KIPS Tr. Software and Data Eng.
Vol.3, No.6 pp.219~230 pISSN: 2287-5905

http://dx.doi.org/10.3745/KTSDE.2014.3.6.219

An Analysis of Korean Dependency Relation by Homograph Disambiguation

Hong-Soon Kim† · Cheol-Young Ock††

ABSTRACT

An analysis of dependency relation is a job that determines the governor and the dependent between words in sentence. The dependency relation of predicate is established by patterns and selectional restriction of subcategorization of the predicate.

This paper proposes a method of analysis of Korean dependency relation using homograph predicate disambiguated in morphology analysis phase. The disambiguated homograph predicates has each different pattern. Especially reusing a stage transition training dictionary used during tagging POS and homograph, we propose a method of fixing the dependency relation of {noun+postposition, predicate}, and we analyze the accuracy and an effect of homograph for analysis of dependency relation.

We used the Sejong Phrase Structured Corpus for experiment. We transformed the phrase structured corpus to dependency relation structure and tagged homograph. From the experiment, the accuracy of dependency relation by disambiguating homograph is 80.38%, the accuracy is increased by 0.42% compared with one of undisambiguated homograph. The Z-values in statistical hypothesis testing with significance level 1% is $|Z| = 4.63 \geq z_{0.01} = 2.33$. So we can conclude that the homograph affects on analysis of dependency relation, and the stage transition training dictionary used in tagging POS and homograph affects 7.14% on the accuracy of dependency relation.

Keywords : Dependency Relation, Homograph, Pattern of Predicate, UTagger, Stage Transition Model, Z-values of Statistical Hypothesis Testing

동형이의어 분별에 의한 한국어 의존관계 분석

김 홍 순† · 옥 철 영††

요 약

의존관계 분석은 문장의 어절 간에 의존소-지배소를 결정하는 작업이다. 용언은 문형 및 하위범주화 정보의 선택제약에 의해 다른 어절과의 의존관계를 형성한다.

본 논문은 형태소 분석 단계에서 동형이의어 분별된 용언의 문형을 이용하여 용언의 의존관계를 분석하는 방법을 제안한다. 특히, 형태소분석 단계에서 품사 및 동형이의어 태깅을 위해 사용하는 단계별 전이모델의 학습사전을 재활용하여 {명사+격조사, 용언} 간의 의존관계를 확정하는 방안을 제안하고 그의 정확률 및 영향을 분석한다.

동형이의어가 부착되고 의존관계로 변경된 21개의 세종구문분석말뭉치를 이용하여 실험한 결과, 동형이의어 분별된 의존관계 분석 정확률이 80.38%로, 동형이의어가 분별되지 않은 의존관계분석에 비해 0.42%의 정확률 향상이 있었으며, 유의수준 1%의 검정통계량 Z는 $|Z| = 4.63 \geq z_{0.01} = 2.33$으로 동형이의어 분별이 의존관계 분석에 영향이 있음을 보였다. 또한, 단계별 전이모델이 의존관계 분석 정확률에 약 7.14% 영향을 미치는 것을 알 수 있었다.

키워드 : 의존관계, 동형이의어, 용언 문형, UTagger, 단계별 전이모델, 검정통계량 Z

1. 서 론

한국어는 SOV(S:주어, O:목적어, V:용언) 어순을 나타내는 언어로 구문론적으로 다음의 특징이 있다.

- 어떤 어형의 문법적 기능은 그 어형 발단의 요소(조사 등)에 의해 결정되는 첨가어이다.

※ 이 논문은 2010년 2012년 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2010-32A-H00006, 2012R1A1A2006906).
† 준 회 원 : 울산대학교 정보통신공학과 석사
†† 종신회원 : 울산대학교 컴퓨터정보통신공학부 교수
Manuscript Received : February 13, 2014
First Revision : March 31, 2014; Second Revision : April 29, 2014
Accepted : May 13, 2014
* Corresponding Author : Cheol-Young Ock(okcy@ulsan.ac.kr)

- 어순이 비교적 자유로운 언어이다. 그러나 한 어절 내의 형태소적 어순은 엄격하게 존재 한다.
- 한국어는 지배소가 항상 의존소(dependent)의 뒤에 나타나는 지배소 후위(governor-final) 언어이다.
- 한 문장에서 필수적인 요소의 생략이 자주 발생한다.

이러한 특성으로 인하여 한국어를 구 구조 문법(Phrase Structure Grammar)으로 분석한다면 매우 많은 규칙을 필요로 하고 그 처리과정이 복잡하다. 반면, 의존 문법(Dependency Grammar)은 문장 구성성분 사이의 의존관계(의존소, 지배소)에 중심을 두는 문법으로 어순의 제약을 거의 받지 않기 때문에 어순의 도치나 주요 성분의 생략이 일어나는 한국어 문장에 대한 분석에 적합하다.

의존문법에 의한 기존의 의존 구문 분석에서는 구축된 말뭉치(세종구문분석말뭉치)들을 기준으로 학습하거나 자질을 추출하여 사용하는데 그 구축된 말뭉치들은 대부분 형태소/품사만 태깅되어 있다. 즉 형태소가 동형이의어일 경우 동형이의어 분별이 되어 있지 않다.

입력문장의 형태소 분석 결과에서 특히 용언이 동형이의어일 경우 동형이의어가 분별되지 못함으로써 구문 구조 분석 시에 많은 중의성이 발생했다. 동일 어근의 용언(예, '타다')일 경우라도 동형이의어에 따라 문형이 달라 필수 논항이 다르다. 예를 들어 아래 두 문장의 '타다'를 비교해 보자.

(1) 쓰레기를 타는 것, 타지 않는 것 등으로 구분했다.
(2) 썰매를 타려면 꼭 장갑을 끼어야 한다.

위 문장 (1)에서의 "불씨나 높은 열로 불이 붙어 번지거나 불꽃이 일어나다."의 의미를 가진 자동사 '타다__01[1]'이며, 문장 (2)에서는 "도로, 줄, 산, 나무, 바위 따위를 밟고 오르거나 그것을 따라 지나가다." 뜻의 타동사 '타다__02'로 【…을】의 문형[2]을 요구한다. 위 두 문장에서의 형태소 분석 결과,

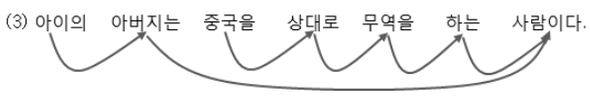
(1') 쓰레기/NNG+를/JKO 타__01/VV+는/ETM 것__01/NNB+, /SP 타__01/VV+지/EC 않/VX+는/ETM 것__01/NNB 등__05/NNB+으로/JKB 구분하__03/VV+었/EP+다/EF+./SF
(2') 썰매/NNG+를/JKO 타__02/VV+려면/EC 꼭__03/MAG 장갑__01/NNG+을/JKO 끼__01/VV+어야/EC 하__01/VX+ㄴ다/EF+./SF

로 동형이의어 '타다'가 분별된다면, 문장 (1)의 의존관계 분석 시에 "쓰레기를"과 "타는"은 의존관계를 맺지 않게 처리할 수 있으며, 문장 (2)에서 "썰매를"은 "타려면"과 의존관계를 맺게 될 것이다.

용언뿐만 아니라 명사도 동형이의어 분별된다면 구문 분석 과정의 중의성을 해소할 수 있는 경우가 많다. 예를 들어,

(3) 아이의 아버지는 중국을 상대로 무역을 하는 사람이다.

의 문장 (3)에서 {상대, 무역}의 명사는 서술성 명사로 용언 {상대하다, 무역하다}의 성질(문형 및 필수 논항)을 내포하고 있어, 용언으로 간주해야만 다음과 같이 정확히 의존관계를 분석할 수 있다.



이러한 서술성 명사는 {상대__04, 무역__02}로 모두 동형이의어 명사이며, 이들 명사들이 (3')과 같이 동형이의어 분별된다면 의존관계 분석 과정에서 서술성 명사의 용언 성질을 활용할 수 있을 것이다.

(3') 아이__01/NNG+의/JKG 아버지/NNG+는/JX 외국__02/NNG+을/JKO 상대__04/NNG+로/JKB 무역__02/NNG+을/JKO 하__01/VV+는/ETM 사람/NNG+이/VCP+다/EF+./SF

이와 같이 형태소 분석 단계에서 동형이의어가 분별된다면 구문 분석 시 많은 중의성을 해소할 수 있으며 정확한 의존관계를 분석할 수 있다. 본 논문에서는 형태소 분석 단계에서 동형이의어를 분별하는 UTagger[3]의 결과를 이용하여 의존관계를 분석하는 방법을 제안한다. 특히 UTagger에서 사용하는 단계별 전이 학습말뭉치를 재활용하여 {명사+격조사 용언} 간의 의존관계를 확정하는 방안을 제시한다. 이번 논문의 연구는 동형이의어 용언에만 한정하며, 동형이의어 서술성 명사에 대한 의존관계 분석 연구는 추후 연구로 미룬다.

