백과사전 질의응답을 위한 격틀 기반 의존관계 분석

임수종, 정의석, 장명길
한국전자통신연구원 음성/언어정보연구부
isj, eschung, mgjang@etri.re.kr

Dependency Relation Anlaysis using Case Frame for Encyclopedia Question-Answering System

Lim Soojong, Jung euisuk, Jang Myoung Gil
Speech/Language Information Research Department ETR

요 약

백과사전에서 정답을 찾기 위한 정보 중의 하나로 구조분석 정보를 이용하기 위하여 의존 관계 분석을 통해 정확한 구조분석에 대한 연구를 하였다.
정답을 찾기 위한 대상이 되는 용언과 논항의 관계를 파악하기 위해 먼저 의존관계 분석의 모호성 정도를 줄이기 위해 문장을 구묶음으로 나누었고 나눠진 구묶음에서 중심어와 중심어에 해당하는 의미코드를 추출하였다.
이렇게 구분된 구묶음 간의 의존관계를 파악하기 위하여 주로 격틀과 의미코드에 의존하는 의미자질, 거리 자질, 격관계 자질, 절형태 자질을 이용하여 의존관계 모호성을 해소하였다. 백과사전의 특성상 생략되는 성분과 연속 동사 처리를 하여 보다 정확하게 백과사전 QA시스템에서 정답을 찾을 수 있는 정보를 제공하도록 하였다.
실험결과 동사구와 명사구의 의존관계는 89.43의 성능을 보였고 의존관계에 격을 부여한 경우는 78.40%의 정확율, 백과사전 후처리에 해당하는 복원은 68.23의 성능을 보인다.

1. 서 론

한국어 문장을 분석하기 위한 과정 중에서 문장의 구조와 문법적인 기능을 분석하는 완전한 구문 분석(full parsing)은 성능 문제로 인해 실제 시스템에 널리 적용되지 못 하고 있다. 그러나 다음의 질의응답 시스템에서 답을 찾는 예와 같이 정확한 문법적인 기능을 제공한다면 부분적으로나마 실제 시스템에 적용할 수 있다.

2000년에 노벨평화상을 받은 사람은?
받다(subj : 사람, obj : 노벨평화상, adv : 2000년)

Title : 김대중
공로로 2000년 노벨평화상을 받았다.
받다(subj : 김대중, obj : 노벨평화상, adv : 2000년, 공로)

위와 같이 문장의 모든 구조를 파악하지 않고 주어, 목적어, 부사어, 보어만으로도 원하는 정보를 얻을 수 있다.

한국어 문장을 구문 분석 하기 위해서는 구구조, 범주 문법, 자질연산 문법 등 여러 가지 문법이 적용되지만 그 중에서도 자유 어순인 한국어에 가장 적합한 문법으로는 의존 문법이 있다.

의존 문법을 적용하여 문장 구조를 분석했을 때 성능을 저하시키는 이유 중 하나는 너무 많은 의존관계 후보가 생성된다는 점이다. 이러한 문제점을 해결하고 앞에서 언급한 문장 구조 분석 결과를 적용할 질의응답 시스템에서 요구하는 정도의 구조 분석을 위해서 본 연구에서는 먼저 문법적으로 연관이 있는 단어들을 하나의 묶음으로 인식하여 의존관계의 복잡도를 줄이고 애매성을 줄인다.

본 연구는 백과사전 질의 응답 시스템에 응용되기 위해 입력된 문장에 대해서 용언을 중심으로 주어, 목적어, 부사어, 보어 관계를 파악하는 것을 목적으로 한다. 이러한 연구 목적을 위해 입력된 문장의 의존 관계 파악을 위해 문장을 구묶음으로 인식한다. 동사구와 명사구의 의존관계를 파악하기 위해서는 번역용으로 구축된 격틀을 사용하는데 격틀과 매칭을 하기 위해서는 인식된 구묶음 중에서 중심어를 추출하여 중심어의 의미코드를 구묶음의 대표 의미코드로 이용한다. 의미코드와 격틀에만 의존하지 않고 거리, 격 관계, 문장 형태 정보

를 이용하여 의미정보의 부족한 점을 보충할 수 있도록 한다.

전통적인 구문 분석과는 다르게 백과사전을 대상으로 삼기 때문에 생기는 문제를 극복하기 위해서도 격틀을 사용한다. 백과사전 문서의 특성과 질의응답 시스템에서 정답을 찾기 위한 과정에서 해결해야 할 문제는 다음과 같다.

