인지과학, 제17권 제2호, 2006, 119~138.

한국어 구문분석을 위한 구묶음 기반 의존명사 처리*

Processing Dependent Nouns Based on Chunking for Korean Syntactic Analysis

박 의 규** (Eui-Kyu Park)　　나 동 열*** (Dong-Yul Ra)

요 약 구묶음 작업은 문장의 분석을 보다 용이하게 해주는 것으로 알려져 있다. 본 논문에서는 한국어 문장의 구조 분석에 유용한 구묶음의 한 기법을 소개한다. 의존명사는 한국어 문장을 매우 복잡하고 길게 만드는 특성이 있다. 의존명사와 그 주변의 관계되는 단어에 대한 구묶음 작업을 통하여 문장의 복잡도를 낮출 수 있으며 이는 다음 분석 단계인 구문분석 작업을 보다 용이하게 만든다. 본 논문에서는 이러한 목적을 달성하기 위한 의존명사와 관련된 구묶음 처리에 대해서 자세히 알아보았다. 우리는 의존명사의 종류에 따라 매우 다양한 형태의 구묶음 방식을 제안하였다. 실험을 통하여 본 논문에서 제안한 의존명사 관련 구묶음 처리 기법이 구문분석 시스템의 성능을 크게 향상시키는 것을 확인하였다.

주제어 구묶음, 의존명사, 구문분석

Abstract It is widely known that chunking is beneficial to syntactic analysis. This paper introduces a method of chunking that is useful for structural analysis of sentences in Korean. Dependent nouns in Korean usually tend to make sentences complex and long. By performing chunking operations related with dependent nouns, it is possible to reduce sentence complexity and thus make syntactic analysis easier. With this aim in mind we investigated techniques for chunking related with dependent nouns. We proposed a variety of chunking schemes according to the types of dependent nouns. The experiments showed that carrying out chunking leads to significant improvement of performance in syntactic analysis for Korean.

Keywords Chunking, Dependent nouns, Syntactic analysis

---

* 이 논문은 정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임 (과제번호: R05-2003-000-12125-0).

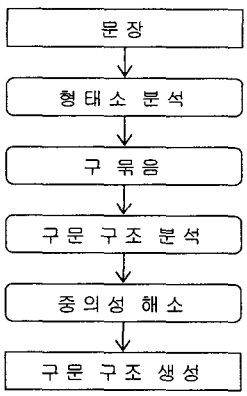
** 연세대학교 컴퓨터정보통신공학부

*** 연세대학교 컴퓨터정보통신공학부, 연구분야: 자연어처리 및 정보검색
강원도 원주시 홍업면 매지리, 연세대학교 컴퓨터정보통신공학부, E-mail: dyra@yonsei.ac.kr

서 론

구문분석이란 문장을 분석하여 문장을 구성하는 요소들 간의 관계를 밝히는 작업이라 할 수도 있다. 구문분석은 컴퓨터에 의한 자연어의 처리에서 매우 중요한 작업으로 간주되고 있다. 따라서 높은 성능의 구문분석 시스템을 개발하기 위해 세계적으로 많은 연구가 수행되어 왔다.

한국어에 대해서도 구문분석 연구와 이론들이 많이 만들어졌고 실험되어 왔다(나동열, 1994; 윤덕호, 1993; 홍영국 등, 1993). 그러나 한국어의 잦은 문장 요소 생략과 자유로운 어순은 완전한 한국어 구문분석 시스템 개발에 걸림돌이 되어 왔다. 또한 문장이 길어지면 길어질수록 구문분석 시스템의 결과로써 제시되는 구문구조의 수가 너무 많아서 좋지 않은 결과를 내는 원인이 되어 왔다. 따라서 본 논문에서는 긴 문장에 대해서도 좋은 결과를 낼 수 있는 구문분석 시스템의 달성을 위한 구묶음 기법을 제안하고자 한다. 전체시스템 안에서 구묶음 모듈의 위치는 그림 1과 같다.



(그림 1) 구문분석 시스템 구성

Abney는 구문분석의 첫 단계(frontend stage)로 구묶음을 두는 것을 제안하였다(Abney, 1991). 그는 구묶음이란 문장에서 겹쳐지지 않는 덩어리로서 구문트리에서 하위 트리를 형성하는 것이라고 정의하였다. 구묶음 된 덩어리를 하나의 단어처럼 본다면 구묶음 후의 문장은 원래의 문장에 비해서 단어의 수가 감소하며 복잡성이 감소하는 경향이 있다. 구묶음을 한 후에도 원래의 구는 계속 그 성질을 유지한다. 다만 구를 구성하던 여러 단어들이 구묶음 속에 묻혀지고 구묶음은 대표적인 단어로 외부에 보여진다. 따라서 구묶음 단계를 두면 구문분석 과정의 부담을 경감시킬 수 있다. 이와 관련하여 다음 예를 보자.

(1) 유럽 중앙 은행은 유로의 공식 통용에 앞서 지폐와 동전을 발행할 수 있었다.

위의 예문 (1)은 모두 12개의 어절로 이루어져 있다. 구문분석으로 많이 사용되는 CYK 기법을 예로 설명해 보자. 자연어 처리에서 사용하는 CYK 구문분석 기법의 경우 중의성 현상 때문에 시간복잡도는 $O(n^3)$이다 (Aho & Ullman, 1972). 여기에서 n은 어절의 수이다. 예문 (1)의 시간복잡도는 $12^3 = 1728$ 이다. 예문 (1)을 구묶음한 결과는 (1')과 같다.

(1') ┌>유럽 중앙 은행은 유로의 공식 통용에 앞서 지폐와 동전을 발행하였다. ┌>공식 ┌>수 있

예문 (1)은 구묶음을 통하여 (1')과 같이 7개의 어절로 이루어진 것처럼 된다. 예문 (1')의

시간복잡도는 $7^3 = 343$ 이다. 이는 예문 (1)의 복잡도의 약 1/5 수준이다. 이와 같이 구묶음을 이용하면 구문분석 시스템에서 처리해야하는 단어의 수가 줄어들게 되어 구문분석의 시간 및 공간 복잡도를 줄일 수 있다.

구묶음과 관련하여 지금까지 잘 알려진 것으로 복합명사, 본용언/보조용언과 관련한 처리가 있다 (황이규 등, 2000). 본 논문에서는 이러한 지금까지의 일반적인 구묶음에서 더 나아가 의존 명사에 대한 구묶음 처리를 제안한다[1]. 의존명사는 문장의 구조를 복잡하게 하는 성질이 있어서 이의 효과적인 처리 기법의 중요성이 매우 크다. 따라서 우리는 의존명사에 대한 고찰을 통하여 의존명사를 위한 세밀한 구묶음 방식을 도입할 것을 제안한다. 그리고 이는 한국어 구문분석 시스템의 성능 향상을 가져 온다는 것을 실험을 통하여 보이고자 한다.

의존명사란 자립성이 없는 특수한 명사를 일컫는다 (남기심, 고영근, 2004; 서정수, 1994). 즉 그 앞에 어떤 한정 성분이 나타나지 않으면 홀로 쓰일 수 없는 비자립적인 명사라는 것이다. 그 앞에 나타나는 성분은 관형어 다시 말하면 관형사나 관형사 기능을 가지는 한정어이다. 본 연구에서는 의존 명사는 자립적으로 사용할 수 없다는 사실에 주목한다. 이 사실은 구문분석에 있어서 의존명사 어절을 그대로 두어서는 그 형태가 너무 다양하여 견고한 구문분석을 하기 어렵다는 것을 의미한다. 의존 명사를 일반 명사처럼 처리하면 많은 중의성과 다양한 문형을 유발하게 된다. 그렇다고 해서 기존의 단순한 방식의 의존명사 처리 기법을 이용하면 다양한 형태로 나타나는 의존 명사에 대한 처리가 불완전하게 된다. 따라서 견고하고 정확한 구문분석을 위해서는 의존명사에 대한 세밀한 구묶음 처리가 필요하다. 예를 들면 다음과 같다.

(2) 길수는 사과를 10개 먹었다.
(2') 길수는 사과를 먹었다.
　　└>10개

예문 (2)에서 어절 "10개"는 사과의 개수를 나타낸다. 여기에서 "개"는 단위 의존명사이다. 즉 사과에 갯수라는 의미를 추가할 수 있기만 하다면 "10개"를 문장에서 제외해도 되는 것이다. 따라서 우리는 "10개"를 앞 어절에 구묶음이 되도록 하여 문장에서 제외시키고 다만 (2')에서와 같이 "사과를" 어절의 내부 성질로서 "10개"라는 정보를 부착해 놓는다.

