새로운 어절 해석에 기반한 한국어 의존관계 파서

강호관, 이종혁, 이근배
포항공과대학교 전자계산학과

Korean Dependency Parser Based on New Interpretation of Eojeol

HoGwan Kang, Jong-Hyeok Lee, Geunbae Lee
Dept. of Computer Science and Engineering, POSTECH

요 약

본 논문에서는 기계번역과 의미분석의 전단계로서의 구문분석에 대하여 논한다. 의존문법에 기반을 둔 구문분석의 효율성을 위하여 한국어 어절에 대한 새로운 해석을 시도하며, 이를 기반으로 한국어 의존관계 파서의 새로운 기본 단위(SynN: Syntactic Node)를 제시한다. 또한 새로운 기본 단위를 구문분석 과정에 적용하는 방법과 그 결과를 보인다. 마지막으로, 구현된 구문분석기를 중간언어 방식 시스템인 한-중 기계번역 시스템에 채용하여 그 성능을 검증한다.

1. 서 론

기계번역은 해당 언어에 대한 분석의 정도에 따라서 대략적으로 직접방식, 변환방식, 중간언어방식 등으로 나뉜다[8]. 한국어와 일본어의 경우와 같이 그 언어적 유사성이 매우 큰 경우에 선호되는 직접방식의 경우를 제외한다면 원시 언어에 대한 구문분석은 필수적이다.

그러나 이러한 기계번역의 전단계로서의 구문분석 결과는 범용의 구문분석 결과와는 매우 다른 요구 조건을 충족시켜야만 된다. 대표적으로 선어말어미들에 대한 처리를 예로 들 수 있다. 구문분석에서 선어말어미는 사실상 별다른 의미를 가지지 못한다. 선어말어미는 다분히 의미를 중점으로 두는 상황에서 가치를 가진다. 즉, 그에 대한 처리가 제대로 이루어지지 않을 경우 구문분석 단계에서는 아무런 문제를 일으키지 않겠지만, 해당 정보들(시제, 존경, 겸양 등)이 누락됨으로써 의미분석이나 혹은 목표 언어로의 변환 과정에서 그 정보들을 얻기 위한 원시 문장의 재분석이 요구된다.

이와 같은 과정을 거치는 것은 매우 비효율적이며, 또한 독립적인 구문분석 단계의 존재 의미를 희석시킨다. 또한 이후에 얻어질 수 있는 구문분석 단계에서의 성능향상이 의미분석 등의 단계로 제대로 전파되기 힘들게 하는 것이기도 하다.

따라서 본 논문에서는 의미분석과 기계번역 등을 위한 효율적인 한국어 구문분석기에 대하여 논하고자 한다. 이를 위하여 2 장에서는 기존의 연구들이 가졌던 한계점들을 검토하고, 3 장에서는 구문분석 과정을 실제적인 예를 통하여 설명하며, 4 장에서는 구문분석기를 이용한 실험 결과를 분석하고, 마지막으로 5 장에서 결론을 내리고자 한다. 특히, 2.1 장에서는 어절에 대한 새로운 해석을 내리고, 이를 기반으로 하여 의존문법을 재정리한다. 또한 4 장에서 설명될 결과에는 본 구문분석기를 의미분석에 기반한 한-중 기계번역 시스템에 채용하여 그 성능을 검증한 바가 포함되어 있다.

2. 한국어 어절에 대한 고찰

2.1. 한국어 어절과 의존문법의 구문단위

한국어는 그 어순의 자유로움 때문에 구절구조규칙(PS rule)에 기반하는 구절구조문법(PSG) 등으로는 기술할 수 없는 표현들이 많다. 이러한 한계를 극복하는 의존문법에서는 단순히 두 단어들 간의 관계성(relation)만을 기술할 뿐, 그 관계성이 설정된 단어쌍들에 대한

강제적인 구절화 연산(grouping operation)들이 없다. 때문에 의존문법은 어순이 자유로운 언어에 매우 적합하다[7].

그러나 의존문법에 대해 꾸준히 거론되어 온 문제점들도 분명히 존재하며, 대표적인 예로서 지나치게 많은 의존관계의 형성을 들 수가 있다.

이러한 현상은 기본적으로 두 단어 사이의 관계성들에 기반하여 문장의 구조를 파악하고자 하는 데에서 온 것[3]이라는 것이 지금까지의 일반적인 생각이다.

