

국립국어원 2019-01-51

발 간 등 록 번 호
11-1371028-000768-01

구문 분석 말뭉치 구축

사업 책임자
임 성 모

제 출 문

국립국어원장 귀하

국립국어원과 체결한 연구용역 계약에 따라 ‘구문 분석 말뭉치 구축’에 관한 연구 보고서를 작성하여 제출합니다.

■ 사업기간: 2019년 08월 ~ 2020년 02월

2020년 02월 06일

사업 책임자: 임성모(주식회사 마인즈랩)

사업 수행자 주식회사 마인즈랩

주식회사 딥네츄럴

연세대학교 산학협력단

사업 책임자 임성모

사업 참여자 서상원, 이석준, 이원문, 박영선, 송혜원, 윤서영
박지원, 이예준, 김한샘, 유현경, 김재훈, 이공주
김유섭, 류범모, 김학수, 신서인, 나승훈, 봉미경,
김선혜, 김수경, 이찬영, 박혜진, 장연지, 신아영,
정주연, 정진경, 강혜린, 김교연, 김상민, 김지영,
정해윤, 천성호, 박진현, 이수현, 이한범, 전상호,
김유미, 이지원, 김재균, 남궁영, 윤호, 최민석,
최용석, 박천용, 오병두, 허탁성, 민진우, 이영훈,
홍승연, 박성식, 신영진, 강일민, 박요한, 정혜지,
오세은, 황석주, 강동찬, 이종현, 최형준, 김담린,
김보은, 김홍진, 오신혁, 박상원, 박정수, 허민강,
박연호, 정동호, 최진혁, 김예진, 이규덕, 임선민

〈사업 수행자〉 주식회사 마인즈랩 · 주식회사 딥네츄럴 ·
연세대학교 산학협력단

사업 책임자	임성모(주식회사 마인즈랩)
사업 참여자	서상원(주식회사 마인즈랩)
	이석준(주식회사 마인즈랩)
	이원문(주식회사 마인즈랩)
	박영선(주식회사 마인즈랩)
	송혜원(주식회사 마인즈랩)
	윤서영(주식회사 마인즈랩)
	박지원(주식회사 마인즈랩)
	이예준(주식회사 마인즈랩)
	김한샘(연세대학교)
	유현경(연세대학교)
	김재훈(한국해양대학교)
	이공주(충남대학교)
	김유섭(한림대학교)
	류범모(부산외국어대학교)
	김학수(강원대학교)

사업 참여자	신서인(한림대학교)
	나승훈(전북대학교)
	봉미경(연세대학교)
	김선희(연세대학교)
	김수경(연세대학교)
	이찬영(연세대학교)
	박혜진(연세대학교)
	장연지(연세대학교)
	신아영(연세대학교)
	정주연(연세대학교)
	정진경(연세대학교)
	강혜린(연세대학교)
	김교연(연세대학교)
	김상민(연세대학교)
	김지영(연세대학교)
	정해윤(연세대학교)
	천성호(연세대학교)
	박서윤(연세대학교)

사업 참여자	박진현(한림대학교)
	이수현(한림대학교)
	이한범(한림대학교)
	전상호(한림대학교)
	김유미(한림대학교)
	이지원(한림대학교)
	김재균(한국해양대학교)
	남궁영(한국해양대학교)
	윤호(한국해양대학교)
	최민석(한국해양대학교)
	최용석(충남대학교)
	박천용(충남대학교)
	오병두(한림대학교)
	허탁성(한림대학교)
	민진우(전북대학교)
	이영훈(전북대학교)
	홍승연(전북대학교)
	박성식(강원대학교)

사업 참여자	신영진(한국해양대학교)
	강일민(충남대학교)
	박요한(충남대학교)
	정혜지(충남대학교)
	정영석(한림대학교)
	오세은(한림대학교)
	황석주(한림대학교)
	강동찬(전북대학교)
	이종현(전북대학교)
	최형준(전북대학교)
	김담린(강원대학교)
	김보은(강원대학교)
	김홍진(강원대학교)
	오신혁(강원대학교)
	박상원(주식회사 딥네츄럴)
	박정수(주식회사 딥네츄럴)
	허민강(주식회사 딥네츄럴)
	박연호(주식회사 딥네츄럴)

사업 참여자	정동호(주식회사 덩네츄럴)
	최진혁(주식회사 덩네츄럴)
	김예진(주식회사 덩네츄럴)
	이규덕(주식회사 덩네츄럴)
	임선민(주식회사 덩네츄럴)

구문 분석 말뭉치 구축

본 사업은 인공지능 산업 발전을 위한 대규모 우리말 자원 수요를 위한 구문 분석 말뭉치 구축을 목적으로 한다. 또한 이 과정에서 요구되는 구문 분석 말뭉치 지침을 수립하는 것 또한 본 과제의 목적에 포함된다.

본 사업의 범위는 크게 두 부분으로 나눌 수 있다. 첫째는 구문 분석 말뭉치 지침 수립으로, 한국정보통신기술협회(TTA) 등 관련 분야의 구문 분석 지침을 참고하여 비교·연구하였으며 한국전자통신연구원(ETRI)의 한국어 의존 구문 분석 가이드라인 등 관련 분야의 지침 등을 바탕으로 기존 지침의 문제점을 수정·보완한 의존 구문 분석 지침을 수립하였다. 둘째는 구문 분석 말뭉치(200만 어절) 구축으로, 구문 분석 말뭉치 구축 지침을 바탕으로 말뭉치를 구축하였다.

○ 구문 분석 말뭉치 지침 수립

구문 분석 지침을 수립하기 위하여 ‘한국정보통신기술협회(TTA)’ 등 관련 분야의 구문 분석 지침을 검토하여 기존 지침의 문제점을 분석, 보완된 지침을 제시하였다.

기존의 의존 구문 분석 말뭉치 지침은 주로 <21세기 세종계획>에서 구축된 구 구조 기반 구문 분석 말뭉치를 의존 구문 분석 말뭉치로 변환하는 것에 초점을 맞추어 개발되었으며 이에 따라 의존 구문 분석 말뭉치를 구축하기 위한 지침은 그 양적·질적 측면에서 부족하였다고 할 수 있다. 본 사업에서는 한국정보통신기술협회(TTA)의 ‘의존 구문 분석 말뭉치 구축을 위한 의존 관계 태그 세트 및 의존 관계 설정 방법’을 기반으로 하여 의존 구문 분석 말뭉치를 구축하기 위한 실용적인 지침을 개발하고 기존 의존 구문 분석 지침에서 부족하다고 판단되는 내용과 예시를 보완하였다.

○ 구문 분석 말뭉치(200만 어절) 구축

구문 분석 말뭉치의 구축 절차는 다음과 같다.

구문 분석 지침 수립 > 수작업 검수 도구 커스터마이징 > 작업자 교육 > 자동 구문 분석 > 작업자 분석(1차 검수) > 팀장 및 교수진 검수(2차 검수) > 딥 러닝 기반 구문 분석 말뭉치 검증 > 최종 결과물 산출

의존 구문 분석 지침을 바탕으로 200만 어절 규모의 구문 분석 말뭉치를 구축하였다. 말뭉치는 문어 200만 어절 규모로 신문 기사 텍스트로 구성되어 있다.

본 사업에서는 국립국어원에서 제공한 문어 말뭉치를 대상으로 문장 단위 구문 분석을 진행하였다. 또한 각 어절별로 구문 분석 정보를 부착하였으며 제이슨(JSON) 형식의 최종 산출물을 제출하였다.

구축 과정에서 다수의 자동 구문 분석기를 활용하여 작업의 편의성과 일관성을 확보하고자 하였으며 자동으로 분석된 결과를 작업자들이 수작업으로 전수 검토하도록 하였다. 상세한 과정은 다음과 같다. 먼저 국립국어원에서 제공한 신문 기사 말뭉치를 문장 단위를 기준으로 자동 구문 분석을 시행한다. 이때 복수의 자동 분석기 결과가 전체 일치하는 문장에 대해서는 최종 검수자의 검수만으로 검수가 완료되며 자동 분석 결과가 전체 일치하지 않는 문장에 대해서는 1차 검수와 2차 검수를 거쳐 검수가 완료된다. 작업자는 담당 교수, 팀장을 포함하여 총 4개 조로 편성되었으며 각 조의 담당 교수 및 팀장이 나머지 작업자의 1차 검수 결과물을 검수하는 방식으로 2차 검수를 실시하였다. 2차 검수 후에는 후처리 및 파일 형식 변환을 거쳐 최종 산출물을 제출하였다.

본 사업에서는 자동 구문 분석을 위하여 한국전자통신연구원(ETRI), 강원대학교, 전북대학교, 충남대학교의 의존 구문 분석기를 활용하여 자동 분석을 진행하였고, 다수의 기관에서 자동으로 분석한 결과를 교차 검증 및 통합하여 신뢰도를 높이하고자 하였다. 자동 구문 분석의 과정에서는 딥 러닝을 통한 일관성 검증을 진행하였다.

주요어: 구문, 구문 분석, 의존 구문 분석, 말뭉치, 구문 분석 말뭉치

차 례

제1장 서론

1. 사업의 목적	2
2. 사업의 범위	2
2.1. 구문 분석 말뭉치 구축 지침 수립	2
2.2. 구문 분석 말뭉치 구축	3

제2장 구문 분석 말뭉치의 구성 및 구축 절차

1. 구문 분석 말뭉치의 구성	5
2. 구문 분석 말뭉치 구축 절차	5
2.1. 구문 분석 지침 수립	5
2.2. 수작업 검수 도구 커스터마이징	7
2.3. 작업자 교육	15
2.4. 자동 구문 분석	15
2.4.1. 엑소브레인(Exobrain) 언어분석기(WiseNLU)	16
2.4.2. 강원대학교 딥 러닝 기반 의존 구문 분석 모델	17
2.4.3. 전북대학교 딥 러닝 기반 자동 구문 분석 모델	18
2.4.4. 충남대학교 자동 의존 구문 분석 모델	23
2.5. 작업자 분석(1차 검수)	27
2.6. 팀장 및 교수진 검수(2차 검수)	28
2.7. 딥 러닝 기반 구문 분석 말뭉치 검증	28
2.7.1. 해양대학교 구문 분석 말뭉치 검증 모델	28
2.7.2. 전북대학교 딥 러닝 기반 구문 분석 말뭉치 검증 모델	33
2.8. 전문가 집단 심층 면접(Focus Group Interview)	36
2.9. 최종 결과물 산출	38

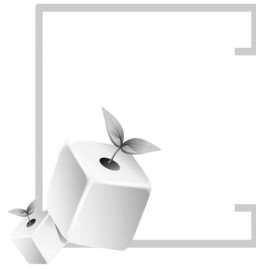
차 례

제3장 구문 분석 말뭉치 구축 지침 수립

1. 지침 수립 과정	41
1.1. 기본 원칙	41
1.2. 의존 관계 태그 세트 설정 방법	41
1.3. 문장 유형별 의존 관계 설정 방법	41
1.4. 세부 구별 태깅 가이드라인	43
1.5. 세부 유형별 가이드라인	48
2. 구문 분석 말뭉치 구축 지침	53

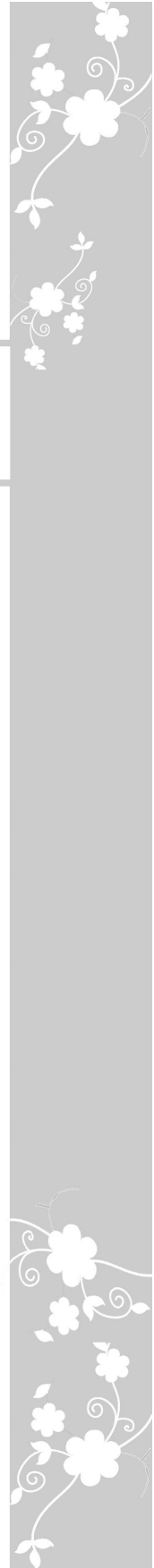
제4장 결론

Abstract	95
<부록 1> 제이슨(JSON) 형식의 기본 구조	98
<부록 2> 제이슨(JSON) 형식의 예시	100
<부록 3> 원시 말뭉치 엑스엠엘(XML) 형식의 예시	101



제 1 장

서 론



1. 사업의 목적

- 구문 분석 말뭉치 지침 수립
- 구문 분석 말뭉치(200만 어절) 구축

본 사업은 인공 지능 산업 발전을 위한 대규모 우리말 자원 수요를 위한 구문 분석 말뭉치 구축을 목적으로 한다. 또한 이 과정에서 요구되는 구문 분석 말뭉치 지침을 수립하는 것 또한 본 과제의 목적에 포함된다.

본 사업은 인공 지능 산업 발전을 위한 대규모 고품질 우리말 자원 수요의 증대의 필요성에 대한 이해를 바탕으로 4차 산업 혁명 대비 대규모 말뭉치 구축으로 국어 자원의 활용도와 가치 제고를 위하여 수행되었다. 기존의 구문 분석 말뭉치는 양적·질적으로 부족하였으며 다양한 산업의 기반이 될 대규모 언어 자원의 부족으로 인하여 인공 지능 등의 기술 개발 수준이 지체되어 있었다. 이에 공공재 말뭉치를 구축하여 이를 바탕으로 4차 산업 혁명의 기반 기술 개발을 도모할 필요성이 나타났으며 이러한 자료를 질적·양적으로 수준 높게 개발하기 위하여 다양한 말뭉치 전문가 인력을 활용하여 구문 분석 말뭉치를 구축하였다.

본 사업에서는 한국정보통신기술협회(TTA) 등 관련 분야의 구문 분석 지침을 참고하여 세부 분석 지침을 수립하고 문어 200만 어절 규모의 구문 분석 말뭉치를 구축하였다. 한국전자통신연구원(ETRI)의 한국어 의존 구문 분석 가이드라인 등 관련 분야의 지침 등을 바탕으로 기존 지침의 문제점을 수정·보완한 의존 구문 분석 지침을 수립하였다.

2. 사업의 범위

2.1 구문 분석 말뭉치 구축 지침 수립

구문 분석 지침을 수립하기 위하여 ‘한국정보통신기술협회(TTA)’ 등 관련 분야의 구문 분석 지침을 검토하여 기존 지침의 문제점을 분석, 보완된 지침을 제시하였다.

기존의 의존 구문 분석 말뭉치는 주로 <21세기 세종계획>에서 구축된 구 구조 기반 구문 분석 말뭉치를 의존 구문 분석 말뭉치로 변환하는 것에 초점을 맞추어 개발되었으며 이에 따라 의존 구문 분석 말뭉치를 구축하기 위한 지침은 그 양적·질적 측면에서 부족하였다고 할 수 있다. 본 사업에서는 한국정보통신기술협회의 의존 구문 분석 가이드라인을 기반으로 하여 의존 구문 분석 말뭉치를 구축하기 위한 실용적인 지침을 개발하고 기존 의존 구문 분석 지침에서 부족하다고 판단되는 내용과 예시를 보완하였다.

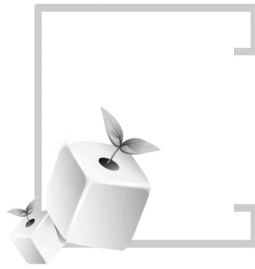
2.2. 구문 분석 말뭉치 구축

위의 과정에서 수립한 의존 구문 분석 지침을 바탕으로 200만 어절 규모의 구문 분석 말뭉치를 구축하였다. 말뭉치는 문어 200만 어절 규모로 신문 기사 텍스트로 구성되어 있다.

본 사업에서는 국립국어원에서 제공한 문어 말뭉치를 대상으로 문장 단위 구문 분석을 진행하였다. 또한 각 어절별로 구문 분석 정보를 부착하였으며 제이슨(JSON) 형식의 최종 산출물을 제출하였다.

구축 과정에서 다수의 자동 구문 분석기를 활용하여 작업의 편의성과 일관성을 확보하고자 하였으며 자동으로 분석된 결과를 작업자들이 수작업으로 전수 검토하도록 하였다. 상세한 과정은 다음과 같다. 먼저 국립국어원에서 제공한 신문 기사 말뭉치를 문장 단위를 기준으로 자동 구문 분석을 시행한다. 이때 복수의 자동 분석기 결과가 전체 일치하는 문장에 대해서는 최종 검수자의 검수만으로 검수가 완료되며 자동 분석 결과가 전체 일치하지 않는 문장에 대해서는 1차 검수와 2차 검수를 거쳐 검수가 완료된다. 작업자는 담당 교수, 팀장을 포함하여 총 4개 조로 편성되었으며 각 조의 담당 교수 및 팀장이 나머지 작업자의 1차 검수 결과물을 검수하는 방식으로 2차 검수를 실시하였다. 2차 검수 후에는 알고리즘을 통한 후처리 및 파일 형식 변환을 통해 최종 산출물을 제출하였다.

본 사업에서는 자동 구문 분석을 위하여 한국전자통신연구원, 강원대학교, 전북대학교, 충남대학교의 의존 구문 분석기를 활용하여 자동 분석을 진행하였고, 다수의 기관에서 자동으로 분석한 결과를 교차 검증 및 통합하여 신뢰도를 높이하고자 하였다. 자동 구문 분석의 과정에서는 딥 러닝을 통한 일관성 검증이 수반되었으며 작업자의 1차 검수와 2차 검수에 사용되는 작업 환경 또한 본 사업 과정에서 보완되었다.



제 2 장

구문 분석 말뭉치의 구성 및 구축 절차



1. 구문 분석 말뭉치의 구성

이 사업에서 구축한 구문 분석 말뭉치의 총 규모는 200만 어절(문어)이다. 구문 분석 말뭉치의 기반이 된 자료는 국립국어원에서 제공한 <2018년 국어 말뭉치 연구 및 구축> 사업 결과물의 일부로서, 신문 기사 말뭉치로 구성되어 있다. 기반 말뭉치의 주제별 구성은 다음과 같다.

	경제	과학	국제	기획	문화	사람들	사회	스포츠	오피니언	정치	지역	합
어절 (천 어절)	265	27	184	61	243	42	379	169	136	396	96	2,000
비율(%)	13.3	1.3	9.2	3.1	12.2	2.1	18.9	8.5	6.8	19.8	4.8	100

<표 1> 구문 분석 말뭉치의 기반 말뭉치 주제별 구성

2. 구문 분석 말뭉치 구축 절차

구문 분석 말뭉치의 구축 절차는 다음과 같다.

구문 분석 지침 수립 > 수작업 점수 도구 커스터마이징 > 작업자 교육 > 자동 구문 분석 > 작업자 분석(1차 점수) > 팀장 및 교수진 점수(2차 점수) > 딥 러닝 기반 구문 분석 말뭉치 검증 > 전문가 집단 심층 면접(Focus Group Interview) > 최종 결과물 산출

2.1. 구문 분석 지침 수립

본 사업에서는 구문 분석 말뭉치를 구축하기 위한 첫 단계로 기존의 구문 분석 지침을 수정·보완하였다. 기존의 의존 구문 분석 말뭉치는 한국전자통신연구원 엑소브레인 세종 변환 말뭉치(73,059 문장), 충남대학교 세종 변환 말뭉치(55,000 문장), 전북대학교 세종 변환 말뭉치(59,648 문장) 등이 구축된 바 있으나 21세기 세종계획에서 구축한 구구조 기반 구문 분석 말뭉치를 의존 구문 분석 말뭉치로 변환한 것으로 사실상 동질적인 언어 자원이다. 이에 의존 문법을 기반으로 구문 분석 자원을 새로이 구축하고자 한국정보통신기술협회(TTA)에서 제공하는 구문 분석 지침을 기반으로 의존 구문 분석 지침을 수립하였다.

티티에이(TTA) 의존 구문 분석 지침에서 수정 및 보완이 필요한 사항은 다음과 같다.

- 신문 텍스트의 특성
 - ▶ 명사로 종결되거나 성분이 생략되는 예가 많음.
 - ▶ 문장 이상의 단위, 즉 한 문장 안에 여러 개의 문장이 포함되는 경우에 대한 지침 필요(특히 인용절).
 - ▶ 띄어쓰기, 특히 명사구의 띄어쓰기에 소극적임.
- 서술성 명사
 - ▶ 서술성 명사가 지배소가 되거나 서술성 명사로 종결되는 등의 문장에 대한 지침 보완 필요
- 의사 보조 용언
 - ▶ 국어에서 문법적인 기능을 하는 의사 보조 용언의 구성이 다양하므로 목록 보완 필요
- 동사 연속 구성
 - ▶ 다양한 동사 연쇄와 보조 용언 구성을 구분할 필요 있음.
- 장형 사동 구문 문제
 - ▶ 장형 사동의 경우 보조 용언 구성이나 구문 분석에 영향을 줌.
 - ▶ 장형 사동을 비롯하여 문장의 논항 실현에 영향을 주는 구성에 대한 지침 마련 필요
- ‘에서’ 주어 인정 문제
 - ▶ 현재 자동 분석 결과 문장의 주어 역할을 하는 ‘에서’ 성분이 정확하게 분석되지 않음.
 - ▶ 단체 명사 주어 등에 대한 지침 수정 및 보완 필요
- 부사절(특히 인용절)에 대한 구문 분석 표지의 세밀도 문제
 - ▶ 부사절 및 인용절이 일괄적으로 VP로 처리되고 있는데 이에 대한 태그 세분화 검토 필요
- 용언의 명사형 처리 문제
 - ▶ 용언의 명사형의 경우 이를 별도로 표시할 수 있는 구문 태그가 부재함.
 - ▶ 용언의 명사형 등 태그 세분화가 필요한 경우에 대한 검토 필요
- 체언 수식 부사 문제
 - ▶ 체언 수식 부사를 인정하여 부사가 명사구를 수식할 수 있도록 할지 고려 필요
 - ▶ 이에 대한 기능 태그 논의 필요
- 기반 말뭉치의 오류
 - ▶ 원 말뭉치의 오타, 띄어쓰기 문제 등이 자동 분석 결과에 영향을 미침.
 - ▶ 원본 오류를 국어원에 보고하기 위한 시스템 구축 필요.

본 단계에서는 위의 주요 쟁점을 비롯하여 다양한 쟁점을 논의하여 기존의 티티에이(TTA) 의존 구문 분석 지침을 수정·보완하였다. 또한 예시를 정제하고 추가하여 작업자들이 예시를 통해 구문 분석 지침을 손쉽게 이해할 수 있도록 하였다. 이 단계는 실제 작업 및 자동 분석의 이전 단계에 이루어져야 하는 것이나 실제 작업에서 발생하는 오류 및 문제점들을 반영하여 지속적으로 보완하였다. 또한 전문가 집단 심층 면접 자문회의, 국어원과의 협의 등을 통하여 분석의 세부 지침을 수정하였다. 수정된 최종 지침은 3장에서 구체적으로 제시한다.

2.2. 수작업 검수 도구 커스터마이징

구문 분석 지침이 수립된 다음 단계는 자동화된 구문 분석 결과를 확인하면서 분석 오류를 효율적으로 수정할 수 있도록 지원하는 수작업 검수 도구를 준비하는 것이다.

본 사업에서는 딥네추럴 에이아이(DeepNatural AI) 클라우드소싱 플랫폼에서 기개발되어 운영 중이던 구문 분석 작업 도구를 커스터마이징하여 최종 지침의 구체적인 내용과 주요 보완 사항들이 반영되도록 했다. 웹 기반의 수작업 검수 도구는 윈도우즈(Windows), 맥오에스(Mac OS), 리눅스(Linux) 등의 운영체제에 상관없이 컴퓨터에 설치된 크롬 브라우저(Chrome Browser)를 통해 구동 가능하다. 이를 통해 다수의 검수자들이 온라인으로 동시 접속하여 병렬적으로 수작업 검수를 수행할 수 있도록 지원하고, 수십 명으로 구성된 검수팀 내부에서 작업을 분배하고 통합할 때에 발생하는 오류들을 사전에 방지하여 검수 과정의 효율성을 높였다. 또한, 작업을 전달하고 그 결과를 수집하는 과정도 검수 도구가 동작하는 플랫폼 내에서 암호화되어 이루어지므로 보안성도 한층 강화되었다.

체계적인 검수 과정을 위해 역할에 따라 검수 팀원과 팀장을 구분하여 도구 사용 권한을 부여하였으며, 검수 팀원들은 1차 검수에 참여하고 팀장들은 2차 검수 및 진행 현황 감독이 가능하도록 검수 도구를 설정하여 활용하였다.

인공지능 비서 ON / OFF
0

NWRW1800000033-0069-0012 실패로 선토림의 기세가 10개 주에서 경선이 동시에 치러지는 다음 달 6일 '슈퍼 화요일'까지 이어지면 뽑히 대세론은 완전히 무너질 수 있다.