- 표제어 성분 생략
- 문장내 생략 성분 복원
- 연속동사 처리

2. 관련 연구

한국어 문장에 대한 의존관계 분석은 통계 정보를 이용하거나 규칙을 이용한 방법으로 분류할 수 있다. 통계 정보와 규칙을 이용한 두 가지 방법 대부분은 KAIST 구문 말뭉치를 사용하였다.

통계자질로 품사, 어휘, 가중치로 거리를 사용한 연구[3], 품사, 어휘, 거리 정보와 조사 구문 요소를 이용하여 통계 자질을 확장한 연구[5], 단어 의존 확률과 지역 의존 확률을 분리하여 적용한 연구[11, 12]가 있다. 구문 말뭉치가 아닌 원시 말뭉치로부터 추출한 공기 정보와 휴리스틱 제약 규칙을 사용한 연구[6], 상호 정보와 격 관계를 고려한 연구[10]이 있다.

규칙을 사용한 연구로는 의존 문법을 사용하여 규칙을 이용한 챠트 파싱을 이용한 연구[2], 지역적 어순 제약을 이용한 인접 의존관계 규칙과 지배가능경로를 이용한 2단계 구문 분석을 한 연구[4], 초기 트리를 인접으로 가정하여 규칙을 사용하여 올바른 트리 생성을 시도한 연구[7], 하위범주 정보와 개념 유사도를 규칙으로 사용한 연구[8], 구문 분석된 말뭉치를 사용하여 격틀을 구축하고 구축된 격틀에서 격명사, 조사 쌍으로 구성된 격확률을 이용한 연구[9], 모호성을 줄이기 위해 구간 분할 기법을 사용하고 규칙 기반 트리를 생성한 후에 통계 정보를 이용하여 모호성을 해소한 연구[1]등이 있다.

본 연구에서는 번역용 격틀을 이용하여 의존관계를 분석한다. 그러나 격틀을 이용할 경우에 모든 언어현상에 대해 격틀을 구축할 수 없고 설령 그렇다 하더라도 격틀을 사용할 때 생기는 모호성은 불가피하게 발생하기 때문에 보조적으로 거리자질, 격 관계 자질, 절 형태 자질을 사용하여 모호성을 해소한다. 절 형태 자질을 제외하고는 기존 연구에서 사용된 자질이지만 이러한 특성들을 조합으로 사용하여 유기적으로 혼용하여 모호성을 해소하는데 효율을 기하였고 특히 의미자질을 이용하여 문장 내에서 생략된 성분을 복원하였다.

3. 구묶음(chunk)과 중심어 인식

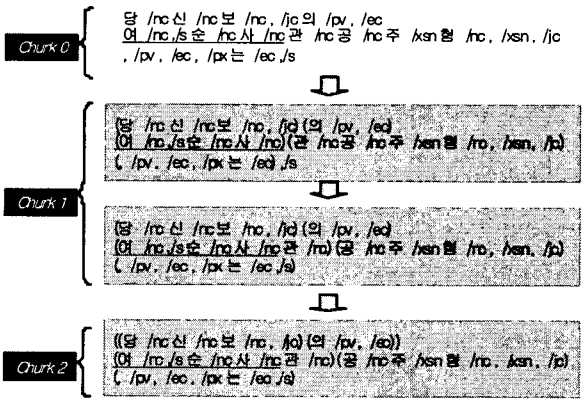
기존 연구에서 의존 문법을 적용한 경우 의존관계의 기본 단위를 어절 혹은 형태소 단위로 보았기 때문에 너무나 많은 종류의 의존관계 후보가 발생하였다. 이러한 문제점을 해소하기 위해서 문법적으로 연관이 있는 단어들을 묶어서 의존관계의 기본 단위로 처리하고 묶음 중에서 중심어를 추출하여 격틀 매칭등에 이용한다.

추출되는 중심어는 명사구의 경우는 일반명사나 개체명에 해당하는 체언이고 용언구의 경우는 용언의 어간이 된다.