그러나, 이러한 지나치게 많은 의존관계의 형성은 단순히 의존문법의 한계 때문만이 아니다. 적어도 지금까지 제시되어 왔던 방법론들에서 언급된 정도의 복잡도에 대해서는 의존문법 자체보다는, 의존문법을 한국어에 적용할 때 한국어에 적합한 의존문법에서 구문단위(word)를 제대로 정립하지 못한 것이 더 큰 이유이다.

일반적으로 한국어 분석시에는 의존문법의 구문단위인 단어를 어절에 대응시키고 있다. 하지만, 이러한 대응은 한국어의 특징을 충분히 고려하지 않은 것이다.

한국어에서는 띄어쓰기 단위의 어절로는 여러 개이지만 그들 간의 관계성들은 고정되어 있는 용례가 많다. '~ㄹ 수 있~', '~지 않~' 등이 그것이다. 이들 경우에 각각의 어절을 하나의 독립된 구문단위로 간주하는 것은 매우 비효율적이다. '~에 대하~'의 경우와 같은 불완전동사(일명 불구동사)와 조사의 결합의 경우도 마찬가지이다. 따라서 이러한 어절들은 하나의 단위로 다루어져야 할 것이며, 그런 작업은 이미 시도되어 그 성과를 입증하였다[3]. 하지만, [3]의 경우에는 의존문법을 기반으로 한 것이 아니었으며, 무엇보다도 어절들을 어떻게 하나의 단위로 묶어 나가는가에 대한 체계적인 기술이 없었다.

앞의 경우와는 반대로, 한국어에서는 하나의 어절이 상이한 특질을 모두 가지거나, 2 가지의 자릿수정보들(valency information)을 가지는 현상 등도 존재한다. 즉, '명사+~이다/답다'의 경우나, '명사형+~이다' 등의 경우들이 그것이다. 일단 '명사+~이다/답다'의 경우에는 "좋은 학생이다" 등과 같이 용언이 수식어로 관형어를 취하는 형태가 된다. 이는 '학생이다'라는 어절이 체언(학생)으로서의 특성과 용언(~이다)으로서의 특성을 모두 가짐으로써 생기는 현상이다. 또한 '명사형+~이다'의 경우는 "문제는 네가 그것을 했음이다"에서 볼 수 있듯, '했음이다'라는 어절이 2 개의 주어('문제는', '네가')와 하나의 목적어를 취하게 되지만, 이는 이중주어의 경우가 아니다. '하다'(타동사)에 의한 주어와 목적어, 그리고 '이다'에 의한 주어 등이 복합된 것이다.

이러한 문제점들을 체계적으로 해결하기 위하여 의존문법을 한국어에 적용함에 있어서 올바른 구문단위가 가져야 할 특징들은 다음과 같다. 첫째, 하나의 자릿수 정보만을 가져야 한다. 이를 위하여 '명사형+~이다' 등의 경우에 있어서는 '명사형'과 '이다'를 독립적인 단위로 분리하여 고려한다. 둘째, 선행하는 어절들과 후행하는 어절들에 대하여 각각 하나의 특성만을 가져야 한다. 따라서 '명사+~이다'의 경우에는 '명사'와 '~이다'를 각각 독립적인 단위로 분리한다. 이러한 단일 어절에서의 분리는 언뜻 형성되는 의존관계의 수를 증가시키는 것처럼 보이지만, 그 단위들 간에 독점적 관계(현재는 보어)를 확정시켜 줌으로 해서 그러한 문제들을 방지할 수 있다. 이와 같은 맥락에서 '~ㄹ 수 있~'의 경우에도 이들 사이에 독점적 관계를 설정하여 하나의 단위로 고려할 수 있게 되는 것이다. 셋째, 구문단위 내부의 성분들에 의하여 그 구문단위 외부의 성분과의 의존관계가 영향을 받아서는 안 된다. 이는 앞의 2 가지 특징들이 보장되면 당연히 성립하게 되는 특징이다. 본 논문에서는 이 단위를 Syntactic Node(이하 SynN)라 정의하겠으며, 이와 같은 SynN 를 구문단위(word)로 대응시켜 이들 간의 의존관계를 기술하는 것으로 의존문법을 정의하겠다.

