Journal of Korea Multimedia Society Vol. 16, No. 3, March 2013(pp. 366-374)
http://dx.doi.org/10.9717/kmms.2013.16.3.366

구간 분할과 논항정보를 이용한 구문분석시스템 구현에 관한 연구

박용욱†, 권혁철††

요 약

본 논문에서는 한국어 구문분석에서 발생하는 중의성을 해결하기 위하여 구간분할 방법과 논항정보를 사용하여 개선한 구문분석시스템을 소개한다. 본 논문에서 제안하는 구문분석 시스템은 어절대신 형태소를 입력으로 사용하고, 또한 주어진 형태소에 대하여 가능한 모든 구문 분석 구조를 생성하는 알고리즘을 사용한다. 따라서 많은 중의성을 포함한 구문 분석 결과를 생성한다. 이러한 중의성 구조 결과를 해결하기 위하여 세 가지 방법을 사용했다. 첫째 방법은 형태소분석 결과에서 중의성을 제거하는 방법이고 두 번째는 구문분석시 구간 분할하는 방법, 세 번째 방법은 논항정보를 이용하는 것이다. 이러한 방법을 사용하여 많은 중의성을 제거할 수 있었다. 실험을 통하여 약 53%의 중의성을 제거할 수 있었음을 보여준다.

A Study of Parsing System Implementation Using Segmentation and Argument Information

Yong Uk Park†, Hyuk Chul Kwon††

ABSTRACT

One of the most important problems in syntactic analysis is syntactic ambiguities. This paper proposes a parsing system and this system can reduce syntactic ambiguities by using segmentation method and argument information method. The proposed system uses morphemes for the input of syntax analysis system, and syntactic analysis system generates all possible parse trees from the given morphemes. Therefore, this system generates many syntactic ambiguity problems. We use three methods to solve these problems. First is disambiguation method in morphological analysis, second is segmentation method in syntactic analysis processing, and the last method is using argument information. Using these three methods, we can reduce many ambiguities in Korean syntactic analysis. In our experiment, our approach decreases about 53% of syntactic ambiguities.

Key words: Segmentation Method(구간분할법),Syntactic Ambiguity(구문중의성), Argument Information (논항정보)

1. 서 론

언어는 인간만이 사용하고 있다. 말을 하고 글을 쓰고 이것을 이해하는 것은 인간만이 지닌 능력이다. 그런데 이러한 인간이 사용하는 언어를 기계 즉 컴퓨터가 이해할 수 있도록 처리하는 것을 자연언어처리

※ 교신저자(Corresponding Author) : 박용욱, 주소 : 울산시 남구 삼산동 평창3차 (680-770), 전화 : 052)230-0693, FAX : 052)230-0670, E-mail : yupark@uc.ac.kr
접수일 : 2013년 1월 2일, 수정일 : 2013년 3월 5일
완료일 : 2013년 3월 28일

† 정회원, 울산과학대학교 컴퓨터정보학부 교수
(E-mail : yupark@uc.ac.kr)
†† 정회원, 부산대학교 정보컴퓨터공학부 교수
(E-mail : hckwon@pusan.ac.kr)
※ 이 논문은 2011년 울산과학대학교 교내학술연구비 지원에 의해 수행되었음.

라고 한다[1]. 이러한 자연언어를 이해하는 데는 두 가지 측면이 있다. 첫째는 문법적 이해이다. 각 단어의 품사를 인식하여 한 문장 안에서 단어가 어떻게 구성되었는가를 인식하는 것이다. 둘째는 의미적 이해이다. 문장을 문법적 지식만으로는 이해할 수 없다. 예를 들어 “아름다운 영희의 동생”라는 문구에서 “아름답다”가 “영희”를 수식하는지 “동생”을 수식하는가는 문법적 이해만으로는 해결할 수 없고, 문맥에 대한 이해를 바탕으로 해결해야 한다. 따라서 자연언어를 올바르게 처리하기 위해서는 문법적 이해와 더불어 의미적 이해를 수반해야 한다. 그러나 의미적 이해 분야는 결코 쉽지 않다. 따라서 현재 자연언어 처리수준은 대부분 문법적 이해 수준에 중점을 두고 있다[1-3]. 문법에 바탕을 두고 한국어 구문을 처리하는 한국어 구문 분석기의 목표는 한글로 된 단어들의 선형적 나열인 문장으로부터 그 문장에 내포되어 있는 문법적 구조를 찾아내는 것이라 할 수 있다. 이러한 문법적 구조를 바탕으로 정확한 뜻을 컴퓨터가 이해할 수 있도록 하는 것이다.

최근 한국어 분석에서 주로 사용되는 문법은 의존문법이다[2,4]. 한국어의 특성인 중심어 후행, 어순의 자유로움, 빈번한 생략 등에 의존문법이 비교적 잘 맞기에 이에 대한 연구가 많이 진행되었다[4,7,11]. 의존문법은 지배소와 의존소의 관계를 파악하여 결합시키는 문법으로 단순하면서 강력하다. 이러한 특성으로 인하여 지금까지의 많은 연구에서 의존문법을 한국어 구문분석에 적용한 시스템을 연구하여 왔다[2,7,9,11]. 구문 분석시에 인접하는 서로 다른 형태소 사이에 의존소와 지배소라는 문법적인 결합관계가 성립하면 결합할 수 있음으로 인하여 구문분석 결과에서 중의성을 증가시키는 문제점을 가지고 있다. 하지만 그 동안 의존문법을 이용한 한국어 구문분석에 많이 사용해오고 있다[6,7,12].

