우리는 철수에게 밥을 먹으라고 명령했다.
- 우리는 철수에게 밥을 먹자고 제안했다.

3.2.1.2 'ㅁ/기' 명사절

용언에 명사화 접미사 'ㅁ/기'가 붙어 형성된다. 이 명사절은 용언을 명사 의미로 전성하여 체언구의 기능을 갖게 한다. 따라서 용언이라는 정보를 먼저 이용하여 문형 정보에 의한 단문 분할을 수행한 후에 분할된 단문에 명사화 접미사가 붙어서 체언구의 기능을 가지도록 처리한다.

- 영어를 배우기가 어렵다. <-> S+[-ㅁ/-기]

3.2.1.3 '지/냐' 명사절

문미에 '-지/-냐'가 붙어 명사절을 이룬다. 내포문은 완전한 문장을 이루며 체언구의 기능을 한다.

- 철수가 학교에 가느냐가 문제이다.
- 철수가 밥을 먹었는지 묻다.

3.2.2 관형절

관형사형 어미(-는, -은, -을, -ㄹ, -ㄴ, -던)가 이끄는 절로 그 뒤에 따르는 명사구를 수식하는 수식절이다. 관형절에 후행하는 명사가 관형절의 한 성분인 경우 관계 관형절이라 하며 동격인 경우는 동격 관형절이라 한다. 또한 관형절의 범위가 관형형 어미를 가지는 용언까지이며 관형사의 구실을 한다[14, 15]. 그러나 본 논문[6]에서는 관형절에 후행하는 체언구의 역할에 의해 관계, 동격, 의존 관형절로 분류하며 관형절의 범위도 후행하는 체언구까지로 한다. 따라서 관형절은 관형사의 역할이 아니라 체언구로서의 기능을 가진다.

3.2.2.1 관계 관형절

내포문(안긴 문장)의 필수 성분 중 하나가 탈락되어 있어 내포문에 후행하는 체언구를 수식하는 절이다. 체언구 "영희가"가 용언 "먹다"의 필수 성분이 된다.

- 철수가 밥을 먹은 영희를 보았다.

3.2.2.2 동격 관형절

내포문이 모든 성분을 완전하게 갖추고 있다. 관형절이 수식하는 명사는 '사실, 소식, 보도, 사건, 냄새, 소문, 결심, ...'등의 보문 명사이다[15].

- 철수가 영희가 집에 간 사실을 안다.

3.2.2.3 의존 관형절

동격 관형절과 마찬가지로 내포문이 모든 성분을 완전하게 갖추고 있지만 관형형 어미 뒤에 의존 명사인 '것/줄/수/데' 등이 붙어 이루어진 관형절로 국어학에서는 명사절로 취급한다[14,15]. 의존 관형절은 추상적인 '사실'의 의미로 관형절을 받아 명사절을 형성하여 주절의 주어나 목적어 기능을 한다.

- 지구가 둥글다는 것은 오래 전에 증명되었다.

3.2.3 부사절

용언의 부사성 활용어미는 부사절을 형성하여 주절의 서술 용언을 수식하는 부사어 역할을 한다[15]. 의미에 따른 부사성 활용어미는 [표 5]와 같다.

- 바람이 소리가 없이 분다.
- 배가 고파서 나는 밥을 먹겠다.

---

5) N은 체언구, S는 인용명사절이 가지는 완전한 문장, V는 용언

6) 일반 국어학에서는 의존 관형절을 '것' 명사절로 취급하지만 본 논문에서는 관형사형 어미에 의해 생성되는 모든 절을 관형절의 범주에 포함한다. 또한 뒤에 따르는 체언구가 관형절의 한 성분이냐 아니냐에 따라 관형절의 범위가 결정되므로 관형절의 범위를 관형절에 후행하는 체언구까지로 한다. 따라서 관형절은 관형사가 아니라 체언구의 기능을 가지게 되며 명사절과 같은 원리로 취급된다.