실패로

2

←NP_MOD

선토림의

3

←NP

기세가

4

←NP

10개

5

←NP

주에서

6

←NP

경선이

7

←NP

동시에

8

←NP

치러지는

9

←NP

다음

10

←NP

달

11

←NP

6일

12

←NP

'슈퍼'

13

←NP

화요일'까지

14

←NP

이어지면

15

←NP

뽑히

16

←NP

대세론은

17

←NP

ARG1

ARG0

ARGM-LOC

ARG1

ARGM-MNR

UF-000100

지표 어 지 는

<<ARGM-TMP>>

ARGM-TMP>>

NP-02

이어지면

외문소

지목소

형태소 분석

구문 분석

의미역 분석

ID	FORM	LEMMA	XPOS	ID	FORM	DEPREL
1	실패로	실패로	MAG	14	이어지면	AP
2	선토림의	선토림 의	NNP+JKG	3	기세가	NP_MOD
3	기세가	기세 가	NNG+JKS	8	치러지는	NP_SBJ
4	10개	10 개	SN+NNB	5	주에서	NP
5	주에서	주 에서	NNG+JKB	8	치러지는	NP_AJT
6	경선이	경선 이	NNG+JKS	8	치러지는	NP_SBJ
7	동시에	동시 에	NNG+JKB	8	치러지는	NP_AJT
8	치러지는	치르 어 지 는	VV+EC+VX+ETM	13	화요일'까지	VP_MOD
9	다음	다음	NNG	10	달	NP
10	달	달	NNG	11	6일	NP
11	6일	6 일	SN+NNB	13	화요일'까지	NP
12	'슈퍼'	'슈퍼'	SS+NNG	13	화요일'까지	NP
13	화요일'까지	화요일' 까지	NNG+SS+JK	14	이어지면	NP_OBJ
14	이어지면	이어지 면	VV+EC	18	무너질	VP

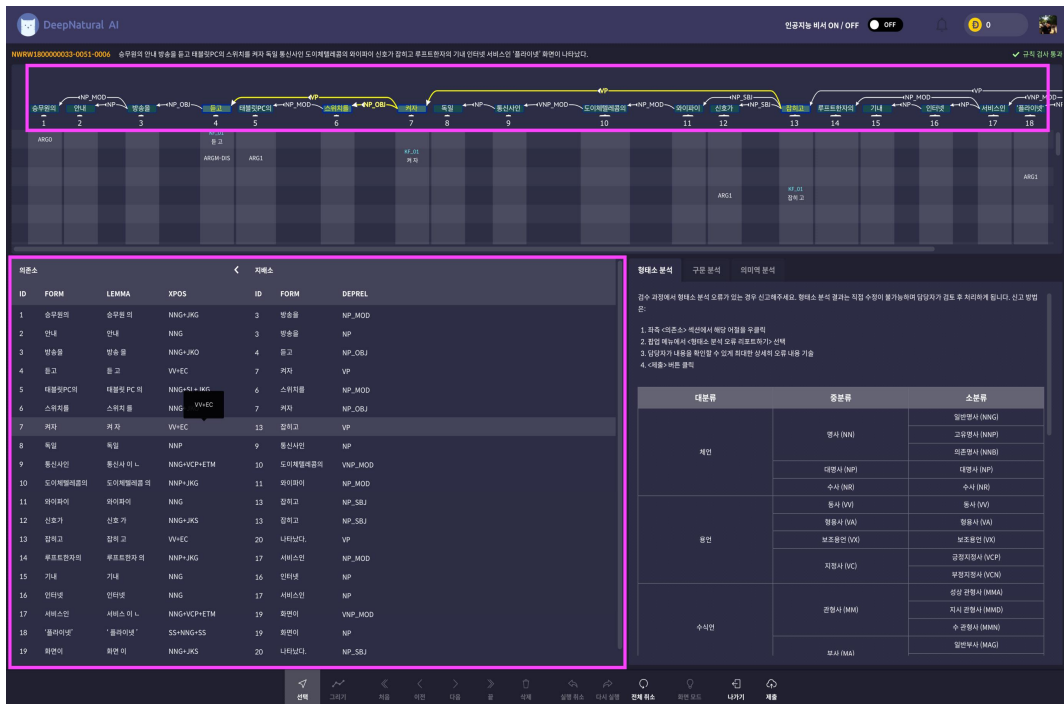
검수 과정에서 형태소 분석 오류가 있는 경우 신고해주세요. 형태소 분석 결과는 직접 수정이 불가능하며 담당자가 검토 후 처리해 드립니다. 신고 방법은:

- 좌측 <외문소> 섹션에서 해당 어절을 우클릭
- 팝업 메뉴에서 <형태소 분석 오류 리포트하기> 선택
- 담당자가 내용을 확인할 수 있게 최대한 상세히 오류 내용 기술
- <제출> 버튼 클릭

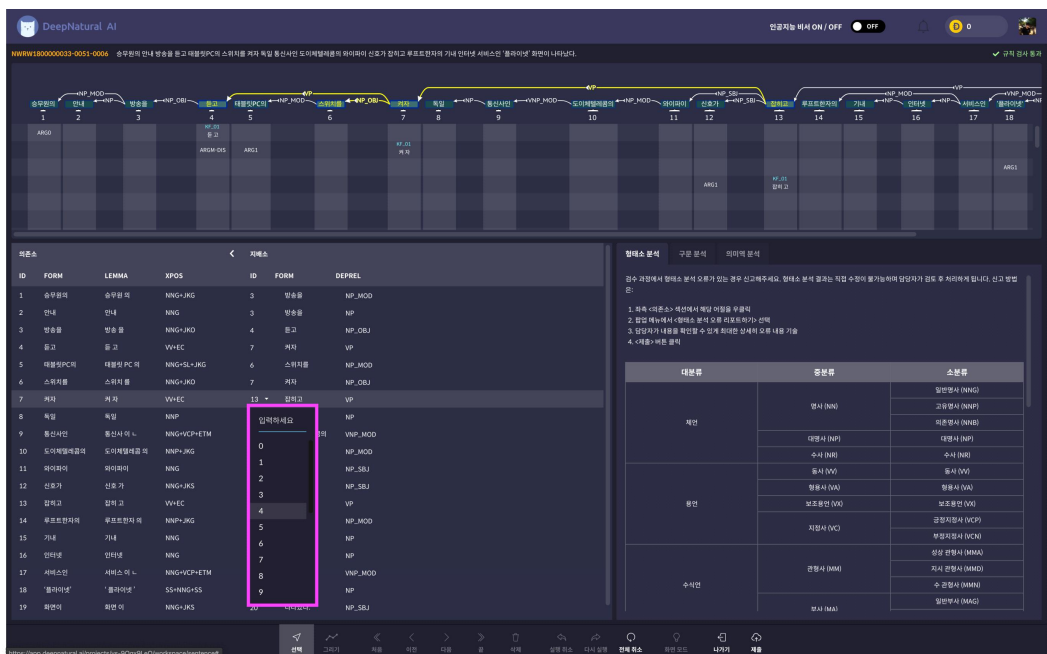
대분류	중분류	소분류
제언	명사 (NN)	일반명사 (NNG)
		고유명사 (NNP)
	대명사 (NP)	의존명사 (NNB)
		대명사 (NP)
응언	동사 (VV)	동사 (VV)
	형용사 (VA)	형용사 (VA)
	보조응언 (VX)	보조응언 (VX)
		보조응언 (VX)

<그림 1> 구문 분석 검수 도구 초기 화면

1차 검수를 수행하는 수작업 검수 팀원들은 원시 문장에 대한 자동 구문 분석 결과를 통합하여 보여주는 검수 도구에서 분석 내용을 확인하고 발견한 오류를 수정하는 작업을 하고, 2차 검수 과정에서는 팀장들이 1차 검수 결과에 문제가 없는지 다시 한번 점검한다.



<그림 2> 선택한 어절에 대한 구문 의존 관계 결과



<그림 3> 지배소 어절 번호 선택 창

먼저 검수자는 현재 작업 문장의 모든 어절이 이루는 구문 의존 관계를 트리 형식으로 시각적으로 나타낸 상단의 결과를 참고하여 각 어절이 올바른 지배소(Head)를 참조하는지 확인한다. 어절의 의존소, 지배소 정보를 나타낸 아래 항목에서 특정 어절을 선택하면, 상단의 구문 의존 관계 트리에서 선택한 어절과 구문 의존 관계를 이루는 어절과의 화살표가 강조 표시되며, 검수자가 지배소의 오류를 발견하면 왼쪽 아래 창에서

지배소 어절 번호를 선택하거나 직접 입력하여 수정한다.

의존소

ID	FORM	LEMMA	XPOS	ID	FORM	DEPREL
1	승무원의	승무원	NNG+JNG	3	방송을	NP_MOD
2	안내	NNG		4	방송을	NP
3	방송을	방송	NNG+JNG	4	듣고	NP_OBJ
4	듣고	듣	VH+EC	7	커자	VP
5	태블릿PC의	태블릿 PC	NNG+SL+JNG	6	스위치를	NP_MOD
6	스위치를	스위치	NNG+JNG	7	커자	NP_OBJ
7	커자	커	VH+EC	13	잡히고	VP
8	독일	독일	NNG	9	통신사인	NP
9	통신사인	통신사	NNG+VCP+ETM	10	도이체텔레콤의	VNP_MOD
10	도이체텔레콤의	도이체텔레콤	NNG+JNG	11	와이파이	NP_MOD
11	와이파이	와이파이	NNG	13	잡히고	NP_SBJ
12	신호가	신호	NNG+JKS	13	잡히고	NP_SBJ
13	잡히고	잡	VH+EC	20	나타났다.	VP
14	루프트한자의	루프트한자	NNG+JNG	17	서비스인	NP_MOD
15	가내	가내	NNG	16	인터넷	NP
16	인터넷	인터넷	NNG	17	서비스인	NP
17	서비스인	서비스	NNG+VCP+ETM	19	화면이	VNP_MOD
18	'플라이아웃'	'플라이아웃'	SS+NNG+SS	19	화면이	NP
19	화면이	화면	NNG+JKS	20	나타났다.	NP_SBJ

의존관계 분석

의존 관계가 존재하지 않으면, 다음 테이블에서 각 어절에 대한 지배소(ID)와 의존관계태그(DEPREL)를 클릭하여 수정할 수 있습니다.

- 의존관계를 수정하면 상단에 보이는 의존구문트리가 실시간으로 업데이트 됩니다.
- 어절을 선택하면 여러 기법의 자동 분석 결과를 볼 수 있습니다.
- 규칙 검사를 통과하도록 ID, DEPREL을 정확하게 입력해주세요.

구문 태그	의미	가능 태그	의미
NP	제언 (명사, 대명사, 수사)	SBJ	주어
VP	동언 (동사, 형용사, 보조동언)	OBJ	목적어
AP	부사구	MOD	관형어 (제언 수식어)
VNP	공정 지칭사구 (명사 + 이다)	AJT	부사어 (동언 수식어)
DP	관형사구	CMP	보어
IP	절단사구 (호칭 및 대칭 등의 표현)	CNJ	접속어 (-와)
X	의사 구 (pseudo phrase, 조사 단독 어절 또는 기호 등)		
L	부호 (관속 문호 및 따옴표)		
R	부호 (오른쪽 문호 및 따옴표)		

<그림 4> 구문 분석 의존 관계 태그 세트 표

의존소

ID	FORM	LEMMA	XPOS	ID	FORM	DEPREL
1	승무원의	승무원	NNG+JNG	3	방송을	NP_MOD
2	안내	NNG		4	방송을	NP
3	방송을	방송	NNG+JNG	4	듣고	NP_OBJ
4	듣고	듣	VH+EC	7	커자	VP
5	태블릿PC의	태블릿 PC	NNG+SL+JNG	6	스위치를	NP_MOD
6	스위치를	스위치	NNG+JNG	7	커자	NP_OBJ
7	커자	커	VH+EC	13	잡히고	VP
8	독일	독일	NNG	9	통신사인	NP
9	통신사인	통신사	NNG+VCP+ETM	10	도이체텔레콤의	VNP_MOD
10	도이체텔레콤의	도이체텔레콤	NNG+JNG	11	와이파이	NP_MOD
11	와이파이	와이파이	NNG	13	잡히고	NP_SBJ
12	신호가	신호	NNG+JKS	13	잡히고	NP_SBJ
13	잡히고	잡	VH+EC	20	나타났다.	VP
14	루프트한자의	루프트한자	NNG+JNG	17	서비스인	NP_MOD
15	가내	가내	NNG	16	인터넷	NP
16	인터넷	인터넷	NNG	17	서비스인	NP
17	서비스인	서비스	NNG+VCP+ETM	19	화면이	VNP_MOD
18	'플라이아웃'	'플라이아웃'	SS+NNG+SS	19	화면이	NP
19	화면이	화면	NNG+JKS	20	나타났다.	NP_SBJ

의존관계 분석

의존 관계가 존재하지 않으면, 다음 테이블에서 각 어절에 대한 지배소(ID)와 의존관계태그(DEPREL)를 클릭하여 수정할 수 있습니다.

- 의존관계를 수정하면 상단에 보이는 의존구문트리가 실시간으로 업데이트 됩니다.
- 어절을 선택하면 여러 기법의 자동 분석 결과를 볼 수 있습니다.
- 규칙 검사를 통과하도록 ID, DEPREL을 정확하게 입력해주세요.

대분류	중분류	소분류
제언	명사 (NN)	일반명사 (NNG)
	대명사 (NP)	그림명사 (NNP)
	수사 (NR)	의존명사 (NNB)
동언	동사 (VV)	대명사 (NP)
	형용사 (VA)	수사 (NR)
	보조동언 (VX)	동사 (VV)
관형사	관형사 (JC)	관형사 (NP)
	관형사 (JC)	수사 (NR)
	관형사 (JC)	관형사 (NP)
접속어	접속어 (JC)	관형사 (NP)
	관형사 (NP)	관형사 (NP)
	관형사 (NP)	관형사 (NP)
부호	부호 (P)	부호 (P)
	부호 (P)	부호 (P)
	부호 (P)	부호 (P)

<그림 5> 의존 관계 선택 창

다음으로 검수자는 의존 관계가 올바르게 분석되었는지 검수를 진행한다. 우측 하단의 구문 분석 의존 관계 태그 세트 표를 참고하여 작업을 진행하며, 의존 관계 태그의 오류를 발견하면 마찬가지로 의존 관계를 선택하여 나타나는 의존 관계 선택 창에서 올바른 의존 관계를 선택하거나 직접 입력하여 오류를 수정한다.

ID	FORM	LEMMA	XPOS	ID	FORM	DEPREL
1	승무원의	승무원 의	NING-JNG	3	방송을	NP_MOD
2	안내	안내	NING	3	방송을	NP
3	방송을	방송 을	NING-UJO	4	듣고	NP_OBJ
4	듣고	듣 고	V+EC	7	켜지	VP
5	태블릿PC의	태블릿 PC 의	NING+SL-JNG	6	스위치를	NP_MOD
6	스위치를	스위치 를	NING-UJO	7	켜지	NP_OBJ
7	켜지	켜 지	V+EC	13	잡히고	VP
8	독일	독일	NP	9	통신사인	NP
9	통신사인	통신사 의	NP	10	도이체텔레콤의	NP_MOD
10	도이체텔레콤의	도이체텔레콤 의	NP	11	와이파이가	NP_MOD
11	와이파이가	와이파이	NP	13	잡히고	NP_SBJ
12	신호가	신호 가	NP	13	잡히고	NP_SBJ
13	잡히고	잡 히고	VP	20	나타났다.	VP
14	루프트한자의	루프트한자 의	NING-JNG	17	서비스인	NP_MOD
15	기내	기내	NING	16	인터넷	NP
16	인터넷	인터넷	NING	17	서비스인	NP
17	서비스인	서비스 인	NING+VCP+ETM	19	화면이	VNP_MOD
18	'플라이넷'	'플라이넷'	SS+NING+SS	19	화면이	NP
19	화면이	화면 이	NING-UJS	20	나타났다.	NP_SBJ

<그림 6> 구문 태그 우클릭 기능 - 문의하기, 팀장 확인 요청

구문 분석표에서 부착된 태그를 우클릭하면 <그림 6>과 같은 툴팁 창이 나타나며 “원시 말뭉치 오류 신고”, “형태소 분석 오류 신고”, “구문 분석 관련 문의”, “문장 분할 오류 신고”, “구문 분석 검수자 확인 필요” 기능을 사용할 수 있다. 수작업 검수 팀원들은 검수 과정에서 질문을 하거나 논의가 필요한 경우 “구문 분석 관련 문의”를 클릭하여 효율적으로 질문과 논의를 시작할 수 있다. 이 기능을 통해 새로운 사안을 등록하면 수작업 검수팀 전원에게 실시간으로 알림이 전송되고, 알림을 받은 팀장은 어떤 문장, 어절, 분석 결과에 대한 내용인지 맥락을 확인할 수 있으며, 바로 이어서 답변을 하거나 논의를 진행할 수 있다. 이러한 효율적인 소통 경로를 통해 수십 명의 검수팀은 보다 효과적으로 검수 과정을 추진할 수 있다.



[구문 분석 관련 문의]

APP Jan 3rd at 6:11 PM

사용자

young_ng

문장 ID

NWRW1800000033-0188-0001

어절

[5] 브레이커'

문장

"올림픽아, 유혹마라"... 마지막 '수능 브레이커' 남았다

구문/의미역 분석결과

1	"올림픽아,	"	올림픽	아	,	SS+NNG+JKV+SP
2	NP_INT	-				
2	유혹마라"	...	유혹	말	아라	" ... NNG+VV+EF+SS+SE
6	VP	-				
3	마지막	마지막	NNG	6	NP_AJT	-
4	'수능	'	수능	SS+NNG	5	NP -
5	브레이커'	브레이커'	NNG+SS	6	NP	-
6	남았다	남	았	다	VV+EP+EF	0 VP -

내용

이런 경우 NP_SBJ로 태깅해도 될까요?

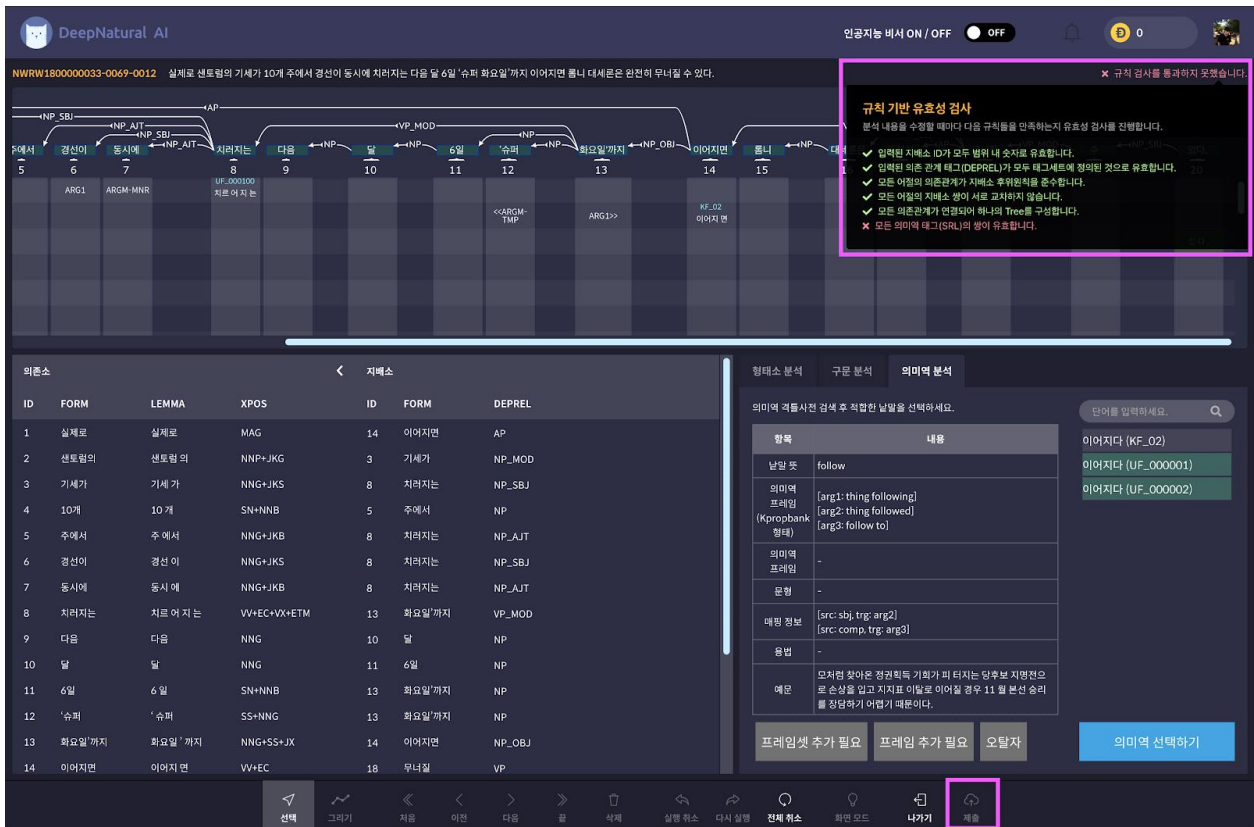
1 reply



18 days ago

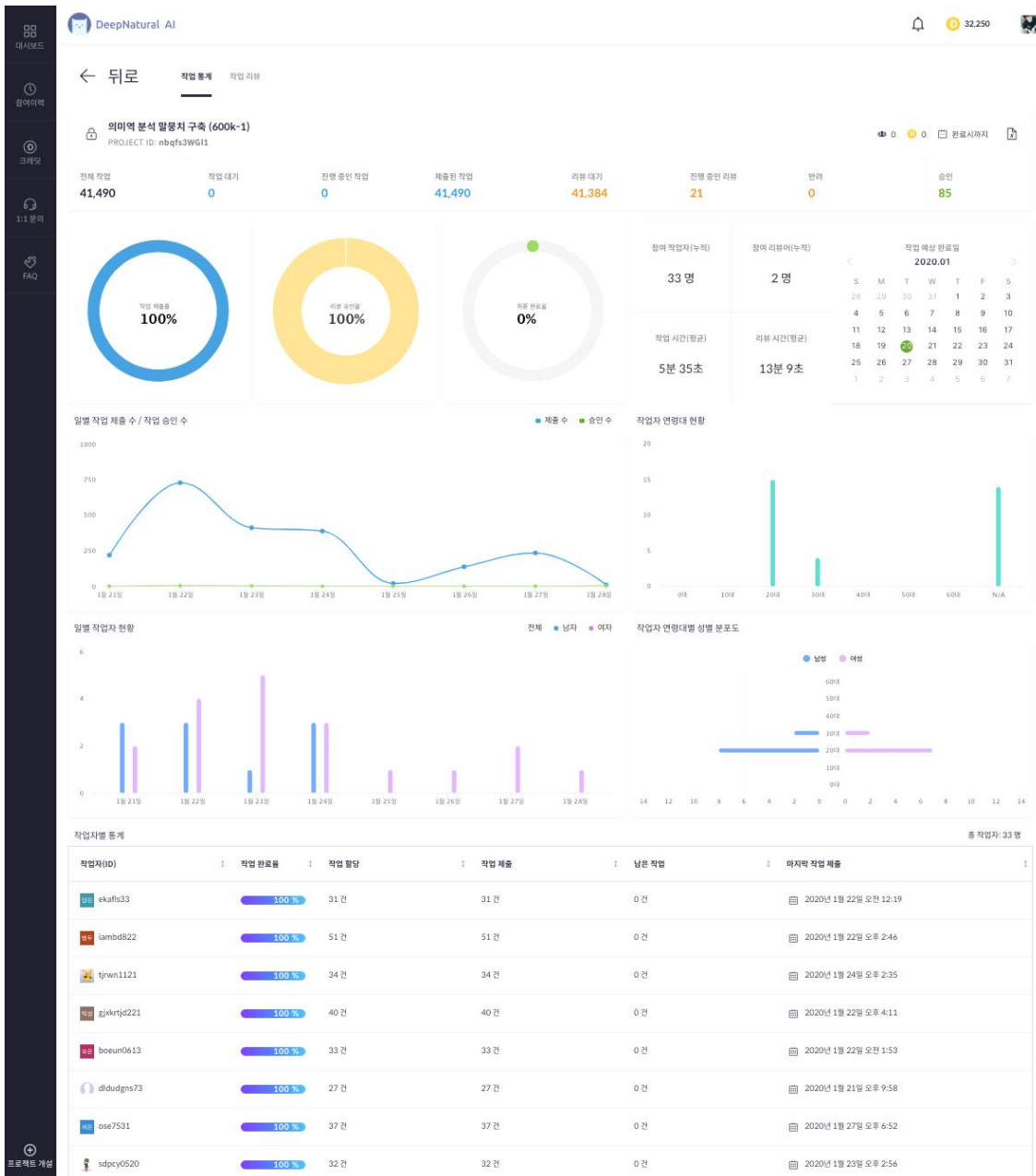
네 맞는 것 같습니다.

<그림 7> 수작업 검수 팀장에게 전달되는 문의 내용과 답변
작성 예



<그림 8> 규칙 기반 유효성 검사

검수 도구는 규칙 기반 유효성 검사 기능을 지원하며 본 사업의 지침을 기반으로 커스터마이징하여 적용되었다. 검수자가 구문 분석 결과를 수정할 때마다 유효성 검사 스크립트가 실시간으로 실행되며 그 결과를 검수자에게 곧바로 제공한다. 모든 유효성 검사를 통과하지 못한 경우에는 제출 버튼이 비활성화되어 부정확한 분석 결과가 말뭉치에 포함되는 것을 사전에 방지하고 검수자가 오류를 바로잡을 수 있도록 유도하였다.



<그림 9> 검수 진행 상황 모니터링 기능 활용

딥네츄럴 에이아이 플랫폼에서는 프로젝트 진행 현황을 확인할 수 있는 기능을 제공하는데, 본 사업에서는 수작업 검수 팀장이 진행 현황을 확인하면서 검수자들의 진도를 관리하는 데 활용했다. 현재 총 몇 건의 구문 분석 결과를 검수해야 하는데 각각의 검수자에게 몇 건의 검수가 할당되었으며 몇 건이 완료되었는지 등의 진행 상황을 실시간으로 확인 가능하기 때문에 보다 효과적으로 일정 관리가 가능했다. 이러한 관리 도구의 활용으로 수작업 검수팀은 고품질 구문 분석 말뭉치 구축 과정 본질에 더 집중할 수 있었다.

2.3. 작업자 교육

구문 분석 말뭉치 구축의 세 번째 단계로 첫 단계에서 수립한 구문 분석 지침을 작업자들에게 교육하였다. 이 과정에서는 구문 분석 지침을 숙지하는 것뿐만 아니라 구문 분석을 위한 말뭉치 자료의 이해, 수작업 검수 도구에 대한 사용법 숙지 등이 포함되었다. 이에 사업 참여자 중 실제 구문 분석 작업자를 대상으로 하여 구문 분석 말뭉치 지침, 수작업 검수 도구 사용법 등에 대한 교육을 실시하였다.

구문 분석의 경우 지침을 숙지하는 것을 비롯하여 말뭉치 분석을 실제로 해 보는 것이 중요하므로 작업 교육에서는 실제 신문 기사 말뭉치를 바탕으로 한 작업 시연 및 실습이 포함되었다. 또한 교육 과정에서 이루어진 질의응답 내용을 통해 지침을 수정·보완하였다. 또한 1차 검수, 2차 검수, 분석 문의 등을 통해 나타난 빈발 오류의 유형 등을 재교육하고 사업 도중 수정된 지침 내용을 각 팀별로 전달하는 등 구문 분석 말뭉치의 질적 향상을 도모하였다. 아래는 실시한 교육의 일정이다.

교육 회차	날짜	대상	교육자	비고
1차	8/25-26	연세대 작업자	김선훈, 이찬영, 박혜진	
2차	10/25	연세대 작업자	김선훈, 이찬영	
3차	10/26	강원대, 한림대 작업자	김한샘	
4차	10/29	전북대 작업자	이찬영	
5차	11/2	해양대 작업자	이찬영	
6차	11/15	연세대 작업자	이찬영, 박혜진	온라인
7차	1/22	충남대 작업자	박혜진	

<표 2> 교육 일정

또한 업무 협업 도구 슬랙(Slack)과 수작업 검수 도구를 통한 문의, 조별 회의, 카카오톡 그룹채팅방 등을 통한 문의 등 다양한 채널을 통하여 작업자들의 문의 사항이나 오류 보고를 확인·관리하였으며 각 조의 팀장은 각 조원의 분석 내용을 주기적으로 검수하여 작업자별 피드백을 제공하였다.