이러한 중의성이 증가하는 문제를 해결하기 위한 연구가 여러 가지 방법으로 진행되어 오고 있다. 그 방법들 중의 하나가 긴 문장을 여러 개의 구간으로 분할하여 구간단위로 구문 분석하는 방법이 있다[4,5,8]. 구간분할 방법으로는 절(phrase) 단위로 분할하는 방법과 용언의 문형정보를 이용한 분할방법이 연구되었다. 구간을 분할하여 구간 내에서 구문분석을 함으로서 중의성을 감소시키는 효과가 있다. 본 논문에서도 이러한 구간분할법을 사용한다. 그러나 구간을 분할하는 방법 및 분할된 구간에서 분석된 결과들을 합치는 방법은 새로운 방법을 사용하였다. 자세한 내용은 본론에서 설명한다.

한국어 분석기를 구현함에 있어서 고려해야 할 요소 중 또 하나는 구문분석의 기본단위이다. 구문분석시에 사용되는 기본단위로는 형태소와 어절이 있다. 최근 많은 연구에서 어절을 구문분석의 기본단위로 사용하고 있다. 어절을 기본단위로 사용하게 되면, 기본단위의 수가 형태소를 사용할 때 보다 많이 줄어들게 된다. 이를 통해 구문분석 과정이 보다 단순해진다. 그러나 하나의 어절이 하나 이상의 형태소로 이루어져 있기 때문에, 하나의 어절을 분류하려면 하나 이상의 형태소 분류의 조합이 필요해진다. 이것은 곧 형태소 기본단위보다 그 분류 체계가 커짐을 의미하고, 이를 바탕으로 구축되는 문법규모 역시 커지게 된다[6,7]. 본 논문에서는 문장 분석 단위로 형태소를 사용한다. 한 어절이 거의 하나의 형태소로 이루어진 영어와 달리 한국어는 하나의 어절이 여러 개의 형태소로 이루어져 있으며, 이 때문에 지역 중의성이 많이 존재한다[2]. 즉 하나의 어절에 대한 형태소 분석 후보가 여러 개가 나올 수 있다. 본 논문에서 제안하는 구문분석기는 지역 중의성을 인정하고, 그에 따른 모든 가능한 구문분석구조를 결과로 출력하므로 어절단위를 사용하는 분석시스템보다 많은 분석결과 트리를 낼 수 있다. 이를 해결하기 위하여 문장 구성성분 단위의 구간분할 방법 및 논항정보를 이용하였다.

2. 한국어 특성

여러 언어들의 구조적 특성을 대비 고찰하고 서로 비슷한 주요 특성을 가진 것끼리 한데 묶어 갈래를 짓는 언어유형론에 따르면, 자연언어를 분류하는 기준이 네 가지가 있다. 모음체계, 자음체계, 음소의 수 등에 따라 분류하는 음운유형, 문법적인 요소인 접사 등의 발달 유무에 따라 분류하는 형태유형, 어순 등 문장 구조의 특성에 따라 분류하는 구문유형, 수사, 친족어 등 어휘적 특성에 따라 분류하는 의미유형이다[13]. 이들 중 실질적 의미부와 문법적 형태부의 결합관계를 바탕으로 분류하는 형태유형에 언어를 분류하면 다시 네 가지 유형으로 나눈다. 중국어와 같이 실질적 형태만 있고 문법형태가 발달 되어 있지

않은 고립형 언어, 실질적 의미부분과 문법적인 기능을 나타내는 부분이 서로 분리할 수 없을 정도로 밀접하게 결합된 굴절형으로 영어, 불어 등이 속한다. 그리고 문장을 구성하는 모든 요소가 한데 엉겨 붙어서 마치 한 단어처럼 여겨지는 에스키모어와 같은 형태인 포용형 언어, 그리고 한국어가 속하는 교착형 언어이다. 교착형 언어의 특징은 의미부분인 실질 형태에 문법 형태인 조사나 어미를 첨가하여 문법관계나 기능을 나타내는 언어로서 한국어, 터키어, 일본어 등이 여기에 속한다[13].