3. 1 구묶음

구묶음 인식은 정답유형 인식 결과와 규칙 기반 접근방법으로 문장을 구묶음 단위로 분리하는 방식으로 진행된다. 구묶음 인식 과정은 모두 3단계로 구분할 수 있다. 정답유형 인식 결과를 Chunk0이라 하고, 형식형태소 단위별 구묶음 분리인 BaseNP, BaseVP, BaseADP, BaseIP를 찾는 과정과 이들의 예외사항을 처리한 후의 구묶음 인식 결과를 Chunk1이라 한다. 기존 구묶음 인식 연구는 대부분 Chunk1을 대상으로 한다. 이후 추출된 구묶음을 규칙기반 접근방법을 취하여 속격 NP와 어미 확장 사전을 통한 VP확장인식 후의 결과를 Chunk2 결과라 하였다.

구묶음 인식 예제는 〈그림 1〉과 같다



[그림 1] 구묶음 인식 예제

3.2 중심어 인식

중심어를 인식하는 대상 구묶음은 명사구(NP)와 용언구(VP, AP)이다. 명사구에서 중심어는 격틀을 이용하여 용언과의 의존 관계 여부를 판단하기 위하여 중심어를 추출하여 그에 맞는 의미코드를 설정하기 위함이다.

용언구의 중심어는 용언에 해당하는 어간을 추출하여 해당 격틀을 추출할 수 있도록 하기 위함이다.

3.2.1 명사구 중심어 인식

명사구의 경우에는 격조사가 부착된 일반 명사를 중심어로 인식한다. 중심어를 인식하는 목적이 명사구묶음을 대표하는 의미코드를 추출하기 위해서이기 때문에 개체명에 대해서는 4번의 규칙을 이용하여 중심어로 인식한다.

1. 일반명사+{격조사, 보조사}
노다지/nc+를/jc
2. 일반명사 의존명사+{격조사, 보조사}
J/nc+./s+레넌/nc+,/s P/nc+./s+매카트니/nc+,/s J/nc+./s+해리슨/nc+,/s R/nc+./s+스타/nc 등/nb+으로/jc..
3. 일반명사+접미사+{격조사, 보조사}
한국/nc+인/xsn+으로/jc+는/jx
4. 개체명+{격조사, 보조사}
<<DURATION : 1980/nn+~/s+1988/nn+년/nb>>+까지/jx

3.2.2 용언구 중심어 인식

용언구의 경우는 형태소 분석 결과를 중심으로 분석한다. 일반적으로 형태소 분석기에 의해 동사, 형용사만을 대상으로 삼고 보조용언은 구별한다. 그리고 지정사를 인식할 때 명사 부분과 함께 용언으로 인식한다.

1. {동사, 형용사}+어미
얻/pv+었/ep+던/etm
2. 일반명사+{동사, 형용사 파생 접미사}+어미
졸업/nc+하/xsv+고/ec
3. {일반명사, 의존명사}+지정사+어미
선수/nc+이/co+다/ef

4. 의존관계 인식

격틀을 이용한 의존관계 인식은 일반적인 의존 문법을 적용하여 의존관계 후보를 생성하고 의존 관계 중에서 모호성이 발생하는 부분에 대해서 격틀을 이용하여 모호성 해소를 시도한다.

4.1 격틀과 의미코드

본 연구에서 사용하는 격틀은 한중 번역을 위해서 구축된 격틀을 변형하여 사용한다. 격틀은 다음과 같은 형태로 구성된다.

A=의미코드! 격조사 용언!다 > 대역어 > 예문
A=사람!가 B=인공적장소!로 가!다 > A 0x53bb : v B [그[A]가 바다[B]로 가다]

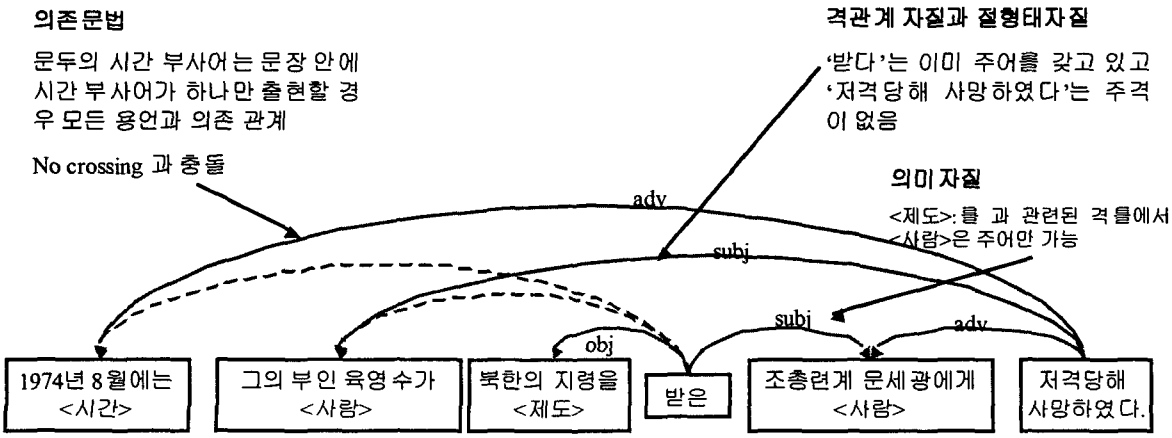
1만8천여개의 용언, 약 11만5천여개의 격틀을 사용하였다.