본 논문의 구성은 다음과 같다. 2장에서는 기존의 한국어 구문분석에 관한 관련 연구들을 살펴보고, 3장에서는 의존관계 분석을 위한 의존규칙 및 동형이의어 분별됨으로써 고려할 수 있는 여러 의존관계들에 대해 설명한다. 다음으로 4장에서는 전체시스템의 구성 및 의존관계 분석 과정, 그리고 세종구문분석말뭉치를 이용한 결과를 살펴본다. 마지막으로 5장에서 결론과 향후 연구에 관해 논하였다.

2. 관련 연구

최근의 한국어 의존 구문 분석에 대한 연구는 보다 실용적인 구문분석기의 개발에 필요한 여러 가지 의존 파싱 방법론에 대한 연구가 주요 관심사항이다. 의존 구문 분석에 사용되는 방법은 규칙 기반 처리 방법과 기계 학습 방법으로 간략히 나누어 볼 수 있다.

[1]에서는 트랜지션 방법의 장점에 문장 성능 향상을 위한 새로운 방법으로 문장 구조를 인식하기 위해 '키어절'이라는 개념을 제안하였다. 인식된 키어절을 구문분석의 자질

1) 동형이의어 어깨번호는 표준국어대사전을 기준으로 하였다.
2) 문형도 표준국어대사전을 기준으로 하였다.
3) UTagger는 연구용으로 무료 기술이전하고 있음(http://nlplab.ulsan.ac.kr).

로 사용하는 방법, 키어절을 위한 수정 모델 실험을 하였다. 세종 구문 코퍼스를 사용하여 실험한 결과 3%의 문장성능 향상을 이룰 수 있다고 하였다. 키어절을 사용하는 것이 문장 구조를 인식하고 오류전파를 막는 데 효과적임을 보였다.

[2]에서는 한국어 구문분석에서 발생하는 중의성을 해결하기 위하여 구간분할 방법과 논항정보를 사용하여 개선한 구문 분석시스템을 소개하였다. 이 논문에서 제안하는 구문 분석 시스템은 어절대신 형태소를 입력으로 사용하고, 또한 주어진 형태소에 대하여 가능한 모든 구문 분석 구조를 생성하는 알고리즘을 사용하였다. 실험을 통하여 약 53%의 중의성을 제거할 수 있었음을 보였다.

[10]에서는 각 절들의 서술부의 의존관계를 서술부의 논항 유무로 보고, 이진 분류 문제로 의존관계를 분석하였다. 서술부의 관련 정보는 간단한 문법 규칙을 기반으로 CKY 차트 파서를 통하여 추출하였고, 기계 학습 방법으로는 SVM을 사용하였다. 실험 결과 어휘 정보들 중에서 어미의 정보만 사용하였을 경우는 64.4%의 정확도를 보였고 문법적인 정보인 동적 자질을 사용한 경우는 73.5%의 정확도를 보였다. 이 논문에서는 어휘 정보 및 문법적인 정보들만 사용하여 의미적인 정보들이 추가되어야 할 것으로 보인다.

[14]에서는 한국어 구성성분은 내용어와 기능어의 결합형태로 구성되고 임의 구성성분 기능어와 임의 구성성분 내용어 간의 의존관계가 의미가 있다는 사실을 반영한 의존문법 학습방법을 제안하였다. KAIST의 트리 부착 코퍼스 31,086개 문장에서 추출한 30,600개 문장의 Tagged Corpus를 가지고 학습한 결과 초기문법을 64%까지 줄인 1,101개의 의존문법을 획득했고, 실험문장 486개 문장을 Parsing한 결과 73.81%의 Parsing 정확도를 보였다. 이 논문에서는 실험 데이터가 너무 작고, 의존 문법이 1,101개로 의존 문법의 개수를 줄여야 할 것으로 보인다.

[4]에서는 의존 파싱에서 최소한의 의존관계를 생성하기 위하여 후보 의존소가 지배가능경로 상에서 술어 지배소와의 의존관계 검사 시에 술어의 하위범주화 정보를 이용하는 의존 파싱 방법을 제안한다. 이 방법으로 의존 파싱 과정에서 후보 의존관계의 과생성으로 인한 비효율성을 미리 차단함으로써 의존 규칙과 지배가능경로만에 의한 의존 파싱에 비하여 보다 정확하고 향상된 파싱 결과를 얻을 수 있었다. 술어의 하위범주화 정보를 사용하는 점에서 본 논문과 비슷하지만 하나의 절이 가지는 의존관계의 수가 1개 이상으로 정확한 의존관계를 찾기 어렵다.

[8]에서는 투사성의 원칙을 이용해 결정적 한국어 의존 구문분석을 보완하는 방법을 제안하였다. 투사성의 원칙을 이용하여 의존 구문분석의 오류를 찾아내고 의존관계를 재탐색한다. 제안한 의존 구문분석 모델이 비결정적 의존 구문 분석 모델보다 뛰어난 성능을 제공함을 실험으로 보였다. 여기서 말하는 투사성의 원칙이 본 논문에서 제안하는 규칙 중 하나인 투영의 원칙을 말한다. 이 투영의 원칙을 이용하여 높은 성능을 가져올 수 있는 것을 보였다.

[9]에서는 술어 중심 제약 만족 알고리즘을 이용한 한국어 의존 파서를 제안하였다. 술어 중심의 제약전파 알고리즘은 술어의 결합가 정보를 바탕으로 초기 구문 의존관계 그래프에 존재하는 비문법적인 의존관계를 제거함으로써 정확한 파싱 결과를 얻을 수 있다고 하였다.

이상의 기존 연구들은 통계적인 방법을 사용하여 한국어 의존 구조를 분석하였고, 입력문장의 형태소 분석된 결과만을 이용한 구문 분석으로 동형이의어 분별된 정보를 이용하지 못하였다. 본 논문에서는 의존 규칙을 기반으로 하여, 형태소 분석 단계에서 동형이의어 분별을 위해 사용된 UTagger의 학습 사전 중에서 AF 전이 모델을 이용하여 의존관계를 분석하는 방법을 제안한다. 의존관계는 세 단계에 걸쳐 분석하는데, 첫 번째 단계에서는 UTagger에서 사용하는 학습 말뭉치를 재활용하여 격조사를 가진 어절과 용언 간의 의존관계를 확정하고, 두 번째 단계에서 동형이의어 분별된 용언 및 서술성 명사 용언의 문형 정보를 이용하여 필수 논항을 결정한다. 마지막 세 번째 단계에서는 지배소와 의존소의 세부규칙에 따라 지배소가 확정되지 않은 어절들의 의존관계를 분석한다.

3. 동형이의어 분별에 의한 한국어 의존관계 분석

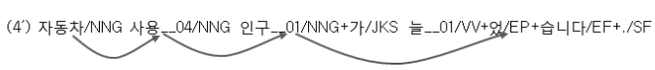
이 장에서는 한국어 의존관계 분석을 위해 적용한 의존 문법을 설명하고, 형태소 분석 단계에서 분별된 동형이의어 정보를 이용하여 의존 관계를 분석하는 방법에 대해 설명한다.

3.1 한국어 의존관계 및 제약 규칙

의존 문법에서는 어떤 두 어절이 결합할 때에 지배소(governor)가 중심이 되고, 의존관계에 있는 형태소, 즉 의존소(dependent)와 결합한다. <표 1>은 한국어 의존 문법을 적용한 지배소와 의존소의 관계를 나타내고 있다.

이러한 의존 규칙을 이용하여 간단히 의존관계를 추출할 수 있다. 예를 들어,

(4) “자동차 사용 인구가 늘었습니다.”의 문장은



의 의존관계를 가진다.