이것들 중에서 한국어와 같은 교착형 언어는 굴절형 언어인 영어보다 복잡한 결합구조를 가지며, 컴퓨터로 처리하기 어렵다. 굴절어인 영어의 경우는 대부분 하나의 어절이 하나의 형태소로 이루어져 있어 형태소를 분리하는 문제가 복잡하지 않지만 한국어는 하나의 어절이 여러 개의 형태소로 이루어져 있기에 영어보다 복잡하고 다양한 중의성 문제를 야기한다[1]. 예를 들어 “감기는 나쁜 병이다”라는 문장에서 “감기는”은 명사 “감기”와 보조사 “는”으로 분리될 수 있고, 동사 “감기다”와 관형형전성어미 “ㄴ”으로도 분석가능하다. 형태소 분석 시에 나타나는 중의성은 그대로 구문분석의 분석구조의 복잡성으로 이어진다. 자연언어를 분석함에 있어 발생할 수 있는 중의성에는 어휘적 중의성, 품사의 중의성, 구조적 중의성, 의미적 중의성 등으로 분류될 수 있다. 어휘적 중의성은 행태소의 원형을 분리하고 복원하는 과정에서 발생하며, 품사의 중의성은 하나의 형태소가 여러 가지의 품사로 사용될 때 발생한다. 구조적 중의성은 한 문장이 다수의 통사구조로 분석될 때 발생하며, 의미적 중의성은 단어의 해석이 두 가지 이상으로 가능할 때 나타난다. 형태소 분석의 결과로 나타나는 어휘적 중의성과 품사의 중의성은 구문분석의 구조적 중의성을 증가시키는 결과로 이어진다. 본 논문에서는 형태소 분석 시 나타나는 어휘적 중의성을 형태소의 앞뒤관계를 조사하여 가능한 중의성을 제거한 후 구문분석 시스템으로 넘겨주도록 한다. 본 논문에서는 구문분석의 단위를 형태소로 사용하기에 이 과정은 구문분석 결과로 나타나는 중의성을 해결하는데 꼭 필요하다. 그리고 구문 분석 시에 형태소 단위로 사용하기 때문에 많은 구조적 중의성을 발생시킨다. 이 문제를 해결하기 위하여 본 논문에서는 구간분할과 논항정보를 이용하였다.

3. 시스템 구현

본 논문의 시스템 구성은 그림 1과 같다. 입력문장에 대하여 형태소를 추출해내는 형태소 분석과정, 분석된 형태소 리스트에 대하여 몇몇 형태소들에 대하여 결합 등을 실시하는 전처리 과정, 어휘적 중의성 해결과정, 구간분할 및 통합과정을 통한 구문을 분석하는 과정, 논항정보를 이용한 완성된 구문트리에 대하여 정리하는 과정으로 시스템이 구성되어져 있다. 본 시스템은 주어진 입력 문장에 대하여 의존문법을 적용하여 가능한 모든 파서트리를 찾아내는 알고리즘을 사용하고 있다. 그러므로 비교적 많은 파서트리를 출력하게 되는데 특히나 문장이 길어지면 매우 많은 중의성을 갖게 된다. 본 연구에서는 이러한 구문분석 중의성을 해결하기 위하여 형태소분석 결과에서 가능한 어휘적 중의성을 해결하도록 하고, 또한 구문분석 시에는 구간을 분할하는 방법을 사용한다. 그리고 최종적으로는 논항정보를 이용하여 완성된 파서트리에서 올바르지 못한 트리는 제거하는 시스템을 구현한다.

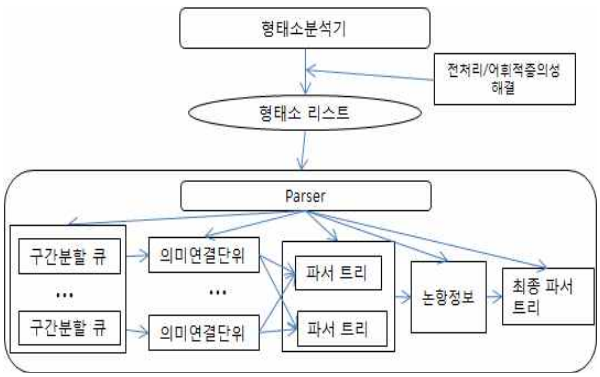

그림 1. 시스템 구성도

3.1 어휘적 중의성 해결

형태소 분석의 결과로 얻어지는 형태소들을 그대로 구문분석 시스템에 넘겨주면 구문분석 시스템의 입력 양이 많아지므로 결과로 출력되는 구문구조의 수가 많아지고 복잡하게 된다. 선어말어미 등과 같은 통사적 의존관계를 가지지 않은 형태소들은 미리 하나로 합쳐 구분분석 시스템 의 후보를 줄여 계산의 복잡도 및 중의성을 해소하는데 많은 도움이 된다[7]. 본 연구에서 사용하는 구문분석 시스템의 후보를 줄이는 방법에는 세 가지가 있다. 첫째, 이웃하는

어떤 형태소들은 하나로 합치는 방법이다. 선어말어미 등과 같은 통사적으로 의존관계를 가지지 않는 형태소들은 구문분석 시스템 후보로 그대로 넘겨주지 않고 미리 하나로 합친다. 또 “ㄹ 수 있다”와 같은 것은 용언관용구로 “관형형 전성어미+수+있다”로 하나로 합친다. “눈코 뜰 새 없이”, “시도 때도 없이”와 같이 숙어처럼 사용되는 형태소들도 하나로 합친다. 둘째는 일부 형태소들을 제거하는 방법이다. 보조용언들은 독립적인 의미를 가지지 않으므로 본용언에 추가되어도 무방하므로 제거가능하고, 또 일부 의존명사가 분석후보에 홀로 나타나는 경우에 좌우 형태소를 보고 제거할 수 있는 분석후보를 제거하는 방법이다. 그리고 세 번째 방법은 하나의 어절이 여러 가지로 형태소 분석이 될 때 앞뒤 어절과의 연관관계 및 문법관계 이용하여 올바르지 않는 후보는 미리 제거하는 방법이다. 여기에 사용되는 규칙에는 다음과 같은 것들이 있다.