(3) 한국은 2002년 월드컵에서 4강에 올랐다.
(3') 한국은 월드컵에서 4강에 올랐다.
　　└>2002년

예문 (3)에서 어절 "2002년"은 월드컵이 개최된 해를 나타낸다. 즉 월드컵에 개최된 해라는 의미를 추가하는 것이다. 여기에서 "년"은 의존명사이다. 이 의존명사 포함 어절을 뒤 어절에 구묶음이 되도록 함으로써 (3')를 얻게 된다.

예문 (2), (3) 모두 의존명사의 구묶음에 대한 것이다. 그러나 그 처리는 서로 다르다. 예문 (2)의 경우에는 앞의 명사에 구묶음을 하였으며, 예문 (3)의 경우에는 뒤의 명사에 구묶

1) 일반적으로 명사구 특히 기본적인 명사구(base noun phrase)를 구묶음의 단위에 넣는 경우가 있으나 본 논문에서는 이러한 경우는 제외한다.

음을 하였다. 이렇듯 의존명사에 따라서 구묶음을 하는 방식이 달라져야 한다. 본 논문에서는 의존명사의 구묶음에 대한 이러한 세밀한 처리를 제안하였다.

의존명사는 그 앞에 나타나는 단어 즉 관형어에 어떤 미세한 의미를 더하지만 전체적인 관점의 구문분석에서는 무시할 수 있는 것이라 생각된다. 본 연구에서는 의존명사와 이 의존명사를 한정하는 관형어 사이의 관계와 의존명사가 관형어에 부여하는 의미를 파악하여 이에 맞는 처리 방법을 각 의존명사에 대해서 제안한다.

본 논문에서 제안한 의존명사 구묶음 처리 방식을 사용한 구문분석 시스템을 구현하고 이를 이용하여 실험한 결과 의존명사 구묶음을 처리하지 않은 시스템보다 재현율과 정확률이 모두 많이 향상되었으며, 에러감소율이 약 43.6%임을 알 수 있었다.

관련 연구

구묶음이란 개념은 Abney의 영어에 대한 구문 분석기에서 맨 처음 비롯되었다(Abney, 1991). 종래의 형태소분석과 구문분석으로 이루어진 문장 분석과정에서 구문분석을 다시 구묶음 모듈(chunker)과 연결 모듈(attacher)로 나누었다. 사람이 문장을 읽을 때 숨쉼을 넣어 끊어 읽는 덩어리를 구묶음과 대응하였다. 그는 이렇게 두 단계로 구문분석을 나눌 경우 애매성의 해결 면에서 많은 도움을 받을 수 있다고 주장하였다. 구묶음 내의 단어들 사이의 연결 중의성(ambiguity)은 그 내부에 국한되므로 상위의 전역적인 구조 분석 작업인 attacher 가 감당해야 하는 연결 중의성의 정도를 감소시킨다는 것이다. 그러나 Abney의 구묶음들을 살펴보면 본 논문에서 다룬 것과는 달리 주로 명사구와 동사구에 대응되는 것들이다. 예를 들면,

"the effort", "such a conclusion", "two foci",
"the study", "of the rocks"
"to establish", "of course will have".

본 논문에서 다루는 구묶음 기법은 영어에는 없는 의존명사와 관련된 것으로서 Abney의 아이디어를 한국어의 상황에 맞추어 응용한 기술이라고 볼 수 있다.

구묶음과 관련된 영어권에서의 주요한 연구로 Ramshaw와 Marcus의 자동학습 기법에 기반을 둔 구묶음 인식에 관한 것이 있다(Ramshaw & Marcus, 1995). 그들은 Brill 이 제안한 transformation-based learning 기법을 구묶음의 인식에 적용하였다. 그들의 시스템은 구묶음 태깅이 된 것으로 간주되는 학습 말뭉치로 학습기를 학습시킨 후에 이를 새로운 말뭉치에서의 구묶음 인식기로 활용한 것이다. 그들의 연구에서는 주로 비순환 기본 명사구(non-recursive basic noun phrase)를 구묶음으로 인식하는 것이었다. 그 외에 동사와 이에 연결된 단어로 이루어진 덩어리(V-type chunk)를 인식하는 실험도 실행하였다. 명사구의 경우에는 93%, 동사구의 경우에는 88%의 성능을 갖는 시스템을 구현할 수 있었다. 그들 연구의 특징은 Abney가 이용한 문맥자유문법에 기반하여 구묶음을 인식하는 대신 높은 성능을 갖는 구묶음 모듈을 자동학습 기법으로 개발할 수 있게 한 것이다. 이들의 연구도 Abney 에

서와 마찬가지로 한국어의 의존명사 관련 구묶음 기법과는 거리가 있다.

그 외에도 Bourigault(1992) 및 Kupiec(1993)은 유한 상태 오토마타(finite state automata)를 이용하여 구묶음을 인식하는 기법을 개발하였는데 이들 역시 영어에서의 명사구 인식에 국한된 것들이다. Voutilainen(1993)은 품사 태그와 유사한 구묶음 태그를 도입하여 문장 내의 구묶음의 구조를 나타낼 수 있도록 하였다. 사전에 있는 각 단어에 가능한 구묶음 태그 및 제한된 문법 패턴을 붙여 놓고 이를 이용하여 명사구를 인식할 수 있도록 하였다. 그의 시스템의 성능은 98.5%의 재현율 및 95%의 정확률을 갖는 것으로 보고하였으나 그의 시스템은 많은 다양한 종류의 명사구를 처리하지는 못한다는 단점을 가지고 있다.

한국어와 관련한 구묶음에 관한 연구로 규칙에 기반하여 명사구를 인식한 후 그 다음 이를 기반으로 동사구를 인식하는 시스템을 개발하였다(신효필, 1999). 윤준태(1999)의 연구에서는 먼저 문맥자유문법으로 명사구를 인식하고 연어 패턴 정보를 이용하여 동사구를 인식하는 시스템을 소개하였다. 이들의 연구에서는 다양한 형태의 명사구나 동사구를 처리하지 못하는 단점을 가지고 있다. 이들보다 보다 정교한 구묶음 시스템으로 김미영의 연구가 있다(김미영 등,2000). 여기에서는 명사구의 인식을 위해 전이망(transition network)을 이용하였다. 여기에서의 특징은 관형형 어미를 가진 용언을 명사구 구묶음에 포함시키고 있는데 타동사 용언이 관형형 용언인 경우에는 구묶음을 할 수 없는 상황이 자주 나타날 수 있다는 문제점이 있다. 동사 구묶음과 관련해서는 유한 상태 오토마타를 이용하여 인접한 동사들을 하나로 묶는 경우와, 논항과 동사를 구묶음하는 경우가 있다. 여기서의 문제점은 후자의 경우는 구묶음 단계보다는 구문분석 단계에서 수행하는 것이 더 적절하다는 데 있다. 이 연구에서 사용한 전체 실험 문장의 수는 200 문장으로서 신뢰성있는 실험 결과를 얻기에는 부족한 점이 있다.

지금까지 수행된 한국어와 관련된 구묶음에 관한 연구들에서는(영어에 관한 연구들과 마찬가지로) 본 논문에서 소개하는 한국어 의존명사와 관련된 구묶음에 관한 것은 찾아 볼 수 없다. 본 연구에서는 한국어 구문분석의 성능을 향상시키는데 있어서 한국어 의존명사와 관련된 구묶음도 매우 효과가 있음을 처음으로 제시하였다.

기존에 알려진 한국어 구묶음 처리

여기서는 기존에 이미 알려진 단순한 구묶음 방법을 설명한다. 이러한 기법은 과거부터 많이 이용되는 기법으로서 본 연구에서도 구문분석 시스템의 성능향상을 위해 그대로 도입하여 이용한다.

복합명사에서의 구묶음 처리

복합명사들을 구성하는 단어들을 합하여 하나의 단위로 만든다. 단위의 마지막 명사가 헤드가 되어 복합명사를 대표하게 된다. 여기에서 복합명사란 일반 명사나 고유 명사들이 조사가 붙지 않은 형태로 연속해서 나타나는 것을 말한다. 마지막 명사에는 조사나 동사화 접미사가 붙는다. 예를 들면 다음과 같다.