Table 1. Rules of Korean Dependency Relation

규칙	관계	지배소	의존소
1	수식	명사	관형사, 관형격조사, 관형형어미, 명사, 부사
2	수식	대명사	관형사, 관형격조사, 관형형어미, 부사
3	부가	동사, 형용사	격조사(주격, 목적격, 보격, 부사격, 호격, 인용격), 보조사, 부사, 연결형어미, 부사형전성어미
4	강조	부사, 관형사	부사

여기서 “자동차 사용”은 규칙 1에 의해 ‘사용’은 앞의 ‘자동차’라는 명사(NNG)를 의존소로 가지는 명사(NNG) 지배소가 된다. 특히 ‘사용’은 서술성 명사로 용언의 성질을 내포(“자동차를 사용하다”)하고 있다. “사용 인구가”도 규칙 1에 의해 “사용(NNG) 인구(NNG)”의 연속 두 명사 간의 의존소-지배소의 관계를 가진다. 그리고 “인구가 늘었습니다”는 규칙 3에 의해 ‘늘었습니다’는 지배소인 동사(VV)이고, 주격조사(JKS)를 가진 ‘인구가’가 의존소이다.

한국어 의존규칙에 의한 의존관계를 추출하는 과정에서 대체로 다음과 같은 의존제약 규칙이 적용된다.

① **지배소 후위의 원칙** : 지배소는 의존소보다 문장 내에서 뒤에 위치한다.
② **투영의 원칙** : 임의의 의존관계 A, B에 대해서 A에 대한 아크와 B에 대한 아크는 서로 겹치지(crossing) 않아야 한다.
③ **지배소 유일의 원칙** : 하나의 의존소는 오직 한 개의 지배소만 갖는다.
④ **격틀/의미정보 제약** : 의존소 A가 격 c1을 나타낼 때, 지배소 B의 c1격에 대한 의미제약(semantic constraint)을 의존소 A가 만족해야 의존관계가 성립한다.
⑤ **필수 성분 제약** : 필수성분을 가져야 하는 어절이 필요한 성분을 갖지 못하고서는 다른 어절의 의존소로 사용될 수 없다.

용언이 동형이의어(혹은 다의어)이고 동형이의어(혹은 다의어)별로 다른 격틀/의미정보 및 필수성분을 요구한다면, 격틀/의미정보 제약 및 필수성분 제약을 적용하기 위해서는 동형이의어(혹은 다의어)가 분별되어야 한다(그림 1 표준국어대사전에서의 ‘차다’ 참조). 본 논문은 형태소 분석 시에 동형이의어 분별이 가능한 UTagger의 결과를 이용하여 동형이의어 분별된 용언의 의존관계를 분석한다.

3.2 UTagger의 단계별 전이모델을 이용한 동형이의어 용언의 의존관계 확정

UTagger는 HMM 기반의 한국어 품사 및 동형이의어 동시 태깅시스템으로, 약 1,100만 어절의 세종형태의미주석말뭉치에서 인접 두 어절 간의 형태소/품사 전이확률을 이용한다[21]. 예를 들어, UTagger는 다음 두 문장

(5) 자동차가 기름을 태워 달린다.
(6) 자동차가 사람을 태워 달린다.

를

(5‘) 자동차/NNG+가/JKS 기름__01/NNG+을/JKO **태우__01/VV+어/EC** 달리__04/VV+ㄴ다/EF+./SF
(6‘) 자동차/NNG+가/JKS 사람/NNG+을/JKO **태우__02/VV+어/EC** 달리__04/VV+ㄴ다/EF+./SF

로 태깅한다. 위 두 문장에서 ‘태우다’는 ‘타다’의 사동사로 (5)에서는 “불에 타다”는 의미이고, (6)에서는 “탈것에 몸을 얹다”는 의미로, “태워”의 좌우 인접 어절에 따라 ‘태우다’의 의미가 결정된다.

연구 [21]에서는 90% 학습말뭉치에 대해 10%의 테스트 집합의 인접 두 어절이 모두 출현한 비율(AA 전이모델)은 33.66%이여서, 인접 두 어절 간의 학습 자료부족 문제를 해결하기 위해 단계별 전이모델을 제안하였다. 그 첫 번째로 인접 두 어절 A와 B에서 어절 A의 형태소 분석 전체 결과와 어절 B의 첫 번째 형태소 간의 전이모델 AF(All morphemes of a word, First morpheme of the next word)을 제안하였다. AF 전이모델은 인접 두 어절에서 앞 어절의 의미와 뒤 어절의 의미는 대체로 서로 관련이 있어, 뒤 어절의 어근(어휘형태소)만으로도 두 어절의 의미 관계를 파악할 수 있다는 점을 반영한 전이모델이다. AF 전이모델의 예로

차__06/NNG+를/JKO	타__02/VV	188
차__09/NNG+를/JKO	타__03/VV	2

를 제시하였다. 여기서 각 동형이의어는 ‘차__06(car)’, 차__09(tea)’이며, ‘타__02(ride)’, ‘타__03(mix)’의 의미를 가진다. 실험에서 AF 전이모델의 적용비율은 18.42%이었다.

연구 [21]에서는 위 AF 전이모델 외에도 어절 A의 마지막 형태소(문법형태소)와 어절 B의 첫 번째 형태소 간의 전이모델 EF(End morpheme of a word, First morpheme of the next word)도 제안하였으나, 본 논문에서는 EF 전이모델은 사용하지 않아 추가 설명은 생략한다. 연구 [21]에서는 각 전이모델마다 다른 가중치를 적용하며, 부분적으로 전이빈도가 0인 경우를 위해 최소 전이점수를 계산하는 예외 처리 루틴을 포함하였다. 세종형태의미말뭉치 중 90%를 학습하고 10%를 테스트 집합으로 사용한 결과, 품사와 동형이의어 둘 다에 대해 96.44%의 태깅 정확률을 보여, 단계별 전이모델이 한국어와 같은 교착어의 품사 및 동형이의어 태깅에 적합한 방법임을 보였다.

한국어는 격조사가 발달하여 어순이 비교적 자유롭다. 예를 들어, 예문

(7) 버스를 타고 빨리 학교에 갔다.
(7-1‘) 버스__02/NNG+를/JKO 타__02/VV+고/EC 빨리/MAG 학교/NNG+에/JKB 가__01/VV+았/EP+다/EF+./SF

가 학습되었다면 {버스__02/NNG+를/JKO 타__02/VV, 타__02/VV+고/EC 빨리/MAG, 빨리/MAG 학교/NNG, 학교/NNG+에/JKB 가__01/VV}의 AF 전이모델이 학습되어 있다. 그렇다면, 문장 (7-2’)에서

(7-2“) 버스__02/NNG+를/JKO 타__02/VV+고/EC 학교/NNG+에/JKB 빨리/MAG 가__01/VV+았/EP+다/EF+./SF

{명사+격조사 용언}의 의존관계를 분석할 때, AF 전이모델로 학습된 정보를 이용한다면 어순이 바뀐 경우에도 “학교/NNG+에/JKB 가__01/VV”의 {명사+격조사 용언} 간의 의존관계를 확정할 수 있다. 본 논문에서는 단계별 전이모델 중 AF 학습사전만을 활용하며, 이 중에서도 {명사+격조사 용언} 간의 의존관계에 확정시에만 사용한다.

AF 학습사전은 1,100만 어절의 세종형태의미말뭉치에서 단순히 인접 두 어절(bigram)의 빈도만 가지고 있다. 본 논문의 의존관계 분석시에는 AF 학습사전의 빈도정보는 활용하지 않고 단순히 학습여부만을 활용한다. 그리고 모든 형태소의 AF 전이를 보는 것이 아니라 AF 전이 중에서 {명사+격조사 용언}의 전이만을 사용하여 의존 관계를 확정 짓는데 사용한다. 따라서 위 (5-2“)의 문장에서 “학교에 갔다”의 두 어절이 인접되지 않았다 하더라도 AF 학습사전에서 발견된다면 이 두 어절 간에 의존관계를 설정할 수 있다.

3.3 동형이의어 분별된 용언의 문형을 이용한 의존관계 분석

<표 1>의 의존 규칙에 의한 기존의 한국어 의존관계 분석 방법은 단순히 어절간의 형태소 및 품사만을 보고 판단하므로 여러 중의성이 발생하고 정확한 의존관계 분석에 실패하는 경우가 많았다. 예를 들어 동사 ‘차다’의 경우도, 동형이의어에 따라 다른 필수 논항을 요구한다. 다음 [그림 1]은 표준국어대사전에 등재된 ‘차다’의 동형이의어별 뜻풀이와 문형을 제시하고 있다. ‘차다’는 크게 동사 {차다01, 차다02, 차다03}와 형용사 {차다04, 차다05}로 구분되며, 동사의 경우도 자동사 {차다01}와 타동사 {차다02, 차다03}으로 구분되며 의미별로 요구하는 문형이 다르다.

차다01 〔차, 차니〕
「동사」
[1]【…에】【…으로】
일정한 공간에 사람, 사물, 냄새 따위가 더 들어갈 수 없이 가득하게 되다.
[2]【…에】
「1」 감정이나 기운 따위가 가득하게 되다.
「2」 어떤 대상이 흡족하게 마음에 들다.
「3」 어떤 높이나 한도에 이르는 상태가 되다.
[3] 「1」 정한 수량, 나이, 기간 따위가 다 되다.