2.2. 어절구조

SynN 는 관련 어절들이 가지는 정보들로부터 초기화된다. 따라서, 효율적인 초기화를 위해서는 어절들이 가지는 일반적인 구조를 파악하는 것이 중요하다. 이는 SynN 가 결국 문장에서 나타나는 어절들을 보다 정규화된 형태로 재정의한 것이므로, 그 구조가 매우 유사할 수밖에 없기 때문이다.

일반적인 어절들의 구조, 즉 어절구조는 이미 형태소분석 단계에서 사용되고 있다. 하지만 이 경우에는 좌우접속, 즉 인접한 성분들 사이에서만 의미를 가지는 것으로, 어절 전체의 순서 정보와는 상관이 없는 것이다.

그러나 어절내부에서 각 성분들의 순서가 의미를 가짐은 이미 알려진 바이다. [9]에서 밝혀 둔 바대로 선어말어미 '-겠-'(미래/추정)의 경우는 나타나는 위치에 따라서 그 기능이 달라지게 된다. 이를 단순히 구문적인 수준에서만 본다면 전혀 고려할 사항이 되질 않는다. 의존문법에서 구문구조를 파악할 때 관심의 초점은 자릿수정보(valency information)와 문형정보 등이기 때문이다[10].

본 논문에서 제시하는 어절구조는 자릿수정보의 효과적인 활용은 물론, 각 성분들이 가지는 의미적 성향

$$\begin{bmatrix}\text{체언+조용보조어간}\\(\text{보조})\text{용언어간}\end{bmatrix}+(-ㅁ)+\text{뿐}+\text{만}+\text{이}+\begin{bmatrix}\text{보조적연결어미(부사전성어미포함)}\\\text{선어말어미}+\begin{bmatrix}\text{어말어미}\\\text{전성어미}\end{bmatrix}+\text{격조사}\end{bmatrix}+\text{보조사}$$

그림 1. 어절구조

들에 대한 정보를 의미분석 단계로 온전히 전달하는 것에 대해서도 충분히 고려하여 설정하였다.

일단 한국어 어절들은 대부분 [그림 1]과 같은 기본 골격을 가진다. 참고로 [그림 1]은 순서적 정보를 중심으로 기술되어 있다. 물론 이러한 구조를 벗어나는 어절들도 형성할 수는 있으나, 그러한 어절들은 그 출현빈도와 어절구조 기술의 편이성을 고려할 때 무시할 만하다.

이러한 어절구조를 설정함으로써 주어진 어절에서 구문분석 단계뿐만 아니라, 이후의 의미분석 단계에서 필요로 하는 정보들을 효율적으로 추출할 수 있게 된다.

일단 구문분석 단계에서 사용되는 정보들로는 실질어휘의 구문분류와 기능어휘의 문법적 특성들이다. 실질어휘의 구문분류는 그 어절이 가지게 되는 자릿수 정보와 문형에 대한 정보를 주며, 기능어휘의 문법적 특성은 문장에서의 순서상 뒤에 위치하게 되는 어절들과의 관계에 대한 정보를 준다. 즉, 하나의 어절이 가지는 양방향의 관계들에 대한 정보들을 나타내게 되는 것으로, 특히 앞에 나타나는 어절들에 대해서는 자릿수 정보를, 뒤에 나타나는 어절들에 대해서는 조사, 어미 등과 같은 기능어휘와의 어울림에 대한 정보를 나타내게 되는 것이다. 물론 관형절 등의 경우에 있어서는 자릿수 정보가 해당 관형절의 수식을 받는 어절에도 적용된다는 예외는 존재한다.

의미분석 단계에서 필요로 하는 정보들로는, 선어말어미로부터 얻을 수 있는 시제, 존경, 겸양, 회상, 추측 등의 정보와 보조용언들을 처리함으로써 얻을 수 있는 피동, 사동, 강조, 봉사 등의 정보 등이 있다. 보조용언의 처리과정은 이후에 설명될 것이며, 이 때 얻어지는 정보에 대한 결정은 구문분석 단계가 아니라, 의미분석 단계에서 이루어 진다. '~ㄹ 수 있~'의 경우를 예로 들어 보자. "해결할 수가 있다"라는 문장에서 '~ㄹ 수 있~'의 구문적인 연관성들은 고정되어 있지만, 의미적으로는 '가능'을 나타낼 수도 있지만, 실제적으로 '해결의 방안이 있다'는 뜻이 될 수도 있으며, 이에 대한 선택은 구문분석에서 할 수 있는 것이 아니다. 따라서, 이러한 경우에는 '~ㄹ 수 있~'의 관계성이 존재한다는 정보만을 의미분석 단계로 전달하게 된다.