(4) **유럽 중앙 은행**은 유로의 **공식 통용**에 앞서 ...

--> **은행**은 유로의 **통용**에 앞서 ...

└>유럽 중앙 └>공식

(5) 월드컵과 **부산 아시아 경기**의 성공은 ...

--> 월드컵과 **경기**의 성공은 ...

└>부산 아시아

(6) **동아 일보 신년 여론 조사**에서 **한나라당 이회창 총재**가 **양자 대결**할 경우...

┌>한나라당 이회창

--> **조사**에서 **총재**가 **대결**할 경우...

└>동아 일보 신년 여론 └>양자

위의 예문 (6)과 같이 명사에 동사화 접미사가 붙은 경우에도 앞의 명사와 함께 복합명사 처리를 한다. 복합명사 내부 단어 간의 관계는 본 논문에서는 분석하지 않는다. 왜냐하면 복합명사 내부의 단어 간의 관계에 대한 분석은 단어들의 의미를 고려하여야 하며 구문분석보다 더 상위 수준의 분석 단계에서 처리하는 것이 좋다.

본용언/보조용언에 대한 구묶음 처리

보조용언은 자체적으로 독립적인 의미를 갖지 않고 본용언의 의미에 추가적인 성질을 부여하는 역할만 함으로 본용언에 부속시킨다 (황이규 등, 2000). 본용언과 보조용언을 구묶음 할 때 어간은 본용언의 어간을 사용하고 어미는 보조용언의 어미를 사용한다. 예를 들면 다음과 같다.

(7) 사회는 불공정하다고 생각하고 있음을 보여주고 있다

사회는 불공정하다고 **생각하/VV+고/EE 있/VX+음/ETN+을/JK 보여주/VV+고/EE 있/VX+다/EE**[2)]

┌>있/VX

—> 사회는 불공정하다고 **생각하/VV+음/ETN+을/JK 보여주/VV+다/EE**

└>있/VX

(8) 심각한 정치적 진공 상태로 치닫고 있다

심각한 정치적 진공 상태로 **치닫/VV+고/EE 있/VX+다/EE**

┌>있/VX

--> 심각한 정치적 진공 상태로 **치닫/VV+다/EE**

(9) 무정부 상태부터 해결해야 할 총체적인 위기에 빠졌다

무정부 상태부터 **해결하/VV+아야/EE 하/VX+ㄹ/ETM** 총체적인 위기에 빠졌다

--> 무정부 상태부터 **해결하/VV+ㄹ/ETM** 총체적인 위기에 빠졌다 └>하/VX

(10) 남북 관계를 발전시켜 나가도록 하겠다

남북 관계를 **발전시키/VV+어/EE 나가/VX+도록/EE 하/VX+겠/EP+다/EE**

--> 남북 관계를 **발전시키/VV+다/EE**

└>나가/VX 하/VX

2) 본 논문에서 사용하는 품사 태그셋은 다음과 같다. NNG(명사), NNB(의존명사), VV(동사), VA(형용사), VX(보조용언), ETM(관형형어미), ETN(명사형전성어미), JK(조사), EE(어미), EP(선어말어미), SN(숫자).

위의 예문 (10)과 같이 보조용언이 연이어 나오면 마지막에 나오는 보조용언의 어미를 구묶음의 어미로 사용한다.

단순한 경우의 의존명사 관련 구묶음 처리

의존명사와 관련하여 구묶음을 하는 것은 과거에도 단순한 경우에 대하여 많이 이용되어 왔다. 이러한 경우는 보통 의존 명사가 그 앞에 나타나는 용언에 구묶음 되는 단순한 경우만을 고려하였다. 여기에서 취급되는 의존 명사로는 다음과 같은 것들이 있다.

단순 구묶음 의존명사	수, 리, 것

이들과 관련한 구묶음의 예를 들면 다음과 같다.

(11) 그 사과는 먹을 수 있는 사과가 아니다
그 사과는 **먹/VV+을/ETM 수/NNB 있/VX+는/ETM** 사과가 아니다
--> 그 사과는 **먹/VV+는/ETM** 사과가 아니다
└>수/NNB 있/VX

(12) 그런 말을 할 리가 없다
그런 말을 **하/VV+ㄹ/ETM 리/NNB+가/JK 없/VX+다/EE**
--> 그런 말을 **하/VV+다/EE**
└>리/NNB 없/VX

(13) 국운 융성의 발판이 될 것이다
국운 융성의 발판이 **되/VV+ㄹ/ETM 것/NNB+이/VC+다/EE**
--> 국운 융성의 발판이 **되/VV+다/EE**
└>것/NNB

위의 예문 (11)에서 의존 명사 "수"는 용언 "먹다"의 의미를 좀 더 강조한다고 할 수 있다. 따라서 의존 명사 "수"를 앞의 용언 "먹다"에 구묶음을 해도 전체 문장에 영향을 주지 않는다. 이 예문에서는 의존 명사 "수" 다음에 보조용언 "있다"가 있기 때문에 "먹을+수+있는"으로 구묶음이 된다. "먹을+수+있는"은 구문분석 시스템이 볼 때에는 "먹+는"으로 보이도록 한다.

위의 예문 (13)에서 의존 명사 "것"은 의존 명사 "수"와 마찬가지로 앞의 용언 "되다"의 의미를 좀 더 강조한다고 할 수 있다. 그러나 의존 명사 "수"와는 형태가 다르게 지정사 "이"가 붙어 있는 형태이다. 의존 명사에 지정사 "이"가 붙어서 서술성을 갖는 경우에는 보조용언과 같은 쓰임새를 갖는다. 따라서 보조용언의 구묶음과 같은 방식으로 처리한다. 위의 예문 (12)와 같이 의존 명사 "리"도 의존 명사 "수"와 같은 형태로 구묶음 처리가 된다.

의존명사와 관련한 새로이 제안되는 구묶음 기법

우리는 본 장에서 구문분석 시스템의 성능 향상을 위해서 새로이 추가적으로 제안하는 구묶음 기법을 소개한다. 우리는 되도록 많은 의존명사들에 대하여 이를 앞이나 뒤에 나오는 내용어(체언 또는 용언)의 일부로 구묶음 되도록 함으로써 구문분석 시스템의 부담을

경감시키도록 노력하였다.

의존 명사란 자립성이 없는 특수한 명사를 일컫는다. 곧 그 앞에 어떤 한정 성분이 나타나지 않으면 홀로 쓰일 수 없는 비자립적인 명사라는 것이다. 그 앞에 나타나는 성분은 관형어 곧 관형사나 관형사 기능을 지닌 한정어이다. 모든 명사는 관형어와 어울릴 수 있는 구문론적 특성을 지니게 되는데 특히 의존명사도 그 앞에 관형어를 반드시 수반하게 되는 것이다.

의존명사는 일반 명사에 비해 그 어휘적 의미가 비교적 불분명하고 문법적인 기능면이 두드러지는 경향이 있다. 예를 들면, "것"은 본디 '물건'이라는 뜻이 있기는 하나 특정한 사물을 지시하는 경우보다는 "좋은 것", "먹을 것" 에서와 같이 사물을 두루 가리키는 대명사적 기능이 짙다. 더구나 여기에서의 의존명사 "것"은 "갈 것이다", "들어가지 말 것" 등과 같은 쓰임에서는 문법적인 기능을 나타내는 데에 깊이 관련되어 있다. 이런 사정은 다른 의존명사의 경우에도 비슷하다. 의존명사는 크게 두 갈래로 나누어 볼 수 있다(서정수, 1994).

• 단위 의존명사: "마리, 자루, 장" 등과 같이 수효나 분량을 헤아릴 때에 수사 아래에 쓰이는 것이다. 단위 의존명사는 선행 관형어인 수사와 어울려 수량을 나타내는 구실을 한다. 단위 의존명사는 선행어가 수사 또는 수량 표시어라는 한정성이 있는 점에서 비단위 의존명사와 다르다. 대개 단위 의존명사들은 수량을 나타내는 동량형의 단위 체계를 이루는 특징이 있다.

• 비단위 의존명사: "것, 만큼, 나름, 노릇" 등과 같이 흔히 의존명사로 알려진 것들이다. 의존명사라 하면 대개 이 무리를 가리킨다. 이것은 위의 단위 의존명사가 아닌 의존명사를 모두 가리킨다.