차다02 〔차, 차니〕
「동사」
【…을】
「1」 발로 내어 지르거나 발아 올리다.
「2」 발을 힘껏 뻗어 사람을 치다.
「3」 허끝을 입천장 앞쪽에 붙였다가 떼어 소리를 내다.
「4」 발로 힘 있게 밀어젖히다.
「5」 (속되게) 주로 남녀 관계에서 일방적으로 관계를 끊다.
「6」 날쌔게 빼앗거나 움켜 가지다.

차다03 〔차, 차니〕
「동사」
[1]【…에 …을】
「1」 물건을 몸의 한 부분에 달아매거나 끼워서 지니다.
「2」 수갑이나 차꼬 따위를 팔목이나 발목에 끼우다.
[2]【…을】
(속되게) 애인으로 삼아 데리고 다니다.
【<ᄎ다 <용가>】

차다04 〔차, 차니〕
「형용사」
[1] 「1」 몸에 닿은 물체나 대기의 온도가 낮다.
「2」 인정이 없고 쌀쌀하다.
「3」 『한의학』 약재(藥材)나 약제(藥劑)에 사람의 몸을 차갑게 하는 성질이 있다. ≒냉하다 「3」.
[2] 『북한어』 성격이 곧으면서도 냉철하다.
【<ᄎ다 <올곡>】

차다05
「형용사」 『방언』
‘짜다03’의 방언(제주).

Fig. 1. ‘차다’ in Korean Standard Great Dictionary

따라서 용언이 동형이의어 분별되면 용언별로 요구하는 문형에 따라 정확한 필수논항-용언의 의존관계를 분석할 수 있다. 본 논문에서는 동형이의어별 용언의 문형정보를 <표 4>와 같이 등록하고 이를 활용한다. <표 4>의 문형 정보는 표준국어대사전에서 추출(30,865개)하였다.

동형이의어별 용언이 다의어이고 다의어별로 다른 문형을 요구할 수 있지만(예, [그림 1]에서 ‘차다01’의 [1]과 [2]), 현재의 UTagger는 다의어를 분별할 수 없기 때문에 동형이의어 단위로 모든 문형을 등록하였다. 또한, ‘차다03’의 [1]【…에 …을】 문형과 [2]【…을】를 따로 구분하기 어려워 모두 “를 에 을”을 등록하였다. 여기서 ‘을’과 ‘를’, ‘로’와 ‘으로’, ‘과’와 ‘와’는 이형태 격조사이며 편의상 이형태들도 모두 등록하였다.

Table 2. Pattern of Predicate

관련하/VV	과 에 와
먹__01/VV	를 을
먹__02/VV	를 에 을
무역하/VV	과 를 와 을
상대하/VV	과 를 와 을
일어나/VV	에서
차__01/VV	로 에 으로
차__02/VV	를 을
차__03/VV	를 에 을

이렇게 <표 2>의 동형이의어별 문형정보를 이용하면, 문장 구성성분의 도치로 인한 연속된 두 개 이상의 본용언의 의존관계를 정확히 분석할 수 있다. 예를 들면,

(8) 일어나서 사과를 먹었다.
(9) 사과를 일어나서 먹었다.

의 두 문장에서 문장 (9)에서 “사과를”은 문장 (8)에서 도치된 경우로,

(9‘) 사과__05/NNG+를/JKO 일어나/VV+아서/EC 먹__02/VV+었/EP+다/EF+./SF

로 형태소/동형이의어 분석된다. <표 2>의 문형 정보를 이용하면 “사과를”은 목적격 조사 “를”을 문형으로 가지는 “먹__02/VV”와 의존관계를 형성할 수 있다(‘일으키다’는 목적격 조사를 필수논항으로 가지지 않는 자동사이다).

AF 전이모델로 학습된 경우의 {명사+격조사 용언} 간의 의존관계를 확정할 경우에도 용언의 문형정보는 다시 확인되어야 한다. 이는 앞 예문 (1‘)

(1‘) 쓰레기/NNG+를/JKO 타__01/VV+는/ETM 것__01/NNB+,/SP 타__01/VV+지/EC 않/VX+는/ETM 것__01/NNB 등__05/NNB+으로/JKB 구분하__03/VV+었/EP+다/EF+./SF

Table 3. Detail Rules used for Sejong Phrase Structured Corpus

규칙	규칙 설명 예문
1	격조사를 가진 체언은 가장 가까운 오른쪽 용언을 지배소로 가진다. (문형 확인) 시력에 따라 23종이 있다. (시력__01/NNG+에/JKB -> 따르__01/VV+아/EC)
2	“SS”를 제외한 기호의 경우 다음 어절을 지배소로 가진다. 렌즈 굴절력 - 내구성 잘 살펴야 (-/SO -> 내구성/NNG)
3	부사(MAG)에 대해서 바로 다음 어절이 부사이면 다음 부사를 지배소로 가지고, 그렇지 않으면 가장 가까운 오른쪽 용언을 지배소로 가진다. 배부른 것을 감추기 위해 너무 꼭 끼는 속옷을 입거나…(너무->꼭, 꼭->끼는)
4	부사 뒤에 명사가 왔을 때 부사가 {꼬박, 더, 바로, 순, 오래, 오직 등}이면 명사를 지배소로 가진다. 두 지역은 오래 전부터 심각한 분열을 겪어왔다. (오래__02/MAG->전__08/NNG+부터/JX)
5	관형격조사(JKG), 명사(NNG), 명사형전성어미(ETN), 명사파생접미사(XSN), 숫자(SN)의 경우 가장 가까운 오른쪽 명사를 지배소로 가진다. (예외, “결국”, 기호와 명사가 함께 나온 경우) 하지만 그의 가게는 장벽으로 두 동강이 나버린다. (그__01/NP+의/JKG->가게/NNG+는/JX)
6	관형사(MM), 관형형어미(ETM)는 다음 체언(명사/대명사/의존명사) 어절을 지배소로 가진다. 변해가는 한 아파트에 두 남녀가 산다. (변하/VV+어/EC+가__01/VX+는/ETM->아파트/NNG+에/JKB)
7	관형격조사(JKG)는 명사들이 연속으로 나올 경우 마지막 명사를 지배소로 가진다. 세기말의 인간 소외를 그려낸다. (세기말/NNG+의/JKG->소외/NNG+를/JKO)
8	명사 뒤에 기호(SO, SS, SP)가 올 경우 기호를 지배소로 가진다. 렌즈 굴절력 - 내구성 잘 살펴야 (굴절력/NNG -> -/SO)
9	명사에 기호(,/SP)가 붙어 연속으로 나오는 경우, 마지막 명사를 지배소로 가진다. 목욕가운부터 탁자보, 냅킨, 앞치마까지 그가 디자인한… (탁자__01/NNG+보__16/NNG+,/SP->앞치마/NNG+까지/JX)
10	연결어미가 연속으로 나오는 경우 다음 동사를 지배소로 가진다. 식사까지 할 수 있는 욕실이 나와야 한다고 주장할 정도다. (나오/VV+아야/EC->하__01/VX+ㄴ다고/EC)
11	JC의 경우 다음에 나오는 격조사(JKS, JKO, JKG, JKB, JX)를 지배소로 가진다. 팔순의 아버지와 두 딸 고등학생인 막내 아들은… (아버지/NNG+와/JC -> 아들/NNG+은/JX)
12	JX의 경우 주용언(제일 오른쪽 동사)을 지배소로 가진다. (예외, 바로 다음이 동사인 경우, 문장이 “,/SP”로 나누어진 경우) 이 단체는 일본군이 지배하는 … 발족시킨 항일 비밀 결사이다. (단체__02/NNG+는/JX -> 결사__04/NNG+이/VCP+다/EF+./SF)
13	“SW”가 여러 개 있을 경우 동일한 “SW”의 앞 어절을 지배소로 가진다. 동창회장단은 ▲ 일부 귀족학교화로 국민 위화감 조성 ▲ 공립고와 많… (▲/SW -> 조성__04/NNG)
14	접두사(XPN) 다음에 숫자(SN)가 오는 경우 SN을 지배소로 가진다. 수사(NR) 다음 관형사(MM)이나 부사(MAG)가 오는 경우 MM 혹은 MAG를 지배소로 가진다. 이들은 제 4회 부천시의회 임시회의가 열린 24일 오전에도… (제__21/XPN -> 4/SN+회__08/NNB) 우린 둘 다 더 깊은 외로움에 … (둘__01/NR -> 다__03/MAG)
15	한 어절 안에 형태소가 동사와 “SP”가 같이 사용될 경우 문장의 분리로 보고 주용언(문장 내에서 마지막에 온 용언)을 지배소로 가진다. 군 개혁은 … 사기를 떨어뜨렸고, 그런 분위기는 … 풍조로까지 번졌다. (떨어뜨리/VV+었/EP+고/EC+,/SP -> 번지__01/VV+었/EP+다/EF+./SF)
16	문장 시작이 “MAJ”인 경우 가장 오른쪽 용언을 지배소로 가진다. 그러나 미국의 경우 … 지어 집단 대응하는 일은 없다. (그러나/MAJ -> 없__01/VA+다/EF+./SF)
17	위의 모든 규칙에 해당하지 않으면 가장 오른쪽 어절을 지배소로 가진다. (용언이 없는 경우) 의상서 실내 장식품으로… (의상__01/NNG+서/JKB -> 장식품/NNG+으로/JKB+…/SE)