의미코드는 ETRI 명사개념망에서 상위 노드 436개를 정하여 사용하였다. 의미코드에 대한 정의는 표준국어대사전을 이용하였다.

4.2 의존관계 모호성 해소

의존관계 모호성은 명사구-용언구 간의 관계에서 발생하는데 이러한 모호성을 해소하기 위해서 사용되는 자질은 다음과 같다.

의미자질 : 명사구에서 추출된 중심어의 의미코드를



[그림 2] 의존관계 모호성 해소 예

사용하여 격틀에서 해당 용언이 논항으로 의미 코드를 취하는지 여부를 판단

거리자질 : 인접한 논항과 비인접 논항으로 구분하여 인접한 논항의 경우 의미자질에 위반되지 않으면 논항으로 채택.

격 관계자질 : 이중주어와 이중 목적어를 취하지 않는다는 가정 하에 의미자질과 거리자질을 사용하여도 모호성이 해소되지 않을 경우에 같은 격에 해당하는지를 판단하여 모호성을 해소

절 형태 자질 : 거리자질과 결합하여 내포문의 경우는 거리자질을 적용할 때 비인접 논항의 경우는 부정적으로 작용하고 내포문이 아닌 경우는 거리가 가까울수록 긍정적으로 작용

위의 4가지 자질을 이용하여 규칙을 정의하고 그 규칙을 사용하여 의존관계 모호성을 해소하는 알고리즘은 표1과 같다.

인접 규칙 : 동사구와 명사구가 인접해 있고 동사구의 격틀의 선택제약에 명사구의 중심어의 의미코드가 해당되는 경우는 우선적으로 의존 관계를 설정한다.

비인접 규칙 : 의존관계 후보 중에서 동사구와 명사구가 거리상으로 인접해있지 않은 경우에 적용되는 규칙. 규칙은 다음과 같은 순서대로 적용하여 모호성이 해소될 때까지 적용된다..

1. 의미자질
2. 기인식된 용언의 격관계
3. 절 형태 (내포문, 종속절, 대등절)

[표 1] 의존관계 모호성 해소 알고리즘

문장 S=Chk1Chkn-1Chkn 에서 가능한 모든 의존관계를 생성한다.
2. 모호성이 없는 의존관계에서 격관계를 생성하여 결정한다.
3. 인접의존규칙(adjacency dependency rule)을 사용하여 동사구에 인접한 명사구의 의존 관계 결정.
4. 비인접의존규칙을 적용하여 의존관계 결정

일반적인 명사구와 용언구 간의 의존 관계 이외에도 백과사전에서 요구하는 주어, 목적어, 부사어, 보어 관계를 파악하기 위해서는 관계 관형절을 처리하여 논항 여부를 판단하고 논항으로 채택이 가능할 경우에는 격판단을 하여야 하는데 이 경우에 관계 관형절 채택 여부는 동격 관형절을 취하는 명사의 리스트를 참조하여 판단하고 논항의 격 여부는 의미자질을 사용하여 격틀에서 해당 의미코드에 부착된 조사를 보고 판단한다. 처리 예는 다음과 같다.

(운문[텍스트]/nc_I 으로/jc_O) : NP
(쓰이/pv_I ㄴ/etm_O) : VP
(6/nn_I 권[단위]/nb_I 의/jm_O 철학[학문]/nc_I 시[텍스트]/nc_I 로서/jc_O ,/s_O) : NP

쓰이다(subj : 철학시, adv : 운문)

관계관형절은 용언 바로 다음의 어절을 논항으로 채택하는 것이 일반적이지만 의미자질을 적용할 경우 좀 더 정확한 논항과 격을 인식할 수가 있다.