2.4. 자동 구문 분석

고품질 구문 분석 말뭉치를 구축하기 위하여 본 사업에서는 기계 학습 기반의 자동 구문 분석 결과를 활용한다. 검수자는 자동 구문 분석 결과를 지침에 따라 살펴보면서 오류를 수정하였으며, 따라서 높은 정확도를 갖춘 자동 분석 결과를 사용하는 것이 말뭉치 구축 효율을 높이는 데 중요한 역할을 한다.

본 사업에서는 자동 분석 결과의 정확도를 최대한 높이기 위해 5개의 형태소 분석기(한국전자통신연구원(ETRI), 강원대, 전북대, 충남대, 한국해양대)와 4개의 구문 분석기

(한국전자통신연구원, 강원대, 전북대, 충남대)의 자동 분석 결과를 활용하였다. 분석 결과 통합 알고리즘은 여러 기관의 자동 분석 결과를 비교하여 가장 신뢰도가 높은 분석 결과를 도출하는 형태로 하나의 최종 분석 결과를 생성한다.

통합 알고리즘은 다수 분석 결과 신뢰 규칙(Majority Rule)에 기반을 두고 있으며 판단이 어려운 경우에는 한국전자통신연구원 분석기의 결과에 높은 가중치를 두고 있다. 검수자들은 통합 알고리즘을 통해 산출된 하나의 자동 분석 결과뿐만 아니라 각각의 분석 엔진 결과도 참조하면서 수작업 검수를 진행했다. 커스터마이징된 검수 도구는 이러한 자동 구문 분석 결과를 한눈에 확인하고, 분석 결과에 포함된 오류를 쉽게 수정하는 과정의 효율을 극대화했다.

2.4.1. 엑소브레인(Exobrain) 언어분석기(WiseNLU)

구문 분석 말뭉치 구축 과정 중 기계를 이용하여 분석한 자동 언어 분석이다. 자동 언어 분석으로 작업자의 말뭉치 오류 분석 작업에 대한 효율을 높이게 된다. 본 사업에서는 높은 분석 정확률을 갖춘 엑소브레인(Exobrain)의 언어분석기(WiseNLU)를 사용하였다.

와이즈엔엘유(WiseNLU)는 엑소브레인(Exobrain) 사업을 통해 한국전자통신연구원(ETRI)에서 개발한 언어 분석 엔진이다.

구문 분석을 위한 선행 작업으로 형태소 분석과 개체명 분석, 어휘 의미 분석이 진행된다.

와이즈엔엘유의 형태소 분석기와 개체명 인식기는 국립국어원의 세종 말뭉치 데이터와 한국전자통신연구원에서 분류한 15개의 대분류, 146개의 세부 분류 데이터를 이용하여 구조 서포트 벡터 머신(Structural Support Vector Machine, 이하 SSVM) 기반 음절 단위 분류, 경계 및 대분류 인식, 세부 분류 기술로 만들어진 엔진들이다. 전처리 사전과 후처리 패턴 기반 사전으로 성능이 추가적으로 더 개선되었다.

어휘 의미 분석은 고빈도 의미 기반과 공기 정보 기반의 어휘 의미 분석을 하였으며, 동음이의어 818만 어절, 다의어 377만 어절의 태깅 말뭉치를 사용했다.

구문 분석은 한국어 문법에 기반하여 문장의 구조를 분석하였으며, SSVM 기반으로 분석한다.

성능은 2017년 기준 <그림 10>과 같은 성능을 보이며, 현재도 국내 최고의 수준을 보이고 있다.

본 사업에서는 와이즈엔엘유(WiseNLU)를 사용하여 2만 어절에 대한 샘플 말뭉치 분석을 하여 선제출하였고, 2만 어절 제출 이후에 최종 말뭉치(200만 어절)에 대한 언어 분석 결과를 제출하였다.



<그림 10> 와이즈엔엘유(WiseNLU) 성능 지표

2.4.2. 강원대학교 딥 러닝 기반 의존 구문 분석 모델

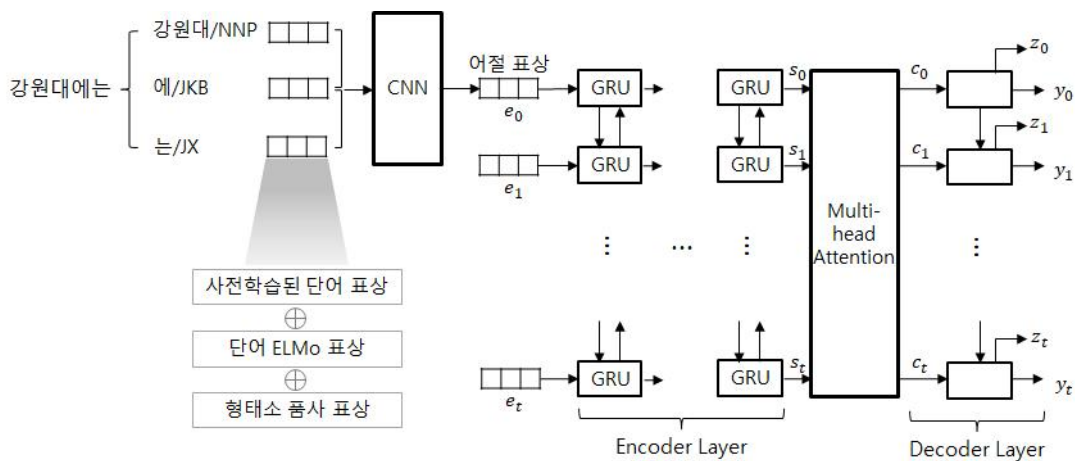
대용량의 고품질 구문 분석 말뭉치를 구축하기 위해선 고성능의 자동 구문 분석기를 활용해 자동 구문 분석 말뭉치를 제작한 후 이를 사람이 직접 검수하는 방식이 가장 효과적이다. 한국어처럼 어순이 자유롭고 단어의 생략이 빈번한 언어의 경우 의존 구문 분석을 적용하는 것이 적합하다. 의존 구문 분석이란 문장을 구성하는 어절들 간의 의존 관계 정보를 분석하여 각 어절의 지배소 위치를 알아내고 의존 관계명을 부착하는 구문 분석의 한 방식이다. 본 사업에서는 자동 의존 구문 분석 말뭉치 구축을 위해 강원대 의존 구문 분석기를 사용하였다. <그림 11>은 강원대 의존 구문 분석기의 구조도를 보여준다.

강원대 의존 구문 분석기는 딥 러닝(Deep Learning) 기반의 분석기로 문장 내의 어절의 의미를 실수 값으로 표현하기 위한 인코더 계층(Encoder Layer)과 그 값을 바탕으로 포인터 네트워크에서 각 어절의 지배소 위치와 의존 관계명을 추정하기 위한 디코더 계층(Decoder Layer)으로 구성된다. 인코딩 계층의 입력은 어절 단위의 표상(Embedding)을 사용했다. 어절을 구성하는 형태소들의 표상을 합성곱 신경망을 통해 하나의 어절을 의미하는 벡터(Vector)로 만들어 미등록어 문제에 강건하도록 설계했다. 형태소의 의미를 실수 값으로 표현하기 위해서 대용량의 말뭉치로 사전 학습한 단어 표상을 사용했다. 또한 언어 모델인 엘모(Embedding from Language Model, 이하 ELMo)를 통해 문맥에 따라 달라지는 형태소의 의미도 어절 표상에 반영하도록 했으며 품사 정보가 중요하게 작용하는 구문 분석의 특성을 고려해 형태소 품사 표상도 함께 사용했다.

인코더 계층에서는 양방향 순환 신경망(Bidirectional Recurrent Neural Network)을 사용해서 양방향의 문맥 정보를 반영했다. 인코딩 계층의 출력은 멀티 헤드 주의 집중 계층(Multi-head Attention Layer)에서 의미 관계 분석에 중요한 어절의 위치와 의미 정보를 부각하고 그 정보를 디코더 계층에 전달하여 의존 구문 분석에 도움을 주도록 하였다. 디코더 계층에서는 포인터 네트워크(Pointer Network)를 통해 각 어절의 지배소

위치를 추정했다. 포인터 네트워크는 주의 집중 방법(Attention Mechanism)을 응용한 기술로서 각 어절의 디코딩 단계마다 문장의 모든 어절과의 연관 관계를 계산한다. 구문 분석에 있어서는 이 연관 관계가 지배소에 해당하는 어절과 높아지도록 학습을 진행했다. 의존 관계명은 인코더 계층 출력과 멀티 헤드 주의 집중 계층의 출력과 디코더 계층의 출력을 종합하여 최적의 결과를 도출하도록 했다.

분석기의 학습은 한국어 의존 구문 분석 말뭉치인 세종 말뭉치를 사용했다. 성능 평가 척도는 지배소 위치 예측 정확도인 유에이에스(Unlabeled Attachment Score, 이하 UAS)와 의존 관계명 예측 정확도인 엘에이에스(Labeled Attachment Score, 이하 LAS)를 사용했다. 결과적으로 강원대 의존 구문 분석기는 UAS 92.85%, LAS 90.65%의 성능을 보였다.



<그림 11> 강원대 의존 구문 분석기 구조도

2.4.3. 전북대학교 딥 러닝 기반 자동 구문 분석 모델

구문 분석은 문장의 구조를 분석하는 자연어 처리 분야로 구 구조 구문 분석과 의존 구문 분석으로 나눌 수 있다. 최근 의존 구문 분석 방법이 주로 연구되고 있는데 의존 구문 분석은 지배소(Head)와 의존소(Modifier)의 관계에 따라 문장의 구조를 분석하여 파스 트리를 생성하는 방법을 말한다. 의존 구문 분석은 전이 기반 방식과 그래프 기반 방식의 두 갈래로 나뉘어 연구되어 왔다. 전이 기반 방식은 스택과 버퍼의 지역적 정보에 의존하여 전이 액션을 결정하는 방식이고 그래프 기반 방식은 문장 내의 모든 단어 쌍의 지배소와 의존소의 점수를 계산하여 구문 분석을 수행하는 방식이다. 대용량의 구문 분석 말뭉치를 구축하기 위해 구문 분석을 수행하는 작업자는 처음부터 모든 의존 구문 분석을 수행하는 것이 아닌 미리 학습된 구문 분석 모델을 통해 자동 구문 분석 결과로부터 오류를 수정하는 방식으로 구축하는데 이러한 방식이 효율적이다. 품질이 높은 자동 구문 분석 결과는 오류를 수정해야 하는 작업량뿐 아니라 구축된 말뭉치의

완성도에도 영향을 미칠 수 있다. 고품질의 자동 구문 분석 결과 도출을 위해 다양한 전이 기반 의존 구문 분석 모델과 그래프 기반 의존 구문 분석 모델에 대한 연구를 진행하였다.

1) 전이 기반 의존 파서

전이 기반 방식은 전이 액션을 수행하는 방식에 따라 아크 이거(Arc-Eager), 아크 스탠더드(Arc-Standard) 방식이 대표적이며 전자는 버퍼의 톱(Top) 노드와 톱의 톱 노드 사이에서 의존성을 결정하고 후자는 스택의 톱 노드 사이에서 의존성을 결정하는 방식이다. 아크 하이브리드(Arc-Hybrid)¹⁾는 아크 이거 방식과 아크 스탠더드를 혼합한 방식이다. 아크 하이브리드 방식은 시프트(SHIFT), 아크레프트(ArcLEFT), 아크라이트(ArcRIGHT)의 3가지 액션으로 구성되며 레프트(LEFT)는 아크 이거와 같이 수행되며 라이트(RIGHT)는 아크 스탠더드와 같이 수행되어 하이브리드 방식으로 부른다. 각 전이 액션별 스택 및 버퍼 정보의 갱신 과정은 다음 표와 같다.

Action	S_t	S_{t+1}
<i>SHIFT</i>	$(\sigma, b_0 B, T)$	$(\sigma b_0, B, T)$
<i>ArcLEFT</i> (l)	$(\sigma s_0 s_1, b_0 B, T)$	$(\sigma s_1, b_0 B, T \cup \{b_0, s_0, l\})$
<i>ArcRIGHT</i> (l)	$(\sigma s_0 s_1, B, T)$	$(\sigma s_1, B, T \cup \{s_1, s_0, l\})$

<표 3> 전이 기반 방식의 액션별 스택 및 버퍼 정보 갱신 과정

<표 3>에서 스택, 버퍼, 아크(arc) 셋을 중심으로 SHIFT 액션이 수행되면 단순히 버퍼의 톱 원소를 스택으로 이동하며 LEFT가 수행되면 스택의 톱 원소를 팝(POP)한 후 지배소를 의존소로 하는 의존 관계를 생성하고 아크 셋에 추가한다. RIGHT가 수행되면 스택의 톱 원소를 POP 한 후 의존 관계를 생성하고 아크 셋에 추가한 후 다시 스택 σ 에 푸시(PUSH)한다.

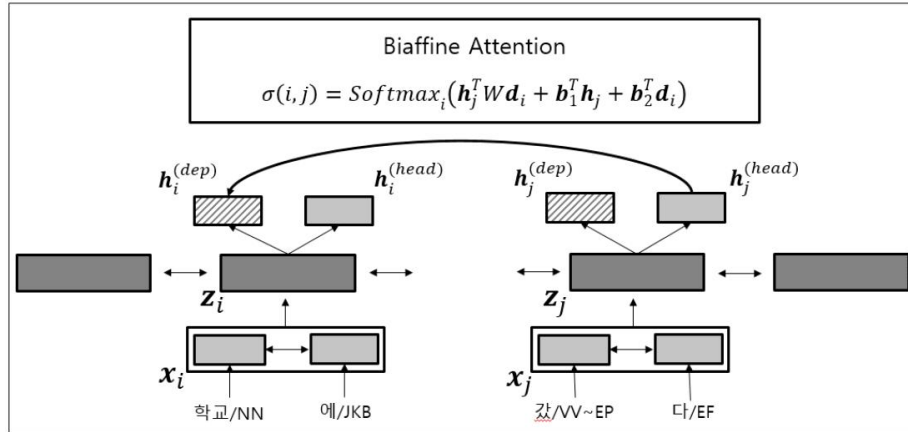
전이 액션을 결정하기 위해 인코더를 통해 인코딩된 은닉열로부터 자질을 추출하게 되는데 본 연구에서는 스택의 최상위 3 층위에 해당하는 은닉 상태와 버퍼의 최상위 단계에 해당하는 은닉 상태를 결합하여 하나의 상태 표상으로 만든 후 다중 링크 절차(multi-link procedure, 이하 MLP)를 거쳐 출력층에서 다음 전이 액션을 결정하게 된다. 최종 상태에 도달할 때까지 전이 액션을 반복하여 파스 트리를 만들고 종료하게 된다.

2) 그래프 기반 파서

그래프 기반 파서는 모든 의존소(Modifier)에 대한 지배소(Head)의 순서쌍의 점수를 나타내는 그래프 스코어 행렬로부터 최대 신장 트리(Maximum Spanning Tree) 알고리

1) Miryam de Lhoneux, Sara Stymne, Joakim Nivre(2017), Arc-Hybrid Non-Projective Dependency Parsing with a Static-Dynamic Oracle, Proceedings of the 15th International Conference on Parsing Technologies

즘을 통해 스코어가 최대가 되는 구문 트리를 반환하는 방식을 사용하다. 그래프 스코어 행렬을 얻기 위해 한국어 구문 분석에서 가장 높은 성능을 보이고 있는 방식인 바이어파인 어텐션 모델(Biaffine Attention Model)을 적용하였다. 아래의 그림은 바이어파인 한국어 구문 분석 모델의 구조를 나타낸다.



<그림 12> 바이어파인 한국어 구문 분석 모델

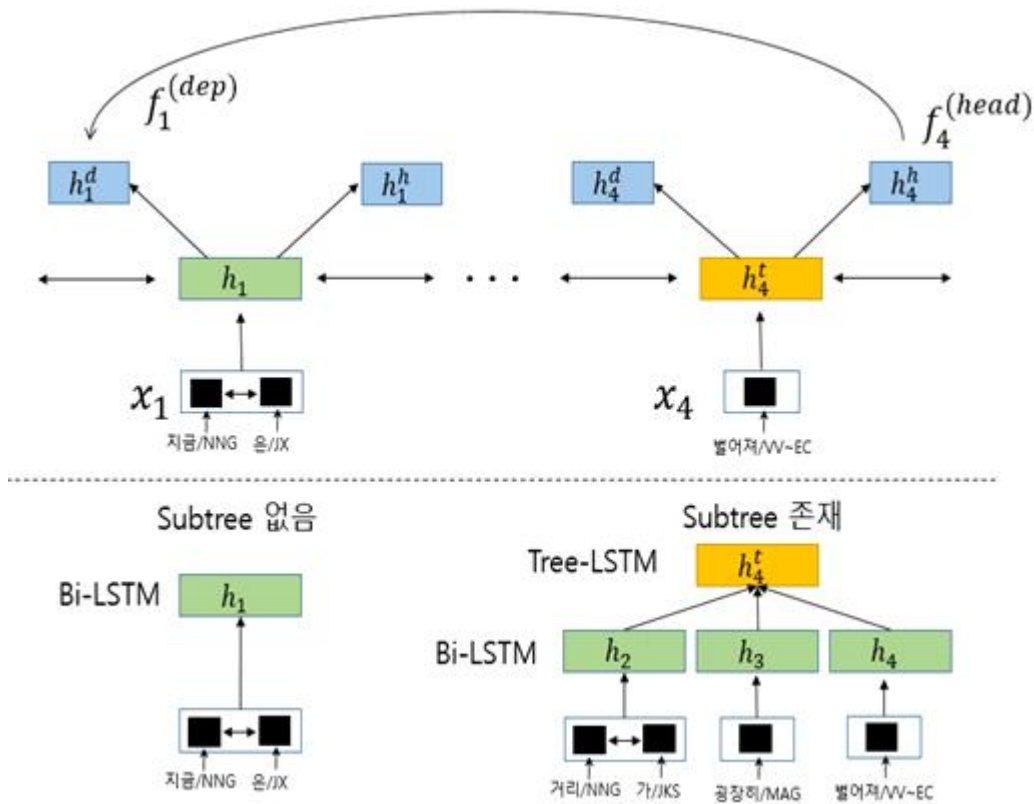
전이 기반 방식은 전이 액션을 수행하는 방식에 따라 아크 이거, 아크 스탠더드 방식이 대표적이며 전자는 버퍼의 톱(Top) 노드와 Top의 Top 노드 사이에서 의존성을 결정하고 후자는 스택의 Top 노드 사이에서 의존성을 결정하는 방식이다. 아크 하이브리드는 아크 이거 방식과 아크 스탠더드를 혼합한 방식이다. 아크 하이브리드 방식은 SHIFT, ArcLEFT, ArcRIGHT의 3가지 액션으로 구성되며 LEFT는 아크 이거와 같이 수행되며 RIGHT는 아크 스탠더드와 같이 수행되어 하이브리드 방식으로 부른다. 인코더를 통해 은닉열 $\{z_1, \dots, z_n\}$ 을 얻고 각각 지배소와 의존소에 대한 MLP를 적용하여 표상을 얻고 이를 그림의 바이어파인 어텐션 함수를 통해 모든 의존소에 대해 지배소가 될 수 있는 점수를 얻게 된다.

3) 이지 퍼스트 딥 바이어파인 어텐션(Easy-First Deep Biaffine Attention) 기반 방법

이지 퍼스트 딥 바이어파인 어텐션 기반 방법은 이지 퍼스트 딥 바이어파인 어텐션을 이용한 한국어 의존 파싱의 논문에서 제안한 방법으로, 기존의 딥 바이어파인 어텐션(Deep Biaffine Attention) 모델을 확장하여 파싱을 진행한 모델이다. 기존의 바이어파인 어텐션 모델은 단어 표상을 바이엘에스티엠(Bi-LSTM)에 적용하여 인코딩한 후 MLP를 적용하여 지배소 표상과 의존소 표상을 얻어 어텐션을 통해 점수를 계산하는 형태를 가지고 있다. 기존의 모델은 지배소 표상과 의존소 표상은 노드 정보만 가지고 있고 트리 특성의 정보를 가지고 있지 않다. 이지 퍼스트 딥 바이어파인 어텐션 기반 방법의 모델은 트리 정보를 반영하기 위하여 쉬운 의존성을 미리 결정하고 결정된 정보를 이용하여 부분 트리를 만들어 트리 정보를 사용 가능하게 만든 모델이다. 쉬운 의존성을 미리 결정하기 위해 선행 학습한 바이어파인 어텐션 모델의 점수 정보를 사용하였고, 높은 점

수를 보이는 의존성을 쉬운 의존성이라고 판단하였다.

아래 그림은 의존성이 결정되었을 때 단어의 부분 트리가 존재하는 경우와 존재하지 않는 경우 표상을 얻는 과정을 보여준다. "지금은"의 단어 표상은 의존성이 결정되었을 때 자식이 없기 때문에 부분 트리를 갖지 않는다. 부분 트리를 갖지 않는 경우 Bi-LSTM을 통해 표상을 얻는다. "벌어져"의 단어 표상은 확실한 의존성이 결정되었을 때 "거리가", "굉장히" 두 단어의 표상을 자식으로 갖는다. 이 정보를 이용하여 부분 트리를 구성하여 트리 엘레먼트(Tree-LSTM)를 통해 표상을 얻는다. 이렇게 얻어진 표상을 MLP를 적용하여 지배소 표상과 의존소 표상을 통해 어텐션을 수행한다. 어텐션 단계에서 바이어파인 어텐션을 적용하여 점수를 얻고 이를 이용하여 가장 높은 점수를 갖는 의존 트리를 결정한다.



<그림 13> Easy-First Deep Biaffine Attention을 이용한 의존 파싱 모델 구조

4) 로버트에이(RoBERTa) 및 로버트에이(RoBERTa) 기반 의존 파싱 결과

버트(BERT)는 대용량의 말뭉치를 이용하여 학습한 트랜스포머(Transformer) 기반 언어 모델로, 최근에 다양한 태스크(Task)에 적용되어 월등한 성능을 보이고 있다. RoBERTa는 기존의 BERT를 개선한 언어 모델로 기존의 BERT가 가지고 있던 문장 예측 부분을 제거하고 마스킹(Masking)되는 단어를 고정하지 않고 다이내믹 마스킹(Dynamic Masking)을 적용한 언어 모델이다. 한국어에 적용하기 위하여 입력은 형태소-태그 단위를 사용하였으며 단어장에 없는 경우에는 BPE단위로 토큰화하여 사용하였다.

한국어 대용량 말뭉치로 위키피디아 코퍼스를 사용하였다.

의존 파싱 모델의 적용은 사전 학습된 RoBERTa의 마지막 레이어 값을 이용하여 각 어절의 마지막 형태소의 출력 값을 어절 정보로 하여 입력으로 들어오는 다른 정보들과 결합하여 RoBERTa를 적용하였다.

원문
나는 휘닉스파크에 갔다
형태소 분석
나/NP 는/JX, 휘닉스파크/NNP, 에/JKB 갔/VV~EP 다.EF
의미역 결정 모델에서의 입력
[CLS] 나/NP, 는/JX, _휘, 닉, 스, 파크, 에/JKB 갔/VV~EP 다.EF [SEP] 갔/VV~EP 다.EF [SEP]

<그림 14> RoBERTa의 입력 예시

RoBERTa를 적용한 결과는 기존의 글로브(Glove)만을 사용한 모델보다 좋은 성능을 보였을 뿐만 아니라 BERT를 적용한 결과를 뛰어넘는 성능을 보였다.

5) 최종 성능

실험 및 평가를 위해 세종 구문 분석 데이터 세트를 사용하였다. 세종 구문 분석 데이터 세트는 총 53,842개 문장의 학습 세트와 5,817개 문장의 평가 세트로 구성되어 있으며 학습 세트에서 1,000개 문장을 별도의 개발 세트로 나누어 학습하였다. 평가 지표로는 지배소가 올바르게 부착되었는지 판별하는 UAS와 지배소와 해당하는 의존 관계 레이블이 모두 올바르게 부착되었는지 판별하는 LAS의 2가지를 제시한다.

모델	UAS	LAS
Easy First-Score50%	91.94 %	91.36 %
RoBERTa + ArcHybrid 전이 기반 파서	94.27 %	92.36 %
RoBERTa + Biaffine 그래프 기반 파서	94.42 %	92.52 %

<표 4> 각 파서의 성능 비교

현재 가장 높은 성능을 보이고 있는 모델은 RoBERTa를 적용한 바이어파인 어텐션 모델로 UAS : 94.42%, LAS : 92.52%의 성능을 보이고 있다.

6) 출력 예제

문장이 주어졌을 때 형태소 분석을 실시하여 어절 단위의 형태소 분석 결과를 얻고 구문 분석 모델을 통해 지배소와 의존 관계 레이블을 부착한다. <그림 15>는 출력 예제이다. 두 번째 열은 입력 문장 어절이고 세 번째 열부터는 분석 정보를 표시한다. 특히 네 번째 열과 다섯 번째 열이 해당 어절에 대해 자동 분석된 지배소와 의존 관계 레이

블을 나타낸다.

1	정치에	정치/NNG 에/JKB	2	NP_AJT		
2	있어서도	있/VV 어서/EC 도/JX	6	VP		
3	정직함이	정직/NNG 함/XSA~ETN 이/JKS	6	VP_SBJ		
4	무한한	무한/NNG 한/XSA~ETM	5	VP_MOD		
5	힘이	힘/NNG 이/JKC	6	NP_CMP		
6	될	될/VV~ETM	7	VP_MOD		
7	수	수/NNB	8	NP_SBJ		
8	있다는	있/VV 다는/ETM	10	VP_MOD		
9	불가능한	불/XPN 가능/NNG 한/XSA~ETM	10	VP_MOD		
10	가능성을	가능/NNG 성/XSN 을/JKO	13	NP_OBJ		
11	우리는	우리/NP 는/JX	13	NP_SBJ		
12	링크에게서	링크/NNP 에게서/JKB	13	NP_AJT		
13	배운다.	배운다/VV~EF . /SF	0	VP		

<그림 15> 출력 예시

2.4.4. 충남대학교 자동 의존 구문 분석 모델

구문 분석은 문장의 구조를 이해하며 구조적 중의성을 해결하는 것이다. 일반적으로 한국어는 어순 배열의 자유도가 높고 문장 성분의 생략이 빈번한 특성이 있기 때문에 여러 구문 분석 방법 중 의존 구문 분석(dependency parsing)이 널리 사용되었다. 의존 구문 구조는 지배소(head)와 피지배소(modifier)로 구성되며 단어 간의 의존 관계로 표현된다. 구문 분석 말뭉치 구축 작업자는 입력 문장에 대해 구문 분석기가 내주는 결과를 토대로 최종 구문 분석 트리를 결정한다. 이를 위해 필요한 자동 구문 분석 모델에 대해 설명한다.