이유/원인	: -어서, -니까, -므로
조건/가정	: -(으)면, -(어)도, -거든, -(이)라면
양보	: -더라도, -(으)ㄹ망정, -(으)ㄹ지언정
목적	: -(으)러, -(으)려고, -고자
전이	: -다(가)
첨가	: -(으)ㄹ수록, -(으)ㄹ뿐더러
유도	: -게, -도록

[표 4]부사성 활용어미의 분류

4. 구문 모호성 해결

4.1 단문 분할 알고리즘

내포문을 포함하는 문장에서 내포문의 범위를 정해서 단문으로 분할하는 과정은 두 단계로 구성된다. 첫 단계는 필수격 결합 단계로 문형이 가지는 필수격들을 최대로 만족할 때까지 체언구나 부사구들을 결합하여 하나의 단위로 한다. 두 번째 단계는 필수격이 모두 채워진 이후의 과정으로 필수격의 충돌이 발생할 때까지 남아있는 체언이나 부사구들을 용언에 부착한다. 이는 지역적으로 최대의 범위를 가지는 단문으로 분할을 하기 위한 것이지만 내포문에 부착된 체언구가 본용언의 필수격으로 사용되는 경우가 있기 때문에 분리해서 처리한다. 이와 같이 본 논문에서는 문형을 제약조건으로 하는 단문분할을 통해 복합문의 형태를 단문으로 변환하여 구문분석을 수행하는 새로운 알고리즘을 제안한다. 본 논문에서 제안하는 단문 분할 알고리즘은 다음과 같다.

1) 용언을 만날 때까지 좌에서 우로 분석을 수행한다.
2) 용언을 만나면 관형형 어미를 가지는지 검사한다.
3) 관형형 어미이면 다음 체언구를 입력한다
4) 관형절의 유형을 파악한다
5) 문형 정보의 필수격을 모두 만족할 때까지 기입력된 정보를 용언과 단일화한다.
   - 필수격을 모두 채운 시점까지를 단문으로 분할
   - 필수격의 충돌이 발생할 때까지 남아있는 정보를 용언과 단일화한다
6) 필수격의 충돌이 발생하면 그 때까지를 내포문의 범위로 한다.

4.2 단문 분할

내포문을 포함하는 문장에서 단문분할 알고리즘을 적용하여 복문의 구조가 단문 형식으로 변환되는 과정을 관형절을 예로 살펴본다. 기술의 편의를 위해 체언구는 'NP', 동사는 'VP', 부사는 'ADV'로 표현하며 입력 순서에 따라 번호를 할당해서 사용한다.

NP1 ADV NP2 NP3 VP1 NP4 VP2

마) 철수가 자주 **영수와 학교에서 싸우는 영희**를 보았다.

문형) 싸우다 : N이 N와 V

보다 : N이 N을 V

바) 철수가 거울로 **영수와 학교에서 싸우는 영희**를 본다.

(a) NP1
(b) NP1 ADV
(c) NP1 ADV NP2
(d) NP1 ADV NP2 NP3
(e) NP1 ADV NP2 NP3 VP1
(f) NP1 ADV NP2 NP3 VP1 NP4
(g) NP1 ADV NP2 NP3 **[VP1 NP4]**
(h) NP1 ADV NP2 **[NP3 VP1 NP4]**
(i) NP1 ADV **[NP2 NP3 VP1 NP4]**
(j) NP1 **[ADV NP2 NP3 VP1 NP4]**
(k) NP1 **[ADV NP2 NP3 VP1 NP4]** VP2
(l) NP1 **[[ADV NP2 NP3 VP1 NP4]** VP2 **]**
(m) **[** NP1 **[ADV NP2 NP3 VP1 NP4]** VP2 **]**

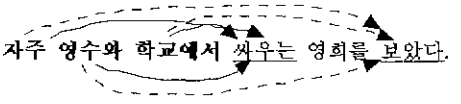
위의 (a)에서 (m)은 마)문장을 구문 분석하는 과정이다. (e)에서는 용언(VP1)이 입력되었어도 관형형 어미를 가지므로 앞의 체언구와 결합하지 않고 다음의 체언구(NP4)를 입력한다. (g)에서 (i)까지는 관형절의 문형 정보를 이용하여 필수격을 채우는 단계이다. (j)에서 보는 바와 같이 문형의 필수격을 모두 채웠어도 필수격의 충돌이 발생하지 않으면 계속해서 가까운 용언에 부착해 간다. 결국에는 필수격의 중복이 발생하기 전까지 최대의 범위를 가지는 관형절을 하나의 단문으로 분할하는 것이다. 그러나 이런 방법은 문제를 가질 수가 있다. 마)의 경우처럼 부사가 사용된 경우는 가까운 용언에 부착이 가능하지만 바)의 경우와 같이 체언구인 경우에는 두 가지 해석이 모두 가능하기 때문이다. 예문 바)에서는 체언구 '거울로'가 용언 '보다'에 부착되어야 의미적으로 타당하다. 그러나 본 논문에서 제안한 방법으로는 '싸우다'에 부착[7]된다. 따라서 이런 문제를 해결하기 위해 필수격을 모두 채운 시점까지를 하나의 단문으로 분할을 하