이 UTagger에서 학습되었다면 {쓰레기/NNG+를/JKO 타__01/VV, 것__01/NNB+,/SP 타__01/VV, 타__01/VV+지/EC 않/VX, 등__05/NNB+으로/JKB 구분하__03/VV}의 AF 전이모델이 학습되었을 것이고, “쓰레기를 타지 않는 …” 등의 문장의 의존관계를 분석할 때 {쓰레기/NNG+를/JKO 타__01/VV}에 의해 잘못된 의존관계를 형성하게 된다. 따라서, AF 전이모델을 적용할 때 해당 용언의 문형에 따른 필수논항인지를 판단하여 문형에 맞지 않는 경우는 의존관계를 설정하지 않아야 한다.

4. 의존관계 분석 시스템(UParser)

4.1 의존관계 분석 과정 및 세부 규칙

입력문장에 대해 형태소/동형이의어 분별된 UTagger의 결과를 대상으로 다음 [그림 2]의 절차에 따라 용언과의 의존관계를 분석한다. 본 논문에서는 용언이 2개 이상 사용된 경우, 의존관계 설정 대상 어절에 대해서 그 어절의 가장 가까운 오른쪽 용언까지를 의존관계 설정 범위로 제한한다.

① 용언(VV, VA, XSV, XSA, VCP, VCN)을 의존관계 분석 기점으로 설정한다.
② 의존관계 분석 기점 용언에 대해 right-to-left 방향의 어절에 대해 다음과 같이 AF 전이모델과 용언의 문형을 확인하여 의존관계를 확정한다.
  ②-1 문장의 첫 어절부터 첫 번째 분석 기점 용언 사이의 어절에 대해 첫 번째 분석 기점 용언과의 AF 전이모델이 학습되었다면 의존관계를 확정한다. (JX, JKS는 예외)
  ②-2 i-1번째 분석 기점 용언 다음 어절부터 i번째 분석 기점 용언 사이의 어절들에 대해 AF 전이모델을 이용한 의존관계를 확정한다.
  ②-3 ②-1과 ②-2에서 의존 관계를 확정할 때, AF 전이모델에서 발견된 격조사가 용언의 문형에 따른 필수 논항에 해당될 때만 의존 관계를 확정한다.
③ 아직 지배소가 결정되지 않은 어절에 대해서 <표 3>의 세부규칙과 의존제약 규칙을 적용하여 의존관계를 설정한다. 이때, 용언의 문형을 이용하여 필수논항에 대해 의존관계를 확정한다.

Fig. 2. Procedure of Analysis of Dependency Relation with Predicate

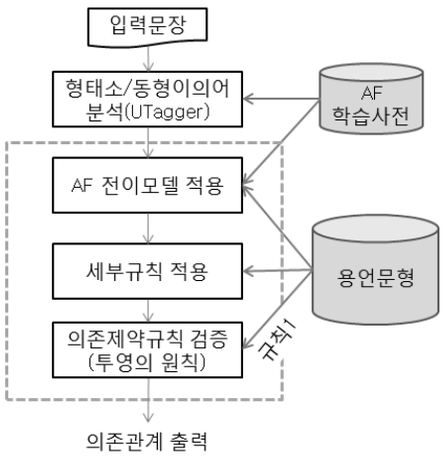


Fig. 3. Process of Analysis of Dependency Relation

세종구문분석말뭉치의 여러 유형의 문장을 분석하기 위하여 <표 1>의 기본 의존 규칙 외에 <표 3>의 17가지 세부규칙을 더 추가하였다. 각 규칙들은 1번부터 17번까지 순서대로 의존관계 분석에 적용된다. 즉 1번 규칙이 적용된 후 2번 규칙도 해당되면 그 어절은 2번 규칙에 의해 의존관계가 재설정된다.

앞에 설명한 과정을 바탕으로 설계한 의존관계 분석시스템(UParser)은 [그림 3]와 같이 진행된다. 입력 문장에 대해 UTagger를 이용하여 형태소 및 동형이의어 태깅을 실시한 후, 동형이의어 분별된 용언에 대해 용언의 문형정보를 이용하여 의존관계를 분석한다.

다음 두 개의 문장이 어떤 과정과 세부규칙이 적용되어 의존관계가 분석되는지 살펴보자.

(10) “오늘은 반드시 미국과 협상할 것을 기업인들이 주장했다.”의 문장은 다음 [그림 4]와 같이 분석과정 및 세부규칙이 적용되며, 최종 분석 결과는 [그림 5]와 같이 출력된다.

[그림 5]의 “3 4 21 미국__03/NNP+과/JKB”에서 “3 4”는 의존소(“미국과”)와 지배소(“협상할”)의 어절의 위치를 나타내며, “21”은 적용된 세부 규칙 번호이다. 세부 규칙 번호 “21”은 3.2절의 AF 전이모델에 의해 [그림 4]의 {명사+격조

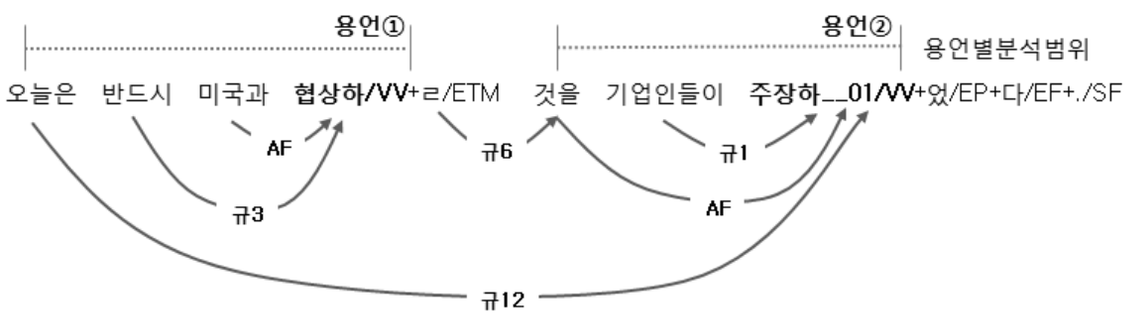


Fig. 4 Analysis Process and Applied Rules in Sentence (10)

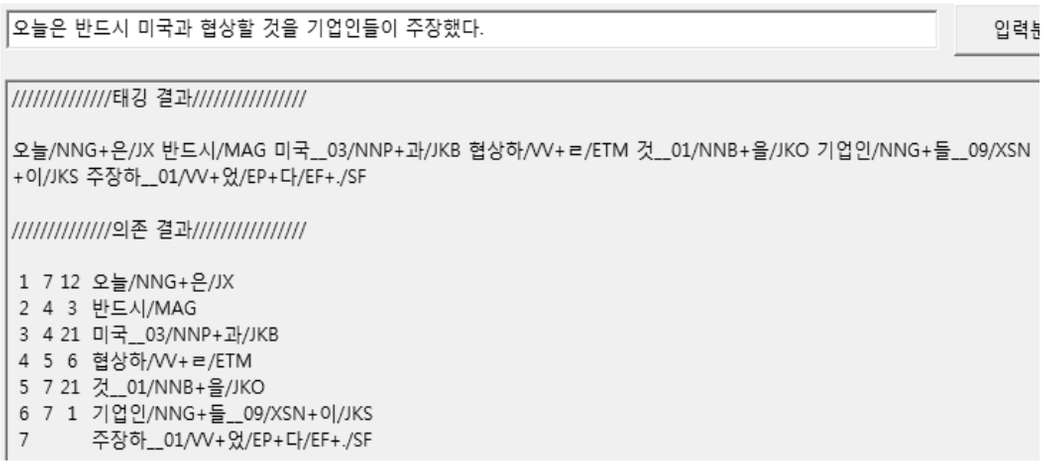


Fig. 5. Output of Analysis of Dependence Relation of Sentence (13)