단위 의존명사 구묶음

단위 의존명사는 선행 관형어인 수사와 어울려 수량이나 시간, 온도 등을 나타내는 구실을 하는 의존명사이다. 본 논문에서 준비한 단위의존명사에 대한 규칙 R-1 ~ R-6 은 모든 단위성 의존명사를 해결할 수 있도록 고안되었다. 단, 모든 단위성 의존명사들이 저희의 의존명사사전에 등록되어 있어야 한다. 우리는 가능한 모든 단위성 의존명사를 수집하기 위해 국어문법 책, 세종말뭉치, 금성사 국어사전을 참조하였다. 현재 우리 시스템의 사전에는 624개 단위성 의존명사가 등록되어 있으며 이들 모두(100%)에 대해서 구묶음 처리가 가능하도록 되어 있다.

수량 단위 의존명사의 처리

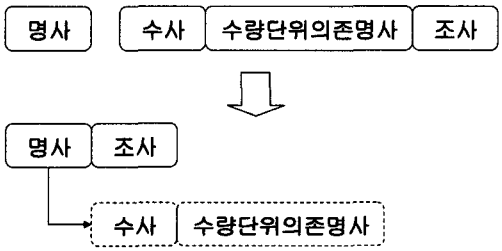
수량 단위 의존명사는 수사와 어울려 앞에 나오는 명사의 수를 나타내는 역할을 한다. 수량 단위 의존명사로는 "개", "척", "벌", "명", "마리" 등 610여 개가 있다.

예문 (14)는 "길수는 사과를 먹었다"라는 것에 부가적으로 몇 개를 먹었는지를 나타내기 위해서 "열개"라는 "수사+수량단위의존명사"를 추가한 것이라 볼 수 있다. 즉 의미를 더하기 위해서 사용한 "열개"라는 어절은 굳이 구문분석을 할 필요가 없는 어절이다. 이러한 의존명사 어절이 구문분석을 복잡하게 한다.

이와 같은 형태는 규칙 R-1과 같이 처리한다.

R-1	명사 $Z+X_c$+조사 --> 명사$(Z+X_c)$+조사 Z: 수사나 숫자 X_c: 수량단위 의존명사

위 규칙에서 스페이스 글자는 어절을 구분하며, +기호는 동일 어절 안의 형태소의 연결을 나타낸다. 괄호는 구묶음되어 숨겨지는 것을 나타낸다.



위의 규칙 R-1을 이용하면 예문 (14)는 예문 (14')과 같이 구묶음이 된다.

(14) 길수는 사과 열개를 먹었다.

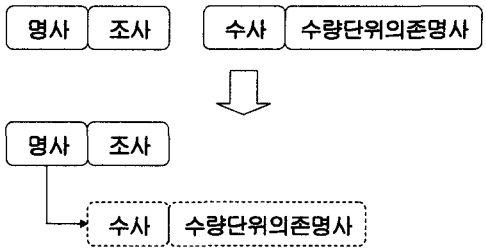
길수는 **사과**/NNG **열**/NR+**개**/NNB+**를**/JK 먹었다.

(14') 길수는 **사과**/NNG+**를**/JK 먹었다.

└>열/NR+개/NNB

예문 (15)는 조사가 앞의 명사에 붙어 있는 경우이다. 이와 같은 형태는 다음과 같이 처리한다.

R-2	명사+조사 $Z+X_c$ --> 명사$(Z+X_c)$+조사 Z: 수사나 숫자 X_c: 수량단위 의존명사



위의 규칙 R-2를 이용하면 예문 (15)는 예문 (15')과 같이 구묶음이 된다.

(15) 철수는 배를 다섯척 가지고 있다.

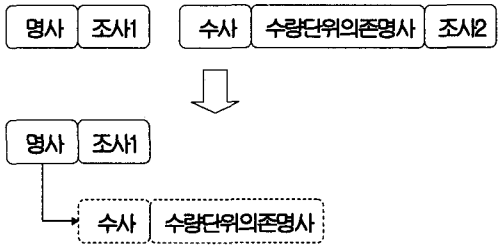
철수는 **배**/NNG+**를**/JK **다섯**/NR+**척**/NNB 가지고 있다.

(15') 철수는 **배**/NNG+**를**/JK 가지고 있다.

└>다섯/NR+척/NNB

예문 (16)은 명사와 의존명사에 모두 조사가 붙어 있는 경우이다. 이와 같은 형태는 다음과 같이 처리한다.

R-3	명사+조사1 $Z+X_c$+조사2 --> 명사$(Z+X_c)$+조사2 Z: 수사나 숫자 X_c: 수량단위 의존명사



위의 규칙 R-3을 이용하면 예문 (16)은 예문 (16')과 같이 구묶음이 된다.

(16) 영희는 운동화를 다섯켤레를 샀다.

영희는 **운동화/NNG+를/JK 다섯/NR+켤레/NNB+를/JK** 샀다.

(16') 영희는 **운동화/NNG+를/JK** 샀다.
└>다섯/NR+켤레/NNB

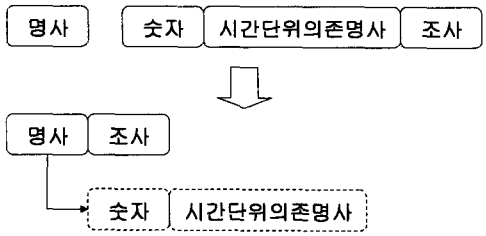
X_c (수량 단위 의존 명사)	자, 치, 푼, 마, 리, 마장, 발, 뼘, 간, 평, 마지기, 정보, 섬, 가마니, 푸대, 말, 되, 홉, 통, 동이, 잔, 병, 접시, 그릇, 양, 돈, 푼, 근, 관, 전, 원, 개, 낱, 가지, 그루, 포기, 자루, 켤레, 채, 대, 척, 장, 권, 편뭇, 짐, 쌈, 두름, 돗, 쾌, 손, 꾸러미, 점, 바퀴, 단, 뭇, 다발, 자, 번, 차, 회, 판, 건, 사람, 분, 명, 인, 마리, 필, 두 등 (총 610여 개)

시간 단위의존명사의 처리

시간 단위 의존명사는 "오전", "오후", "아침", "저녁", "작년" 등 때를 나타내는 명사와 어울려 좀 더 정확한 시간을 나타내는 역할을 한다. 시간 단위 의존명사로는 "년", "월", "일", "시", "분", "초" 등이 있다.

예문 (17)은 "철수는 공항에서 영희를 만났다"라는 것에 부가적으로 언제 만났는지를 나타내기 위해서 "오전 10시에"라는 "오전(NNG) 숫자(SN)+시간단위의존명사(NNB)"를 추가한 것이라 볼 수 있다. 이러한 명사구에서 "10시"는 "오전"이라는 명사에 의미를 더하는 역할을 한다고 할 수 있다. 따라서 이와 같은 형태는 다음과 같이 처리한다.

R-4	Y_t Z+X_t+조사 --> $Y(Z+X_t)$+조사 $Y_t \in$ {오전, 오후, 아침, 저녁} Z: 수사나 숫자 X_t: 시간단위 의존명사



규칙 R-4에서 "오전"이라는 명사 이외에 "오후", "아침", "저녁" 등이 올 수 있다. 위의 규칙 R-4를 이용하면 예문 (17)은 예문 (17')과 같이 구묶음이 된다.

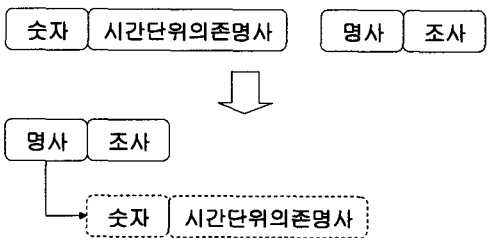
(17) 철수는 오전 10시에 공항에서 영희를 만났다.

철수는 **오전/NNG 10/SN+시/NNB+에/JK** 공항에서 영희를 만났다.

(17') 철수는 **오전/NNG+에/JK** 공항에서 영희를 만났다. └>10/SN+시/NNB

예문 (18)은 "2002년"이라는 연도를 나타내는 구절이 뒤에 나오는 "월드컵"이라는 행사를 수식하는 경우를 나타낸다. 이와 같은 형태는 다음과 같이 처리한다.