2.4.3.1. 스택-포인터 네트워크(Stack-Pointer Networks)

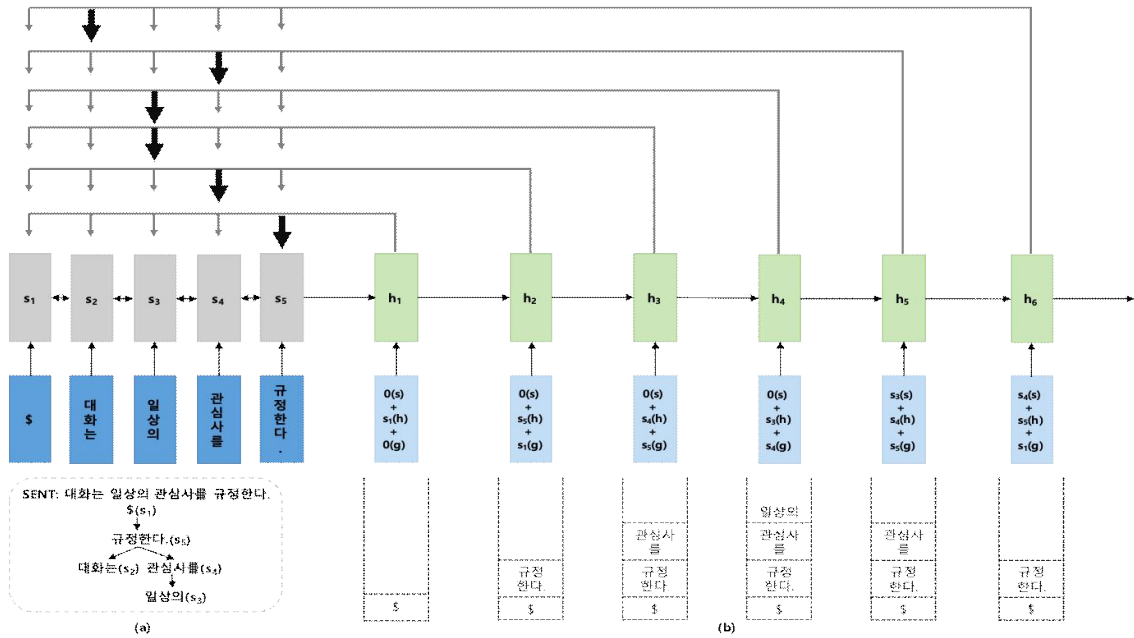
스택-포인터 네트워크²⁾ 모델은 딥 러닝 기반의 의존 구문 분석 모델이다. 스택-포인터 네트워크는 포인터 네트워크³⁾와 디코더의 입력을 관리하는 내부 스택으로 구성되어 있다. 최용석·이공주(2019)⁴⁾의 연구는 스택-포인터 네트워크 모델을 적용하여 한국어 의존 구문 분석을 수행하였다. 한국어 의존 구문 분석은 주로 어절을 분석의 단위로 사용한다. 어절은 여러 개의 형태소로 이루어져 있기 때문에 어절 단위를 그대로 임베딩으로 사용하게 되면 미등록어(Out-of-Vocabulary) 문제가 빈번하게 발생한다. 최용석·이공주(2019)의 연구에서는 미등록어를 해결하기 위해 어절을 구성하고 있는 형태소, 형태소의 음절 그리고 품사 조합의 임베딩을 사용하고 이를 콘볼루션 신경망을 이용하여 어절을

2) Ma, Xuezhe, et al. "Stack-pointer networks for dependency parsing." *arXiv preprint arXiv:1805.01087* (2018).

3) Vinyals, Oriol, Meire Fortunato, and Navdeep Jaitly. "Pointer networks." *Advances in neural information processing systems*. 2015.

4) 최용석·이공주. "고차원 정보와 스택-포인터 네트워크를 이용한 한국어 의존 구문 파서." *정보과학회논문지* 46.7 (2019): 636-643.

표현하였다. <그림 16>은 연구 최용석·이공주(2019)의 스택-포인터 네트워크 모델을 표현한 그림이다.



<그림 16> 스택-포인터 네트워크 (최용석·이공주 2019)

2.4.3.2. 충남대학교 의존 구문 분석 모델

최용석·이공주(2019)의 스택-포인터 네트워크 모델은 형태소, 형태소의 음절, 품사의 임베딩으로 한 어절을 표현한다. 이때, 사용한 임베딩은 한국어 위키피디아와 나무 위키 피디아 문서로부터 워드투벡터(Word2Vec)⁵⁾으로 사전 훈련한 모델을 이용하여 초기화하고 학습을 수행하였다. Word2Vec은 단어의 의미를 표현하기 위해 주변 단어를 고려해서 의미를 표현한다. 이 방법은 해당 단어가 쓰인 문장의 문맥은 고려하지 않고 전체 코퍼스에서 주변에 함께 발생한 단어만을 이용하여 의미를 표현한다. 최근 문장의 문맥을 고려한 언어 표현으로 BERT⁶⁾가 큰 주목을 받고 있다. BERT는 사전 훈련된 언어 표현 모델로 트랜스포머⁷⁾의 인코더 부분을 사용한다. 이를 통해 해당 문장의 양방향 문맥을 고려할 수 있고 문장 수준의 이해를 바탕으로 한 언어 표현을 구할 수 있다. BERT는 대용량의 코퍼스로부터 모델을 사전 훈련시키고 이에 나온 출력을 언어의 표현으로 사용한다.

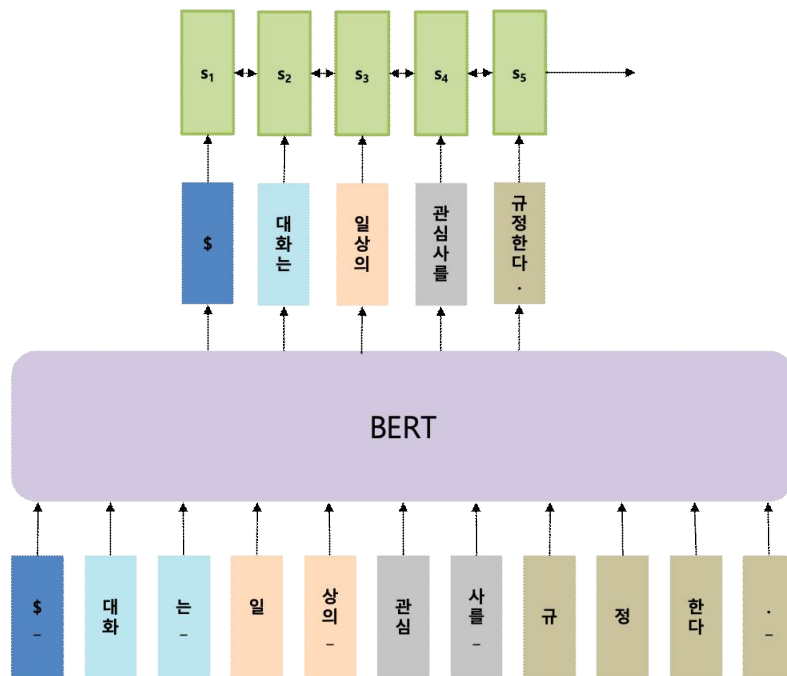
시엔유(CNU) 자동 의존 구문 분석 모델은 인코더 부분의 문장 수준의 이해를 높이기

5) Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781*(2013).

6) Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805*(2018).

7) Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems*. 2017.

위해 최용석·이공주(2019)에서 사용한 단어 임베딩을 BERT로 대체한다. <그림 17>은 CNU 의존 구문 분석기의 인코더 부분을 나타낸 것이다. BERT는 한국전자통신연구원(ETRI)⁸⁾에서 배포한 모델인 코버트(KorBERT)을 사용하였다. KorBERT의 워드피스(WordPiece)⁹⁾의 단위는 한국전자통신연구원(ETRI)에서 제공한 어절 단위의 워드피스인 ‘코버트 워드피스(Korean_BERT_WordPiece)’이다. <그림 17>과 같이 어절 단위의 워드피스는 어절 구분자로 “_”를 사용한다. CNU 자동 의존 구문 분석기는 KorBERT를 이용해 입력 문장을 표현한다. 의존 구문 분석기는 어절을 입력 단위로 수행되기 때문에 KorBERT에서 출력한 워드피스 벡터를 어절 단위의 합(SUM)을 구해서 LSTM의 입력으로 넣어준다.



<그림 17> CNU 의존 구문 분석기에서의 인코더 부분

2.4.3.3. 실험 모델 및 평가

2.4.3.3.1. 실험 모델

CNU 의존 구문 분석기는 스택-포인터 네트워크와 문맥을 고려한 토큰 임베딩으로서 KorBERT를 사용하였다. KorBERT의 토큰 임베딩이 구문 분석기의 성능에 미치는 영향을 확인하기 위해 다음의 모델들을 비교해 보았다. 우선 베이스라인(baseline) 모델은 KorBERT 대신 기존의 Word2Vec의 임베딩을 사용한다. KorBERT의 마지막 출력 값만

8) <http://aiopen.etri.re.kr>

9) D. Bahdanau, K. Cho, Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

을 LSTM의 입력으로 사용하는 모델(버트 라스트(BERT-Last))과 KorBERT의 각 계층에서의 출력 값들을 가중치 합(버트 웨이트(BERT-Weight))으로 계산해서 LSTM 입력으로 사용하는 모델(BERT-Weight)을 비교해 보았다. 비교 모델 중 가장 좋은 성능을 보인 모델을 자동 의존 구문 분석기로 채택한다.

BERT-Weight 모델은 <수식 1>과 같이 KorBERT의 각 계층에서 출력한 값들을 가중치 합으로 계산하여 입력토큰의 임베딩을 정의한다.

$$E_k = \sum_{i=0}^L \lambda_i h_k^i$$

<수식 1>

E는 문장의 토큰을 의미하며, h는 KorBERT의 각 층에서 나온 벡터 값이다. k는 토큰의 인덱스이며, i는 KorBERT의 층수이다. λ 는 각 계층들 사이의 가중치이며, 이는 훈련 과정에서 학습되는 값이다.

2.4.3.3.2. 실험 데이터 및 평가 방법

실험 데이터는 세종 코퍼스이며 총 59,397개 문장 중 53,920개 문장은 학습 데이터로 5,817개 문장은 평가 데이터로 사용하였다. 의존 구문 분석 결과의 평가 척도는 어절 단위로 UAS와 LAS를 사용하였다.

2.4.3.3.3. 모델 매개 변수 설정

모든 모델은 최용석·이공주(2019)의 매개 변수 설정과 동일하다. BERT-Last와 BERT-Weight에서 사용한 KorBERT에 대한 정보는 <표 5>와 같다.

Hyperparameters	Values
number of layers	12
hidden dimension	768
intermediate dimension	3072
number of attention heads	12
activation function	gelu
dropout	0.1
max length	512
vocabulary size	30,797

<표 5> 한국어 BERT 정보(KorBERT)>

2.4.3.3.4. 실험 결과

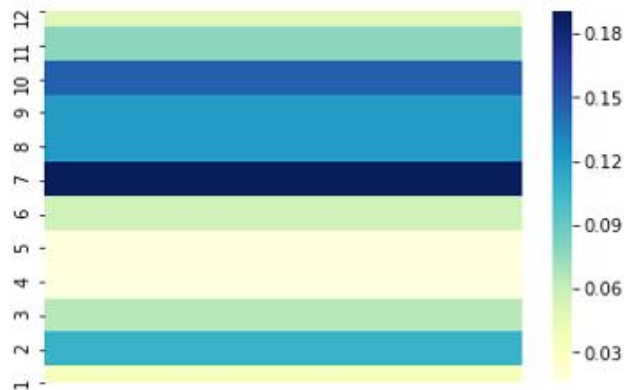
<표 4>는 평가 데이터 5,817개 문장에 대한 성능 결과이다.

Model	UAS	LAS
baseline	91.98	90.24
BERT-Last	93.70	91.80
BERT-Weight	93.83	92.12

<표 6> CNU 의존 구문 분석기 성능

실험 결과로부터 구문 분석기에 KorBERT를 사용하였을 때, 성능이 월등히 향상되는 것을 확인할 수 있었다. BERT-Weight 모델은 베이스라인 모델보다 LAS가 약 1.9% 향상되는 것을 보여주었다. 최종적으로 자동 의존 구문 분석기는 BERT-Weight 모델을 채택하여 구축하였다.

<그림 18>은 BERT-Weight 모델에서 각 계층의 가중치를 히트맵으로 표현한 것이다. 보는 바와 같이 BERT의 마지막 층보다는 중간 계층(7~10)에서 높은 가중치를 보이는 것을 확인할 수 있었다. 영어권의 실험¹⁰⁾에서와 같이 한국어에서도 BERT의 중간 계층이 구문 분석에 가장 많은 영향을 끼치는 것을 확인할 수 있었다.



<그림 18> BERT-Weight 모델에서 계층별 가중치

2.5. 작업자 분석(1차 검수)

자동 구문 분석은 일관성 검증을 통해 보다 신뢰할 수 있는 분석 결과를 얻을 수 있으나 100%의 정답 값을 기대하기 어렵기 때문에 전문가의 수작업 검수를 통하여 이를 보완하였다. 본 사업에서는 자동 구문 분석 결과를 바탕으로 100% 수작업 검수를 통해 구문 분석 말뭉치를 질적으로 향상시킬 수 있도록 작업하였다.

자동 구문 분석 결과는 작업 도구 플랫폼을 통해 작업자들에게 문장 단위로 할당된

10) Jawahar, Ganesh, Benoît Sagot, and Djamé Seddah. "What does BERT learn about the structure of language?." 2019.

다. 작업자들은 작업 도구를 통해 분석 대상 문장을 확인하고 자동 구문 분석 결과를 검수하는 방식으로 작업하였다. 이때 분석에 문제가 있는 경우 작업 도구 등을 통해 보고할 수 있다.

작업자들은 총 4개 조로 편성되었으며 각 조는 담당 교수 1명, 팀장 1명, 작업자 7-8명으로 구성되었다. 팀장을 포함한 작업자들은 작업 도구를 통하여 각 문장에 대한 구문 자동 분석 결과를 검수하며 분석에 어려움이 있는 경우 작업 도구 등을 통하여 팀장에게 보고하였다.

또한 작업 도구에서 발생하는 오류나 작업의 편의를 위하여 개선이 필요한 부분을 작업자들이 팀장들에게 직접 보고할 수 있도록 하여 이를 작업 도구에 반영할 수 있도록 하였다. 그리고 구문 분석 작업 중에 발견되는 형태소 분석 오류, 원시 말뭉치상의 오류 등을 신고할 수 있도록 하여 이를 모아 국어원 측에 전달할 수 있도록 하였다.

2.6. 팀장 및 교수진 검수(2차 검수)

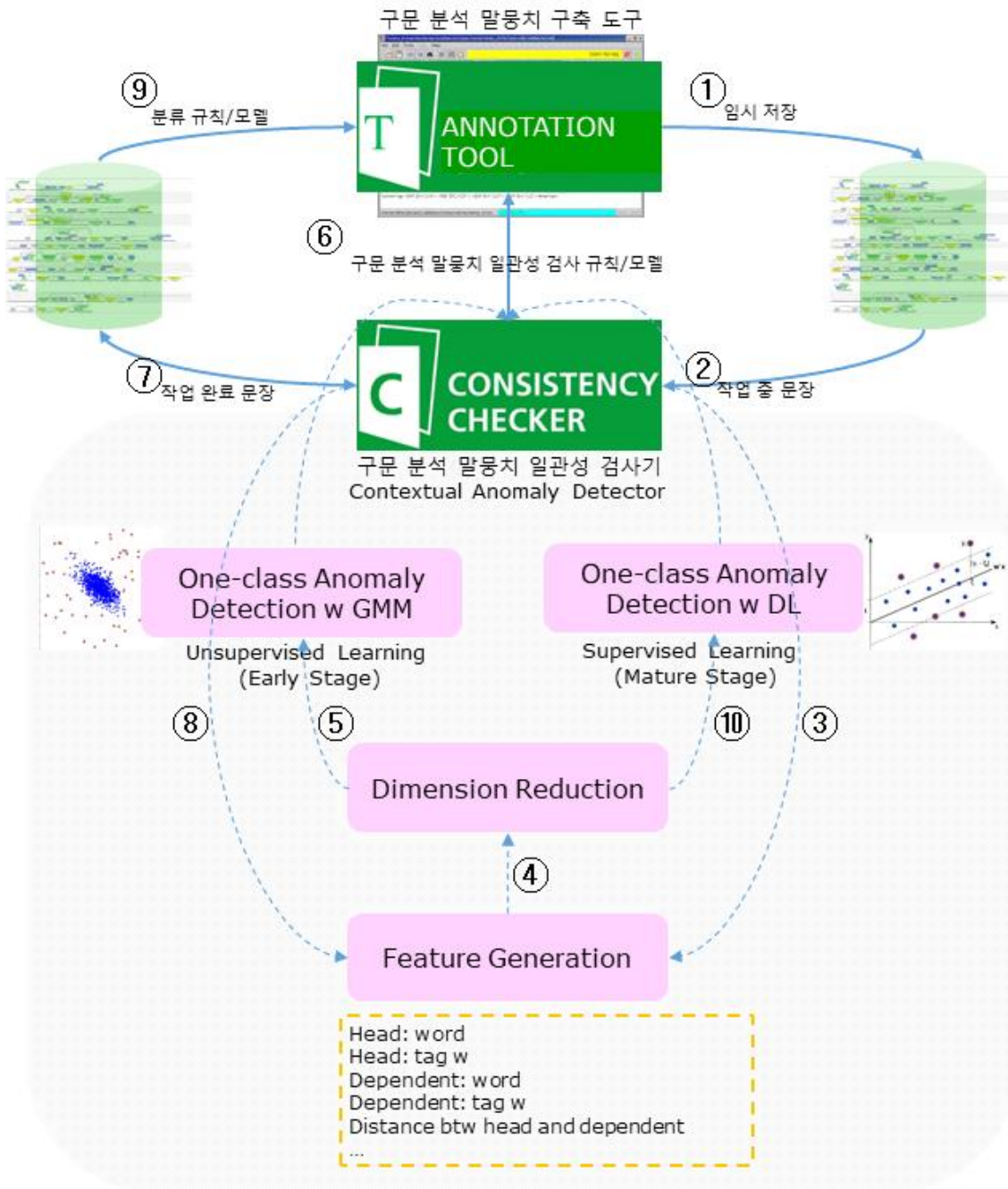
팀장 및 조별 담당 교수는 각 조원이 제출한 분석 결과를 작업 도구를 통해 수시로 검수한다. 이때 각 팀장 및 담당 교수가 전체 작업물에 대하여 검수하는 것에는 시간적·물리적 한계가 있어 각 조원에 대하여 샘플링을 통하여 검수하였다. 또한 검수한 내용을 바탕으로 작업에 피드백이 필요한 경우에는 조원에게 개별적으로 피드백을 제공하여 작업의 질과 일관성을 높였다. 이때 각 팀에서 빈발하는 오류 유형 등을 팀장들이 공유하여 재교육에 포함할 수 있도록 하였다.

1차 검수 과정에서 파악된 질문이나 작업자들이 분석에 어려움을 겪는 유형을 수시로 지침에 반영하였으며 지침에서 수정·보완된 사항을 각 팀원에게 전달하여 작업자들이 작업에 바로 반영할 수 있도록 하였다. 또한 기계 처리를 통해 처리할 수 있는 구문의 유형을 수집하여 자동 구문 분석 및 후처리에 반영함으로써 분석의 일관성을 높이고 검수 환경을 개선하고자 하였다.

2.7. 딥 러닝 기반 구문 분석 말뭉치 검증

2.7.1. 해양대학교 구문 분석 말뭉치 검증 모델

구문 분석 말뭉치 검증의 목적은 크게 두 가지이다. 하나는 말뭉치의 신뢰성 및 일관성을 확보하는 것이고 다른 하나는 말뭉치의 오류를 최소화하는 것이다. 구문 분석 말뭉치 검증 시스템의 구조는 <그림 19>와 같다.

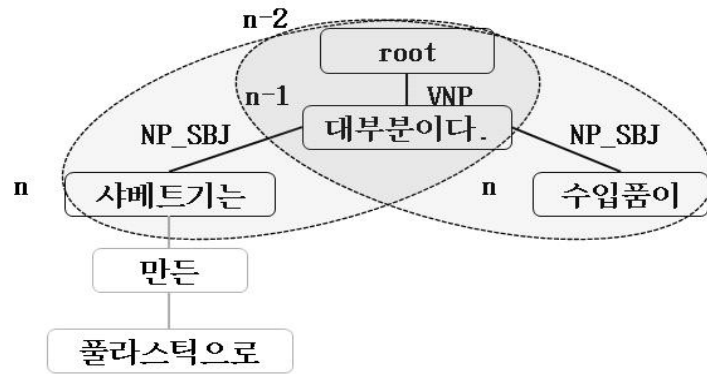


<그림 19> 구문 분석 말뭉치 검증 시스템의 구조도

<그림 19>에서 보는 바와 같이 구문 분석 말뭉치 검증은 일곱 개 단계로 구성되며 이하에서 각 단계에 대해서 구체적으로 설명할 것이다.

- 1) 수동 말뭉치 구축: 초기 학습(early stage) 때에는 전문가가 직접 구문 분석 말뭉치를 구축하여 말뭉치의 신뢰성을 확보한다. 이는 구축 도구에 의해서 구축된다. 그러나 초기 학습 이후에는 분류 모델에 의해서 구문 분석 오류 후보를 제시하여 말뭉치 구축의 생산력을 향상시킬 수 있다.
- 2) 자질 생성(feature generation): 경험적 지식(experiential knowledge)으로 자질을 선택하고 생성하여 구문 분석 말뭉치의 신뢰성을 확보한다. 평가 데이터 집합(test

data set)에 대한 추론(inference) 이후 일관성 검사기에서 나타나는 구문 분석 오류를 분석하여 자질을 추가한다. 본 연구에서는 <그림 20>과 같은 구문 분석 트리를 바탕으로 <그림 21>과 같은 자질을 추출하여 하나의 큰 벡터를 생성하고 이 벡터가 검증 시스템의 입력으로 사용된다.



<그림 20> 구문 분석 트리의 예

NP_SBJ				
$Word_{n-2}$	$DepRel_{n-2 n-1}$	$Word_{n-1}$	$Word_n$	$Dis_{n-1 n}$
root	VNP	대부분이다.	샤베트기는	2
root	VNP	대부분이다.	수입품이	1

660

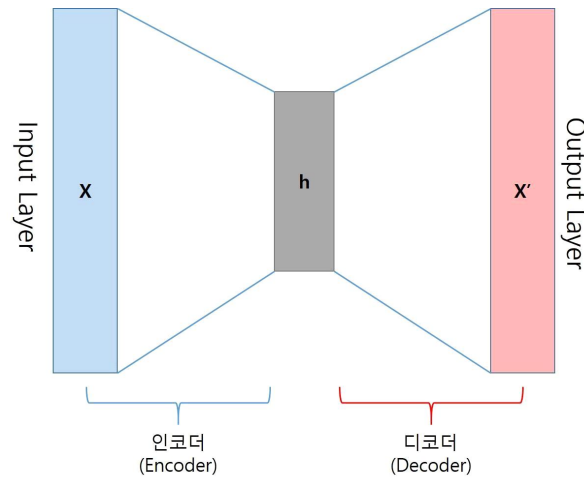
<그림 21> 구문 분석 검증 시스템의 입력 자질 집합

<그림 21>과 같이 구문 분석 문맥 표상에서는 2단계 위($n-2$)의 지배소의 단어 ($Word_{n-2}$)와 1단계 위($n-1$)의 지배소의 단어($Word_{n-1}$), 그리고 두 지배소 간의 의존 관계 (dependency relation)($DepRel_{n-2|n-1}$), 현재 단어($Word_n$), 마지막으로 현재 단어와 $n-1$ 단어와의 거리(distance)정보($dis_{n-1|n}$) 총 5개를 포함한다. 각 단어는 200차원의 크기를 가지고 의존 관계와 거리는 30차원의 크기를 가진다. 결과적으로 구문 분석에서의 문맥 표상의 크기는 총 660이다. 이와 같은 방법으로 구문 분석 말뭉치 전체에서 같은 의존 관계를 가지는 부분 트리에 대해서 자질 벡터를 추출하며 데이터로 사용한다.

3) 자질 축소(dimensionality reduction): 일반적으로 군집화의 복잡도는 $O(NKDI)$ 이다.

여기서 N 은 자질 벡터의 수이고, K 는 군집의 수이고, D 는 자질 벡터의 크기이며, I 는 반복 횟수이다. 따라서 자질 벡터의 차원이 크면 오류 후보 탐지에 많은 시간이 소요된다. 이를 다소 완화하기 위해서 자질 벡터의 크기를 축소하여야 한다. 본 연구에서는 자기부호화기(auto-encoder)를 이용하여 차원 축소를 수행하며 그 구조

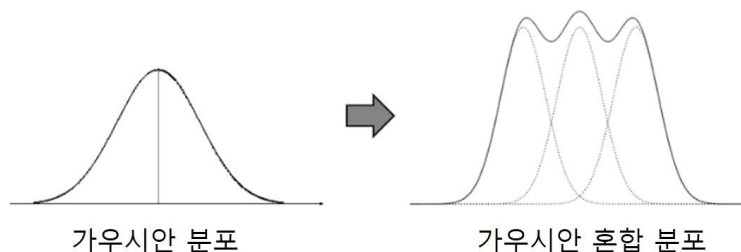
는 <그림 20>과 같다.



<그림 22> 자질 축소를 위한 자기부호화기.