5. 백과사전을 위한 후처리

백과사전 질의응답 시스템에서 의존 관계를 파악하여 정답을 찾는 과정에서는 문장에 존재하는 정보 이외에도 필요한 아래와 같은 후처리가 필요하다. 이와 같은 후처리는 의미자질과 복문의 경우 절 관계자질을 고려하여 처리한다.

5.1 표제어 성분 복원

일반적으로 백과사전 텍스트에서는 표제어는 생략을 한다. 그러나 이렇게 생략된 성분을 고려하지 않고 의존관계 분석 후에 문장 구조 분석을 끝내면 사용자의 질문에 대해서 정답을 찾을 수가 없다.

2000년에 노벨평화상을 받은 사람은?
받다(subj : 사람, obj : 노벨평화상, adv : 2000년)

Title : 김대중
공로로 2000년 노벨평화상을 받았다.
받다(obj : 노벨평화상, adv : 2000년, 공로)

위와 같이 문장 있는 그대로 분석을 할 경우에는 상을 받은 주체인 표제어인 김대중을 알 수가 없어서 정상적으로 문장구조 분석을 마쳤더라도 정답을 제시해 줄 수 없다.

[표 2] 백과사전 카테고리 매핑

백과사전 범주	의미코드
기관과 단체	집단
과학, 사회과학, 철학, 의학	학문
동식물	동물, 식물
문화예술	예술
생활과 레저	생활, 활동
인물	사람
지리, 지역	곳
컴퓨터와 인터넷	장치

표제어를 카테고리별로 분류를 했기 때문에 분류된 카테고리를 격틀에서 사용하는 의미코드로 변환하고 격틀을 참조하여 필수 성분인 주어나 목적어 중에서 표제어로 복원할 수 있는지 여부를 판단한다.

백과사전의 카테고리 정보와 매핑 예는 〈표2〉와 같다.

5.2 문장내 생략 성분 복원

문장내 생략 성분 복원도 표제어 성분 복원과 마찬가지로 같은 문장 내에서는 빈번하게 일어나는 동일 어휘의 반복을 피하는 현상 때문에 발생하여 정답을 찾지 못 하는 경우를 방지하기 위함으로 필수 성분인 주어, 목적어를 대상으로 한다.

복원 방법은 문장의 맨 뒤의 용언구와 대등연결어미로 연결된 용언구의 논항을 보고 상호 생략된 성분을 검사하여 생략된 성분의 경우에는 격틀을 참조하여 논항으로 채택될 수 있는지 여부를 본다. 이 경우 격틀에서 의미 코드 뿐 아니라 자,타동사 여부도 함께 유용한 정보로 활용한다.

5.3 연속동사 처리

연속동사(serial verb)의 경우 보통 같은 용언구로 취급하여 같은 논항을 갖는 것으로 처리를 하지만 이 경우 오류가 발생할 수 있기 때문에 의존관계를 파악하는 용언은 앞 쪽의 용언으로 한정하고 대신에 뒤쪽의 용언의 주어, 목적어, 부사어, 보어 관계를 파악할 경우에는 자, 타동사 여부와 의미코드를 이용한 격틀의 선택 제약 정보를 이용하여 논항으로 취할 수 없는 정보를 배제하여 문장구조를 분석한다.

6. 실험 및 평가

본 연구에 대한 실험은 백과사전 문서에서 임의로 추출한 문장을 대상으로 3가지 종류의 실험을 하였다. 실험 1은 동사구와 명사구의 의존관계의 설정 여부이고 실험2는 설정된 의존 관계에서 격을 부여한 것이고 실험3은 백과사전 문서를 처리하기 위한 후처리에 대한 것이다. 후처리를 위해서 추출된 문장의 표제어와 표제어가 어떤 의미코드에 해당하는지를 수작업으로 태그를 추가하였다.

실험에서 사용된 평가의 척도로는 실험1과 실험 3은 정확률과 재현률을 사용하였고 실험2는 정확률만을 사용하였다.

〈표3〉과 같이 단순한 의존관계 설정에서 있어서는 90% 가까운 성능을 보였다. 오류는 주로 내포문을 포함하는 경우에 발생하였는데 최인접 용언구와 거리가 2 이상인 명사구에 대해서 관형형 어미를 포함하는 동사구와 의존 관계인지 아니면 문장의 맨 끝에 나오는 용언구와 의존 관계인지에 대한 모호성을 해소하는 과정에서 생기는 오류였다. 그리고 백과사전 등에서 빈번하게 사용되는 지정사로 인한 오류도 상당 부분을 차지하였다.