R-5	Z+X_t Y_t+조사 --> $Y_t(Z+X_t)$+조사 Z: 수사나 숫자 X_t: 시간단위 의존명사 $Y_t \in$ {올림픽, 아시안게임, ...}: 행사를 나타내는 명사



규칙 R-5에서 "월드컵"이라는 명사 이외에 "올림픽", "아시안게임" 등의 행사를 나타내는 명사들이 올 수 있다. 규칙 R-5를 이용하면 예문 (18)은 예문 (18')과 같이 구묶음이 된다.

(18) 한국은 2002년 월드컵에서 4강에 올랐다.

한국은 **2002/SN+년/NNB 월드컵/NNG+에서/JK** 4강에 올랐다.

(18') 한국은 **월드컵/NNG+에서/JK** 4강에 올랐다. └>2002/SN+년/NNB

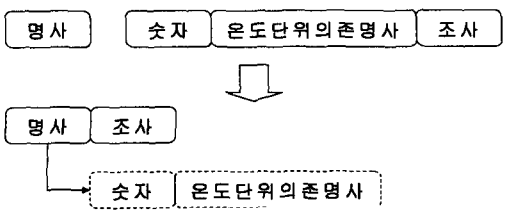
X_t (시간단위 의존명사)	년, 월, 일, 시, 분, 초, 세기

온도 단위의존명사의 처리

온도 단위 의존명사는 "영상", "영하", "섭씨" 등 온도를 나타내는 명사와 어울려 좀 더 정확한 온도를 나타내는 역할을 한다. 온도 단위 의존명사로는 "도"가 있다.

예문 (19)에서 "영하 5도"라는 명사구에서 "5도"라는 어절은 "영하"라는 어절의 의미를 좀더 구체적으로 표현하는 것으로 볼 수 있다. 이러한 경우 "5도"를 "영하"에 구묶음 한다. 이와 같은 형태는 다음과 같이 처리한다.

R-6	Y_o Z+X_o+조사 --> Y_o(Z+X_o)+조사 Y_o ∈ { 영상, 영하, 섭씨, 화씨 } Z: 수사나 숫자 X_o: 온도단위 의존명사



위의 규칙 R-6에서 "영하"라는 명사 이외에 "영상", "섭씨" 등이 올 수 있다. 규칙 R-6을 이용하면 예문 (19)는 예문 (19')과 같이 구묶음이 된다.

(19) 내일 서울의 아침 기온은 영하 5도로 떨어질 것으로 예상된다.

내일 서울의 아침 기온은 **영하/NNG 5/SN+도/NNB+로/JK** 떨어질 것으로 예상된다.

(19') 내일 서울의 아침 기온은 **영하/NNG+로/JK** 떨어질 것으로 예상된다. └>5/SN+도/NNB

X_o	도

단위 의존명사에 대한 처리는 주로 주변에 나타나는 명사에 "숫자(또는 수사)+단위의존명사"를 구묶음을 시키는 형태로 처리된다. 이러한 구묶음 처리를 통해서 단위의존명사에 의해 나타나는 다양한 구문 구조를 단순화 시킬 수 있었다.

비단위 의존명사 구묶음

비단위 의존명사란 의존명사 중 단위 의존명사를 제외한 나머지 의존명사를 말한다. 특히 비단위 의존명사는 문장 내에서 매우 다양한 문형을 일으킨다. 따라서 이러한 다양한 경우에 대한 고찰이 없이 일률적인 처리를 시도하면 올바른 구문 구조를 생성할 수 없게 된다.

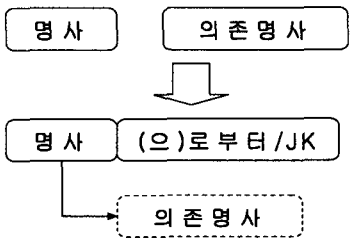
우리는 먼저 모든 가능한 비단위성 의존명사를 수집하는 작업을 수행하였다. 그 결과 총 145개의 비단위성 의존명사가 존재하는 것

으로 판명되었고 이들을 시스템 사전에 수집 및 등록하였다. 그리고 수집된 각 비단위성 의존명사 하나하나에 대하여 구묶음 필요성을 검토하였다. 그 결과 60개에 대하여 구묶음이 필요한 것으로 판명되었고 이들에 대해서 아래의 규칙 R-7 ~ R-12를 구축하게 되었다. 60개에 포함되지 않은 85개의 의존명사들은 (구묶음이 필요 없기 때문에) 일반명사처럼 처리되고 있다. 우리는 구묶음이 필요한 모든 비단위성 의존명사에 대해서 필요한 구묶음 처리가 수행되도록 하였다.

앞 명사의 조사로 처리

의존명사를 앞 명사의 조사로 처리하는 경우이다. 특히 조사 중에서도 부사격 조사로 처리한다. 예문 (20)에서 의존명사 "이래"는 "건국"이라는 단어에 "으로부터"라는 조사의 의미를 더하는 역할을 한다. 따라서 의존명사 "이래"를 앞 명사 "건국"에 구묶음을 시키면서 조사 "으로부터"를 추가한다.

R-7	명사 X_7 --> 명사(X_7)+(으)로부터/JK $X_7 \in$ { 이래, 겸 }



위의 규칙 R-7을 이용하면 예문 (20)은 예문 (20')과 같이 구묶음이 된다.

(20) 건국 이래 처음 있는 일이다.

건국/NNG 이래/NNB 처음 있는 일이다.

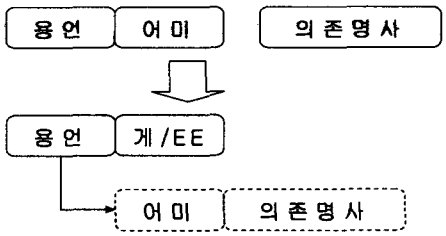
(20') **건국/NNG+으로부터/JK** 처음 있는 일이다. └>이래/NNB

앞 용언을 부사로 처리

의존명사 앞의 용언을 부사화 시킨다. 즉 의존명사가 앞 용언에 구묶음이 되면서 용언이 부사의 성격을 가지게 하는 것이다.

예문 (21)에서 "먹을 만큼"이 "먹었다"를 수식하는 것이다. 따라서 "만큼"을 "먹을"에 구묶음을 하면서 "먹을"을 부사형으로 만든다. 예문 (22)도 마찬가지로 "배운 대로"가 "해라"를 수식하는 것이다. 따라서 "대로"를 "배운"에 구묶음을 하면서 "배운"을 부사형으로 만든다. 부사형으로 만들 때는 부사형 어미 "게"를 이용한다. 이를 정리하면 다음과 같다.

R-8	용언+어미 X_8 --> 용언(X_8)+게/EE $X_8 \in$ {만큼, 대로, 채, 채로, 척, 체, 나위}



예문 (21), (22)에 규칙 R-8을 적용하면 예문 (21'), (22')과 같이 구묶음이 된다.

(21) 철수는 사과를 먹을 만큼 먹었다.

철수는 사과를 **먹/VV+을/ETM 만큼/NNB** 먹었다.

(21') 철수는 사과를 **먹/VV+게/EE** 먹었다.

└>만큼/NNB

위의 예문 (21')에서 부사형 어미 "게"를 붙임으로써 어감이 자연스럽지 못하다. 그러나 이것은 어절 "먹을"이 나중에 나오는 용언을 (여기서는 "먹었다") 부사적으로 수식하도록 하기 위한 조치일 뿐이다. 이렇게 하면 구문 분석기로 하여금 앞 용언이 나중에 나오는 용언을 부사적으로 수식하는 분석을 수행하게 유도하는 효과를 가진다.

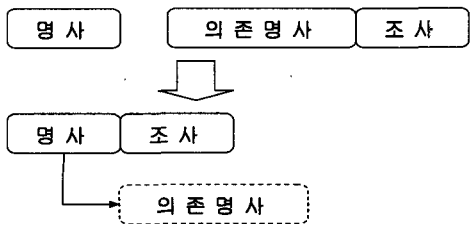
(22) 너는 배운 대로 해라.
너는 **배우/VV+ㄴ/ETM 대로/NNB** 해라.
(22') 너는 **배우/VV+게/EE** 해라.
ㄴ>대로/NNB

앞 명사에 구묶음 처리

앞 명사에 의존명사를 구묶음을 시킨다. 앞 명사에 조사가 없고 의존명사에 조사가 있는 경우에는 의존명사에 붙어 있는 조사를 앞 명사에 전달한다.