(rule 1) 의존명사는 관형어 뒤에 와야 한다.

(rule 2) 수사 의존명사는 수관형사 뒤에 온다.

(rule 3) 관형어 뒤에 관형어나 수식을 받는 명사가 와야 한다.

(rule 4) 보조용언 앞에는 본용언이 와야 한다.

이러한 규칙을 사용하여 “보고 싶다”에 대해 분석하면 다음과 같다.

보고 : 1. 보고(명사)
2. 보다(동사)+고(연결어미)

싶다 : 1. 싶다(보조용언)+다(어미)

(rule 4)에 의하면 보조용언 앞에는 본용언이 와야 한다. 따라서 “보고”에 대한 분석 결과 중 두 번째 것인 “보다(동사)+고(연결어미)”를 취하고, “보고(명사)”는 버리게 된다. 또 다른 예로 “먹은 줄이다”라는 문장을 형태소 분석하면 다음과 같다.

먹은 : 1. 먹다(동사)+은(관형형전성어미)
2. 먹(명사)+은(조사)

줄이다 : a. 줄이다(동사) + 다(어미)
b. 줄(명사) + 이다(지정사)+다(어미)

위의 형태소분석 결과에 대하여 일반화하면 다음과 같다. 즉 어떤 두 어절 사이에 결합관계를 맺을 수 있을 때, 첫 번째 어절에서 형태소가 n개, 두 번째 어절에서 m개의 형태소로 분석된다가 가정한다면 두 어절사이에 결합 가능한 후보의 수는 이론적으로 (n x m)개가 된다. 따라서 “먹은 줄이다”라는 문장에는 2개의 어절로 구성되었고, 각각의 어절로부터 2개 형태소로 분석되므로 총 4개의 결합이 가능하며 이들의 조합은 (1,a),(1,b),(2,a),(2,b) 로 구성된다. 그러나 (rule 3)에 의하여 (1,a) 조합은 제거된다. 따라서 나머지 3개의 조합만이 구문분석 후보로 전달되게 된다.

앞의 설명에서 살펴보았듯이 본 연구의 구문분석기는 모든 형태소 분석결과에 대한 조합으로 된 형태소 리스트에 대해 분석을 시도한다. 그러므로 모든 형태소 분석결과 조합을 사용하지 않는 다른 구문분석 시스템에 비하여 입력문장에서 보다 많은 중의성이 내포되어 있을 가능성 크다고 볼 수 있다. 그러므로 형태소 분석 결과로 나타나는 형태소들을 그대로 구문분석의 입력으로 사용하지 않고 앞에서 설명한 세 가지 방법을 사용하여 형태소를 합치거나 제거한 후 구문분석의 후보를 최소한으로 줄일 수 있도록 구현하였다.

3.2 의존문법 및 구간분할을 이용한 파싱

본 시스템은 구문 분석 시에 의존문법을 사용한다. 의존문법은 문장을 구성하는 구성 성분들 사이에 존재하는 의존관계를 파악함으로써 문장을 분석한다. 여기서 구성 성분이란 어절이 될 수도 있고 형태소가 될 수도 있다. 본 논문에서는 형태소를 사용한다. 의존 관계는 두 구성 성분 사이에 존재하는데 이 중 한 구성요소는 지배소(governor)가 되고 다른 구성 요소는 의존소(dependent)가 된다. 지배소는 의미의 중심이 되는 요소이고, 의존소는 지배소가 갖는 의미를 보완해 주는 요소이다. 한국어를 구문 분석할 때 의존문법을 많이 사용하게 되는데 그 이유는 어순이 자유롭고 문장 구성 요소의 생략이 많은 한국어의 특성을 잘 반영할 수 있기 때문이다. 이와 같은 특성을 갖는 의존 문법을 한국어 구문 분석에 적용하기 위해서는 의존 규칙이 필요하다. 본 연구에서는 구문분석의 기본 구성요소를 형태소로 처리하기 때문에 형태소를 중심으로 한 의존 규칙을 만들었다. 형태소를 중심으로 구문 분석을 처리하지만 의존 규칙에 의해 형태소와 형태소의 결합 후 구(phrase)가 만들어지게 된다. 따라서 본 시스템에서는 이러한 구에

대해서도 의존 규칙에 추가하였다. 본 시스템은 현재 구를 포함하여 112개의 품사 분류와 이를 바탕으로 한 218개의 의존문법 규칙이 만들어져 있으며, 이러한 품사 분류와 규칙들을 계속해서 개선해 나가고 있다. 표 1은 본 시스템에서 사용하는 의존 규칙의 일부이다.