<그림 22>에서 자기부호화기는 입력층의 크기를 660, 은닉층의 크기를 100, 출력층의 크기를 입력층과 동일한 660으로 설정하여 자기부호화기를 학습시켜 사용한다. 이렇게 미리 학습한 자기부호화기의 부호부(encoder) 부분을 사용하여 실시간으로 문맥 표상의 차원 크기를 100으로 축소하여 사용한다.

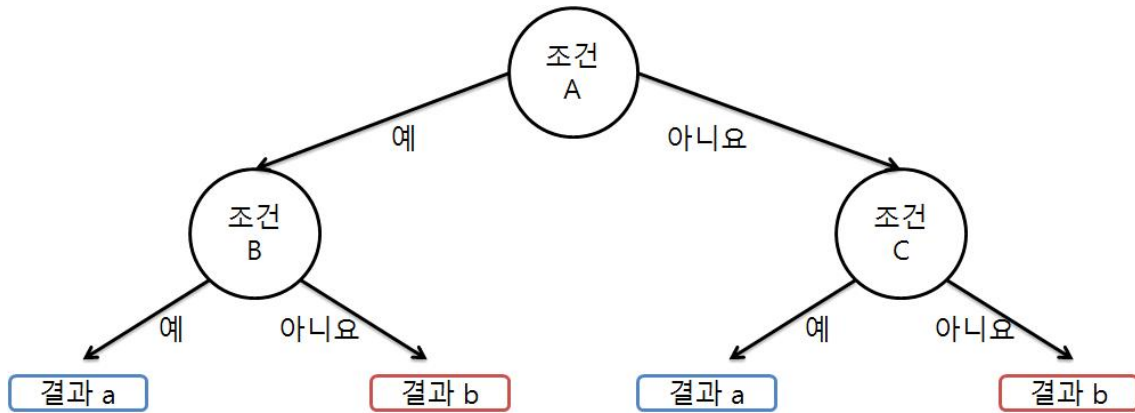
- 4) 가우시안 혼합 모델(Gaussian Mixture Model, GMM): 가우시안 혼합 모델은 군집화 알고리즘 중 하나이며 <그림 23>과 같이 여러 개의 가우시안 분포를 하나로 결합한 모델이며 이 모델은 말뭉치의 양이 충분하지 않을 때 사용한다. 군집화 알고리즘은 데이터에 표지(label)가 붙어있지 않은 학습인 비지도 학습(unsupervised learning)으로 일관성 검사기에서는 하나의 군집(one-class)만 사용한다. 즉 하나의 표지에 대해서 하나의 모델을 만들고 각 모델에서 입력되는 자질이 가우시안 혼합 분포에서 얼마나 벗어났는지를 이용해서 오류 후보를 제시한다.



<그림 23> 가우시안 혼합 분포에 대한 설명

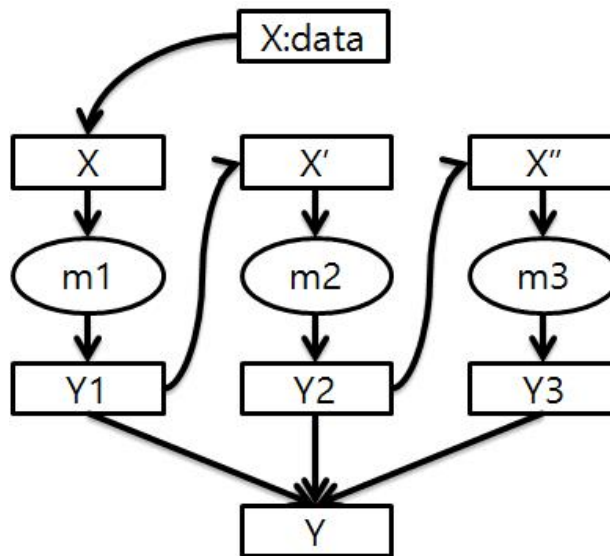
- 5) 의사 결정 나무(Decision Tree): 본 연구에서는 당초 서포트 벡터(support vector)를 이용하여 오류를 추정하려고 했으나 지지 벡터보다 일반적으로 성능이 더 좋은 것으로 알려진 앙상블(ensemble) 모델을 사용하며 그 기본 구조는 <그림 24>와 같은

의사 결정 나무를 바탕으로 한다.



<그림 24> 의사 결정 나무를 이용한 의존 관계 분류 모델

<그림 24>는 의존 관계를 분류하는 심층 학습 모델이다. 일반적으로 오류 검증에서는 하나의 분류 모델만으로는 그 성능이 만족스럽지 못하므로 본 연구에서는 여러 개의 모델을 결합하는 앙상블 모델을 사용한다. 본 연구에서 사용한 앙상블 모델은 엑스지부스트(XGBoost) 모델이며 그 구조는 <그림 25>와 같다.



<그림 25> 오류 검증을 위한 XGBoost 모델

<그림 25>는 여전히 분류 모델이다. 이와 같은 방법으로는 전체 말뭉치의 오류를 검증할 수 없다. 왜냐하면 전체를 학습 말뭉치로 사용하기 때문이다. 따라서 본 연구에서는 말뭉치의 일부를 학습으로 하고 나머지에 대해서 오류를 검증하는 교차 검증(cross validation)을 사용한다. 본 연구에서는 <그림 26>과 같이 세 부분으로 나누어 오류를 검증할 것이다. 교차 검증(cross validation) 1에서는 AB를 학습하고 C에 대해서 오류를 검증하고 교차 검증(cross validation) 2에서는 AC를 학습하고 B에 대해서 오류를 검증하며 교차 검증(cross validation) 3에서는 BC를 학습하고 A를 검증하여 모든 말뭉치에

대해서 오류를 검증한다.

	A	B	C
Cross Validation 1	Train	Train	Test
Cross Validation 2	Train	Test	Train
Cross Validation 3	Test	Train	Train

<그림 26> 오류 검증을 위한 교차 검증

- 6) 점진적 확장(incremental expansion): <그림 19>에서 보인 바와 같이 본 연구에서는 말뭉치가 충분해지면 가우시안 혼합 모델을 이용한 비지도 학습 대신에 심층 학습을 이용한 지도 학습 방법을 사용한다. 이와 같은 방법으로 말뭉치로 학습을 심화하여 구문 분석 말뭉치의 오류를 최소화하고 말뭉치 축적으로 전문가가 직접 검수해야 하는 분량을 최소화할 수 있으므로 생산성을 크게 향상시킬 수 있을 것이다.
- 7) 결과 분석: <그림 26>은 구문 분석 말뭉치 검증 결과의 예시이다. 오류가 있다고 판단되면 첫 줄과 두 번째 줄에는 문장번호(sent_id)와 원문이 적힌다. 그 이후부터는 오류가 있다고 판단하는 의존 관계의 아이디(id), 단어, 구문 분석, 태깅된 의존 관계, 정답으로 판단하는 의존 관계 순으로 적혀진다. 각 의존 관계 옆에는 각 의존 관계로 판단되는 확률이 적혀있으며, 적혀있지 않은 전체 의존 관계의 확률을 더하면 1이 된다.

```
# sent_id = NWPW1800000021-0258-0022
# sent_id = 선정된 학생에게는 고등학교 졸업 때까지 매 학기 모든 참고서를 후원하기도 한다.
2 학생에게는 학생에게는 NP_AJT(0.421) NP_SBJ(0.454)

# sent_id = NWPW1800000021-0277-0012
# sent_id = 무법선은 1895년 10월 8일 주한 일본공사 미우라 고로의 지휘하에 일본 측이 명성황후를 시해(을미사변)할 때 조선훈련대 제2대대장을 맡고 있었다.
4 8일 8일 NP_AJT(0.249) NP(0.395)

# sent_id = NWPW1800000021-0296-0020
# sent_id = 'god' 콘서트나 솔로 콘서트에서야 실수를 해도 팬들이 너그럽게 봐주겠지만, 뮤지컬 무대라면 그렇지 않은 것이다.
1 'god' 'god' AP(0.324) NP(0.395)

# sent_id = NWPW1800000021-0378-0002
# sent_id = 2007년 4월 아시아나항공 승무원 987이 공채로 입사한 미무원 씨(28)의 별명은 한때 '두원먼니' 였다.
11 한때 한때 NP(0.392) NP_AJT(0.471)
```

<그림 26> 구문 분석 말뭉치 검증 결과 예시

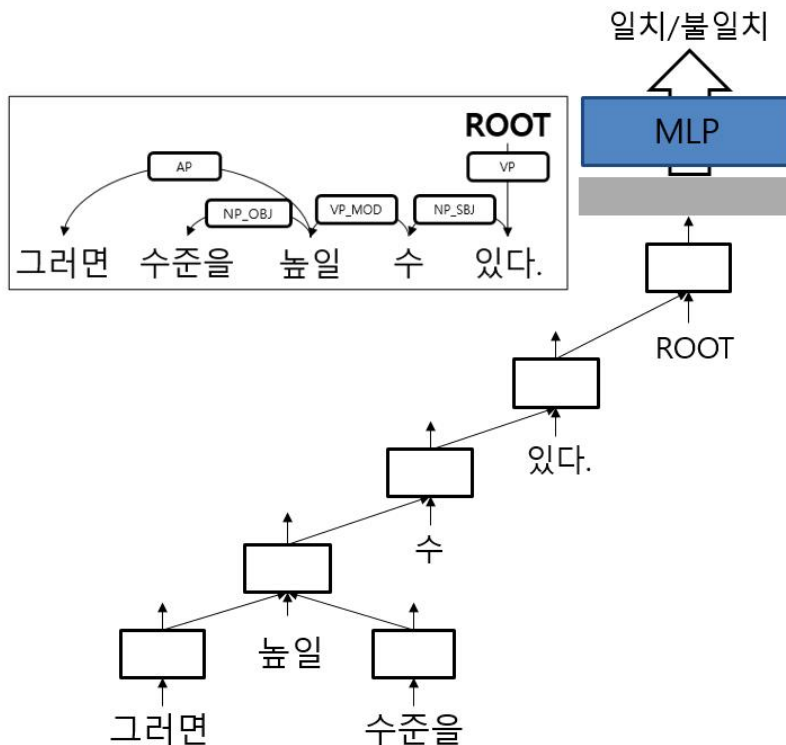
2.7.2. 전북대학교 딥 러닝 기반 구문 분석 말뭉치 검증 모델

구문 분석 말뭉치 검증은 자동 구문 결과 혹은 수작업으로 구축한 구문 분석 데이터 세트에 대해서 문장별 혹은 단어 단위의 태깅 품질을 예측하도록 하는 딥 러닝 기반의 구문 분석 태깅 품질 검증/교정기를 개발하는 것을 목표로 한다. 제안한 딥 러닝 기반 구문 분석 말뭉치 검증 방법은 1) 뉴럴 모델에 기반한 검증 모델과 2) 베이지안 모델 불확실성에 기반한 검증 모델, 두 가지 검증 방법을 제안한다. 그래프 뉴럴 모델에 기반한 검증 모델은 뉴럴 모델을 통해 구문 분석 품질을 예측하고 베이지안 모델 불확실성에 기반한 검증 모델은 엠시 드롭아웃(MC Dropout) 등을 이용하여 반복적인 샘플링을 통해 결과의 일관성을 측정하여 모델의 결과가 어느 정도의 불확실성을 갖는지에 대한 ‘확신도’를 측정한다.

1) 문장 단위 검증 모델

-트리 엘에스티엠(Tree-LSTM)

문장 단위의 검증은 문장 내에 모든 단어 즉, 의존소에 대해 정확한 지배소가 부착되었는지 여부를 검증하는 모델로 본 연구에서는 Tree-LSTM을 통해 인코딩된 트리의 최종 표상인 루트(ROOT)에서의 판별 모델을 제안한다. 제안하는 방식은 아래의 그림과 같다.



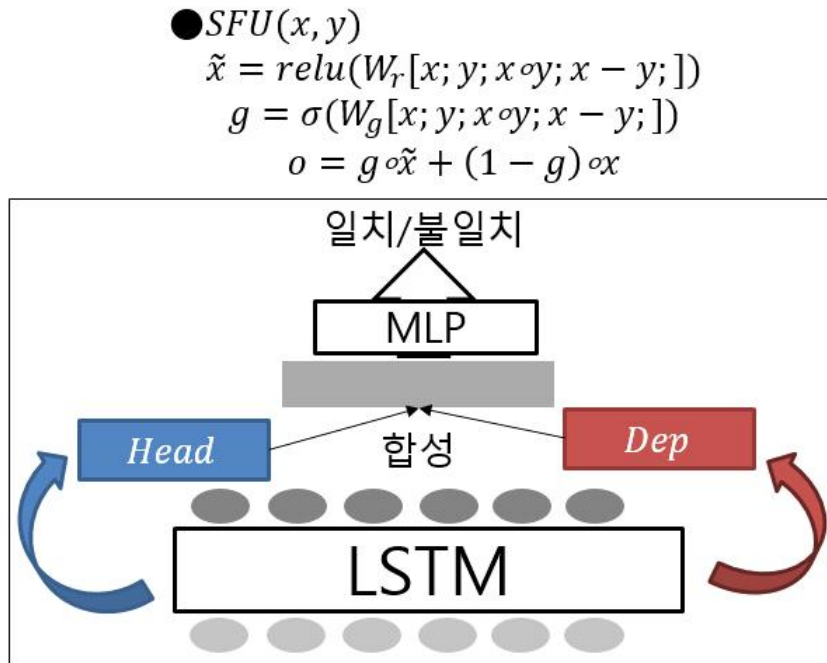
<그림 27> 트리 엘에스티엠 모델

먼저, 예측된 결과로부터 트리를 구성한 후 구성된 트리를 Tree-LSTM을 통해 각 노드를 자식으로부터 부모로 전달하는 방식으로 업데이트한 후 최종 인코딩된 ROOT 표

상에 MLP를 적용하여 구성된 트리가 정답 트리와 일치하는지에 대해 이진 분류를 수행한다.

2) 단어 단위 검증 모델

- 합성 기반



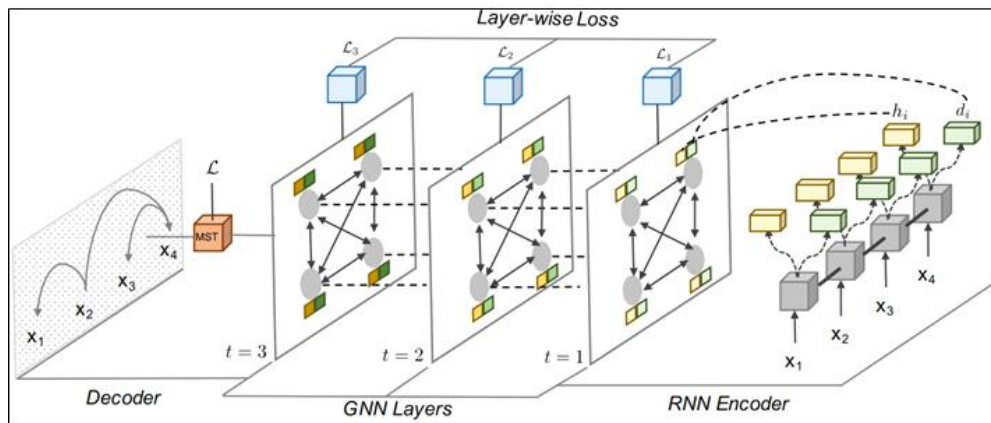
<그림 28> 단어 단위 검증 모델 중 합성 기반 방식

단어 단위 구문 분석 검증 모델은 수작업으로 태깅된 코퍼스 혹은 모델을 통해 예측된 결과가 주어졌을 때 각 단어(의존소)에 대해 예측 지배소가 정답 지배소가 맞는지 판별하는 모델로 위의 그림에서 보듯이 입력 문장에 대해서 LSTM을 통해서 각각 MLP를 통해 의존소(Dep)와 예측 지배소(Head)에 대한 표상을 얻고 이를 합성하여 MLP 층을 통해 각 의존소에 대한 지배소가 올바르게 태깅되었는지 여부를 판별하게 된다. 합성하는 방법은 먼저, 두 표상을 결합(concatenate)한 후 에프엔엔(FNN)을 통해 합성하는 1) 콘캣에프엔엔(ConcatFNN) 방법과 위의 <그림 28>에서의 에스에프유(SFU) 함수를 통해 하나의 벡터로 합성하는 2) SFU 합성 방법 두 가지를 사용한다.

- 그래프 신경망(GNNs) 기반

그래프 신경망(Graph Neural Networks, 이하 GNNs) 기반 모델은 합성 기반의 확장 모델로 최종적으로 의존소와 예측 지배소를 표상을 얻을 때 합성 기반과 같이 LSTM을 통해 인코딩된 결과에 MLP를 적용하는 것이 아닌 <그림 29>와 같은 그래프 신경망 기반 구문 분석 모델의 최종 의존소, 지배소의 표상에 대해 합성을 수행하게 된다. 그래프 신경망 기반 구문 분석 모델의 손실을 L_1 이라 하고 검증 모델의 손실을 L_2 라 할 때 이 두 손실을 더해 최종 손실 L_{final} 을 통해 GNNs 구문 분석 모델과 검증 모델을 동시에

학습하는 멀티 태스크(Multi-task) 방식으로 진행한다.



<그림 29> 단어 단위 검증 모델 중 그래프 신경망 기반 방식

- 베이지안 모델 기반

본 연구에서는 엠시 드롭아웃(MC Dropout)을 적용하여 모델이 불확실성을 가지게 되어 샘플링을 통해 확신도를 측정하는 방식으로 베이지안 모델을 도입하였다. 이렇게 얻은 확신도는 확신한 의존성 결정 여부를 판단하는 지표로 사용하여 샘플링한 결과에 대해서 동일한 결과가 얼마나 나오는지에 대한 확신도를 90%, 80%, 70% 등등 측정하여 전체 결과에 대해서 각각 확신도에 대해서 실제로 정답과 일치하는지 여부를 판단하여 가능한 많은 문장을 포함하면서 일정 수준의 품질을 보장할 수 있는 확신도를 선택한다.

2.8. 전문가 집단 심층 면접(Focus Group Interview)

통사론 및 구문 분석 전문가 집단과 연구진의 심층 면접(Focus Group Interview)을 통해 지침 및 말뭉치 검토 결과에 대해 토론하고 구문 분석 말뭉치 보완 방향에 대해 논의하였다.

1) 집단 심층 면접 참석 전문가

문승철(한국항공대, 통사론)
홍문표(성균관대, 전산언어학)
유혜원(단국대, 통사론/구문 분석)

2) 토론 내용

■ 인용문 처리

① 경계와 기능의 문제

- 원시 말뭉치가 신문 기사라는 특성으로 인하여 ‘OOO은 “~한다”며 “~한다”고 말했다.’와 같은 형식으로 이루어진 문장들이 많음. 이때 서술어들은 모두 VP로 연결되어 있는데, 일반적인 동사구와 동일하게 보기는 어려운 측면이 있음.
- 신문에서 사용되는 따옴표는 온전한 직접 인용이라고 보기 어렵고, 간접 인용의 강조 기능이라고 할 수 있음. 따라서 이들을 일반적인 인용문처럼 취급하는 것은 바람직하지 않음.
- 구문 분석 지침에서 인용문을 VP_CMP, 즉 보문으로 처리하도록 되어 있으나 직관적으로나 언어학적으로나 매우 이질적임.
- 인용문뿐만 아니라 신문 기사라는 장르적 특성으로 인해 나타나는 문제들에 대한 처리를 자세히 검토하고 향후 반영할 필요가 있음.

② 어절 분할

- 현재 구문 분석 말뭉치 분석 작업은 원시 말뭉치의 어절 단위를 반드시 그대로 유지하고 더이상 분석하지 않는다는 대전제를 가지고 있음. 이에 따라 ‘~ 느껴진다”면서’와 같이 상이한 단위가 결합한 어절을 하나의 단위로 처리할 수밖에 없음. 이번 사업이 아니더라도 언어학적으로는 이를 분석할 수 있는 방법론을 고려해야 함.

■ ‘이다, 아니다’의 처리

- ‘아니다’는 서술어로 보고 ‘이다’는 처리하지 않고 있는데, 한국어 문장에서 ‘이다’ 구문을 제외하는 것은 문장의 구조를 파악하는 데에 문제를 야기함.
- ‘이다’를 별도의 서술어로 처리하는 데에 더하여 ‘반대이다’류와 같은 서술어들의 목록을 마련하여 구분해야 함.

■ 부사절의 처리

부사절 태그는 VP로만 되어 있어서, 문장 마지막 서술어와 부사절(접속)이 태그 상으로 구분되지 않는 문제가 있음. 연관된 문제로 아래의 문장처럼 인용문이 두 개 나오고 두 절이 ‘-며’ 등의 어미로 연결이 될 때 앞에 ‘있다”며’는 VP가 되고 ‘방침”이라고’는 VNP_CMP가 되는 등 계사가 아닌 동사가 오면 VP_CMP가 된다는 것이므로¹¹⁾ 이런 유형의 처리에 대해 재고할 필요가 있음.

예) 황시영 부사장은 "이러한 IT 경쟁력이 최근 맹추격하는 중국 조선업을 따돌릴 수 있는 핵심 능력으로 인정받고 있다"며 "세계 최초의 와이브로 조선소가 앞으로 1~2년간 성공적으로 운영되면 다른 회사에 노하우를 전수하는 방안도 추진

11) 실제 구축 지침에서는 ‘~다며’ 등의 인용절은 부사절과 동일하게 처리하는 것이 원칙이고 ‘~라고’의 경우 인용절로 처리하는 것이 원칙이다. 이는 인용격 조사 ‘라고’와 ‘고’에 한정하여 인용절을 인정하고 있기 때문이다. 또한 이때 기호로 분리되어 있는 어절의 경우 기호의 앞부분을 기준으로 구문 태그를, 기호의 뒷부분을 기준으로 기능 태그를 부여하므로 ‘방침”이라고’의 구문 태그는 ‘방침”을 기준으로 NP를 부여한다.

할 방침"이라고 말했다.

■ 지배 관계의 적절성

- ① 체언 수식 부사의 경우, 언어학적으로는 부사가 체언을 수식하도록 분석되어야 하지만 본 구문 분석 말뭉치 구축 과제에서는 분석의 일관성 및 경제성을 기하기 위하여 부사는 체언을 직접 수식할 수 없는 것으로 일괄 처리하는 방식을 채택하였는데 이렇게 기능을 변별하지 않는 문제에 대해 향후 더 고민할 여지가 있음.
- ② 본 사업에서 구축하는 말뭉치가 신문 기사인 관계로 다수의 문장이 따옴표 안에 직접 인용되고 이렇게 구성된 여러 인용문이 하나의 긴 문장을 구성하는 경우가 다수 발생함. 여러 층위의 주석을 같은 대상에 대해 부착하는 통합 분석 말뭉치이기 때문에, 층위별 통일성을 기하기 위해 문장 분할이 잘못되어 여러 문장이 구문 분석의 단위가 되어도 더 쪼개지 않고 문장 간의 의존 관계를 표시하는 방식으로 처리하고 있는데 일반적인 의존 관계 설정과 변별하거나 처리 방식을 변경할 가능성은 없을지 면밀히 검토해야 함.

■ 괄호나 기호 등으로 붙어서 나타난 어절의 처리

신문 기사의 특성이면서 어절 분할이 제대로 되지 않은 결과이기도 하나, 괄호나 기호 앞뒤에 공백이 없어 한 어절로 묶여 처리되는 경우 구문, 의미역 분석에 모두 어려움이 있고 작업자마다 처리하는 방식이 다를 것으로 보임. 이러한 경우의 주석에 대한 구체적인 지침을 마련할 필요가 있음.

■ 용언의 명사형 처리 지침 구체화

용언의 명사형의 경우 관형형, 부사형 등과는 달리 명확한 태깅 원칙이 제시되지 않아 작업자별로 다르게 적용할 가능성이 높음.

용언의 명사형 유형을 검토하여 지침에 반영할 수 있을지에 대한 논의가 필요함.

■ 단위의 문제

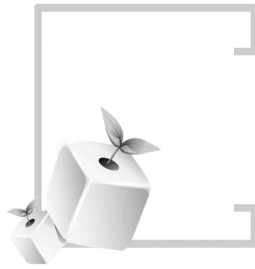
합성 용언과 구 구분 문제, 띄어쓰기 문제를 지침상에서 분리하고 구체적인 기준을 마련해야 할 필요 있음.

2.9. 최종 결과물 산출

구문 분석 말뭉치 구축의 최종 단계는 2차 검수를 마친 구문 분석 말뭉치를 제이슨(JSON) 형식으로 변환하여 최종 결과물을 산출하는 것이다. 이때 2차 검수 과정에서 수집된 후처리가 필요한 데이터를 수정하는 과정이 포함된다. 제이슨(JSON) 형식의 기본

구조와 예시는 부록에서 제시하였다.

또한 본 사업의 최종 납품 후 국어원의 피드백을 반영하여 하자 보수 기간이 진행될 예정이다.



제 3 장

구문 분석 말뭉치
구축 지침 수립



1. 지침 수립 과정

구문 분석 말뭉치 구축 지침은 기본적으로 한국정보통신기술협회(TTA)에서 구축한 ‘의존 구문 분석 말뭉치 구축을 위한 의존 관계 태그 세트 및 의존 관계 설정 방법’(이하 ‘티티에이(TTA) 지침’)을 기반으로 하되, 실제 작업에서 나타날 수 있는 문제점들을 해결하기 위한 사항 및 구체적인 예시 등을 반영하는 방식으로 이루어진다.

티티에이 지침은 한국어를 대상으로 하여 제시된 표준적인 지침으로서의 지위를 가지고 있으며, 특히 형태 분석 및 의미역 분석 태깅과 원활하게 호환될 수 있도록 함을 목적으로 한다는 점에서 본 사업에서 구축하는 말뭉치의 활용을 용이하게 한다.

그러나 기존 티티에이 지침에는 수정이 필요한 오류가 적지 않게 존재할 뿐만 아니라, 실제로 원시 말뭉치를 대상으로 하여 구문 분석 작업을 하는 데 있어서 결정이 어려운 사항들에 대한 가이드라인을 제시함으로써 일관된 작업이 이루어지도록 할 필요가 있다.

이에 본 사업에서는 우선 티티에이 지침을 기반으로 하되, 일부 내용을 수정하고 항목별 설명을 상세화하며 다양한 사례를 포함함으로써 실제 구축 과정에서의 정확성과 일관성을 기하고자 하였다. 여기에서는 본 사업의 구축 지침의 각 장에서 추가·보완된 사항을 중심으로 살펴보도록 한다.