예문 (23), (24)에 나타난 의존명사 "따위"나 "등"은 앞에 나열된 명사들을 강조하는 역할을 한다. 따라서 앞 명사에 구묶음이 되어도 전체적인 문장 구조에 영향을 미치지 않는다. 이에 대한 처리는 다음과 같다.

R-9	명사 X_9+조사 --> 명사(X_9)+조사 X_9∈ {따위, 등, 중, 가량, 간, 내, 들, 등등, 등지, 말, 조, 짜리, 쯤, 측, 나마, 등속, 초, 태미, 템, 가운데, 편, 폭, 바람, 무렵, 나절, 줄, 물}



위의 규칙 R-9를 이용하면 예문 (23), (24)은 예문 (23'), (24')과 같이 구묶음이 된다.

(23) 쌀, 보리, 밀 따위가 곡식이다.
쌀, 보리, **밀/NNG 따위/NNB+가/JK** 곡식이다.
(23') 쌀, 보리, **밀/NNG+가/JK** 곡식이다.
ㄴ>따위/NNB

(24) 가축에는 소, 말, 돼지 등이 있다.
가축에는 소, 말, **돼지/NNG 등/NNB+이/JK** 있다.
(24') 가축에는 소, 말, **돼지/NNG+이/JK** 있다.
ㄴ>등/NNB

앞 용언에 구묶음 처리

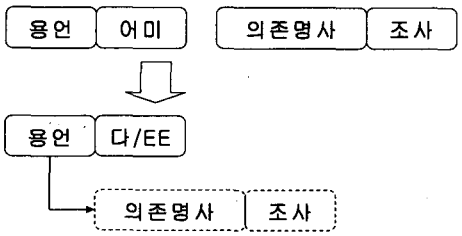
의존명사를 앞 용언에 구묶음을 하는 것이다. 의존명사에는 조사가 붙어 있을 수도 있고, 또는 지정사와 어미가 붙어 있을 수도 있다. 각각에 대한 처리 방법은 다르다.

예문 (25)의 경우에는 의존명사 "즈음"을 용언 "떠나"에 구묶음을 한다. 또한 "집을 떠날 즈음에"가 문장 내의 종속절로 해석되어야 하므로 이를 위해 용언 "떠나"의 어미를 종결형 어미 "다"로 바꾼다. 처리 방법은 다음과 같다.

R-10	용언+어미 X_{10}+{조사|ε} --> 용언(X_{10})+다/EE X_{10} ∈ {마련, 뿐, 십상, 일쑤, 때문, 즈음, 참, 통, 따름, 즘, 겸, 만큼, 채, 채로, 척, 체, 편, 폭, 바람, 무렵, 나절, 나위, 데}

위 규칙에서 기호 '|' 는 OR 연산의 의미를 나타내며 ε 은 아무 것도 나타나지 않음을

나타낸다.



예문 (25)를 규칙 R-10을 이용하여 처리하면 예문 (25')과 같다.

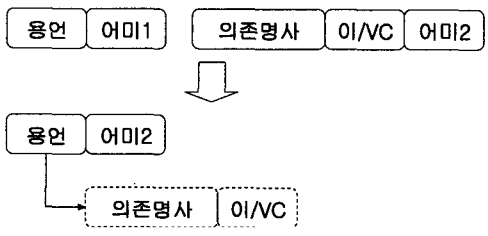
(25) 집을 떠날 즈음에 철수에게서 전화가 왔다.

집을 **떠나/VV+ㄹ/ETM 즈음/NNB+에/JK** 철수에게서 전화가 왔다.

(25') 집을 **떠나/VV+다/EE** 철수에게서 전화가 왔다. ㄴ>즈음/NNB(종속절화)

예문 (26)의 경우에는 의존명사에 지정사가 붙어 있는 경우이다. 이 경우에는 본용언/보조용언의 처리 방식과 유사하게 처리한다. 즉, 의존명사를 앞 용언에 구묶음을 하면서 앞 용언의 어미를 의존명사가 갖고 있는 어미로 바꾼다. 처리 방법은 다음과 같다.

R-11	용언+어미1 X_{11}+지정사+어미2 --> 용언(X_{11}+지정사)+어미2 X_{11} ∈ {마련, 뿐, 십상, 일쑤, 때문, 즈음, 참, 통, 따름, 즘, 꺼, 것, 게, 거, 셈}



예문 (26)을 규칙 R-11을 이용하여 구묶음을 하면 예문 (26')과 같다.

(26) 학교 공부만 열심히 했을 뿐이다

학교 공부만 열심히 **하/VV+았/EP+을/ETM 뿐/NNB+이/VC+다/EE**

(26') 학교 공부만 열심히 **하/VV+다/EE**

ㄴ>뿐/NNB+이/VC

다음 규칙 R-12 는 본 용언과 보조 용언 사이에 나오는 의존명사들을 처리하는 구묶음을 위한 것으로서 앞의 3.3절에서 소개한 기존의 의존명사 구묶음 작업과 같은 역할을 한다.

R-12	용언+어미1 X_{12}+{조사|ε} {보조용언|지정사}+어미2 --> 용언(수사+X_{12})+어미2 X_{12}∈ {수, 리, 것, 꺼, 게, 거, 만, 성, 성상, 줄, 데}

충돌 해소 규칙

지금까지 소개한 규칙들이 충돌이 일어나는 경우가 있다. 즉 주어진 의존명사에 대하여 2개 이상의 규칙의 조건부가 매칭되어 어느 규칙을 선택할지 중의성이 나타나는 경우가 있다.

(27) 철수는 사과를 **먹/VV+을/ETM 만큼/NNB** 먹었다.

(28) 발이 **크/VA+ㄴ/ETM 만큼/NNB** 신도 크다.

위의 예문 (27)이나 (28)에서는 의존명사 "만

큼"으로 인하여 규칙 R-8 과 규칙 R-10 의 조건이 동시에 매칭된다. 이 경우 둘 중 하나를 선택하여 수행하여야 하는 데 이것을 규칙 충돌의 해소라 부르자. 예문 (27)에서는 규칙 R-8 이 수행되어야 하지만 (28)의 경우에는 규칙 R-10 이 선택되어 수행되어야 한다. 이러한 규칙 충돌 현상은 여러 의존명사에서 발생하는데 이의 해소를 위하여 우리는 충돌해소-규칙(collision resolution rule; CRR)을 이용하도록 한다. 충돌해소-규칙은 표 1과 같다.

위에서 설명한대로 위의 예문 (27)에서는 R-8, R-10 두 개의 규칙의 조건부가 매칭에 성공한다. 그러나 의존명사 "만큼" 뒤에 용언이 왔으므로 충돌해소 규칙 CRR-1에 의하여 R-8 이 선정되어 수행된다. 그러나 예문 (28)의 경우에는 매칭에 성공한 R-8, R-10 중에서 R-10 이 CRR-1 에 의해 선택된다.

본 논문에서 이용하는 모든 충돌해소 규칙을 표 1에서 볼 수 있다. 지금까지의 조사에 의하면 3개 이상의 규칙의 매칭시키는 의존명사는 없는 것으로 판단된다.

<표 1> 충돌해소 규칙

규칙명	규칙 내용
CRR-1	if 의존명사 ∈ {만큼,채,채로, 척,체} { if (뒤 어절== 용언) [R-8] else if (뒤 어절== 명사) [R-10] }
CRR-2	if 의존명사 ∈ {꺼,것,게,거,셈} { if (뒤 어절==보조용언) [R-12] else if (의존명사뒤에 지정사가 붙음) [R-11] }
CRR-3	if 의존명사 ∈ {나위} { if (뒤 어절== "없이") [R-8] else if (뒤 어절==보조용언) [R-10] }

구묶음의 이점

의존명사는 종류가 매우 다양하다. 이렇게 다양한 종류의 의존명사를 처리하는 방법으로 본 논문에서는 구묶음이란 방법을 제안하고 이를 구문분석 시스템과는 별도로 동작하도록 하였다. 구문분석 시스템에서 이러한 구묶음 작업을 처리할 수도 있으나 그렇게 되면 구문분석 시스템이 복잡하게 되고 새롭게 추가되거나 처리방식이 바뀌게 되는 의존명사에 대해서 구문분석 시스템을 변경해야 하는 단점이 있다. 이러한 이유로 구묶음을 구문분석 시스템과 별도로 동작하도록 하였다.