[표 3] 실험결과

	실험1	실험2	실험3
정확율	88.75	78.40	76.31
재현율	91.02		61.70
F-measure	89.43		68.20

설정된 의존 관계에 대해 격을 부여하는 것은 격조사가 부착된 경우에는 쉽게 격을 파악할 수 있었지만 보조사가 부착되었거나 관계관형절에서 격 관계를 파악해야 하는 경우에는 격틀이 존재하지 않을 경우에는 많은 오류가 발생하였다.

복원의 경우에는 격틀만을 참조하여 복원을 했을 때 백과사전 카테고리를 의미코드로 매핑하는 경우에 발생한 오류가 많았다. 백과사전에서 표제어를 카테고리별로 분류할 경우에 한 표제어가 하나의 카테고리에만 할당되는 것이 아니라 주로 2개 이상의 카테고리에 할당이 되기 때문에 실험에서는 맨 위의 카테고리를 선택하였기 때문에 발생하는 문제가 아주 많았다. 예를 들어 '박정희'의 경우 인물, 역사와_지리, 지역 3개의 카테고리를 갖는다.

7. 결론 및 향후 연구

백과사전에서 정답을 찾기 위한 정보 중의 하나로 구조분석 정보를 이용하기 위하여 의존 관계 분석을 통해 정확한 구조분석에 대한 연구를 하였다.

정답을 찾기 위한 대상이 되는 용언과 논항의 관계를 파악하기 위해 먼저 의존관계 분석의 모호성 정도를 줄이기 위해 문장을 구묶음으로 나누었고 나눠진 구묶음에서 중심어와 중심어에 해당하는 의미코드를

추출하였다.

이렇게 구분된 구묶음 간의 의존관계를 파악하기 위하여 주로 격틀과 의미코드에 의존하는 의미자질, 거리 자질, 격 관계 자질, 절 형태 자질을 이용하여 의존관계 모호성을 해소하였다. 백과사전의 특성상 생략되는 성분과 연속 동사 처리를 하여 보다 정확하게 백과사전 QA시스템에서 정답을 찾을 수 있는 정보를 제공하도록 하였다.

향후 연구로는 본 연구가 주로 격틀 정보를 이용하고 있기 때문에 격틀 정보가 존재할 경우 보다 정확한 의존관계 분석이 가능하지만 격틀 정보가 없을 경우에는 거리, 격 관계, 절 형태만을 참조할 수 밖에 없기 때문에 통계 정보를 활용하면 격틀 정보의 부재를 만회할 수 있을 것이다. 그리고 용언구와 용언구와의 의존 관계로 파악할 수 있는 원인, 목적 등의 관계도 의미 정보를 이용하여 분석할 수 있도록 해야 할 것 이다.

참고문헌

[1] 김광백, "구간 분할 기반 한국어 구문 분석", 연세대 석사논문, 2003.

[2] 김창현, "한국어 구문 분석을 위한 오른쪽 우선 챠트 파서", KAIST 석사논문, 1993.

[3] 김형근, "확률적 의존 문법과 한국어 구문 분석", KAIST 석사논문, 1995.

[4] 류법모, "한국어 파서에서의 지역 의존관계의 이용", 제8회 한글 및 한국어 정보처리, 1996

[5] 서광준, "조상관계에 기반한 한국어의 확률적 의존구조 분석", KAIST 박사논문, 1999.

[6] 윤준태, "공기 관계 기반 어휘 여관도를 이용한 한국어 구문분석", 연세대학교 박사학위논문, 1997

[7] 이성욱, "변환 규칙 학습기를 이용한 한국어 의존구조 분석기", 제9회 한글 및 한국어 정보처리, 1997.

[8] 장명길, "통계/의미 정보를 이용한 한국어 의존 파싱", 제9회 한글 및 한국어 정보처리, 1997.

[9] 전은희, "한국어 동사의 격틀 정보를 이용한 구문분석 후처리기", 제13회 한글 및 한국어 정보 처리, 2001.

[10] 정석원, "격 관계와 상호정보를 이용한 한국어 의존파서", 제13회 한글 및 한국어 정보 처리, 2001.

[11] 정후중, "지역 의존 확률과 단어 의존 확률을 이용한 통계적 한국어 구문분석 모델", 고려대 박사논문, 2004.

[12] Yoshihide Kato etal. "Efficient Incremental Dependency Parsing", NLPRS 2001.