1.1. 기본 원칙

본 지침의 ‘기본 원칙’은 티티에이 지침의 ‘5. 의존 관계 태그 세트 및 의존 관계 설정 방법’의 기본 원칙 부분을 기반으로 하며, 이로부터 큰 수정 사항은 없다.

1.2. 의존 관계 태그 세트 설정 방법

본 지침의 ‘기본 원칙’은 티티에이 지침의 ‘5. 의존 관계 태그 세트 및 의존 관계 설정 방법’의 ‘의존 관계 태그 세트 설정 방법’ 부분을 기반으로 하며, 이로부터 큰 수정 사항은 없다.

1.3. 문장 유형별 의존 관계 설정 방법

문장 유형별 의존 관계 설정 방법에서는 특히 복문, 즉 내포문의 분석이 핵심적인 부분이 된다. 티티에이 지침에서는 내포문을 명사절 내포문, 관형절 내포문, 부사절 내포문, 서술절 내포문, 직접 인용절 내포문, 간접 인용절 내포문의 여섯 가지 유형으로 분류하였고, 본 지침에서도 이 여섯 가지 유형에 해당하는 문장들의 분석 방법을 지침에

포함하여 기술하였다.

- 내포문은 명사절, 관형절, 부사절, 인용절 내포문으로 크게 분류하여 분석한다.

내포문의 유형별 예문은 아래와 같다.

- 명사절: 우리는 시민의 관전 태도도 그만큼 성숙했음을 잊지 말아야 한다.
- 관형사절: 내가 좋아하는 꽃은 들국화다.(관계 관형절)
내가 사진을 좋아하는 사실을 친구들은 다 안다.(동격 관형절)
- 부사절: 멜라닌은 자외선을 차단해서 자외선으로부터 피부를 보호해 준다.
비가 와서 땅이 미끄럽다.
- 인용절: 그는 그녀가 그 일을 해냈다고 밝혔다.(간접 인용절)
그는 “그녀가 그 일을 해냈어.” 라고 말했다.(직접 인용절)

모문과 내포문의 주어가 다르고, 서술어가 동일한 경우에는 명사구 접속 구조와 혼동할 수 있는 유형을 제시함으로써 작업에 유의하도록 하였다.

- 명사구 접속 구조와 혼동되는 경우가 있는데, 비슷한 구조에서 ‘와/과’로 이어진 접속 명사구만 ‘CNJ’ (명사구 접속)로 처리함에 유의한다.

[예시] 2006년 연구와 2008년 연구에서는 연구팀의 주장이 유지되었다.

2006년	→ NP	연구와
연구와	→ NP_CNJ	연구에서는
2008년	→ NP	연구에서는
연구에서는	→ NP_AJT	유지되었다.
연구팀의	→ NP_MOD	주장이
주장이	→ NP_SBJ	유지되었다.
유지되었다.	→ VP	ROOT

티티에이(TTA) 지침에서는 모문-내포문 관계가 하나만 존재하는 경우만 제시되어 있는데, 실제 원시 말뭉치에서는 내포문이 3개 이상 포함되면서 중의적으로 해석되는 경우가 적지 않게 존재한다. 이를 해결하기 위한 지침을 추가하였다.

3.2.3. 내포문이 3개 이상 포함되어 중의적으로 해석되는 경우

- 문장 좌측에서 우측 방향으로 순서대로 의존 관계를 설정하여 분석한다.
- 복문의 해석이 중의적일 때에는 가능한 의미 중에서 가장 가까이에 후행하는 절에 의존한다.

[예시] 김민이 돈을 잘 써서 멋진 선배로 통한다.

김민이	→ NP_SBJ	써서
돈을	→ NP_OBJ	써서
잘	→ AP	써서
써서	→ VP	멋진
멋진	→ VP_MOD	선배로

선배로	→ NP_AJT	통한다.
통한다.	→ VP	ROOT

- 복문의 해석이 단일한 때에는 해당 해석에 따라 구조를 분석한다.

[예시] 김민은 신작에서 판타지의 비중은 줄이고 사회를 비판한 내용을 부각했다.

줄이고	→ VP	부각했다.
-----	------	-------

한국어 문장 구문 처리에서 큰 문제가 되는 것 중 하나는 바로 이른바 이중 주어문이다. 다양한 양상으로 나타나는 이중 주어문을 일관적으로 처리할 수 있도록 관련 지침을 추가하여 제시하였다.

3.2.5. 이중 주어문 분석 방법		
- 이중 주어문은 각 구성 성분별로 동일한 서술어에 의존하도록 분석한다.		
- 이때 일반적인 이중 주어문뿐만 아니라 [NP1이 NP2가 VP]의 격틀 구조를 가지는 일부 용언 분석에도 동일하게 적용한다.		
- 예문은 아래와 같다.		
[예시] 나는 그 시계가 필요했다.		
나는	→NP_SBJ	필요했다.
그	→DP	시계가
시계가	→NP_SBJ	필요했다.
필요했다.	→VP	ROOT
[예시] 그는 그녀가 자랑스러웠다.		
그는	→NP_SBJ	자랑스러웠다.
그녀가	→NP_SBJ	자랑스러웠다.
자랑스러웠다.	→VP	ROOT

1.4. 세부 구별 태깅 가이드라인

‘[관형어+명사+명사] 유형 분석’에서는 여러 개의 명사로 이루어진 명사구가 병렬적으로 이어지는 경우와 수식 구조를 지니는 경우로 나뉘므로, 이들을 구분할 수 있는 지침을 추가할 필요가 있다.

- 여러 개의 명사로 이루어진 명사구의 내부에 수식 구조가 존재하는 경우에는, 각 명사구의 수식 구조를 고려하여 의존 관계를 설정한다.		
[예시] 프랑스 파리 루브르 박물관		
프랑스	→ NP	파리
파리	→ NP	박물관

루브르	→ NP	박물관
박물관	→ NP	ROOT

[예시] 김철수 OO당 대표가 정계에 복귀했다.

김철수	→ NP	대표가
OO당	→ NP	대표가
대표가	→ NP_SBJ	복귀했다.
정계에	→ NP_AJT	복귀했다.
복귀했다	→ VP	ROOT

티티에이(TTA) 지침에서는 주 서술어 다음에 후행하는, 보조 용언은 아니지만 서법을 나타내는 언어 단위들을 ‘의사 보조 용언’이라 하여 별도로 처리하고 있는데, 본 지침에서는 이를 인정하지 않는다. 따라서 티티에이(TTA) 지침에서 의사 보조 용언 구성으로 제시한 것들을 ‘본용언+보조 용언 구성’과 ‘의존 명사 구성’으로 나누어서 재구성하여 제시한다.

4.3.2. 본용언+보조 용언

- 본용언과 보조 용언이 연속하여 두 개 이상 나올 때는 주어를 본용언에 연결하고 본용언은 보조 용언에 연결한다.
- 주어 외에도 의존소들을 연결하고 있는 필수 성분들이 일차적으로 본용언에 연결되고, 본용언이 보조 용언에 의존하는 방식으로 처리한다.
- 본용언과 보조 용언이 붙여쓰기로 제시되어 있는 경우에는 해당 형식을 하나의 용언으로 처리한다.
- 보조 용언 구성이 두 개 이상 연속될 때에도 마찬가지로 본용언 → 보조 용언1 → 보조 용언2의 순으로 연결한다.

[예시] 멜라닌은 물에는 용해되지 않는다.

멜라닌은	→ NP_SBJ	용해되지
물에는	→ NP_AJT	용해되지
용해되지	→ VP	않는다.
않는다.	→ VP	ROOT

[예시] 만 원의 기부로 매일 세 명의 어린이가 살아 가고 있는 것으로 보고됐다.

만	→ NP	원의
원의	→ NP_MOD	기부로
기부로	→ NP_AJT	살아
매일	→ AP	살아
세	→ DP	명의
명의	→ NP_MOD	어린이가
어린이가	→ NP_SBJ	살아
살아	→ VP	가고
가고	→ VP	있는
있는	→ VP_MOD	것으로

것으로	→ NP_AJT	보고됐다.
보고됐다.	→ VP	ROOT
- 문장부사 또한 본용언이 아닌 보조 용언에 연결한다. 의존 명사 구성의 경우에도 동일하게 적용한다.		
[예시] 그러나 이번 분기는 생각보다 순조롭게 진행되고 있었다.		
그러나	→ AP	있었다.
4.3.3. 의존 명사 구성(의사 보조 용언 구성)		
- 주 서술어 다음에 보조 용언은 아니지만 서법을 나타내는 의존 명사가 포함된 구성이 오는 경우, 해당 서술어 및 의존 명사 구성을 별개의 단위로 처리하여 분석하고 의존 관계를 연결한다.		
- 의존 명사에 ‘이다’, ‘하다’ 등이 결합해 있는 경우에는 그 자체를 각각 VNP, VP 등으로 처리한다.		
- 이때 문장의 주어와 의존 명사의 관계는 해당 의존 명사가 실질 명사로 대체 가능한 경우와 불가능한 경우로 구분하여 분석한다. 의존 명사가 주어와 공지시(co-reference)되는 경우에는 주어와 의존 명사 어절을 연결하고, 그렇지 않은 경우에는 주어와 서술어를 연결한다.		
- 표면적으로는 ‘-ㄴ 것이다’ 와 같이 의사 보조 용언 구성에 해당되더라도, 의존 명사나 명사가 서법을 나타내지 않는 경우는 의사 보조 용언 구성에 해당되지 않는다.		
[예시] 그는 일어날 수 없었다.		
그는	→ NP_SBJ	일어날
일어날	→ VP_MOD	수
수	→ NP_SBJ	없었다.
없었다.	→ VP	ROOT
[예시] 나는 곧 밥을 먹을 것이다.		
나는	→ NP_SBJ	먹을
곧	→ AP	먹을
밥을	→ NP_OBJ	먹을
먹을	→ VP_MOD	것이다.
것이다.	→ VNP	ROOT
[예시] 이 사료는 가축들이 먹는 것이다.		
이	→ DP	사료는
사료는	→ NP_SBJ	것이다.
가축들이	→ NP_SBJ	먹는
먹는	→ VP_MOD	것이다.
것이다.	→ VNP	ROOT
→ 이때의 ‘것’ 은 ‘사료, 식품’ 을 뜻하므로 ‘사료는’ 이 ‘것이다’ 에 의존함.		
[예시] 내가 놀란 것은 철수가 천재라는 것이다.		
내가	→ NP_SBJ	놀란
놀란	→ VP_MOD	것은
것은	→ NP_SBJ	것이다.(것 = 사실 -> 사실 = 사실)
철수가	→ NP_SBJ	천재라는
천재라는	→ VNP_MOD	것이다.

것이다.	→ VNP	ROOT
→ 이때의 ‘것’은 ‘사실’을 뜻하므로 ‘것은’이 ‘것이다’에 의존하는 것으로 처리함.		
[예시] 곧 비가 올 것 같다.		
곧	→ AP	올
비가	→ NP_SBJ	올
올	→ VP_MOD	것
것	→ NP	같다.
같다.	→ VP	ROOT

티티에이(TTA) 지침에는 부호에 대한 별도의 지침이 없는데, 실제 말뭉치에서는 다양한 유형의 부호가 출현한다. 특히 띄어쓰기 등의 문제로 부호가 별도의 어절로 처리될 경우에는 이에 대한 태깅 가이드라인이 필요하다.

4.4. 부호

- 따옴표 (‘ ’ / “ ”) 및 괄호 (< > / []) 등과 같이 좌우 짝이 있는 부호는 각각 L, R 표지를 부착하고, 그 외의 경우에는 일괄적으로 X를 부착한다(괄호가 수리 기호나 층위 표시로 쓰일 때에도 X로 분석함).

- L은 R에 의존하도록 분석한다.

[예시] “ 나는 너를 좋아해 ” 라고 말했다.

“	→ L	”
”	→ R	라고

[예시] “ 나는 너를 좋아해 ” 라고 말했다.

“	→ L	좋아해 ” 라고
좋아해 ” 라고	→ VP_CMP	말했다.

[예시] 과일 (사과, 배 등)의 등급은

과일	→ NP)의
(사과,	→ NP_CNJ	배
배	→ NP	등
등	→ NP)의
)의	→ X_MOD	등급은

[예시] 과일 (사과, 배 등)의 등급은

과일	→ NP	의
(→ L)
사과,	→ NP_CNJ	배
배	→ NP	등
등	→ NP)
)	→ R	의
의	→ X_MOD	등급은

- 단락 기호는 기호가 속하는 최상위 행에 의존하도록 분석한다.

[예시] 철수의 버릇 : 다리 꼬기, 이갈기

버릇 → NP :
: → X 이갈기

[예시] 철수의 버릇: 다리 꼬기, 이갈기

버릇: → NP 이갈기
꼬기, → NP_CNJ 이갈기

[예시] - 철수의 버릇 : 다리 꼬기, 이갈기

- → X 이갈기
버릇 → NP :
: → X 이갈기

[예시] ● 내년 주요 핵심안은 예산 결의 문제

● → X 문제

- 그 외: 후행 어절에 의존하도록 분석한다.

[예시] 세종(1418 ~ 1450)은 조선전기 제4대 왕이다.

세종(1418 → NP ~
~ → X 1450)은
1450)은 → NP_SBJ 왕이다.

- 삽입구가 복수의 어절로 구성되어 있을 경우, 기호로 결합된 복합 형태소는 선행 성분을 기준으로 구문 태그를 결정하고, 후행 성분을 기준으로 기능 태그를 결정한다.

[예시] 전문위원의 임기는 3년을 보장한다(1차에 한하여 연임 가능).

보장한다(1차에 → VP_AJT 한하여

4.5. 외국 문자/외국어 처리 방법

- 외국 문자, 숫자를 비롯한 기능을 알 수 없는 미등재어의 구문 태그는 NP이다.

[예시] “닥쳐(Shut up)!”

“닥쳐(Shut → VP up)!”
up)! → NP ROOT

- 외국 문자/외국어는 바로 다음 요소에 의존하도록 분석한다.

[예시] 아이 러브 유

아이 → NP 러브
러브 → NP 유

[예시] I love you

I → NP love
love → NP you

4.6. 띄어쓰기 오류 처리 방법

- 어절 내부가 분할되어 있는 경우, 구문 태그와 기능 태그는 형태 분석 결과를 기준으로 정하고, 의

존 관계는 바른 띄어쓰기를 기준으로 정한다. ('19 국립국어원 형태 분석 말뭉치 참조)

- 절단 어절의 형태 분석 결과와 구문 분석 태그의 대응표는 다음과 같다.

형태 분석 태그	구문 분석 태그
NNG, NNP, NNB, NP, NR, XSN, XR, NF	NP
VV, VA, VX, VCN, EP, EF, EC, ETN, XSV, XSA, NV	VP
MMA, MMD, MMS	DP
MAG, MAJ	AP
IC	IP
JKS, JKC, JKG, JKO, JKB, JKV, JKQ	X
SF, SP, SS, SE, SO, SW, SL, SH, NA	NP

- 기능 태그는 맨 마지막 분할 어절에 부여한다. (조사 생략 경우도 마찬가지임.)
- 원래 어절 내부에서는 다음 분할 어절에 의존하고, 원래 어절의 분할 어절은 지배소 어절(의 최종 분할 어절)에 의존한다.

[예시] 마음 씨 가 중요하 다.

마음	→ NP	씨
씨	→ NP	가
가	→ X_SBJ	다.
중요하	→ VP	다.
다	→ VP	ROOT

1.5. 세부 유형별 가이드라인

티티에이(TTA) 지침에서는 ‘의존 관계 태그 부착 세부 유형 가이드라인’, ‘의존 관계 설정 세부 유형 가이드라인’, ‘부사구 연결 세부 유형 가이드라인’, 그리고 ‘장형 사동 구문 유형 세부 가이드라인’ 등 한국어의 특성으로부터 기인한 문제점들에 대한 태깅 원칙을 제시하고 있다. 그런데 실제로 작업을 수행하다 보면 굉장히 다양한 영역에서 태깅의 기준이 모호한 상황이 발생한다. 따라서 지침의 5장에서는 작업 중 확인되는 부분들에 대한 태깅 원칙을 지속적으로 추가해 나가고 이를 작업자들에게 꾸준히 교육함으로써 구문 분석 검토 작업이 일관적으로 이루어질 수 있도록 한다.

5.1. 의존 관계 태그 부착 세부 유형 가이드라인

5.1.1. 보조사적 쓰임을 보이는 ‘이/가’, ‘을/를’의 주석

- 본용언에 화용적 기능을 가지는 조사 {가/를}이 붙은 경우는 기능 표지를 부착하지 않는다. 즉, 명사형(-음, -기)을 제외한 용언의 활용형에 붙은 조사 {가/를}은 무시한다.

[예시] 철수가 밥을 이틀내 먹지를 않았다.

철수가	→ NP_SBJ	먹지를
밥을	→ NP_OBJ	먹지를
이틀내	→ NP_AJT	먹지를

먹지를	→ VP	않았다.
않았다.	→ VP	ROOT
[예시] 그 산은 그리 높지가 않다.		
그	→ DP	산은
산은	→ NP_SBJ	높지가
그리	→ AP	높지가
높지가	→ VP	않다.
않다.	→ VP	ROOT

- ‘-기 바라다’, ‘-기 시작하다’ 등의 ‘-기’ 절을 요구하는 서술어의 경우 뒤에 격 조사가 붙지 않는 경우와 붙지 않는 경우 모두 기능 태그를 부착하여 VP_OBJ로 태깅한다.

[예시] 나는 네가 빨리 오기(를) 바라다.		
나는	→ NP_SBJ	바라다.
네가	→ NP_SBJ	오기
빨리	→ AP	오기
오기(를)	→ VP_OBJ	바라다.
바라다.	→ VP	ROOT

5.1.2. ‘~즈음’, ‘~쯤’ 부사구의 주석

- “~즈음”, “~쯤” 등과 같이 의미적으로 시간과 공간을 의미하고 조사가 없는 경우 AJT 기능 태그를 부착한다.

[예시] 광복절 즈음 해서 독립기념관을 찾았다.		
광복절	→ NP	즈음
즈음	→ NP_AJT	해서
해서	→ VP	찾았다.
독립기념관을	→ NP_OBJ	찾았다.
찾았다.	→ VP	ROOT

- 그러나 ‘~ 즈음’, ‘~ 쯤’ 뒤에 다른 격 조사가 쓰이는 경우 격 조사를 고려하여 주석한다.

[예시] 다섯 명 즈음의 학생이 길에 서 있었다.		
다섯	→ DP	명
명	→ NP	즈음의
즈음의	→ NP_MOD	학생이
학생이	→ NP_SBJ	서
길에	→ NP_AJT	서
서	→ VP	있었다.
있었다.	→ VP	ROOT

5.1.3. 격 조사가 붙은 수량 관련 표현의 주석

- 시간이나 거리 등 수량이나 단위를 나타내는 명사구에 목적격 조사 ‘을/를’ 이 붙은 경우에는 목적어로 분석한다.

[예시] 나는 학교까지 다섯 시간을 걸었다.		
--------------------------	--	--

나는	→ NP_SBJ	걸었다.
학교까지	→ NP_AJT	걸었다.
다섯	→ DP	시간을
시간을	→ NP_OBJ	걸었다.
걸었다.	→ VP	ROOT

- ‘을/를’ 이 결합하지 않았더라도 ‘을/를’ 이외의 다른 격조사가 결합할 수 없는 개체/수량/횟수/시간/거리 표현의 경우 OBJ로 분석한다.

[예시] 노동자들의 삶을 담기 위해 2년에 걸쳐 5번(→ NP_OBJ 방문했다.) 방문했다.

[예시] 운동장을 세 바퀴(→NP_OBJ 뛰었다.) 뛰었다.

5.1.4. 품사와 문장 성분이 일치하지 않는 경우

- 품사로는 부사나 활용형 등이 쓰였지만 문장 내에서 인용이 된 것처럼 쓰인 경우에는 해당 문장 성분을 기준으로 기능 태그를 부여하되, 구문 태그 분석은 단어의 본래 품사를 기준으로 함. 다만 형태소의 일부가 잘린 경우 미등재어로 보아 NP를 부여한다.

[예시] ‘거꾸로’ 는 ‘거꾸’ 와 ‘로’ 로 분석할 수 있을까?

‘거꾸로’ 는	→ AP_OBJ	분석할
‘거꾸’ 와	→ NP_CNJ	‘로’ 로
‘로’ 로	→ NP_AJT	분석할
분석할	→ VP_MOD	수
수	→ NP_SBJ	있을까?
있을까?	→ VP	ROOT

5.2. 의존 관계 설정 세부 유형 가이드라인

5.2.1. 서술어의 역할(~이다. 그리고)을 하는 ‘으로’ 의 주석

- “~으로” 가 의미적으로 “~이다. 그리고” 와 같이 사용되었다면, 예외적으로 서술어로 인정한다. 즉, 주어 논항을 가질 수 있다.

[예시] 모나리자는 레오나르도 다빈치가 그린 초상화로, 현재 프랑스 파리 루브르 박물관에 전시되어 있다.

모나리자는	→ NP_SBJ	초상화로,
초상화로,	→ NP_AJT	있다.

- * “~으로” 가 겹표로 연결되어 있지 않아도 의미역으로 “~이다. 그리고” 에 대응된다면 마찬가지로 서술어로 인정한다.

[예시] 모나리자는 레오나르도 다 빈치가 그린 초상화로 현재 프랑스 파리 루브르 박물관에 전시되어 있다.

모나리자는	→ NP_SBJ	초상화로
초상화로	→ NP_ATJ	있다.

5.2.2. 부사 ‘없이’, ‘같이’의 주석

- “없이”나 “같이”와 같은 부사(용언의 활용형이 아님에 유의)는 서술어와 마찬가지로 논항을 취할 수 있다. 특히 부사 ‘같이’는 앞에 ‘~와’에 해당하는 명사구가 나타나는 경우 ‘~와’ 명사구를 ‘같이’의 부사어로 처리한다.

[예시] 철수는 아무 생각도 없이 길을 나섰다.

철수는	→ NP_SBJ	나섰다.
아무	→ DP	생각도
생각도	→ NP_SBJ	없이
없이	→ AP	나섰다.
길을	→ NP_OBJ	나섰다.
나섰다.	→ VP	ROOT

[예시] 예상한 바와 같이 주가가 크게 떨어졌다.

예상한	→ VP_MOD	바와
바와	→ NP_AJT	같이
같이	→ AP	떨어졌다.
주가가	→ NP_SBJ	떨어졌다.
크게	→ VP_AJT	떨어졌다.
떨어졌다.	→ VP	ROOT

5.3. 연결된 부사구 세부 유형 가이드라인: ‘~부터 ~까지’, ‘~에서 ~으로’의 주석

- ‘~부터’, ‘~까지’(또는 ‘~에서’, ‘~로’)와 같은 부사구는 각각을 지배소에 연결한다.

[예시] 그는 작년부터 지금까지 열심히 일했다.

그는	→ NP_SBJ	일했다.
작년부터	→ NP_AJT	일했다.
지금까지	→ NP_AJT	일했다.
열심히	→ AP	일했다.
일했다.	→ VP	ROOT

[예시] 부산에서 서울로 가는 표 있나요?

부산에서	→ NP_AJT	가는
서울로	→ NP_AJT	가는
가는	→ VP_MOD	표
표	→ NP_SBJ	있나요?
있나요?	→ VP	ROOT

- 그러나 만일 ‘~부터 ~까지를’, ‘~에서 ~로의’ 등과 같이 ‘~부터’, ‘~까지’(또는 ‘~에서’, ‘~로’) 부사구가 하나의 단위로 묶이는 경우 선행 성분을 후행 성분의 부사어(AJT)로 연결한 후, 후행 성분을 지배소에 연결한다.

[예시] 평균 90점부터 100점까지를 모두 금상으로 처리한다.

평균	→ NP	90점부터
90점부터	→ NP_AJT	100점까지를
100점까지를	→ NP_OBJ	처리한다.
[예시] 이것은 바로 개인주의에서 단체주의로의 전환을 의미했다.		
이것은	→ NP_SBJ	의미했다.
바로	→ AP	의미했다.
개인주의에서	→ NP_AJT	단체주의로의
단체주의로의	→ NP_MOD	전환을

5.4. 장형 사동 구문 유형 세부 가이드라인

- ‘~게 하다’의 장형 사동 구문은 다른 보조 용언과 동일하게 처리한다. 즉, ‘~게 하다’에서 선행하는 용언(‘~게’)에 다른 성분들을 모두 연결한다.
- 또한 ‘A가 B가/B를/B에게 V-게 하다’와 같은 장형 사동 구문의 B 성분은 격 조사에 의존하여 기능 태그를 부착한다. 즉, ‘B가’는 SBJ, ‘B를’은 OBJ, ‘B에게’는 AJT로 처리한다.

[예시] 그가 철수를 집에 가게 하였다.		
그가	→ NP_SBJ	가게
철수를	→ NP_OBJ	가게
집에	→ NP_AJT	가게
가게	→ VP	하였다.
하였다.	→ VP	ROOT
[예시] 그가 철수에게 집에 가게 하였다.		
그가	→ NP_SBJ	가게
철수에게	→ NP_AJT	가게
[예시] 그가 철수가 집에 가게 하였다.		
그가	→ NP_SBJ	가게
철수가	→ NP_SBJ	가게

5.5. 여러 개의 문장 처리

- 여러 개의 문장이 분할되지 않은 상태로 제시되어 있는 경우, 각 문장의 서술어가 순차적으로 의존하도록 처리한다.

[예시] “철수는 집에 갔어. 영화는 모르겠어. 민지는 집에 있겠지.”라고 말했다.		
갔어.	→ VP	모르겠어.
모르겠어.	→ VP	있겠지.”라고
있겠지.”라고	→ VP_CMP	말했다.

5.6. 명사-부사 통용어 또는 체언 수식 부사의 처리

- 명사-부사 통용어 중 뒤에 조사가 붙지 않고 서술어에 의존하는 경우(‘오늘’ 등) 또는 본래 부사이지만 체언을 수식하는 용법으로 쓰이는 경우에는(‘가장’, ‘아주’, ‘바로’ 등) 이들의 품사적 지위를 기준으로 하여 ‘AP’로 처리한다.

[예시] 친구네 집은 우리 집 바로 뒤에 있어,		
----------------------------	--	--

친구네	→ NP	집은
집은	→ NP_SBJ	있어.
우리	→ NP	집
집	→ NP	뒤에
바로	→ AP	뒤에
뒤에	→ NP_AJT	있어.
있어.	→ VP	ROOT

2. 구문 분석 말뭉치 구축 지침

이 장에서는 본 사업에서 구문 분석 말뭉치를 구축하기 위하여 적용한 세부 지침의 전체를 보이고자 한다. 지침의 구성은 다음과 같다.