구묶음을 하게 되면 구문분석 시스템의 복잡성을 줄일 수 있다. 예를 들면 다음과 같다.

(29) 길수는 사과를 10개 먹었다
길수/NNP+는/JK 사과/NNG+를/JK 10/SN+개/NNB 먹/VV+었/EP+다/EE

(30) 길수는 사과 10개를 먹었다
길수/NNP+는/JK 사과/NNG 10/SN+개/NNB+를/JK 먹/VV+었/EP+다/EE

위의 예문 (29)를 구묶음을 하지 않은 상태로 처리하기 위해서 구문분석 시스템은 용언 앞에 오는 의존명사가 조사를 갖지 않는 경우에 대해서 처리하는 방식이 정의되어야 한다. 또한 예문 (30)을 구묶음을 하지 않은 상태로 처리하기 위해서 구문분석 시스템은 용언 앞에 오는 의존명사가 조사를 갖는 경우에 대해서 처리하는 방식이 정의되어야 한다. 또는 일반명사와 의존명사가 결합하는 방식에 대한 정의가 필요하다. 이러한 처리 방식을 모든

의존명사에 대해서 정의하여야 하는데 이를 구문분석 단계에서 정의하게 되면 구문분석 시스템이 너무 복잡하게 된다.

위의 예문 (29), (30)을 본 논문에서 제안하는 방식으로 구묶음을 하면 다음 예문 (29'), (30')과 같다.

(29') 길수는 사과를 먹었다.
└>10개

(30') 길수는 사과를 먹었다.
└>10개

위의 예문 (29'), (30')을 보면 알 수 있듯이 예문 (29), (30)을 구묶음을 한 결과가 서로 같음을 알 수 있다. 예문 (29), (30)을 구묶음을 한 후에는 구문분석 시스템에 입력되는 형태가 같기 때문에 같은 방식으로 처리가 된다. 즉, 구문분석 시스템은 의존명사가 사용된 형태에 영향을 받지 않음을 알 수 있다.

위와 같이 구묶음을 하게 되면 서로 다른 형태로 사용된 의존명사에 대해서 동일한 형태로 변환시켜 주기 때문에 구문분석 시스템이 의존명사에 따라서 복잡해질 필요가 없다.

```
;철수는 천벌을 받을 놈이다.
55 4
1 4 S
2 3 O
3 4 S
;저기 뛰어오는 놈이 제 막내아들입니다.
56 5
1 2 A
2 3 S
3 5 S
4 5 M
```

(그림 2) 의존 관계 태깅 문서 예

실험 및 성능 평가

실험 환경

국내에서 공식적으로 개발된 한국어 구문트리 부착 말뭉치로 세종 계획에서 구축된 것이 있다. 이는 본 연구의 품사 태그 셋과 격 관계의 표시 방법 등에 차이가 있다. 따라서 세종 계획의 구문트리 부착 말뭉치는 본 연구에서 개발하는 시스템의 실험 및 평가에 적용하기 어려운 점이 많다. 따라서 본 연구에서는 개발하는 시스템에 적용하기 적합한 구문트리 말뭉치를 구축하였다. 본 연구에서 사용한 실험 데이터 즉, 말뭉치는 동아일보 2002년 기사이다. 현재까지 1,000문장에 대해서 의존 관계 태깅을 수행하였다. 이 테스트 말뭉치 안에는 1,150 개의 의존명사가 출현하였다 (표 2). 구묶음 수가 의존명사 수 보다 큰 이유는 복합명사나 본용언/보조용언과 관련된 구묶음이 많이 있기 때문이다.

의존 관계 태깅 문서는 그림 2와 같은 형식으로 되어 있다. 그림 2에서 ';'로 시작하는 행은 실험 문장이다. 그 다음 행의 두 개의 숫자는 문장번호와 어절의 수를 나타낸다. 그

<표 2> 테스트 데이타

문장 수	1,000
의존명사 수	1,150
구묶음 수	3,658

다음 행부터는 실험 문장에 나타난 의존 관계를 나타낸다. 의존 관계의 첫 번째 숫자는 수식을 하는 어절의 번호이고, 두 번째 숫자는 수식을 받는 어절의 번호이다. 세 번째에 있는 문자는 어절 간의 격 관계를 나타낸다. 격 관계로는 S(주격), O(목적격), A(부사격), M(관형격), X(무격)이 있다.

실험은 그림 2와 같이 의존관계가 태깅된 1,000문장에 대하여 실험하였다. 먼저 품사가 부착되지 않은 1,000문장을 형태소 및 품사 태깅 시스템을 이용하여 품사가 부착된 형태의 문장으로 변환한다. 품사가 부착된 문장들은 구묶음 시스템을 거치면서 복합명사, 본용언/보조용언, 의존명사 등 구묶음이 필요한 부분들에 대해서 구묶음이 된 형태로 변환된다. 이렇게 구묶음이 된 문장들이 구문분석 시스템의 입력으로 주어진다. 구문분석 시스템은 구묶음이 된 문장들의 각 어절들에 대해서 의존관계를 알아낸다. 분석결과는 위의 의존관계 태깅 문서와 같은 형태로 저장된다.

실험결과는 정답이 태깅된 문서와 구문분석 시스템이 내준 결과 문서를 이용하여 측정하였다. 실험결과로는 재현율, 정확률, F-score, 에러감소율 등을 측정하였다. 실험결과 측정은 어절의 의존관계가 올바른지 아닌지에 대하여 수행하였다.

이러한 실험을 의존명사 구묶음을 하지 않는 구문분석 시스템, 기존의 의존명사 구묶음 처리 방식만 사용하는 구문분석 시스템, 본 논문에서 제안한 모든 의존명사 처리 방식을 사용하는 구문분석 시스템에 대하여 각각 수행하였다.

실험 시스템

본 논문에서 실험을 위해 사용한 구문분석 시스템은 의존문법에 기반한 시스템이다. 의존문법은 단어 사이의 의존관계에 중심을 두는 문법이다. 주어진 문장 안의 단어 사이의 의존관계를 파악하는 작업이 의존문법에 의한 언어 분석의 중요한 작업이다. 어순의 자유성, 생략 등 한국어의 특성을 잘 처리할 수 있다고 생각되어 의존문법은 지금까지 한국어의 분석에 많이 이용되어 왔다(김미영 등, 2000; 김미영 등, 2002; 류범모 등, 1996).

구문분석 기법으로는 CYK 구문분석 기법을 이용하였다. CYK 구문분석 기법에 의해 생성된 구문구조 중에서 가장 적합한 것을 선택하여 문장의 구문구조로 한다. 가장 적합한 구문구조를 선택하는 기법으로는 상호정보를 이용한 통계적인 기법을 이용하였다.

본 논문에서 사용한 구문분석 시스템은 재현율 86.96%, 정확률 88.13%의 성능을 갖는 시스템이다.

실험 결과

실험은 다음과 같은 3가지 방식을 실험하였다.

- 방식1: 기존의 구묶음 중에서 복합명사 및 본 용언/보조 용언과 관련된 것만 사용함(즉 3.1, 3.2절에서 설명된 구묶음을 사용하는 방식).
- 방식2: 방식1에 추가적으로 기존에 사용되던 의존명사 구묶음 처리 방식만 사용함(즉 3.3절에서 소개한 것을 추가함.).
- 방식3: 본 논문에서 설명한 모든 구묶음

처리를 사용한 방식. (즉 방식2의 것에 4장에서 소개한 것들을 추가함.)

표 3에 나타난 성능 측정은 다음과 같은 메트릭을 사용하였다. 성능의 측정에서 구묶음된 덩어리는 한 단어로 취급하였다.

$$\text{재현율}(R)=\frac{\text{제안한 의존관계 중 올바른 것의 수}}{\text{정답 문서가 제안한 의존관계 수}}$$

$$\text{정확률}(P)=\frac{\text{제안한 의존관계 중 올바른 것의 수}}{\text{구문분석기가 제안한 의존관계 수}}$$

$$F-score=\frac{2RP}{R+P}$$

방식1은 모든 구문분석기가 쉽게 채용할 수 있는 단순한 구묶음 기법만을 사용한 것으로서 성능 비교의 기준 (baseline)으로 간주할 수 있다. 방식2에 대한 결과는 기존의 단순한 의존명사 관련 구묶음 만을 사용한 것으로서 방식1에 비하여 아주 약간의 성능 향상을 가져오는 것으로 밝혀졌다. 그러나 방식3 즉 우리가 본 논문에서 새로이 제안하는 의존명사 관련 구묶음 기법을 모두 사용하는 경우 상당한 성능 향상을 가져올 수 있음을 실험을 통하여 확인하였다.