앞에서 설명한 바와 같이 표 1의 의존 문법 적용 규칙에 구(phrase)를 사용한다. 명사구, 동사구, 형용사구와 같은 구를 문법규칙에 사용하게 되면 품사 분류의 수와 규칙의 수는 늘어난다. 그러나 원래 하나의 형태소일 때와 그것이 다른 형태소와 결합한 후 성격 변화로 인하여 규칙 적용이 보다 확장될 수 있는데 이를 처리하기가 용이하다. 예를 들어, '이다'와 같은 지정사는, 단순히 지정사 일 때는 바로 앞의 명사 등만 지배할 수 있지만, 그것과 결합하여 지정사구가 되면 논항을 지배할 수 있게 된다.

다음으로 구간분할 알고리즘을 사용한 구문 분석 과정을 살펴보겠다. 앞에서 설명하였던 내용과 같이 본 논문은 구문 분석의 기본단위로 형태소를 사용한다. 하나의 어절 속에도 어휘적 중의성 등으로 인하여 어절을 사용하는 구문분석 시스템보다 구분분석에 입력되는 요소가 많아진다. 이에 따라 구문 분석의 결과로 얻어지는 구문 구조의 중의성이 증가하게 된다. 이것을 해결하기 위한 하나의 방법으로 구간분할 방법을 사용한다. 구간 별로 부분 구문 분석을 실시한 결과를 가지고 문장의 전체를 분석함으로서 구문트리의 중의성을 줄일 수 있다[4,8]. 문장을 구성하는 주요 구성성분에는 주어, 목적어, 부사어, 관형어, 술어 등이 있다. 구문구조 분석의 결과로 나타나는 분석트리를 보면 술어를 중심으로 주어, 목적어, 부사어 등이 그에 맞는 의존규칙에 의해 결합된 결과를 보여준다. 따라서 본 논문에서는 구간분할을 주어, 목적어, 부사어가 있는 곳에서 분할을 실시한다. (예문1)은 이 기준에 따라 분할된 곳을 ∨기호를 사용하여 표시한 것이다.

(예문1)철수가∨착한 영희의 동생을∨좋아한다.

실제적으로 본 시스템은 주어진 문장에 대하여 미리 구간을 분할하지 않는다. 주어진 문장에 대한 형태소 분석의 결과로 생성된 형태소들에 대하여 처음부터 마지막 형태소까지 순서대로 입력받아 처리하는 과정 중에 구간분할의 기준을 만나면 그때 구간분할을 실시한다. 그리고 이때 분할 된 구간에 대하여 의미연결단위(semantic connection unit) 노드를 생성한다. 의미연결단위노드란 하나 구간에 존재하는 모든 형태소들이 의존규칙에 의해서 모두 결합된 결합노드를 말한다. 본 논문에서는 구간별로 의미연결단위 노드를 구하고 이들을 결합하는 과정을 거쳐 구문분석이 이루어진다. 의미연결단위 노드는 그 구간의 형태소 구성에 따라 한 개 또는 그 이상의 의미연결단위가 생성될 수 있다. 다음은 앞에서 설명한 내용에 대하여 정리한 구문분석 알고리즘이다.

1) 형태소 분석기를 통하여 형태소 리스트 생성

표 1. 의존문법 적용 규칙 일부분

Relation	Governor	Dependent	Result
수식	명사	관형사	명사구
수식	명사	명사	명사구
수식	동사	동사수식부사	동사구
수식	동사	부사구	동사구
수식	형용사	형용사수식부사	형용사구
수식	형용사	부사구	형용사구
품사전성	관형형전성어미	형용사	관형사구
논항	동사	격조사구	동사구
논항	형용사	격조사구	형용사구
격부여	주격보격조사	명사	격조사구
격부여	관형격조사	명사/명사구	관형사사구
종결	종결어미	동사/동사구	종결구
종결	종결어미	형용사/형용사구	종결구
완성	온점	종결어미/종결구	문장

2) 형태소 리스트로부터 왼쪽에서 시작하여 3)단계부터 실시
3) if(형태소가 {주어/목적어/부사어}중 하나에 속함)
then 3-1) 새로운 구간을 생성 및 현재의 형태소를 새로운 구간에 넣음
3-2) 이전구간에 대하여 semantic connection unit 노드를 생성
else if (형태소가 술어(predicate)에 속함) then
3-3) 이 predicate를 이전구간의 의미연결단위 노드와 결합
else 현재 구간에서 이전의 형태소노드들과 의존규칙을 이용하여 결합하여 결합노드를 생성
4) 3)단계를 형태소리스트의 끝까지 반복수행
5) 모든 구간의 의미연결단위노드를 합하여 분석결과 트리 생성
6) 논항정보를 이용하여 잘못 결합된 분석트리 제거하여 최종 분석결과 트리 생성

(예문1)의 두 번째 구간인 “착한 영희의 동생을”에 대하여 의미연결단위를 구하는 과정을 보면 다음과 같다.