1. 기본 원칙
2. 의존 관계 태그 세트 설정 방법
2.1. 구문 태그 세트
2.2. 기능 태그 세트
3. 문장 유형별 의존 관계 설정 방법
3.1. 홀문장(단문) 분석 방법
3.2. 겹문장(복문) 분석 방법
3.2.1. 명사절, 부사절 및 간접 인용절 내포문 분석 방법
3.2.1.1. 모문과 내포문의 주어 및 서술어가 각각 다른 경우
3.2.1.2. 모문과 내포문의 주어가 다르고, 서술어가 동일한 경우
3.2.1.3. 모문과 내포문의 주어가 같고, 서술어가 다른 경우
3.2.2. 관형절 내포문 분석 방법
3.2.2.1. 모문과 내포문의 주어 및 서술어가 다른 경우
3.2.2.2. 모문과 내포문의 주어가 다르고, 서술어가 동일한 경우
3.2.3. 내포문이 3개 이상 포함되어 중의적으로 해석되는 경우
3.2.4. 인용절 분석 방법
3.2.4.1. 간접 인용의 처리
3.2.4.2. 직접 인용의 처리
3.2.5. 이중 주어문 분석 방법
4. 세부 구별 태깅 가이드라인
4.1. [관형어+명사+명사] 유형 분석
4.2. 명사구 접속 유형 분석
4.3. [용언+용언] 유형 분석
4.3.1. 본용언+본용언
4.3.2. 본용언+보조 용언
4.3.3. 의존 명사 구성
4.3.4. ‘NP 중이다’ 구문

- 4.4. 부호
- 4.5. 외국 문자/외국어 처리 방법
- 4.6. 단독 어절로 구성된 부호 처리 방법
- 4.7. 띄어쓰기 오류 처리 방법
- 5. 세부 유형별 가이드라인
 - 5.1. 의존 관계 태그 부착 세부 유형 가이드라인
 - 5.1.1. 보조사적 쓰임을 보이는 ‘이/가’, ‘을/를’의 주석
 - 5.1.2. ‘~즈음’, ‘~쯤’ 부사구의 주석
 - 5.1.3. 격 조사가 붙은 수량 관련 표현의 주석
 - 5.1.4. 문장 첫머리에 나오는 ‘은/는’ 결합 어절의 주석
 - 5.1.5. 품사와 문장 성분이 일치하지 않는 경우
 - 5.2. 의존 관계 설정 세부 유형 가이드라인
 - 5.2.1. 서술어의 역할(~이다. 그리고)을 하는 ‘으로’의 주석
 - 5.2.2. 부사 ‘없이’, ‘같이’의 주석
 - 5.3. 연결된 부사구 세부 유형 가이드라인: ‘~부터 ~까지’, ‘~에서 ~으로’의 주석
 - 5.4. 장형 사동 구문 유형 세부 가이드라인
 - 5.5. 여러 개의 문장 처리
 - 5.6. 명사-부사 통용어 또는 체언 수식 부사의 처리
 - 5.7. 술어 생략 세부 유형 가이드라인

1. 기본 원칙

- (1) 자연 언어 처리를 위한 일관성 유지와 효율성 제고에 초점을 두되, 일반 언어학적 관점에 서도 크게 벗어나지 않도록 한다.
- (2) 문장의 표층 구조를 중시하여 분석한다.
- (3) 의존 관계 분석의 기본 단위로 어절을 사용한다.
- (4) 지배소 후위 원칙에 따라 각 어절의 지배소는 자신보다 뒤에 위치하도록 분석한다.
- (5) 각 어절은 1개의 지배소를 가진다.(Single-Head Constraint)
- (6) 각 어절 및 지배소 쌍은 서로 교차하지 않는다.(Projective Constraint)
- (7) 보어와 부가어를 구분하되 보어의 범위를 엄격히 제한한다.
 - 보어 CMP는 보격 조사가 부착된 NP, 용언구, 절, 그리고 인용절 보문의 용언구와 절에 한해서 분석한다.
 - 조사가 생략되거나 보조사 JX로 표시된 명사구 또는 이에 상응하는 용언구와 절도 서술어 구 문들에 따라 보격 조사로 대치 가능하면 CMP로 분석한다.
 - (가) 그가 돌아왔다고(VP_CMP) 그녀가 알려줬어.(TTA, 9쪽)
 - (나) 그녀가 그 일을 했다고(VP_CMP) 스스로 말했다.(TTA, 10쪽)
 - (다) 비평가 칼라일이 “인도와도 바꿀 수 없다.” 고(VP_CMP) 말하였다.(TTA, 13쪽)
 - (라) 물이 얼음이(NP_CMP) 되었다.
 - (마) 철수가 발이 아프다고(VP) 훈련을 빠졌다(VP).
 - (바) 마법사가 와인을 물로(NP_AJT) 바꾸었다.
 - (사) 철수가 영화가(NP_SBJ) 보고 싶다.
- (8) 원칙적으로 접속과 내포를 구별하지 않으며, 접속절은 모두 부사절로 분석한다.(다만, 명사구 접속은 인정한다.)
- (9) 하나의 성분이 모문과 내포문 모두에 관련되어 있으면 내포절의 유형에 따라 해당 주어의 지배소를 결정한다.

2. 의존 관계 태그 세트 설정 방법

- 각 어절은 자신 어절과 지배소 어절 사이의 관계를 표현하는 의존 관계 태그를 가진다.
- 의존 관계 태그는 아래 구문 태그와 기능 태그를 결합하여 사용한다.(예:NP_SBJ, VP_MOD 등)

2.1. 구문 태그 세트

구문 태그	의미	예시
NP	체언(명사, 대명사, 수사) 또는 외국 문자, 숫자를 비롯한 미등 제어.	-비가(NP) 와서 우산을(NP) 샀다. -공신의 딸을(NP) 부인으로(NP) 삼았다. -그는(NP) 셋을(NP) 세었다. - “닥쳐(Shut up)!” (NP)
VP	용언(동사, 형용사, 보조 용언)	-비가 와서(VP) 우산을 샀다(VP). -피부가 몹시 건조하다(VP).
AP	부사구	-전화 잠깐(AP) 써도 될까요? -피부가 몹시(AP) 건조하다.
VNP	긍정 지정사구(명사+이다)	-이게 우리집이야(VNP). -할머니는 걱정이셨다(VNP).
DP	관형사구	-벌써 새(DP) 학기가 되었다. -한(DP) 마흔(DP) 살쯤 되어 보인다.
IP	감탄사구(호칭 및 대답 등의 표현)	-아이고(IP) 이 일을 어쩔까. -선생님(IP) 질문이 있습니다.
X	의사 구(pseudo phrase, 조사 단독 어 절 또는 기호 등, 영어 등의 외국어)	
L	부호(왼쪽 괄호 및 따옴표)	
R	부호(오른쪽 괄호 및 따옴표)	

2.2. 기능 태그 세트

기능 태그	의미	예시
SBJ	주어	-비가(NP_SBJ) 와서 우산을 샀다. -그는(NP_SBJ) 셋을 세었다.
OBJ	목적어	-비가 와서 우산을(NP_OBJ) 샀다. -그는 셋을(NP_OBJ) 세었다.
MOD	관형어(체언 수식어)	-그(DP) 시계가 필요하다. -낮익은(VP_MOD) 자동차 한 대가 내려왔다. -어제 본(VP_MOD) 영화는 무척 재미있었다.
AJT	부사어(용언 수식어)	-나는 서울에(NP_AJT) 산다. -과일을 칼로(NP_AJT) 잘랐다.
CMP	보어	-나는 학생이(NP_CMP) 아니다. -얼음이 물이(NP_CMP) 되었다. -나는 싫다고(VP_CMP) 말했다.
CNJ	접속어(~와)	-철수와(NP_CNJ) 영희는 친구이다.

		-개와(NP_CNJ) 고양이의 관계.
--	--	----------------------

- 의존 관계 설정 및 의존 관계 태그 부착 결과는 아래와 같다.

[예시] 멜라닌은 사람의 피부색을 결정하는 중요한 요소이다.

멜라닌은	→NP_SBJ	요소이다.
사람의	→NP_MOD	피부색을
피부색을	→NP_OBJ	결정하는
결정하는	→VP_MOD	요소이다.
중요한	→VP_MOD	요소이다.
요소이다.	→VNP	ROOT

[예시] 역린은 용 목에 거꾸로 난 비늘을 의미한다.

역린은	→NP_SBJ	의미한다.
용	→NP	목에
목에	→NP_AJT	난
거꾸로	→AP	난
난	→VP_MOD	비늘을
비늘을	→NP_OBJ	의미한다.
의미한다.	→VP	ROOT

3. 문장 유형별 의존 관계 설정 방법

- 의존 관계 설정을 위한 문장 유형 구분은 일반 언어학 통사론의 기준을 따라 서술어가 한 개인 홀문장과 서술어가 2개 이상인 겹문장으로 문장을 분류하여 분석한다.
- 겹문장은 다시 접속문과 내포문으로 구분되지만, 기본 원칙 (8)에 따라 접속문을 따로 분류하지 않고, 접속문을 부사절 내포문으로 분류하여 분석한다.
- 내포문은 명사절, 관형절, 부사절, 인용절 내포문의 4가지 유형으로 분류하여 분석한다.

내포문의 유형별 예문은 아래와 같다.

- 명사절: 우리는 시민의 관전태도도 그만큼 성숙했음을 잊지 말아야 한다.
- 관형사절: 내가 좋아하는 꽃은 들국화다.(관계 관형절)
내가 사진을 좋아하는 사실을 친구들은 다 안다.(동격 관형절)
- 부사절: 멜라닌은 자외선을 차단해서 자외선으로부터 피부를 보호해 준다.
비가 와서 땅이 미끄럽다.
- 인용절: 그는 그녀가 그 일을 해냈다고 밝혔다.(간접 인용절)
그는 “그녀가 그 일을 해냈어.” 라고 말했다.(직접 인용절)

3.1. 홀문장(단문) 분석 방법

- 홀문장을 이루는 문장의 구성 성분은 크게 주어, 목적어, 관형어, 부사어, 보어, 서술어로 구분한다.

- (1) 주어는 SBJ 기능 태그를 가지고, 홀문장 서술어에 의존하도록 분석한다.
- (2) 목적어는 OBJ 기능 태그를 가지고, 홀문장 서술어에 의존하도록 분석한다.
- (3) 관형어는 MOD 기능 태그를 가지고, 수식하는 명사구에 의존하도록 분석한다.
- (4) 부사어는 AJT 기능 태그를 가지고, 홀문장 서술어에 의존하도록 분석한다.

- 단문 내에서 부사어로 기능하는 용언구의 경우 VP_AJT로 분석한다. 즉, ‘-게’ 부사형 어미가 나타난 어절이 보조 용언 구성(‘-게 하다’, ‘-게 되다’)이 아닌 경우 단문 구조로 보아 VP_AJT로 태깅한다. 이때 ‘-게’ 부사형 어미가 나타난 구성이 하나의 어절일 경우에만 (작게) VP_AJT로 처리하고, 여러 개의 어절로 구성된 절인 경우(눈에 띄게) VP로 처리한다.

[예시] 현주는 작게 한숨을 내쉬었다.

현주는	→ NP_SBJ	내쉬었다.
작게	→ VP_AJT	내쉬었다.
한숨을	→ NP_OBJ	내쉬었다.
내쉬었다.	→ VP	ROOT

[예시] 최근에는 눈에(> NP_AJT 띄게) 띄게(> VP 늘고) 늘고 이러한 추세가 가속화되는 경향이 있다

- ‘-게 되다’는 보조 용언 구성으로 처리하지 않고, 다른 ‘-게’ 부사절과 동일하게 ‘-게’ 성분이 단일 어절이면 VP_AJT, ‘-게’ 성분이 다어절이면 VP로 처리한다.

[예시] 저는 이런 자리에(→NP_AJT 참여하게) 참여하게(→ VP 되어서) 되어서 매우 영광입니다.

[예시] 결과가 이렇게(→VP_AJT 돼서) 돼서 참

[예시] 이번 쿠키는 정말(→AP 예쁘게) 예쁘게(→VP_AJT 됐어요~) 됐어요~

→ ‘정말’, ‘아주’ 등의 부사가 ‘-게’ 성분을 수식하는 경우는 ‘-게’ 성분을 VP_AJT 로 분석한다. 즉, ‘-게’ 성분에 SBJ, OBJ, AJT가 의존하는 경우에는 이를 부사절로 보아 VP 로 처리하고 부사가 의존하는 경우에는 부사어로 보아 VP_AJT로 처리한다.

(5) 보어는 CMP 기능 태그를 가지고, 홀문장 서술어에 의존하도록 분석한다.(보어의 범위는 기본원칙 (7)을 참조)

(6) 홀문장의 경우, 서술어는 문장의 가장 뒤에 위치하며, ROOT 어절에 의존하도록 분석한다.

- 홀문장의 구문 분석 예는 아래와 같다.

[예시] 멜라닌은 사람의 피부색을 결정한다.

멜라닌은	→NP_SBJ	결정한다.
사람의	→NP_MOD	피부색을
피부색을	→NP_OBJ	결정한다.
결정한다.	→VP	ROOT

3.2. 겹문장(복문) 분석 방법

- 겹문장은 명사절 내포문, 관형절 내포문, 부사절 내포문, 인용절 내포문의 4가지 유형으로 분류하여 분석한다.

3.2.1. 명사절, 부사절 및 간접 인용절 내포문 분석 방법

3.2.1.1. 모문과 내포문의 주어 및 서술어가 각각 다른 경우

- 모문의 주어는 모문의 서술어로, 내포문의 주어는 내포문의 서술어로 연결한다.
- 내포문의 서술어는 수식하는 모문의 해당 어절에 의존하도록 분석한다.
- 명사절, 부사절 및 간접 인용절별 예문은 아래와 같다.

[예시] 우리는 시민의 관전태도도 그만큼 성숙했음을 잊지 말아야 한다.

우리는	→NP_SBJ	잊지
시민의	→NP_MOD	관전태도도
관전태도도	→NP_SBJ	성숙했음을
그만큼	→AP	성숙했음을
성숙했음을	→VP_OBJ	잊지
잊지	→VP	말아야
말아야	→VP	한다.
한다.	→VP	ROOT

[예시] 그가 돌아와줘서 우리는 정말 고마웠어.

그가	→NP_SBJ	돌아와줘서
돌아와줘서	→VP	고마웠어.
우리는	→NP_SBJ	고마웠어.
정말	→AP	고마웠어.
고마웠어.	→VP	ROOT

[예시] 그가 돌아왔다고 그녀가 알려줬어.

그가	→NP_SBJ	돌아왔다고
돌아왔다고	→VP_CMP	알려줬어.
그녀가	→NP_SBJ	알려줬어.
알려줬어.	→VP	ROOT

3.2.1.2. 모문과 내포문의 주어가 다르고, 서술어가 동일한 경우

- 모문과 내포문의 주어가 다르고 서술어가 동일하며 내포문의 서술어가 생략된 경우, 내포문의 성분은 내포문의 가장 마지막 어절에 연결하고, 내포문의 가장 마지막 성분을 모문의 서술어에 연결한다.

[예시] 나는 과자를, 동생은 빵을 먹었다.

나는	→NP_SBJ	과자를
과자를	→VP_OBJ	먹었다.
동생은	→NP_SBJ	먹었다.
빵을	→VP_OBJ	먹었다.
먹었다.	→VP	ROOT

[예시] 티셔츠는 2002년 1500만장, 2006년 1000만장이나 팔려나갔다.

티셔츠는	→ NP_SBJ	1500만장,
2002년	→ NP_AJT	1500만장,
1500만장,	→ NP_SBJ	팔려나갔다.
2006년	→ NP_AJT	팔려나갔다.
1000만장이나	→ NP_SBJ	팔려나갔다.
팔려나갔다.	→ VP	ROOT

- 명사구 접속 구조와 혼동되는 경우가 있는데, 비슷한 구조에서 ‘와/과’로 이어진 접속 명사구만 ‘CNJ’ (명사구 접속)로 처리함에 유의한다.

[예시] 2006년 연구와 2008년 연구에서는 연구팀의 주장이 유지되었다.

2006년	→ NP	연구와
연구와	→ NP_CNJ	연구에서는
2008년	→ NP	연구에서는
연구에서는	→ NP_AJT	유지되었다.
연구팀의	→ NP_MOD	주장이
주장이	→ NP_SBJ	유지되었다.
유지되었다.	→ VP	ROOT

3.2.1.3. 모문과 내포문의 주어가 같고, 서술어가 다른 경우

- 기본 원칙 (9)에 따라 모문과 내포문의 관계에 따라 주어의 지배소를 결정한다.
- 명사절, 부사절의 경우, 주어가 내포문의 서술어를 지배소로 가지도록 분석한다.
- 내포문의 서술어는 수식하는 모문의 어절에 의존하도록 분석한다.
- 명사절, 부사절 예문은 아래와 같다.

[예시] 그는 목표를 이루었음에 매우 감사했다.

그는	→NP_SBJ	이루었음에
목표를	→NP_OBJ	이루었음에
이루었음에	→VP_AJT	감사했다.
매우	→AP	감사했다.
감사했다.	→VP	ROOT

[예시] 멜라닌은 자외선을 차단해서 자외선으로부터 피부를 보호해 준다.

멜라닌은	→NP_SBJ	차단해서
자외선을	→NP_OBJ	차단해서
차단해서	→VP	보호해
자외선으로부터	→NP_AJT	보호해
피부를	→NP_OBJ	보호해
보호해	→VP	준다.
준다.	→VP	ROOT

- ‘~다”며’의 경우 ‘~다고 하며’가 줄어든 꼴로 보아 부사절과 동일하게 처리한다. 즉, 모문의 주어는 ‘~다”며’ 어절에 의존하고, ‘~다”며’ 어절은 모문의 서술어에 VP로 의존한다.

[예시] 위원장은 “아직 준비가 미흡하다”며 “빠른 시일 안에 준비를 마치겠다”고 강조했다.

위원장은	→ NP_SBJ	미흡하다”며
“아직	→ AP	미흡하다”며
준비가	→ NP_SBJ	미흡하다”며
미흡하다”며	→ VP	강조했다
“빠른	→ VP_MOD	시일
시일	→ NP	안에
안에	→ NP_AJT	마치겠다”고
준비를	→ NP_OBJ	마치겠다”고
마치겠다”고	→ VP_CMP	강조했다.
강조했다.	→ VP	ROOT

- 특히, ‘대하여/대해’, ‘통하여/통해’, ‘관하여/관해’, ‘위하여/위해’에 선행하는 주어 성분은 이들 서술어에 의존하는 것으로 분석하여야 함에 유의해야 한다.(그러나 교차 금지 제약에 해당하는 경우 모문 서술어에 의존할 수 있음.)

[예시] 김 의원은(→ 통해) 15일 시정 질의를 통해 “어려운 경제여건과 1000억 원 이상 드는 야구장 신축이 어렵다면 기존 무등경기장 야구장을 리모델링하는 방안을 검토해야 한다”고 주장했다.

3.2.2. 관형절 내포문 분석 방법

3.2.2.1. 모문과 내포문의 주어 및 서술어가 다른 경우

- 명사절, 부사절 및 간접 인용절 내포문 분석 방법과 동일하게 분석한다.
- 모문의 주어는 모문의 서술어로, 내포문의 주어는 내포문의 서술어로 연결한다.
- 내포문의 서술어는 수식하는 모문의 어절에 의존하도록 분석한다.

[예시] 내가 좋아하는 꽃은 들국화이다.

내가	→NP_SBJ	좋아하는
좋아하는	→VP_MOD	꽃은
꽃은	→NP_SBJ	들국화이다.
들국화이다.	→VNP	ROOT

[예시] 내가 사진을 좋아하는 사실을 친구들은 다 안다.

내가	→NP_SBJ	좋아하는
사진을	→NP_OBJ	좋아하는
좋아하는	→VP_MOD	사실을
사실을	→NP_OBJ	안다.
친구들은	→NP_SBJ	안다.
다	→AP	안다.
안다.	→VP	ROOT

3.2.2.2. 모문과 내포문의 주어가 같고, 서술어가 다른 경우

- 관형절은 모문 내의 체언을 수식하는 성격이 더욱 강하기 때문에, 동일 주어를 모문의 서술어로 연결하도록 한다.
- 일반적으로 종차 개념에 의한 정의 구문 유형([A는 ~~한 B이다.])이 이 유형에 속한다.
- 예문은 아래와 같다.

[예시] 세포벽은 식물 세포의 가장 바깥층을 에워싸고 있는 약간 두꺼운 막이다.

세포벽은	→NP_SBJ	막이다.
식물	→NP	세포의
세포의	→NP_MOD	바깥층을
가장	→AP	바깥층을
바깥층을	→NP_OBJ	에워싸고
에워싸고	→VP	있는
있는	→VP_MOD	막이다.
약간	→AP	두꺼운
두꺼운	→VP_MOD	막이다.
막이다.	→VNP	ROOT

[예시] 멜라닌은 사람의 피부색을 결정하는 주요 요소이다.

멜라닌은	→NP_SBJ	요소이다.
사람의	→NP_MOD	피부색을
피부색을	→NP_OBJ	결정하는

결정하는	→VP_MOD	요소이다.
주요	→NP	요소이다.
요소이다.	→VNP	ROOT
[예시] 이 사료는 가축들이 먹는 음식이다.		
이	→DP	사료는
사료는	→NP_SBJ	음식이다.
가축들이	→NP_SBJ	먹는
먹는	→VP_MOD	음식이다.
음식이다.	→VNP	ROOT

3.2.3. 내포문이 3개 이상 포함되어 중의적으로 해석되는 경우

- 문장 좌측에서 우측 방향으로 순서대로 의존 관계를 설정하여 분석한다.
- 복문의 해석이 중의적일 때에는 가능한 의미 중에서 가장 가까이에 후행하는 서술어에 의존한다.

[예시] 김민이 돈을 잘 써서 멋진 선배로 통한다.		
김민이	→ NP_SBJ	써서
돈을	→ NP_OBJ	써서
잘	→ AP	써서
써서	→ VP	멋진
멋진	→ VP_MOD	선배로
선배로	→ NP_AJT	통한다.
통한다.	→ VP	ROOT

- 복문의 해석이 단일한 때에는 해당 해석에 따라 구조를 분석한다.

[예시] 김민은 신작에서 판타지의 비중은 줄이고 사회를 비판한 내용을 부각했다.		
줄이고	→ VP	부각했다.

3.2.4. 인용절 분석 방법

- 기본적으로 문장 부호에 의하여 구분된 단위를 준수하여 분석한다.
- 인용 부호가 있는 직접 인용문은 모문과 내포문을 인용 부호에 의해서 구분할 수 있으므로, 모문은 모문 구조 내의 주어와 서술어 간의 의존 관계를 연결하고 내포문은 내포문 내의 주어와 서술어 간의 의존 관계를 연결하여 분석한다.
- 보어의 범위를 엄격하게 제한하고 있기 때문에(기본 원칙 (7) 참조) 기능 태그로 CMP를 부여하는 인용절의 범위에 유의하여야 한다.
- 인용 술어(모문 서술어)가 <표준국어대사전>에서 ‘-고’ 성분을 격틀로 제시하고 있는 경우: 기능 태그로 CMP를 부여함.
- 인용 술어가 <표준국어대사전>에서 ‘-고’ 성분을 격틀로 가지고 있지 않은 경우: 기능 태그로 AJT를 부여함.
- ‘-다고’ 류 절이 이유를 나타내는 연결 어미인 경우에는(서로 잘 아는 친구 사이라고 무례하게 대하는 안 된다.) 인용절이 아닌 일반 부사절로 처리해야 함에 유의한다. 또한 흔히 속담과 같은 관용구

를 인용하는 경우에도(가지 많은 나무 바람 잘 날 없다고 우리 부모님 마음 편할 날이 없으셨지.) 일반 부사절로 처리한다.

3.2.4.1. 간접 인용의 처리

- 간접 인용은 따옴표 없이 ‘~다고/냐고/자고/라고’ 로 인용된 것으로, 부사절과 동일하게 처리하되 인용 술어(모문의 서술어)가 <표준국어대사전>에서 격틀 정보로 ‘-고’ 를 가지고 있을 경우에만 기능 태그로 CMP를 부여하고, 격틀 정보로 ‘-고’ 를 가지고 있지 않은 경우에는 AJT를 부여한다.
- 예문은 아래와 같다.

[예시] 그녀가 그 일을 했다고 스스로 말했다.

그녀가	→NP_SBJ	했다고
그	→DP	일을
일을	→NP_OBJ	했다고
했다고	→VP_CMP	말했다.(‘말하다’의 격틀에 ‘-고’ 있음)
스스로	→AP	말했다.
말했다.	→VP	ROOT

[예시] 형사가 용의자에게 사건 시각에 어디에 있었느냐고 신문하고 있었다.

형사가	→ NP_SBJ	신문하고
용의자에게	→ NP_AJT	신문하고
사건	→ NP	시각에
시각에	→ NP_AJT	있었느냐고
어디에	→ NP_AJT	있었느냐고
있었느냐고	→ VP_AJT	신문하고(‘신문하다’의 격틀에 ‘-고’ 없음)
신문하고	→ VP	있었다.
있었다.	→ VP	ROOT

3.2.4.2. 직접 인용의 처리

- 직접 인용은 따옴표를 사용하여 뒤에 ‘이라고/라고/고’ 등이 결합한 것으로, 따옴표 밖에 있는 주어는 따옴표 안의 서술어의 주어와 동일하더라도 모문의 서술어에 연결한다.
- 즉, 직접 인용에서 모문의 주어와 내포문의 주어가 일치하는 경우 주어를 모문의 서술어에 연결하여 분석한다.
- 예문은 아래와 같다.

[예시] 비평가 칼라일이 “인도와도 바꿀 수 없다”고 말하였다.

비평가	→NP	칼라일이
칼라일이	→NP_SBJ	말하였다.
“인도와도	→NP_AJT	바꿀
바꿀	→VP_MOD	수
수	→NP_SBJ	없다.” 고
없다.” 고	→VP_CMP	말하였다.(‘말하다’ 격틀에 ‘-고’ 있음)
말하였다.	→VP	ROOT

[예시] 그는 “불이야!” 라고 소리쳤다.