3. 구문분석 과정에서의 어절구조 정보 적용 방법

본 단락에서 설명하고자 하는 구문분석기는 기본적으로는 지금까지 본 연구실에서 계속적으로 연구되어 온 2-단계 의존관계 파서(TFDP-K: Two-Phase Dependency Parser for Korean)[11]를 기반으로 하며, 새로운 기본 단위인 SynN 를 기반으로 술어 중심 제약전파[5]와 지역 의존관계[6]를 효율적으로 적용할 수 있도록 구현되었다. 따라서 기본적인 분석과정은 크게 변화하지 않았으며, 관련 어절들로부터 SynN 를 초기화하는 과정과 그 SynN 들에 대한 정규화 과정(merge 와 split 을 통한)이 첫번째 단계의 맨 처음 부분으로 첨가되었다.

"영희가 먹지 않았을 수도 있다"의 경우를 예로 들어 수행 과정을 설명을 하면, 문장에 속한 각각의 어절들로부터 형성된 SynN 들은 [그림 2]와 같다. 이러한 SynN 들의 list 에 대해서 오른쪽부터 고려를 시작한다. 일단 맨 오른쪽의 어절과 그 앞의 어절을 지배소와 의존소로 생각하며, 이 때 지배소의 실질어휘가 '있'이며 보조용언의 후보에 등록되어 있으므로, 의존소의 실질

그림2

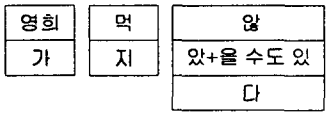

그림3

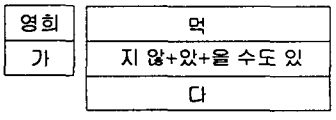

그림4

어휘도 점검하여 '~ㄹ 수 있~'의 형태에 해당함을 결정하게 된다. 이러한 형태가 형성되면 3 개의 SynN 들을 하나로 뭉친다(merge). 결과적으로 SynN 의 수가 5 개에서 3 개로 줄었으며 그 형태는 [그림 3]과 같다.

이번에는 맨 오른쪽 어절의 기능어휘가 '않'이며, 역시 보조용언의 후보에 등록되어 있으므로 의존소의 기능어휘에 '지'가 있는가를 확인, 또다시 SynN 들을 하나로 뭉치게 된다. 따라서 최종적으로 형성된 결과는 [그림 4]와 같으며, 의존관계는 1 개뿐이다.

4. 실험 및 결과

한-중 기계번역 시스템에서 구문분석의 결과는 이후의 의미분석 단계에서 효율적으로 사용될 수 있게 하기 위하여 어절 단위로서는 동일한 결과를 나타내는 것끼리 하나의 단위로 묶여져 전달된다. 이러한 단위를 topology 라고 정의하고 있는데, 이는 이번 논문의 해당 영역이 아니므로 자세한 설명은 생략하겠다.

이와 같은 특징으로 인하여 구문분석 결과는 가능한 모든 것들을 생성하게 된다. 또한 구문분석 결과가 의미해석의 입력으로 쓰이는 것이기 때문에, 구문분석 결과에 대한 평가는 그 결과들의 수와 이상적인 분석 결과의 수가 얼마나 근사한가에 초점을 맞추었다. 물론, 구문분석기가 제시한 결과들 중에 이상적인 분석 결과가 존재해야 함은 당연한 것이다.

일차적인 실험은 한-중 기계번역 시스템의 시연을 위한 실험 문장 101 개를 대상으로 수행하였다. 이 문장들은 [1]에 기반하여 한국어의 대표적인 문형들에 해당하는 것들로서 선정되었다. 이들에 대한 분석의 결과는 매우 흡족한 것으로 평가되었다.