형태소분석결과:
착하다(형용사)/ㄴ(관형사형 어미)/
영희(명사)/의(조사)/동생(명사)/
을(목적격조사)

구간 내 의미연결단위 형성과정 :
▸input:[착하다] -> [착하다]
▸input:[ㄴ] -> [착하다],[ㄴ],[[착하다][ㄴ]]
▸input:[영희] -> [착하다],[ㄴ],[[착하다][ㄴ]],
[영희],[[[착하다][ㄴ]][영희]]
▸input:[의] -> [착하다],[ㄴ],[[착하다][ㄴ]],
[영희],[[[착하다][ㄴ]][영희]],[의],
[[영희][의]],[[[[착하다][ㄴ]][영희]][의]]
▸input:[동생] -> [착하다],[ㄴ],[[착하다][ㄴ]],
[영희],[[[착하다][ㄴ]][영희]],[의],[[영희][의]], [[[[착하다][ㄴ]][영희]][의]],[동생],
[[[영희][의]][동생]],
[([착하다][ㄴ])([[영희][의]][동생])],
[[([[착하다][ㄴ]][영희])[의]][동생]
▸input:[을] -> [착하다],[ㄴ],[[착하다][ㄴ]],
[영희],[[[착하다][ㄴ]][영희]],[의],[[영희][의]],
[[[[착하다][ㄴ]][영희]][의]],
[동생],[[[영희][의]][동생]],
[([착하다][ㄴ])([[영희][의]][동생])],
[[([[착하다][ㄴ]][영희])[의]][동생]],[을],
[[[([[착하다][ㄴ]][영희])[의]][동생]][을]],
[[([착하다][ㄴ])([[영희][의]][동생])][을]],
[[[[영희][의]][동생]][을]],[[동생][을]]

위에서 노드간의 구분은 콤마(,)로 표시하였다. 표시된 것과 같이 “착한 영희의 동생을” 구간에서 최종적으로 생성된 노드 수는 모두 16개이다. 이 중에서 모든 형태소를 가지는 노드는 밑줄 친 두개의 노드이다. 나머지 14개의 노드는 모두 제거된다. 모든 형태소가 연결된 노드 즉 의미연결단위 노드만 그 구간의 최종 노드로 남겨서 구간결합 시에 사용된다. 두 개의 의미연결단위 노드의 구조가 갖는 의미를 해석해 보면 하나는 “영희가 착하다”는 의미를 갖는 구조이고 다른 하나는 “영희의 동생이 착하다”라는 의미를 갖는 구조이다.

이와 같이 분할된 구간에 대하여 구간별로 의미연결단위 노드를 구한 후 최종적으로 구간을 통합하는 과정으로 구분분석이 이루어진다.

3.3 논항정보를 이용한 중의성 해결

앞의 단계를 거쳐 생성된 모든 구문구조 분석 결과트리에 존재하는 중의성을 제거하는 마지막 단계에서 용언의 논항정보를 이용한다. 논항정보는 부산대학교에서 구축한 한국어 어휘의미망(KorLex)[10]을 사용한다. 의존 규칙을 이용한 구문분석은 문법에 중점을 두기 때문에 많은 중의성을 만들어 낼 수 있다. 다음의 (예문2)를 살펴보자.

(예문2) 철수가 영희의 연필을 쓴다.

위의 예문을 앞의 구문 구조 분석과정을 거쳐 분석하면 다음 그림 2와 그림 3과 같은 2개의 분석 구조를 가진다. 그림 1과 그림 2에서 알 수 있듯이 어절 “쓴다”는 다음과 같이 2개의 의미로 분석되었다.

쓴다 => 쓰다 / 쓸다

문법적으로만 분석하면 두 가지 분석이 모두 가능하다. 그림 2는 “쓰다”로 분석된 결과를 보여주고 그림 3은 “쓸다”로 분석된 결과를 보여준다. 그러나 (예문2)의 의미를 고려하면 연필을 사용한다는 의미의 “쓰다”로 분석되는 것이 올바르다.

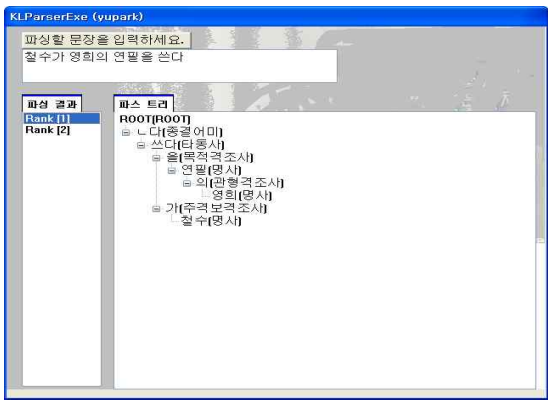

그림 2. (예문2)의 첫 번째 분석결과

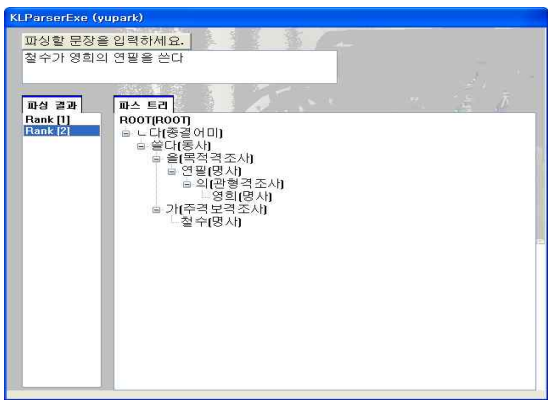

그림 3. (예문2)의 두 번째 분석결과

이와 같은 경우를 처리하기 위하여 용언의 논항정보를 사용한다. “쓰다” 또는 “쓸다”와 같은 중의성을 가지는 용언들에 대하여 각각이 취할 수 있는 논항들을 미리 정의해 두고 이에 맞는 문장은 받아들이고 그렇지 못한 문장은 제거한다. 표 2는 본 논문에서 사용하는 용언에 대한 논항정보 규칙 일부이다.