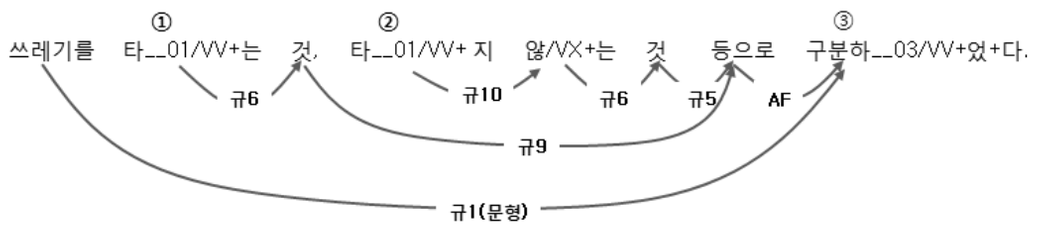


Fig. 6. Analysis Process and Applied Rules in Sentence (1)

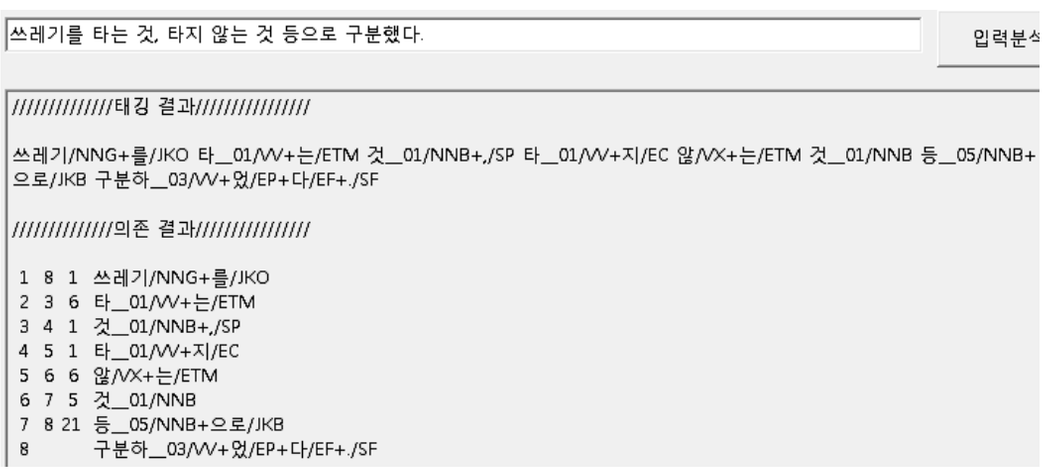


Fig. 7. Output of Analysis of Dependence Relation of Sentence (1)

사 용언}으로 학습된 경우(AF)에 해당되는 규칙번호이다. 마지막 어절인 “주장하__01/VV+었/EP+다/EF+./SF”는 지배소를 가지지 않는다.

[그림 6]은 예문 (1) “쓰레기를 타는 것, 타지 않는 것 등으로 구분했다.”의 의존관계 분석과정에서 적용되는 규칙이며, [그림 7]와 같이 분석결과가 출력된다.

[그림 7]의 “1 8 1 쓰레기/NNG+를/JKO”은 “타__01/VV”의 문형을 고려하지 않으면 2번 어절 “타는”을 지배소로 가지게 되나, 동형이의어 분별된 “타__01/VV”의 문형이 적용되어 “타는”을 지배소로 가지지 못한다. 결국 8번 어절 “구분하__03/VV”을 지배소로 가지게 된다. [그림 7]의 실행 결과에서 세 번째 어절 “것,”은 네 번째 어절 “타지”를 지배소로 가지고 있으나, 문장에서 “,”의 사용법이 다양하여 규칙 9에서 아직 이를 완전히 처리하지 못하고 있다(<표 4> 규칙 9의 정확률 참조). 또한 이 경우 세 번째 어절 “것,”의 지배소를 여섯 번째 어절 “것”으로 할 것인지, 일곱 번째 어절 “등으로”로 할 것인지에 대한 규정이 아직 명확히 정해지지 않았다.

4.2 실험용 세종구문분석말뭉치

21세기세종계획에서는 총 31개 파일, 77,136개 문장, 약 80만 어절의 구구조분석 말뭉치를 구축하였다. 이러한 세종구문분석말뭉치는 텍스트 선별에서 장르별 균형성 및 문장의 복잡도까지 고려하여 구축되었기 때문에 보편적인 구문적 특성을 포착할 수 있는 대표성을 갖춘 언어자료이다. [그림 8]은 세종구문분석말뭉치의 형태이다.

[그림 8]과 같이 세종구문분석말뭉치는 구구조 문법으로 태깅된 결과만을 제공하기 때문에 의존구조 말뭉치로 사용하지 못한다. 본 논문에서는 서강대 SKA[22]를 수정하여 구구조 트리를 의존 구조 트리인 형태로 변형하여 실험 데이터로 사용하였다. 이러한 수정 과정에서 SKA와는 다르게 보조용언과의 의존관계 등이 일부 수정되었다. 의존 구조로 변형된 말뭉치는 [그림 9]와 같다.

본 논문에서는 동형이의어 분별에 의한 의존관계 분석의 개선 효과를 비교하기 위해서, 31개의 세종구문분석말뭉치 파일 중 동형이의어가 부착된 세종형태의미말뭉치에서도 사용된 21개의 파일에 대해서만 실험말뭉치로 사용하였다. 동형이의어가 부착된 의존 구조 말뭉치는 [그림 10]과 같다. [그림 10]에서는 동형이의어가 부착되었을 뿐만 아니라 서술성명사 용언의 경우(11번째 어절 “군림하고”, 15번째 어절 “묵살되고”) 본래 어근형을 모두 어간형으로 수정하였다([그림 9]와 비교). UTagger의 학습말뭉치들은 어간형으로 학습되어 있어 UTagger의 실행결과와 일치시키기 위해서 어간형으로 수정하였다. 실험에 사용된 동형이의어 부착된 의존구조 말뭉치는 21개 파일, 39,300개 문장, 413,184개 어절(문장 당 평균 10.5개 어절), 373,884개의 의존관계를 가지고 있다.

UTagger의 동형이의어 분별된 결과가 의존관계 분석에

```
; 지금 서울에서 열리고 있는 ANOC는 NOC의 총회이지만 그 위에 IOC가 군림하고 있어 ANOC의 제안이 묵살된 경우도 더러 있다.
(S          (S          (NP_SBJ   (VP_MOD  (AP 지금/MAG)
                                    (VP_MOD  (NP_AJT 서울/NNP + 에서/JKB)
                                             (VP_MOD  (VP 열리/VV + 고/EC)
                                                      (VP_MOD 있/VX + 는/ETM))))
                                    (NP_SBJ ANOC/SL + 는/JX))
                        (VNP      (NP_MOD NOC/SL + 의/JKG)
                                    (VNP 총회/NNG + 이/VCP + 지만/EC)))
            (S          (S        (NP_AJT   (DP 그/MM)
                                    (NP_AJT 위/NNG + 에/JKB))
                                  (S        (NP_SBJ IOC/SL + 가/JKS)
                                            (VP       (VP 군림/NNG + 하/XSV + 고/EC)
                                                      (VP 있/VX + 어/EC))))
                        (S        (NP_SBJ   (S_MOD    (NP_SBJ   (NP_MOD ANOC/SL + 의/JKG)
                                                                  (NP_SBJ 제안/NNG + 이/JKS))
                                                      (VP_MOD 묵살/NNG + 되/XSV + ㄴ/ETM))
                                    (NP_SBJ 경우/NNG + 도/JX))
                                  (VP       (AP 더러/MAG)
                                            (VP 있/VV + 다/EF + ./SF)))))
```

Fig. 8. Sejong Phrase Structured Corpus (File BGJ00152)

```
; 지금 서울에서 열리고 있는 ANOC는 NOC의 총회이지만 그 위에 IOC가 군림하고 있어 ANOC의 제안이 묵살된 경우도 더러 있다.
1       3       지금/MAG AP_NONE
2       3       서울/NNP + 에서/JKB NP_AJT
3       4       열리/VV + 고/EC VP_NONE
4       5       있/VX + 는/ETM VP_MOD
5       7       ANOC/SL + 는/JX NP_SBJ
6       7       NOC/SL + 의/JKG NP_MOD
7       18      총회/NNG + 이/VCP + 지만/EC VNP_NONE
8       9       그/MM DP
9       11      위/NNG + 에/JKB NP_AJT
10      11      IOC/SL + 가/JKS NP_SBJ
11      12      군림/NNG + 하/XSV + 고/EC VP_NONE
12      18      있/VX + 어/EC VP_NONE
13      14      ANOC/SL + 의/JKG NP_MOD
14      15      제안/NNG + 이/JKS NP_SBJ
15      16      묵살/NNG + 되/XSV + ㄴ/ETM VP_MOD
16      18      경우/NNG + 도/JX NP_SBJ
17      18      더러/MAG AP_NONE
18      0       있/VV + 다/EF + ./SF VP_NONE
```