표 1. 어절구조 적용시의 의존관계 수의 변화

문장의 길이 (어절)	문장 수	평균 의존관계 수		(B)/(A) (%)
		[6]에서의 결과(A)	어절구조 적용(B)	
3~4	12	3.83	3.61	94.2
5~6	34	9.41	8.52	90.5
7~8	55	16.34	13.42	82.1
9~10	42	24.52	19.47	79.4
11~12	34	29.62	22.90	77.3
13~14	12	52.00	39.94	76.8
15~16	8	55.38	40.59	73.3
17 이상	3	128.67	93.54	72.7
계	200	23.78	18.76	78.89

그러나 이들 문장들은 다분히 문형적이고 구문분석의 복잡도를 야기시키는 요소들이 적기 때문에 보다 명확한 성능 향상을 위하여 [6]에서 실험했던 문장들에 대해서도 실험하였으며, 그 결과도 [표 1]에서 확인할 수 있는 것처럼, 이전의 결과들과 비교할 때 성능의 향상이 있었음을 확인할 수 있다.

5. 결 론

본 논문에서는 의존문법에 기반을 둔 구문분석을 위하여 한국어 어절에 대한 개념을 새로이 정의했으며, 이를 바탕으로 구문분석에서의 새로운 구문단위(SynN)를 정의하고 그 구조를 설정하였다. 이를 통하여 구문분석 과정에서 매우 유용한 어절 단위의 결합 및 분할을 체계적으로 기술할 수 있게 되었을 뿐만 아니라, 또한 의미분석이나 기계번역에서 필수적으로 요구되는 각종 정보들도 용이하게 추출할 수 있게 되었음을 보였으며, 이로써 구문분석을 이용한 의미분석이나 기계번역에서의 성능향상은 물론 그 과정에 대한 기술의 용이성을 가져올 수 있음도 보였다.

그리고 위에서의 결과들을 통하여 본 연구실에서 계속해서 연구해 온 2 단계 한국어 의존관계 파서의 실용적 이용 가능성을 검증하였다.

하지만 본 논문에서 제시한 어절구조는 본 연구실에서 보유하고 있는 말뭉치들을 참고하여 얻어진 것인 만큼, 앞으로 보다 광범위한 말뭉치들에 대한 용례 조사 등을 통한 검증을 수행할 것이다. 또한, 구문분석 결과의 수를 실용 수준으로 감소시키기 위해서 한국어 문장에 대한 보다 더 심도 있는 연구 또한 계속적으로 진행해 나갈 것이다.

참 고 문 헌

[1] 강은국, "조선어 문형연구," 박이정출판사, 1995.

[2] 고영근, 남기심, 박경조, "고등학교 문법 자습서," 탑출판사, 1985.

[3] 김창제, 정천영, 김영훈, 서영훈, "부분적인 이절 결합을 이용한 효율적인 한국어 구문분석기", 1995 년도 한국정보과학회 가을 학술발표논문집 Vol. 22, No. 2, pp. 597~600.

[4] 나동렬, "한국어 파싱에 대한 고찰," 정보과학회지 제 12 권 제 8 호, pp. 33~46, 1994.

[5] 류법모, 이태승, 이종혁, 이근배, "술어 중심 제약 전파를 이용한 2-단계 한국어 의존 파서," 한국정보과학회 봄 학술발표논문집, 제 23 권 1 호, pp. 923~926, 1996.

[6] 류법모, 이종혁, 이근배, "한국어 파서에서의 지역 의존관계의 이용," 1996 년도 제 8 회 한글 및 한국

어 정보처리 학술발표 논문집, pp. 464~468.
[7] Covington, M. A., "A Dependency Parser for Variable-Word-Order Languages," Research Report AI-1990-01, Artificial Intelligence Programs, Univ. of Georgia, 1990.
[8] D. Arnold, et al., "Machine Translation," NCC Blackwell, 1994.
[9] EunJa Kim, Jong-Hyeok Lee, Geunbae Lee, "A Table-driven Modality Generation in COBALT/JK," PRICAI '94(The 3rd Pacific Rim Int'l Conf. on Artificial Intelligence), Beijing China, pp. 759~763 Aug.18, 1994.
[10] Kalevi Tarvainen, "Einführung in die Dependenzgrammatik," Max Niemeyer Verlag Tübingen, 1981.
[11] Tae Seung Lee, et al., "A Two-Phase Dependency Parser of Korean," Natural Language Processing Pacific Rim Symposium '95, Vol. 2, pp. 715~720, 1995.