표 2의 논항정보 규칙을 적용하여 (예문 2)를 분석하면 그림 4와 같은 1개의 결과로만 분석된다.

표 2. 용언에 대한 논항정보 규칙

Predicate	Argument 1		Arguments 2	
	조사	의미범주	조사	의미범주
쓰다	이	인간	을	물건
쓸다	이	인간	을	장소
매다	이	인간	을	논,밭,끈

4. 실험 및 결과

본 논문에서 제시한 세 가지 방법을 적용할 대상

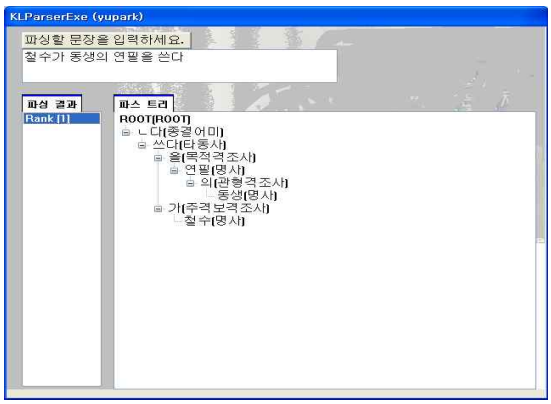

그림 4. (예문2)에 대하여 논항정보 규칙 적용한 결과

문장은 중학교 국어 교과서에서 100문장을 추출하였다. 본 논문에서 제시하는 구문분석 알고리즘은 형태소 분석의 결과로 주어지는 형태소들에 대하여 모든 가능한 결합구조를 찾는 방법을 사용한다[2]. 따라서 기본적으로 많은 분석결과를 도출하게 된다. 하지만 이 결과들 중에는 의존 규칙에는 맞지만 의미적으로 잘못 결합된 분석구조들이 많이 포함되어 있다. 이러한 잘못된 결과들을 제거하기 위하여 제안한 3가지 방법을 적용하여 실험하였다. 실험은 다음과 같은 세 가지 방법으로 하였다.

첫째, 3가지 방법 모두 사용안한 경우

둘째, 어휘적 중의성 해결방법과 구간분할 및 의미연결단위노드 사용

셋째, 어휘적 중의성 해결방법, 구간분할 및 논항정보 3가지 모두 사용한 경우

각각의 방법으로 추출한 문장을 분석하여 추출되는 구문분석 결과의 수와 정확도(precision), 재현율(recall)을 가지고 비교하였다. 표 3은 세 가지 방법으로 비교한 결과를 보여준다. 또한 이 결과들을 다른 구문분석기[11]와 비교해 보았다.

표 3의 결과를 분석해 보면 3가지 방법 즉 어휘적

표 3. 실험결과 비교

	3가지방법 사용 안함	어휘적중의성,구간분할사용	어휘적중의성, 구간분할,논항정보사용	임경업
분석구조수(평균)	10.3	5.7	4.8	9.2
정확도	85.4%	91.7%	93.4%	74.3%
재현율	97.5%	87.8%	85.4%	98.5%

중의성 해결방법, 구간분할법 및 논항정보 모두 사용하지 않는 경우 많은 중의성을 포함하는 것을 알 수 있다. 어휘적 중의성 및 구간분할법을 사용할 경우 약 45%의 중의성을 제거해 주었다. 그리고 3가지 모두 사용하지 않는 경우 대비 3가지 방법 모두를 사용할 경우는 약53% 중의성 감소를 보인다. 재현율 측면에서는 3가지 방법을 모두 사용 안하는 경우가 가장 높게 나타난다. 그 이유는 중의성은 존재하지만 모든 분석 구조를 찾아주기 때문에 정확한 분석 구조를 더 많이 포함하고 있기 때문이다. 하지만 정확도 측면에서는 3가지 방법을 모두 사용하는 경우가 가장 높게 나타난다. 이는 중의성을 제거하는 과정에서 잘못 결합된 많은 구조를 제거하기 때문에 정확률이 더 좋게 나타나는 것으로 분석된다. 본 연구실에서 구현한 기존 구문분석기인 [11]의 임경업(이하 임경업)과 비교분석하면 다음과 같다. 임경업 구문분석기는 구간분할을 실시하지 않는 시스템으로 본 시스템과 직접적으로 비교할 수는 없지만 본 논문의 구문분석시스템은 임경업 구문분석시스템을 개선한 것이므로 이와 비교해보면 다음과 같다. 임경업 구문 시스템은 정확도가 74.3%였으나 본 연구의 시스템에서는 93.4%로 개선된 것을 볼 수 있다. 그러나 재현율에서는 임경업 시스템이 98.5% 이고 본 논문의 시스템은 85.4%로 낮아졌다. 이는 위에서 설명한 이유로 인하여 발생되었음을 알 수 있고, 차후 이것을 개선해야 한다.