Fig. 9. Sejong Phrase Structured Corpus transformed to Dependency Structure

```
; 지금 서울에서 열리고 있는 ANOC는 NOC의 총회이지만 그 위에 IOC가 군림하고 있어 ANOC의 제안이 묵살된 경우도 더러 있다.
1       3       지금__03/MAG
2       3       서울__01/NNP + 에서/JKB
3       4       열리__02/VV + 고/EC
4       5       있__01/VX + 는/ETM
5       7       ANOC/SL + 는/JX
6       7       NOC/SL + 의/JKG
7       18      총회__02/NNG + 이/VCP + 지만/EC
8       9       그__01/NP
9       11      위__01/NNG + 에/JKB
10      11      IOC/SL + 가/JKS
11      12      군림하/VV + 고/EC
12      18      있__01/VX + 어/EC
13      14      ANOC/SL + 의/JKG
14      15      제안__02/NNG + 이/JKS
15      16      묵살되/VV + ㄴ/ETM
16      18      경우__03/NNG + 도/JX
17      18      더러__01/MAG
18      0       있__01/VA + 다/EF + ./SF
```

Fig. 10. Sejong Dependency Structured Corpus tagged Homograph

미치는 영향을 명확히 파악하기 위하여, 실험으로 사용하는 21개의 세종구문분석말뭉치는 제외하고 UTagger용의 학습말뭉치를 새로이 구축하여 실험하였다. UTagger의 학습말뭉치는 318개 파일, 10,570,836개 어절로 구성되어 있으며, 논문[21]의 단계별 전이모델 학습말뭉치도 새로이 구축하였다.

4.3 전체 성능 실험 및 결과

동형이의어 부착되고 의존관계로 바뀐 [그림 10]의 세종구문분석말뭉치를 정답으로 보고 의존관계 분석 시스템의 의존관계 분석 결과의 정확률을 비교하였다.

$$\text{정확률} = \frac{\text{정확하게 분석된 의존관계 수}}{\text{전체 의존관계 수}}$$

실험말뭉치에 대해서 UTagger 결과 14,968 어절이 형태소 및 동형이의어 분석 오류가 발생하여 96.38%의 정확률을 보였으며, 이 오류로 인한 다음 단계의 의존관계 오류도 의존관계 분석 정확률 계산에 포함하였다.

동형이의어 분별 여부가 의존관계 분석에 어느 정도의 영향을 미치는지를 분석하기 위하여 다음과 같이 실험을 진행하였다. 첫째로 UTagger의 결과가 분별된 동형이의어를 제거하고 <표 3>의 규칙만 적용하여 의존관계를 분석하였다. 둘째로 동형이의어 분별된 결과를 이용하여 [그림 3]과 같이 의존관계를 분석하였다. 이 두 개의 결과를 비교하여 동형이의어 분별된 결과가 의존관계 분석에 미치는 정도를 파악하였다. <표 4>는 규칙만 적용한 경우와 동형이의어 분별 경우에 대해 각 규칙별 정답 및 오답 빈도이다.

<표 4>에서 규칙 22는 [그림 3]의 의존관계 분석 과정의 마지막 단계인 의존제약규칙 의 투영의 원칙에 위배되는 경우이다. 현재 이 경우는 다른 규칙에 비해 정확률은 상당히 낮다. 향후 규칙 22의 오류 원인을 면밀히 분석하여 개별 규칙이 적용되는 범위 및 순서 조정 등의 보완이 필요하다.

Table 4. Correct and Error Frequency from Rule and Homograph

규칙 번호	규칙만 적용			동형이의어 분별(AF)		
	정답	오답	정답률	정답	오답	정확률
1	159,824	31,092	83.71	132,499	28,655	82.22
2	176	137	56.23	177	137	56.37
3	5,922	5,007	54.19	6,036	4,974	54.82
4	15	3	83.33	165	77	68.18
5	31,664	7,335	81.19	31,940	7,134	81.74
6	48,266	7,956	85.85	48,276	7,979	85.82
7	16,252	2,368	87.28	16,265	2,372	87.27
8	449	116	79.47	449	112	80.04
9	1,296	1,979	39.57	1,297	1,979	39.59
10	9,333	7,596	55.13	10,100	7,537	57.27
11	3,880	1,806	68.24	3,956	1,733	69.54
12	13,805	6,552	67.81	14,128	6,794	67.53
13	123	31	79.87	126	29	81.29
14	46	30	60.53	46	30	60.53
15	1,850	1,033	64.17	1,870	1,036	64.35
16	2,365	705	77.04	2,462	605	80.27
17	3,163	106	96.76	3,163	105	96.79
AF	0	0		26,713	664	97.57
22	511	1,092	31.88	870	1,393	38.44
합계	298,940	74,944	79.96	300,538	73,345	80.38

<표 4>에서 용언이 동형이의어 분별된 경우(AF) 전체 정확률은 80.38%로 동형이의어가 분별되지 않았을 때보다 0.42%의 의존관계 분석 정확률이 향상하였다. 동형이의어 분별 전후의 의존관계 분석에 대해 유의수준 1%에서 검정통계량 $Z = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ 으로 검정하면,

기각역 $|Z = 4.63466 \geq z_{0.01} = 2.33$이므로 동형이의어 분별이 의존관계 분석에 유의미한 영향을 미치는 것을 알 수 있다. 여기서,

$$p1 = \frac{298,940}{(298,940 + 74,944)} = 0.799553,$$

$$p2 = \frac{300,538}{(300,538 + 73,346)} = 0.803827,$$

$$p = \frac{(298,940 + 300,538)}{(373,884 + 373,884)} = 0.80169,\ n1 = n2 = 373,884$$ 이다.

<표 4>에서 AF 전이모델이 적용된 어절(총 27,377개 어절, 정답 26,713, 오답 664개)에 대해서 규칙만 적용하였을 때의 의존관계 분석 결과와 비교하면 다음 <표 5>와 같다.

Table 5. Effect of Accuracy of Dependency Relation by Disambiguation Homograph

AF 적용 / 규칙만 적용	AF 적용시 정답	AF 적용시 오답	규칙 적용 합계
규칙 적용시 정답	26,351	21	26,372
규칙 적용시 오답	362	643	1,005
AF 적용 결과 합계	26,713	664	27,377

<표 5>를 보면 AF 적용 시 정답의 결과가 나온 총 26,713개 어절 중 26,351개 어절은 규칙만 적용했을 때도 정답으로 분석되었다. 즉 98.64%의 대부분의 어절은 굳이 동형이의어를 분별하지 않고 기존의 의존규칙만으로도 의존관계를 정확하게 분석한다. 이는 본 논문에서 사용하고 있는 근거리 의존관계를 적용(어절의 오른쪽의 가장 가까운 용언과 의존관계 설정)하더라도 대부분 정확한 의존관계를 설정할 수 있음을 알 수 있다. 반면, 규칙만 적용했을 때의 오답 1,005개 어절 중 362개의 어절은 동형이의어가 분별됨으로써 정답으로 분석되었다. 즉 오답에 대해 36.02%(362/1005)의 개선효과가 있다. 이는 다른 문형을 가지는 동형이의어 용언에 대해서는 동형이의어를 분별할 필요가 있음을 나타낸다. 또한, 규칙 적용시에는 정답이었던 것이 동형이의어 분별됨으로써 오히려 오류로 분석된 21가지가 경우가 발생하였으나, 이 중 14개는 정답 자체가 오류이었다.

AF 전이모델에서 학습된 {명사+격조사 용언} 간의 의존관계를 확정할 때 3.4절과 같이 용언의 문형을 고려했을 경우와 그렇지 않았을 경우의 정확률을 비교하였으며 그 결과는 <표 6>과 같다.