우리가 실험에 이용한 말뭉치는 임의의 말뭉치로서 다양한 언어 현상과 많은 단어 수를 가진 문장들을 포함한다. 한국어가 생략과 어순이 자유롭고 어미의 변화가 심한 교착어라는 사실을 고려할 때 위 표에서 얻은 구문분석 시스템의 성능은 상당히 높은 것으로 볼 수 있다. 복잡하고 긴 문장이 많이 포함된 우리의 실험 말뭉치를 이용하는 것과 같은 일반적인 실험에서는 한국어 뿐만 아니라 영어의 경우에도 이와 같은 높은 구문분석 성능을 얻기 어렵다. 본 논문에서 제안한 의존명사 관련 구묶음 기술은 이러한 높은 성능의 시스템을 가능하게 하는데 중요한 역할을 한 것으로 판단된다.

<표 3> 실험방식에 따른 구문분석 시스템의 성능

	재현율	정확률	F-score	에러감소율
방식1	79.31%	76.57%	0.7792	-
방식2	79.74%	77.60%	0.7866	3.35%
방식3	86.96%	88.13%	0.8754	43.57%

우리가 제안한 구묶음 기법은 규칙에 기반한 것으로서 규칙이 적용되어 만들어진 구묶음은 완전히 정확하다고 말할 수 있다. 그 이유는 구묶음이 확실한 경우에만 규칙이 가동되도록 설계되었기 때문이다. 즉 구묶음에 대한 정확률은 완전할 정도로 매우 높다. 그러나 재현률의 경우는 그렇지 못하다. 예를 들면 다음과 같은 경우를 보자.

(31) 사과/NNG+를/JK 큰/VA 놈/NNB+으로/JK 열/NR+개/NNB+만/JK 주시오/VV.

여기에서 "사과를" 에 "큰 놈으로 열개만" 이 구묶음 되어져야 맞다. 그러나 우리 시스템의 경우 이를 수행하지 못하고 있다. 그 이유는 의존명사가 연속해서 나타나는 경우에 대한 규칙을 정의하지 않았기 때문이다. 현재 논문에 있는 규칙으로는 "사과를 열개만" 에 대해서는 구묶음을 잘 수행한다. 그러나 이 두 어절 사이에 "큰 놈으로" 라는 의존명사를 가진 어절이 끼어 들어가서 두 개의 의존명사가 연속으로 나타나게 되면 이들을 처리하지 못하고 있다. 즉 복잡한 의존명사 패턴에 대한 규칙이 모두 마련되어 있지 못하기 때문이다.

품사태거의 오류로 인하여 구묶음이 잘못되는 경우도 발생할 수 있다. 이것은 품사 태거의 성능에 따라서 많은 문제를 야기할 수도 있다. 그리고 우리의 의존명사 사전이 완전하지 못하여 고려하지 못한 의존명사들이 있을 수 있고 이들에 대해서는 구묶음 방식이 제안되지 못했을 수 있다. 현재로서는 구묶음이 태깅된 정답 말뭉치가 없으므로 재현율을 측정할 수 없는 아쉬움이 있다.

결 론

본 논문에서는 문장의 구조를 단순화시키는 구묶음 기법에 대해서 다루었다. 특히 문장에서 매우 다양한 형태로 나타나는 의존명사에 대한 구묶음 처리에 대해서 자세히 알아보았다. 구묶음이란 문장을 구성하는 단어들 중에서 서로 연관된 단어들을 합쳐서 하나의 덩어리로 만드는 것이다. 연관된 단어들을 합칠 때 의존명사가 포함된 경우에 매우 다양한 형태로 구묶음이 이루어진다.

본 논문에서는 의존명사를 단위의존명사와 비단위의존명사로 구분하였다. 단위의존명사는 선행 관형어인 수사와 어울려 수량이나, 시간, 온도 등을 나타내는 구실을 하는 의존명사이다. 단위의존명사를 수량단위의존명사, 시간단위의존명사, 온도단위의존명사로 구분하여 각각에 대해서 처리 방법을 제안하였다. 비단위 의존명사는 의존명사 중 단위의존명사를 제외한 나머지 의존명사를 말한다. 비단위 의존명사는 문장 내에서 매우 다양한 문형을 일으킨다. 본 논문에서는 비단위 의존명사에 의하여 나타나는 다양한 문형을 정리하여 각각에 해당하는 처리 방법을 제안하였다.

본 논문에서 제안한 각 방법을 사용하여 실험한 결과, 기존의 의존명사 구묶음 처리 방식만 사용한 실험은 의존명사에 대한 구묶음 처리를 하지 않은 방식에 대한 에러감소율이 약 3.4%인 것에 비하여 본 논문에서 제안한 모든 의존명사 구묶음 처리 방식을 사용한 실험은 에러감소율이 약 43.6%로써 구문분석 시스템의 성능이 크게 향상되는 것을 알 수 있었다.

참고문헌

김미영, 강신재, 이종혁, "규칙과 어휘정보를 이용한 한국어 문장의 구묶음(Chunking)", 제12회 한글 및 한국어 정보처리 학술대회, pp.103~109, 2000.

김미영, 강신재, 이종혁, "단위(Chunks)분석과 의존문법에 기반한 한국어 구문분석", 한국정보과학회 2002 봄 학술발표논문집, pp.327~329, 2002.

나동열, "한국어 구문 분석에 대한 고찰", 정보과학회지, 12(8), pp.33~46, 1994.

남기심, 고영근, 표준국어문법론, 탑출판사, 2004.

류범모, 이태승, 이종혁, 이근배, "술어중심 제약전파를 이용한 2단계 한국어 의존파서", 한국정보과학회 1996 봄 학술발표논문집, pp.923~926, 1996.

박의규, 조민희, 김성원, 나동열, "구묶음과 구간분할을 이용한 의존 관계 추출 기법", 제16회 한글 및 한국어 정보처리 학술대회, pp.131~137, 2004.

서정수, 국어문법, 뿌리깊은나무, 1994.

신효필, "최소자원 최대효과의 구문분석", 제 11회 한글 및 한국어 정보처리 학술대회, pp.242~248, 1999.

윤덕호, "숙어 정보를 활용한 한국어 파싱", 서울대학교 박사학위 논문, 1993.

홍영국, 이종혁, 이근배, "의존문법에 기반을 둔 한국어 구문분석기", 정보과학회 1993 봄 학술발표 논문집, pp.781~784, 1993.

황이규, 이현영, 이용석, "형태소 및 구문 모호성 축소를 위한 구문단위 형태소의 이용", 한국정보과학회논문지 제27권 7호, pp.784~793, 2000.

S. Abney, "Parsing by chunks", In Berwick, Abney, and Tenny, editors, *Principle-Based Parsing*, Kluwer Academic Publishers, pp.257~278, 1991.

A. Aho and J. Ullman, *The Theory of Parsing, Translation, and Compiling, Vol. 1: Parsing*, Prentice Hall, Englewood Cliffs, NJ, 1972

D. Bourigault, "Surface grammatical analysis for the extraction of terminological noun phrases", In Proceedings of the Fifteenth International Conference on Computational Linguistics, pp.977~981, 1992.

J. Kupiec, "An algorithm for finding noun phrase correspondences in bilingual corpora", In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, pp.17~22, 1993.

L. Ramshaw, M. Marcus, "Text chunking using transformation-based learning", In Proceedings of the Third ACL Workshop on Very Large Corpora, Association for Computational Linguistics, pp.157~176, 1995.

A. Voutilainen, "NPTool, a detector of English noun phrases", In Proceedings of the Workshop on Very Large Corpora, Association for Computational Linguistics, pp.48~57, 1993.

Juntae Yoon, "Three Types of Chunking in Korean and Dependency Analysis Based on Lexical Association", In Proceedings of the 18th International Conference on Computer Processing Languages(ICCPOL'99), pp.59~65, 1999.

1 차원고접수: 2006. 3. 24
2 차원고접수: 2006. 5. 18
최종게재승인: 2006. 6. 13