5. 결론 및 향후연구

본 논문에서 제안한 시스템은 의존문법을 사용하며, 입력되는 요소로는 어절이 아니라 형태소를 사용한다. 그리고 주어진 형태소들 사이에 존재하는 모든 의존구조를 찾아내는 알고리즘을 사용함으로서 많은 중의성을 가지는 분석구조 결과를 출력한다. 이러한 중의성을 해결하기 위하여 제안한 시스템에서는 형태소 분석결과에 대하여 어휘적 중의성 해결방법을 사용하였고, 구문분석 단계에서는 구간분할 방법과 의미연결단위 노드를 사용하여 중의성을 발생시킬 수 있는 노드들을 제거하였다. 또한 논항정보를 사용하여 의미적으로 올바르지 못한 결합구조에 대하여 제거함으로서 중의성을 제거하였다. 이와 같은 3가지 방법을 사용한 결과 사용하지 않는 경우 대비 많은 중의성을 제거할 수 있음을 보였다. 또한 정확률 측면에서도 제안된 방법을 사용하지 않는 경우보다 좋아지는 것을 확인하였다. 그러나 재현율 측면에서는 다소 감소되는 결과로 나타났다. 이것은 중의성을 제거하기 위하여 3가지 방법을 사용함으로서 제거하지 말아야 할 노드가 제거됨으로서 발생한다. 향후 연구에서 어떠한 경우에 이러한 현상이 발생하는지를 연구할 것이다. 또한 더 많은 용언들에 대한 논항정보를 구축하는 것이 필요할 것이다.

참 고 문 헌

[1] 김영택 외 공저, 자연언어처리, 생능출판사, 서울, 2001.

[2] 권혁철, 최준영, “단일화 기반 의존 문법을 이용한 한국어 분석기,” 정보과학회논문지, 제19권 제5호, pp. 467-476, 1992.

[3] 정영임, 조선호, 윤애선, 권혁철, “구문 관계와 운율 특성을 이용한 한국어 운율구 경계 예측,” 제19회 한글 및 한국어 정보처리 학술대회 논문집, pp. 7-14, 2007.

[4] 김광배, 박의규, 나동렬, 윤준태, “구간 분할 기반 한국어 구문분석,” 제14회 한글 및 한국어 정보처리 학술대회, pp. 163-168, 2002.

[5] 이현영, 황이규, 이용석, “문형과 단문 분할을 이용한 한국어 구문 모호성 해결,” 제12회 한글 및 한국어정보처리 학술대회, pp. 116-123, 2000.

[6] 최선화, “형태소 단위의 한국어 확률 의존문법 학습,” 정보처리학회 논문지 B, 제9-B권, 제6호, pp. 791-798, 2002

[7] 박의규, 나동열, “한국어 구문분석을 위한 구묶음 기반 의존 명사 처리,” 인지과학, 제17권, 제2호, pp. 119-138, 2006.

[8] 이성욱, 서정연, “한국어 문법관계에 대한 부분 구문 분석”, 정보과학회논문지, 제32권, 제10호, pp. 984-989, 2005.

[9] 김창제, 정천영, 김영훈, 서영훈, “부분적인 어절 결합을 이용한 효율적인 한국어 구문 분석기,” 한국정보과학회 가을 학술발표논문집, 제22권 제2호, pp. 597-600, 1995.

[10] 이은령, 윤애선, “피동 정보를 통한 한국어 동사 어휘의미망 정제,” 한국어학, 제28권, pp. 139-

166, 2005.
[11] 임경업, 정영임, 권혁철, “한국어 어휘의미망에 기반한 논항 정보를 이용한 의존문법 구문분석기의 구현,” 제19회 한글 및 한국어 정보처리 학술대회, pp. 158-164, 2007.
[12] 김영자, 김현주 “구조 기반 검색을 위한 색인 구조에 대한 분석,” 멀티미디어학회논문지, 제7권, 제5호, pp. 601-616, 2004.
[13] 이익섭, 한국어 문법, 서울대학교출판부, 서울, 2009.
[14] I.A. Mel′cuk, *Dependency Syntax : Theory and Practice*, State Univ. of New York Press, New York, 1988.

박 용 욱

1991년 부산대학교 전자계산학과 석사
1991년 3월～1997년 2월 전자부품연구원(KETI) 전임연구원
2000년 부산대학교 전자계산학과 박사수료
1998년 3월～현재 울산과학대학 컴퓨터정보학부교수
관심분야 : 자연언어처리, 정보검색, 멀티미디어

권 혁 철

1982년 서울대학교 컴퓨터공학과 학사
1984년 서울대학교 컴퓨터공학과 석사
1987년 서울대학교 컴퓨터공학과 박사
1992년～1993년 (미)Stanford 대학 CSLI 방문교수.
1987년～현재 부산대학교 정보컴퓨터공학부 교수
관심분야 : 인간언어공학, 정보검색, 인공지능