7) '보다'의 문형을 N이 N을 V로 제한한 경우이며 실제로는 N이 N로 N을 V라는 문형에 의해 '망원경으로'는 '보다'와 결합한다.

고 높은 가중치를 할당한다. 문형의 모든 필수격이 채워지고 난 이후의 단일화에서는 낮은 점수를 할당한다. 파스트리는 가장 높은 가중치를 가지는 결과만을 출력하고 나머지 정보는 그대로 보관한다. 이렇게 함으로써 "철수가 서울로 영희를 본다"는 의미를 가지는 파스트리도 만들어지므로 구문 모호성을 가지게 되지만 의미적으로 올바르지 못한 결과에서 타당한 결과를 추출하는데 이용할 수가 있다.

4.3 문형과 단문분할을 이용한 구문 모호성 해결

문형정보와 단문분할 원리는 한국어 문장에서 부사구 부착이나 체언구 부착에 의한 부착의 문제를 해결하는 제약 조건으로 사용할 수가 있다. 예를 들어 마)의 문장은 부사구 '자주'나 체언구 '영수와', '학교에서'가 어느 용언에 부착하느냐에 따라 많은 구문 모호성을 가지지만 문형과 단문분할을 사용하면 이런 구문 모호성을 해결할 수가 있다. 본 논문에서의 결과는 4)이다.



마) 철수가 **자주 영수와 학교에서** 싸우는 영희를 보았다.

문형) 싸우다 : N이 N와 싸우다
보 다 : N이 N을 보다

1) 철수가 자주 영수와 학교에서 [싸우는 **영희를**] 보았다.
2) 철수가 자주 영수와 [학교에서 싸우는 영희를] 보았다.
3) 철수가 자주 [영수와 학교에서 싸우는 영희를] 보았다.
4) **철수가 [자주 영수와 학교에서 싸우는 영희를] 보았다.**

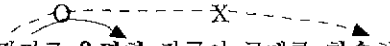
또한 문형을 이용하면 용언이 필수적으로 요구하는 부사나 필수격의 처리가 가능하다. 예문 사)에서는 '군다'가 가지는 문형 "N이 ADV[-게] V"를 이용하고 예문 자)에서는 '싸우다'가 가지는 문형 "N이 N와 V"에 의해 비문을 가려낼 수가 있다. 이와 같이 구문적으로는 바르지만 의미적으로 비문인 구조를 쉽게 처리할 수가 있다.

사) 철수가 성가시게 군다. 사*) 철수가 군다.*
문형) N이 ADV[-게] V
자) 철수가 영수와 싸운다. 자*) 철수가 싸운다.*
문형) N이 N와 V

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

차) 철수가 돈이 모자란다.
문형) N1이 N2이 V
카) 어머니가 철수를 아침을 굶겼다.
문형) N1이 N2을 N3을 V

그러나 한국어는 구문구조만으로는 해결할 수 없는 모호성이 존재한다. 예를 들어 다음의 예문은 문형 정보만을 이용하면 '아동작가로'는 '유명하다' 뿐만 아니라 '철수하다'와도 결합할 수 있다. 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.



타) *아동작가로* **유명한** 장군이 군대를 **철수하였다.**

문형) 유명하다 <-> N이 N로 V
[직업-신분, 구체물]
철수하다 <-> N이 N로 N을 V
[구체적 장소]

또한 체언구의 역할은 조사에 의해서 대부분 결정되지만 보조사인 경우에는 어떤 역할을 하는지 결정하기가 어렵다. 파)에서 '철수'와 '밥'의 의미지표는 '사람'과 '음식'이므로 '먹다'의 문형에 의해 '철수'가 주어로 '밥'이 목적어로 결정될 수가 있다. 또한, 하)와 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.