Table 6. Effect of Pattern of Predicate in AF Transition Model

AF 적용	문형 고려하지 않음	문형 고려함
AF 적용된 관계 수	34,601	27,377
의존관계 적용 비율	9.25%	7.32%
정답 의존관계 수	33,319	26,713
AF전이 적용 정확률	96.29%	97.57%

<표 6>에서 전체 의존관계 중에서 단순히 UTagger의 AF 학습사전에서 발견되어 의존관계를 확정할 수 있는 경우가 총 373,884개 어절 중 34,601개로 약 9.25%를 차지하였으며, 이 중에서도 문형까지 확인하여 의존관계를 확정할 수 있는 경우는 23,777개로 약 7.32%를 차지하였다. 또한 AF 전이모델에 의한 의존관계 확정 시 문형을 고려한 경우의 정확률은 97.57%로 문형을 고려하지 않았을 때 비해 1.28%의 정확률 향상이 있다. 결국 AF 전이모델의 학습사전이 의존관계 분석 정확률에 약 7.14%(0.0732 x 0.9757)의 영향을 미침을 보이고 있다.

5. 결론 및 향후 연구

본 논문은 형태소 분석 단계에서 분별된 동형이의어 정보를 이용하여, 동형이의어 용언별로 다른 문형의 의존관계 분석 방안을 제안하였다. 또한 형태소 분석 단계에서 동형이의어를 분별하는 UTagger에서 사용하는 AF 전이 학습사전을 재활용하여 {명사+격조사 용언} 간의 의존관계를 확정하는 방안을 제안하였다.

제안한 방법의 성능 평가를 위해 31개 파일의 세종구문분석말뭉치 중에서 세종형태의미말뭉치에서도 사용된 21개 파일을 대상으로 실험하였다. 실험말뭉치는 동형이의어를 부착하였고 의존관계 구조로 변환하였으며, 전체 39,300 문장, 413,185 어절, 373,884개의 의존관계로 구성되어 있다.

실험결과 동형이의어 분별된 경우 전체 정확률은 80.38%로, 동형이의어가 분별되지 않았을 때보다 0.42%의 의존관계 분석 정확률이 향상하였다. 이러한 결과는 유의수준 1%에서 검정통계량 Z로 검정하면, 기각역 $|Z = 4.63 \geq z_{0.01} = 2.33$로 동형이의어 분별이 의존관계 분석에 유의미한 영향을 미치는 것을 알 수 있었다.

또한, 실험결과 전체 의존관계 중에서 UTagger의 AF 전이모델이 적용되어 의존관계를 확정할 수 있는 경우가 34,601개로 약 9.25%를 차지하였으며, 이 중에서 문형을 고려하여 정확히 의존관계를 확정할 수 있는 경우는 27,377개로 약 7.32%를 차지하였다. 또한 AF 전이모델에 의한 의존관계 확정 시 문형을 고려한 경우의 정확률은 97.57%로 문형을 고려하지 않았을 때 비해 1.28%의 정확률 향상이 있다. 결국 AF 전이모델로 학습된 학습사전이 의존관계 분석 정확률에 약 7.14%의 영향이 있음을 보이고 있다.

앞으로 보다 정확한 의존관계 분석을 위해서는 각 규칙별 오류 유형을 면밀히 분석하여 개선 방안을 제시하여야 할 것이며, 격조사가 생략된 경우와 보조사에 대한 의존 규칙을 보강할 필요가 있다. 또한, 서술성 명사와의 의존관계 설정에 대해 면밀한 분석이 필요하다.

Reference

[1] J. Y. Oh and J. W. Cha, “Korean Dependency Parsing using Key Eojoel”, Journal of KIISE : Software and Applications, Vol.40, No.10, pp.600-608, 2013.

[2] Y. U. Park and H. C. Kwon, “A Study of Parsing System Implementation Using Segmentation and Argument Information”, Journal of Korea Multimedia Society, Vol.16, No.3, pp.366-374, 2013.
[3] E. K. Park and D. Y. Ra, “Processing Dependent Nouns Based on Chunking for Korean Syntactic Analysis”, The Korean Society for Cognitive Science, Vol.17, No.2, pp.119-138, 2006.
[4] M. G. Jang, H. A. Lee, J. D. Park, and D. I. Park, “Korean Dependency Parser Using Subcategorization Information of Predicates”, Journal of KIISE, pp.452-463, 1996.
[5] S. J. Lim, Y. T. Kim, and D. Y. Ra, “Korean Dependency Parsing Based on Machine Learning of Feature Weights”, Journal of KIISE : Software and Applications, Vol.38, No.4, pp.214-223.
[6] Y. H. Lee and J. H. Lee, “Korean Dependency Parsing Using Online Learning”, Journal of KIISE, pp.299-304, 2010.
[7] S. W. Jung, E. K. Park, D. Y. Ra, and J. T Yoon, “A Study on Korean Dependency Parser Using Case Relation and Mutual Information”, Journal of KIISE, pp.450-455, 2001.
[8] Y. M. Park and J. Y. Seo, “Correction Method for Korean Dependency Parsing using Projectivity and Re-searching”, Korean Journal of Cognitive Science, Vol.22, No.4, pp.429-447, 2011.
[9] P. M. Ryu, T. S. Lee, J. H. Lee, and G. B. Lee, “Two-Phase Dependency Parser of Korean Using Predicate-Driven Constraint Propagation”, Journal of KIISE, pp.923-926, 1996.
[10] S. S. Kim, S. B. Park, and S. J. Lee, “Analyzing Dependency of Korean Subordinate Clauses Using Support Vector Machine”, Journal of KIISE, pp.148-155, 2006.
[11] G. E. Im, Y. G. Jung, and H. C. Kwon, “Implementation of Dependency Parser using Argument Information based on Korean WordNet”, Journal of KIISE, pp.158-164, 2007.
[12] M. Y. Kim, S. J. Kang, and J. H. Lee, “Dependency Parsing by Chunks”, Journal of KIISE, pp.327-329, 2000.
[13] S. W. Lee, “Cascaded Parsing Korean Sentences Using Grammatical Relations”, pp.69-72. 2008.
[14] S. H. Choi and H. R. Park, “Probabilistic Dependency Grammar Induction”, The KIPS transactions, pp.513-515, 2003.
[15] S. S. Kim, S. B. Park, S. J. Lee, and S. Y. Park, “Analyzing dependency of Korean subordinate clauses using a composit kernel”, Korean Journal of Cognitive Science, Vol.19, No.1, pp.1-15, 2008.
[16] P. M. Ryu, J. H. Lee, and G. B. Lee, “Using Local Dependency for Dependecy Parser of Korean”, The KIPS transactions, pp.464-468, 1996.
[17] J. H. Eun, M. W. Jeong, and G. B. Lee, “Korean Dependency Structure Analyzer based on Probabilistic Chart Parsing”, Journal of KIISE, pp.105-111, 2005.
[18] S. B. Park and B. T. Zhang, “A Hybrid of Rule based Method and Memory based Learning for Korean Text Chunking”, Journal of KIISE : Software and Applications, Vol.31, No.3, pp.369-378, 2004.
[19] L. K. Joo and J. H. Kim, “Implementing Korean Partial Parser based on Rules”, pp.389-396, 2003.
[20] Y. M. Woo, Y. I. Song, S. Y. Park, and H. C. Rim, “Modification Distance Model using Headible Path Contexts for Korean Dependency Parsing”, Journal of KIISE : Software and Applications, Vol.34, No.2, pp.140-149, 2007.
[21] J. C. Shin and C. Y. Ock, “A Stage Transition Model for Korean Part-of-Speech and Homograph Tagging”, Journal of KIISE : Software and Applications, Vol.39, No.11, pp.889-901, 2012.
[22] Y. M. Park and J. Y. Seo, “SKA(Sogang Korean dependency Analyzer)”, Competition of Korean Information Processing System, 2011.

김 홍 순

e-mail : rlaghdtns2@ulsan.ac.kr
2011년 울산대학교 컴퓨터・정보통신학부 (학사)
2013년 울산대학교 정보통신공학과 석사
현 재 미디어젠 연구원
관심분야 : 한국어정보처리, 구문분석, 자연어처리

옥 철 영

e-mail : okcy@ulsan.ac.kr
1982년 서울대학교 컴퓨터공학과(학사)
1984년 서울대학교 컴퓨터공학과(석사)
1993년 서울대학교 컴퓨터공학과(박사)
1994년 러시아 TOMSK 공과대학 교환교수
1996년 영국 GLASGOW 대학교 객원교수
2007년~2008년 한국정보과학회 언어공학연구회 위원장
2007년 몽골국립대학교 IT대학 명예박사학위
2008년 국립국어원 객원연구원
1984년~현 재 울산대학교 컴퓨터정보통신공학부 교수
관심분야 : 한국어정보처리, 온톨로지, 지식베이스, 기계학습, 문서분류