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

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

본 논문에서 제안하는 방법은 관형절이나 부사절, 명

사절과 같은 내포문을 포함하는 문장에서 문형정보를 이용하여 내포문을 단문으로 분할하고 그 정보를 이용해서 구문 모호성을 해결하는 것이다. 정문을 위주로 실험을 했기 때문에 이와 같이 의미파악에 의해 구문 모호성을 야기하는 경우는 적었다. 그러나 생략이 빈번하고 도치가 많은 한국어를 의미적으로 정확하게 구문분석하기 위해서는 구문분석 과정에 의미지표를 도입해야 할 필요성이 있다. 따라서 이에 대한 연구는 차후에 진행하고자 한다.

4.4 실험 및 분석

한국어를 구문분석하기 위해서 조건단일화 기반의 CFG를 이용하였다[16, 17]. 문장을 분석하기 위해서 동사, 형용사, 서술격동사로 용언을 세분하여 문법 규칙을 기술했다. 또한 관형형 어미를 가지는 용언은 후행하는 체언구와 먼저 단일화 연산을 수행하도록 하기 위해서 일반 용언과 분류하여 기술되었다. 이렇게 기술된 문법 규칙은 총 79개이다. 일반적인 구구조 기법보다 문법규칙의 수가 적은 것은 본 논문에서는 어절 단위의 분석과 구문 형태소, 자질 정보를 이용하고 문법 규칙을 완전이진 문법으로만 기술했기 때문이다. 본 논문에서 사용하는 문법 규칙의 일부는 다음과 같다.

```
<ANP> <==> <MA> <ANP>   ;;<메우> <예쁜 철수>
<ANP> <==> <NP> <ANP>   ;;<마음이> <예쁜 철수>
<ANP> <==> <PAM> <NP>   ;;<예쁜> <철수>
<MAS> <==> <SV>         ;;<밥을 먹어서>
<NP> <==> <NP> <NP>     ;;<철수의> <생각>
<NP> <==> <SV>          ;;<철수가 밥을 먹기로>
<NP> <==> <VNP>         ;;<밥을 먹은 철수가>
```

[표 5] 완전이진 CFG를 이용한 문법규칙의 예

실험을 위해 KAIST-Corpus에서는 단문분할이 가능한 10어절 이내의 232문장을 추출하고 초등학교 사회교과서에서는 단문을 비롯하여 평균 2.23어절로 이루어진 200문장을 추출하였다. 추출된 문장을 이용하여 문형 정보와 최소거리 원리를 사용하지 않고 결합 가능한 모든 후보를 생성할 수 있도록 구문 분석한 결과(실험1) 평균 56.79개의 파스트리가 생성되었다. 그러나 본 논문에서 제안한 단문분할 방법(실험2)을 이용하여 구문 분석한 결과 파스 트리의 수는 평균 6.97개였다.

실험 문장	평균 용언 수	실험1	실험2
KAIST- Corpus	3.67	66.2	6.88
사회교과서	2.23	47.37	7.06
평 균	2.95	56.79	6.97

[표 6] 문형과 단문분할에 따른 구문모호성의 수

실험 결과를 분석해 보면 비교적 정문에 따르는 산문이나 교과서보다는 소설에서 발췌한 문장에서 오류가 많았다. 이는 소설의 경우 어순의 도치와 생략이 심하고 "철수의 손잡이가 달린 가방"과 같이 의미정보를 요하는 문장이 많았기 때문이었다. 또한 사회교과서에서는 명사들의 나열이나 보조사가 많이 사용되었기 때문에 KAIST-Corpus보다 모호성이 많았다. 명사들의 나열인 경우에도 명사구의 범위에 의해 많은 모호성이 발생하였고 보조사에 의한 불확실한 격정보 때문에도 많은 구문 모호성이 발생하였다. 그러나 본 논문에서 제안한 방법으로 구문 분석을 수행하면 구문모호성의 수가 12.27%로 줄어들기 때문에 자연어 처리의 많은 응용 시스템에 보다 효율적으로 적용할 수 있다.

5. 결론

한국어의 구문분석에서 발생되는 구문 모호성의 원인은 여러 가지가 있다. 그 중에서 한국어의 구문적 특성에 의해 발생되는 체언구 부착이나 부사구 부착에 의한 부착의 문제를 해결하기 위해서 한국어 문장 구성 유형을 분류하고 문형을 설정했으며 단문 분할 원리를 도입하였다. 또한 내포문을 포함하는 문장에서는 내포문의 범위를 정하여 단문으로 분할해서 내포문을 하나의 구문적 범주로 취급하는 것이 부착의 문제를 해결하는데 좋은 제약 조건이 될 수 있음을 보였다.

특히 구문 구조 규칙만으로 정확한 파싱을 한다는 것은 거의 불가능하며 의미의 결합이 반드시 필요하다. 그러나 의미의 반영이 정확하고 충분하게 이루어지지 않으면 오히려 시스템의 성능이 떨어질 가능성이 있다. 따라서 본 논문에서는 구문-의미의 해체 분석을 하였으며 차후에 완성된 의미 처리를 받아들이고자 한다. 따라서 향후 연구과제로는 지금까지 개발된 문형규칙이 한국어의 여러 현상을 처리할 수 있도록 개선하고 의미지식을 활용할 수 있는 방안에 대한 연구가 필요하다.

[참고문헌]

[1] 이희자, "현대 국어 관용구의 결합 관계 고찰", 제6회 한글 및 한국어 정보처리 학술대회, pp. 333-352, 1994.

[2] 강승식, 음절 정보와 복수어 단위 정보를 이용한 한국어 형태소 분석, 서울대학교 박사학위 논문, 1993.

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

[4] 박상규, 정창민, 조준모, 이상조, "최장 묶음을 이용한 효과적인 한국어 구문 분석기", 정보과학회 봄 학술 발표논문집, pp.961-964, 1995.

[5] 안미정, 옥철영, "한국어 구문구조 분석을 위한 복수 동사 처리", 정보과학회 가을 학술 발표논문집, pp. 625-628, 1994.

[6] 황이규, 이현영, 이용석, "형태소 및 구문 모호성 축소를 위한 구문 단위 형태소의 이용", 한국정보과학회 논문지, Vol. 27, No. 7, pp. 784-793, 2000.

[7] 강은국, 조선어 문형 연구, 박이정출판사, 1996.

[8] 인현철, 박광철, 양세라, 이정현, "용언을 중심으로 한 개념 단위 분리기", 정보과학회 가을 학술 발표논문집, pp.609-612, 1995.

[9] 김광진, 송영훈, 이정현, "한국어 내포문을 단문으로 분리하는 시스템의 구현", 제5회 한글 및 한국어 정보처리 학술대회, pp. 333-352, 1993.

[10] 박현재, 이수선, 우요섭, "의미 정보를 이용한 이단계 단문분할 알고리즘", 제 11회 한글 및 한국어 정보처리 학술대회, pp. 237-241, 1999.

[11] 이현아, 이종혁, 이근배, "단문 분할을 통한 명사구 색인 방법". 한국 정보과학회 논문지(B), Vol. 24, No. 3, pp. 302-311, 1997.

[12] 한용가, 황이규, 이용석, "문형정보를 이용한 한국어 구문 분석", 제7회 한글 및 한국어 정보처리 학술대회, pp. 23-29, 1995.

[13] 연세대학교 언어정보개발원, 연세한국어 사전, 두산동아, 1999.

[14] 장석진, 정보기반 한국어 문법, 도서출판 언어와 정보, 1993.

[15] 남기심, 고영근, 표준국어문법론, 탑출판사, 1983.

[16] Kong Joo Lee, Jae-Hoon Kim, and Gil Chang Kim, "An Efficient Parsing of Korean Sentence Using Restricted Phrase Structure Grammar", Computer Processing of Oriental Languages, Vol. 12, No. 1, pp. 49-62, 1997.

[17] 양승원, 박영진, 이용석, "조건 단일화 기반 PATRII 를 이용한 한국어 구문 분석", 한국정보과학회 논문지 Vol. 22, No. 4, pp. 653-662, 1995.