그는	→ NP_SBJ	소리쳤다.
“불이야!” 라고	→ VNP_AJT	소리쳤다.(‘소리치다’ 격틀에 ‘-고’ 없음)
소리쳤다.	→ VP	ROOT

- 또한, 인용된 부분이 명사로 끝나는 경우 VNP가 아닌 NP로 처리해야 함에 유의해야 한다.

[예시] 그는 “○○○는 그룹 내에서 가장 중요한 사업장 중 하나” 라며 “제조 역량이 뛰어나고 개발과 디자인 역량도 충분히 갖추고 있어 그룹의 미래에도 큰 영향을 미칠 것” 이라고(NP_CMP) 평가했다.

- 또한 ‘~다’ 며’ 로 인용된 절은 ‘~다고 하며’ 가 줄어든 것으로 보고 일반적인 부사절과 동일한 원칙으로 처리한다. (3.2.1.3. 참조)

[예시] 앞서 서 판사는 “도망칠 우려가 있다” 며 지난달 11일 구속영장을 발부했다.

앞서	→ AP	발부했다.
서	→ NP	판사는
판사는	→ NP_SBJ	있다” 며
“도망칠	→ VP_MOD	우려가
우려가	→ NP_SBJ	있다” 며
있다” 며	→ VP	발부했다.
지난달	→ NP	11일
11일	→ NP_AJT	발부했다.
구속영장을	→ NP_OBJ	발부했다.
발부했다.	→ VP	ROOT

3.2.5. 이중 주어문 분석 방법

- 이중 주어문은 각 구성 성분별로 동일한 서술어에 의존하도록 분석한다.
- 이때 일반적인 이중 주어문뿐만 아니라 [NP1이 NP2가 VP]의 격틀 구조를 가지는 일부 용언 분석에도 동일하게 적용한다.
- 예문은 아래와 같다.

[예시] 나는 그 시계가 필요했다.

나는	→NP_SBJ	필요했다.
그	→DP	시계가
시계가	→NP_SBJ	필요했다.
필요했다.	→VP	ROOT

[예시] 그는 그녀가 자랑스러웠다.

그는	→NP_SBJ	자랑스러웠다.
그녀가	→NP_SBJ	자랑스러웠다.
자랑스러웠다.	→VP	ROOT

4. 세부 구별 태깅 가이드라인

4.1. [관형어+명사+명사] 유형 분석

- 관형어가 명사구를 수식할 때, 명사구 중 관형어의 지배소는 관형어가 의미적으로 수식하는 어휘를 지배소로 분석한다.
- 여러 개의 명사들이 특별한 수식 관계 없이 나열되어 있는 경우에는, 핵어 명사를 제외한 수식 명사들을 모두 기능 표지 없이 NP로 처리하고, 오른쪽에 있는 명사로 연결시킨다.
- 예문은 아래와 같다.

[예시] 무분별한 포획 문제

무분별한	→VP_MOD	포획
포획	→NP	문제

[예시] 고려의 충신 정몽주

고려의	→NP_MOD	충신
충신	→NP	정몽주

[예시] 조선의 제3대 임금

조선의	→NP_MOD	임금
제3대	→NP	임금

[예시] 변형생성문법은 1987년에 출판된 노엄 촘스키의 저서이다.

변형생성문법은	→NP_SBJ	저서이다.
1987년에	→NP_AJT	출판된
출판된	→VP_MOD	저서이다.
노엄	→NP	촘스키의
촘스키의	→NP_MOD	저서이다.
저서이다.	→VNP	ROOT

[예시] 결명자의 한자 뜻인 ‘눈을 밝게 띄우는 씨앗’이라는 이름대로

결명자의	→NP_MOD	뜻인
한자	→NP	뜻인
뜻인	→VNP_MOD	씨앗’이라는
‘눈을	→NP_OBJ	띄우는
밝게	→VP_AJT	띄우는
띄우는	→VP_MOD	씨앗’이라는
씨앗’이라는	→NP_MOD	이름대로

[예시] 레오나르도 다빈치는 역사상 가장 위대한 천재 중 하나로 기억된다.

레오나르도	→NP	다빈치는
다빈치는	→NP_SBJ	기억된다.
역사상	→NP_AJT	위대한
가장	→AP	위대한
위대한	→VP_MOD	천재
천재	→NP	중
중	→NP	하나로

하나로	→NP_AJT	기억된다.
기억된다.	→VP	ROOT

- 여러 개의 명사로 이루어진 명사구의 내부에 수식 구조가 존재하는 경우에는, 각 명사구의 수식 구조를 고려하여 의존 관계를 설정한다.

[예시] 프랑스 파리 루브르 박물관

프랑스	→ NP	파리
파리	→ NP	박물관
루브르	→ NP	박물관
박물관	→ NP	ROOT

[예시] ○○○ ○○당 대표가 정계에 복귀했다.

○○○	→ NP	대표가
○○당	→ NP	대표가
대표가	→ NP_SBJ	복귀했다.
정계에	→ NP_AJT	복귀했다.
복귀했다.	→ VP	ROOT

- 명사구와 수량사구가 연달아 나타나는 경우, 명사구 앞의 수식어는 명사구에 연결하고 명사구가 수량사구에 의존하는 것으로 처리한다.

[예시] 보유하고 있는 소방헬기는 1995년에 구입한(→VP_MOD 8인승) 8인승 1대뿐

4.2. 명사구 접속 유형 분석

- 복수 개의 명사구가 접속 또는 나열된 경우, 가장 마지막 명사구에 의존하도록 분석한다.
- 이때, 접속 조사로 인정하는 것은 <표준국어대사전> 기준으로 12개이다. (고24/이고5, 과12/와3, 나9/이나2, 니1/이니, 다4/이다4, 량4/이랑2, 며1/이며, 면9/이면3, 예4, 이랴2, 하고5, 하며) 이들 조사가 아닌 경우(‘(이)라든지’, ‘(이)라든가’ 등)는 접속으로 보지 않는다.
- ‘및’, ‘또는’, ‘그리고’ 등에 의해 명사구가 접속 또는 나열된 경우, 이들 접속 부사는 후행하는 명사구에 의존하는 것으로 분석한다. 그리고 이때 구문 표지만 부착하고, 기능 표지는 부착하지 않는다.
- 명사구 접속의 예는 아래와 같다.

[예시] 비단 피부뿐만이 아니라 털, 눈, 귀, 심지어 뇌에도 존재한다.

비단	→ AP	아니라
피부뿐만이	→ NP_CMP	아니라
아니라	→ VP	존재한다.
털,	→ NP_CNJ	뇌에도
눈,	→ NP_CNJ	뇌에도
귀,	→ NP_CNJ	뇌에도

심지어	→ AP	뇌에도
뇌에도	→ NP_AJT	존재한다.
존재한다.	→ VP	ROOT
[예시] 매리너스 협곡과 극관이 존재한다.		
매리너스	→ NP	협곡과
협곡과	→ NP_CNJ	극관이
극관이	→ NP_SBJ	존재한다.
존재한다.	→ VP	ROOT
[예시] 주로 왕과 왕세자의 강론 및 정책토론을 주관했다.		
주로	→ AP	주관했다.
왕과	→ NP_CNJ	왕세자의
왕세자의	→ NP_MOD	정책토론을
강론	→ NP_CNJ	정책토론을
및	→ AP	정책토론을
정책토론을	→ NP_OBJ	주관했다.
주관했다.	→ VP	ROOT

- 조사 ‘와/과’로 나타나는 명사구는 ‘N과 N’ 순서인 경우에 명사구 접속으로, ‘N이 N과’와 같은 순서인 경우에는 부사어로 처리한다.

[예시] 철수와 영희가 만났다.		
철수와	→ NP_CNJ	영희가
영희가	→ NP_SBJ	만났다.
만났다.	→ VP	ROOT
[예시] 철수가 영희와 헤어졌다.		
철수가	→ NP_SBJ	헤어졌다.
영희와	→ NP_AJT	헤어졌다.
헤어졌다.	→ VP	ROOT

- 조사 ‘(이)라든가’, ‘(이)라든지’는 나열 기능을 하는 조사로, ‘명사구+(이)라든가/라든지 명사구+(이)라든가/라든지 하는’ 등의 구성으로 자주 나타난다.
- 이때의 ‘명사구+(이)라든가/라든지’ 성분은 각각 NP_AJT로 서술어에 연결하도록 한다.

[예시] 그는 돈이라든지(>NP_AJT 하는) 명예라든지(>NP_AJT 하는) 하는 것에 연연해 하지 않는다.		
---	--	--

4.3. [용언+용언] 유형 분석

4.3.1. 본용언+본용언

- 본용언이 연속적으로 나타날 경우, 주어를 앞에 위치한 서술어에 연결한다.

[예시] 해구보다는 폭이 넓고 얇다.		
해구보다는	→ NP_AJT	넓고
폭이	→ NP_SBJ	넓고

넓고	→ VP	알다.
알다.	→ VP	ROOT

4.3.2. 본용언+보조 용언

- 본용언과 보조 용언이 연속하여 두 개 이상 나올 때는 주어를 본용언에 연결하고 본용언은 보조 용언에 연결한다.
- 주어 외에도 의존소들을 연결하고 있는 필수 성분들이 일차적으로 본용언에 연결되고, 본용언이 보조 용언에 의존하는 방식으로 처리한다.
- 본용언과 보조 용언이 붙여쓰기로 제시되어 있는 경우에는 해당 형식을 하나의 용언으로 처리한다.
- 보조 용언 구성이 두 개 이상 연속될 때에도 마찬가지로 본용언 → 보조 용언1 → 보조 용언2의 순으로 연결한다.

[예시] 멜라닌은 물에는 용해되지 않는다.

멜라닌은	→ NP_SBJ	용해되지
물에는	→ NP_AJT	용해되지
용해되지	→ VP	않는다.
않는다.	→ VP	ROOT

[예시] 그들의 화려함 속에 감춰진 갈등과 속내가 공개돼 화제를 모으고 있다.

그들의	→ NP_MOD	속내가
화려함	→ VP	속에
속에	→ NP_AJT	감춰진
감춰진	→ VP_MOD	속내가
갈등과	→ NP_CNJ	속내가
속내가	→ NP_SBJ	공개돼
공개돼	→ VP	있다.
화제를	→ NP_OBJ	모으고
모으고	→ VP	있다.
있다.	→ VP	ROOT

- 문장 부사 또한 본용언이 아닌 보조 용언에 연결한다. 의사 보조 용언의 경우에도 동일하게 적용한다.

[예시] 그러나(→AP 있었다.) 이번 분기는 생각보다 순조롭게 진행되고 있었다.

4.3.3. 의존 명사 구성(의사 보조 용언 구성)

- 주 서술어 다음에 보조 용언은 아니지만 서법을 나타내는 의존 명사가 포함된 구성이 오는 경우, 해당 서술어 및 의존 명사를 별개의 단위로 처리하여 분석하고 의존 관계를 연결한다.
- 의존 명사에 ‘이다’, ‘하다’ 등이 결합해 있는 경우에는 그 자체를 각각 VNP, VP 등으로 처리한다.
- 이때 문장의 주어와 의존 명사의 관계는 해당 의존 명사가 실질 명사로 대체 가능한 경우와

불가능한 경우로 구분하여 분석한다. 의존 명사가 주어와 공지시(co-reference)되는 경우에는 주어와 의존 명사 어절을 연결하고, 그렇지 않은 경우에는 주어와 서술어를 연결한다.

- 표면적으로는 ‘-ㄴ 것이다’ 와 같이 의사 보조 용언 구성에 해당되더라도, 의존 명사나 명사가 서법을 나타내지 않는 경우는 의사 보조 용언 구성에 해당되지 않는다.

[예시] 그는 일어날 수 없었다.

그는	→ NP_SBJ	일어날
일어날	→ VP_MOD	수
수	→ NP_SBJ	없었다.
없었다.	→ VP	ROOT

[예시] 나는 곧 밥을 먹을 것이다.

나는	→ NP_SBJ	먹을
곧	→ AP	먹을
밥을	→ NP_OBJ	먹을
먹을	→ VP_MOD	것이다.
것이다.	→ VNP	ROOT

[예시] 이 사료는 가축들이 먹는 것이다.

이	→ DP	사료는
사료는	→ NP_SBJ	것이다.
가축들이	→ NP_SBJ	먹는
먹는	→ VP_MOD	것이다.
것이다.	→ VNP	ROOT

→ 이때의 ‘것’ 은 ‘사료, 식품’ 을 뜻하므로 ‘사료는’ 이 ‘것이다’ 에 의존함.

[예시] 내가 놀란 것은 철수가 천재라는 것이다.

내가	→ NP_SBJ	놀란
놀란	→ VP_MOD	것은
것은	→ NP_SBJ	것이다.(것 = 사실 → 사실 = 사실)
철수가	→ NP_SBJ	천재라는
천재라는	→ VNP_MOD	것이다.
것이다.	→ VNP	ROOT

→ 이때의 ‘것’ 은 ‘사실’ 을 뜻하므로 ‘것은’ 이 ‘것이다’ 에 의존하는 것으로 처리함.

[예시] 곧 비가 올 것 같다.

곧	→ AP	올
비가	→ NP_SBJ	올
올	→ VP_MOD	것
것	→ NP	같다.
같다.	→ VP	ROOT

4.3.4. ‘NP 중이다’ 구문

- ‘NP 중이다’ 구문에서 NP의 논항이 되는 성분은 ‘중이다’ 가 아닌 해당 NP에 의존한다.

[예시] 구치소는 사건경위를 조사한 뒤 손씨에게 규율 위반 행위에 상응하는 징벌을 내리는 방안을(→ NP_OBJ 검토) 검토 중이다.

- 따옴표 (‘ ’ / “ ”) 및 괄호 (< > / []) 등과 같이 좌우 짝이 있는 부호에 한정하여 각각 L, R 표지를 부착하고, 그 외의 경우에는 일괄적으로 X를 부착한다(괄호가 수리 기호나 층위 표시로 쓰일 때에도 X로 분석함).
- L은 R에 의존하도록 분석한다.

[예시] “ 나는 너를 좋아해 ” 라고 말했다.

“ → L ”
 ” → R 라고

[예시] “나는 너를 좋아해” 라고 말했다.

“ → L 좋아해 ” 라고
 좋아해 ” 라고 → VP_CMP 말했다.

[예시] 과일(사과, 배 등)의 등급은

과일(사과,	→ NP_CNJ	배
배	→ NP	등
등	→ NP)의
)의	→ X_MOD	등급은

[예시] 과일 (사과, 배 등)의 등급은

과일	→ NP)의
(사과,	→ NP_CNJ	배
배	→ NP	등
등	→ NP)의
)의	→ X_MOD	등급은

[예시] 과일 (사과, 배 등) 의 등급은

과일	→ NP	의
(→ L)
사과,	→ NP_CNJ	배
배	→ NP	등
등	→ NP)
)	→ R	의
의	→ X_MOD	등급은

- 단락 기호는 기호가 속하는 최상위 행에 의존하도록 분석한다.

[예시] 철수의 버릇 : 다리 꼬기, 이갈기

$$\begin{array}{lll} \text{버릇} & \rightarrow \text{NP} & : \\ : & \rightarrow X & \text{이갈기} \end{array}$$

[예시] 철수의 버릇: 다리 꼬기, 이갈기

버릇:	→ NP	이갈기
꼬기,	→ NP_CNJ	이갈기

[예시] - 철수의 버릇 : 다리 꼬기, 이갈기

-	→ X	이갈기
버릇	→ NP	:
:	→ X	이갈기
[예시] ● 내년 주요 핵심안은 예산 결의 문제		
●	→ X	문제

- 그 외: 후행 어절에 의존하도록 분석한다.

[예시] 세종(1418 ~ 1450)은 조선전기 제4대 왕이다.		
세종(1418	→ NP	~
~	→ X	1450)은
1450)은	→ NP_SBJ	왕이다.

- 삽입구가 복수의 어절로 구성되어 있을 경우, 기호로 결합된 복합 형태소는 선행 성분을 기준으로 구문 태그를 결정하고, 후행 성분을 기준으로 기능 태그를 결정한다.

[예시] 전문위원의 임기는 3년을 보장한다(1차에 한하여 연임 가능).		
보장한다(1차에	→ VP_AJT	한하여

4.5. 외국 문자/외국어 처리 방법

- 외국 문자, 숫자를 비롯한 기능을 알 수 없는 미등재어의 구문 태그는 NP이다.

[예시] “닥쳐(Shut up)!”		
“닥쳐(Shut	→ VP	up)!”
up)!	→ NP	ROOT
[예시] 아이 러브 유		
아이	→ NP	러브
러브	→ NP	유
[예시] I love you		
I	→ NP	love
love	→ NP	you

4.6. 띄어쓰기 오류 처리 방법

- 어절 내부가 분할되어 있는 경우, 구문 태그와 기능 태그는 형태 분석 결과를 기준으로 정하고, 의존 관계는 바른 띄어쓰기를 기준으로 정한다. (* 19 국립국어원 형태 분석 말뭉치 참조)
- 절단 어절의 형태 분석 결과와 구문 분석 태그의 대응표는 다음과 같다.

형태 분석 태그	구문 분석 태그
NNG, NNP, NNB, NP, NR, XSN, XR, NF	NP
VV, VA, VX, VCN, EP, EF, EC, ETN, XSV, XSA, NV	VP
MMA, MMD, MMS	DP
MAG, MAJ	AP
IC	IP
JKS, JKC, JKG, JKO, JKB, JKV, JKQ	X
SF, SP, SS, SE, SO, SW, SL, SH, NA	NP

- 기능 태그는 맨 마지막 분할 어절에 부여한다. (조사 생략 경우도 마찬가지임)
- 원래 어절 내부에서는 다음 분할 어절에 의존하고, 원래 어절의 분할 어절은 지배소 어절(의 최종 분할 어절)에 의존한다.

[예시] 마음 씨 가 중요하 다.

마음	→ NP	씨
씨	→ NP	가
가	→ X_SBJ	다.
중요하	→ VP	다.
다	→ VP	ROOT

5. 세부 유형별 가이드라인

5.1. 의존 관계 태그 부착 세부 유형 가이드라인

5.1.1. 보조사적 쓰임을 보이는 ‘이/가’, ‘을/를’의 주석

- 본용언에 화용적 기능을 가지는 조사 {가/를}이 붙은 경우는 기능 표지를 부착하지 않는다. 즉, 명사형(-음, -기)을 제외한 용언의 활용형에 붙은 조사 {가/를}은 무시한다.

[예시] 철수가 밥을 이틀내 먹지를 않았다.

철수가	→ NP_SBJ	먹지를
밥을	→ NP_OBJ	먹지를
이틀내	→ NP_AJT	먹지를
먹지를	→ VP	않았다.
않았다.	→ VP	ROOT

[예시] 그 산은 그리 높지가 않다.

그	→ DP	산은
산은	→ NP_SBJ	높지가
그리	→ AP	높지가
높지가	→ VP	않다.
않다.	→ VP	ROOT

- ‘-기 바라다’, ‘-기 시작하다’ 등의 ‘-기’ 절을 요구하는 서술어의 경우 뒤에 격 조사가 붙는 경우와 붙지 않는 경우 모두 기능 태그를 부착하여 VP_OBJ로 태깅한다.

[예시] 나는 네가 빨리 오기(를) 바란다.

나는	→ NP_SBJ	바란다.
네가	→ NP_SBJ	오기(를)
빨리	→ AP	오기(를)
오기(를)	→ VP_OBJ	바란다.
바란다.	→ VP	ROOT

5.1.2. ‘~즈음’, ‘~쯤’, ‘정도’ 부사구의 주석

- “~즈음”, “~쯤” 등과 같이 의미적으로 시간과 공간을 의미하고 조사가 없는 경우 AJT 기능 태그를 부착한다.

[예시] 광복절 즈음 해서 독립기념관을 찾았다.

광복절	→ NP	즈음
즈음	→ NP_AJT	해서
해서	→ VP	찾았다.
독립기념관을	→ NP_OBJ	찾았다.
찾았다.	→ VP	ROOT

[예시] 다섯 명 즈음의 학생이 길에 서 있었다.

다섯	→ DP	명
----	------	---

명	→ NP	즈음의
즈음의	→ NP_MOD	학생이
학생이	→ NP_SBJ	서

5.1.3. 격 조사가 붙은 수량 관련 표현의 주석

- 시간이나 거리 등 수량이나 단위를 나타내는 명사구에 목적격 조사 ‘을/를’ 이 붙은 경우에는 목적어로 분석한다.

[예시] 나는 학교까지 다섯 시간을 걸었다.

나는	→ NP_SBJ	걸었다.
학교까지	→ NP_AJT	걸었다.
다섯	→ DP	시간을
시간을	→ NP_OBJ	걸었다.
걸었다.	→ VP	ROOT

- ‘을/를’ 이 결합하지 않았더라도 ‘을/를’ 이외의 다른 격조사가 결합할 수 없는 개체/수량/횟수/시간/거리 표현의 경우 OBJ로 분석한다.

[예시] 노동자들의 삶을 담기 위해 2년에 걸쳐 5번(→ NP_OBJ 방문했다.) 방문했다.

[예시] 운동장을 세 바퀴(→NP_OBJ 뛰었다.) 뛰었다.

5.1.4. 품사와 문장 성분이 일치하지 않는 경우

- 품사로써는 부사나 활용형 등이 쓰였지만 문장 내에서 인용이 된 것처럼 쓰인 경우에는 해당 문장 성분을 기준으로 기능 태그를 부여하되, 구문 태그 분석은 단어의 본래 품사를 기준으로 한다. 다만, 형태소의 일부가 잘린 경우 미등재어로 보아 NP를 부여한다.

[예시] ‘거꾸로’ 는 ‘거꾸’ 와 ‘로’ 로 분석할 수 있을까?

‘거꾸로’ 는	→ AP_OBJ	분석할
‘거꾸’ 와	→ NP_CNJ	‘로’ 로
‘로’ 로	→ NP_AJT	분석할
분석할	→ VP_MOD	수
수	→ NP_SBJ	있을까?
있을까?	→ VP	ROOT

5.2. 의존 관계 설정 세부 유형 가이드라인

5.2.1. 서술어의 역할(~이다. 그리고)을 하는 ‘으로’ 의 주석

- “~으로” 가 의미적으로 “~이다. 그리고” 와 같이 사용되었다면, 예외적으로 서술어로 인정한다. 즉, 주어 논항을 가질 수 있다.

[예시] 모나리자는 레오나르도 다빈치가 그린 초상화로, 현재 프랑스 파리 루브르 박물관에 전시되어 있다.

모나리자는	→ NP_SBJ	초상화로,
레오나르도	→ NP	다빈치가
다빈치가	→ NP_SBJ	그린
그린	→ VP_MOD	초상화로,
초상화로,	→ NP_AJT	있다.
현재	→ AP	전시되어
프랑스	→ NP	파리
파리	→ NP	박물관에
루브르	→ NP	박물관에
박물관에	→ NP_AJT	전시되어
전시되어	→ VP	있다.
있다.	→ VP	ROOT

- “~으로”가 쉼표로 연결되어 있지 않아도 의미적으로 “~이다. 그리고”에 대응된다면 마찬가지로 서술어로 인정한다.

[예시] 모나리자는 레오나르도 다 빈치가 그린 초상화로 현재 프랑스 파리 루브르 박물관에 전시되어 있다.		
모나리자는	→ NP_SBJ	초상화로
초상화로	→ NP_AJT	있다.

5.2.2. 부사 ‘없이’, ‘같이’의 주석

- “없이”나 “같이”와 같은 부사(용언의 활용형이 아님에 유의)는 서술어와 마찬가지로 논항을 취할 수 있다. 특히 부사 ‘같이’는 앞에 ‘~와’에 해당하는 명사구가 나타나는 경우 ‘~와’ 명사구를 ‘같이’의 부사어로 처리한다.

[예시] 철수는 아무 생각도 없이 길을 나섰다.		
철수는	→ NP_SBJ	나섰다.
아무	→ DP	생각도
생각도	→ NP_SBJ	없이
없이	→ AP	나섰다.
길을	→ NP_OBJ	나섰다.
나섰다.	→ VP	ROOT

[예시] 예상한 바와 같이 주가가 크게 떨어졌다.		
예상한	→ VP_MOD	바와
바와	→ NP_AJT	같이
같이	→ AP	떨어졌다.
주가가	→ NP_SBJ	떨어졌다.
크게	→ VP_AJT	떨어졌다.
떨어졌다.	→ VP	ROOT

5.3. 연결된 부사구 세부 유형 가이드라인: ‘~부터 ~까지’, ‘~에서 ~으로’의 주석

- ‘~부터’, ‘~까지’ (또는 ‘~에서’, ‘~로’)와 같은 부사구는 각각을 지배소에 연결한다.

[예시] 그는 작년부터 지금까지 열심히 일했다.		
그는	→ NP_SBJ	일했다.
작년부터	→ NP_AJT	일했다.
지금까지	→ NP_AJT	일했다.
열심히	→ AP	일했다.
일했다.	→ VP	ROOT
[예시] 부산에서 서울로 가는 표 한 장 있나요?		
부산에서	→ NP_AJT	가는
서울로	→ NP_AJT	가는
가는	→ VP_MOD	표
표	→ NP	장
한	→ DP	장
장	→ NP_SBJ	있나요?
있나요?	→ VP	ROOT

- 그러나 만일 ‘~부터 ~까지를’, ‘~에서 ~로의’ 등과 같이 ‘~부터’, ‘~까지’ (또는 ‘~에서’, ‘~로’) 부사구가 하나의 단위로 묶이는 경우 선행 성분을 후행 성분의 부사어(AJT)로 연결시킨 후, 후행 성분을 지배소에 연결한다.

[예시] 평균 90점부터 100점까지를 모두 금상으로 처리한다.		
평균	→ NP	90점부터
90점부터	→ NP_AJT	100점까지를
100점까지를	→ NP_OBJ	처리한다.
[예시] 이것은 바로 개인주의에서 단체주의로의 전환을 의미했다.		
이것은	→ NP_SBJ	의미했다.
바로	→ AP	의미했다.
개인주의에서	→ NP_AJT	단체주의로의
단체주의로의	→ NP_MOD	전환을

5.4. 장형 사동 구문 유형 세부 가이드라인

- ‘~게 하다’의 장형 사동 구문은 다른 보조 용언과 동일하게 처리한다. 즉, ‘~게 하다’에서 선행하는 용언(‘~게’)에 다른 성분들을 모두 연결한다.
- 또한 ‘A가 B가/B를/B에게 V-게 하다’와 같은 장형 사동 구문의 B성분은 격 조사에 의존하여 기능 태그를 부착한다. 즉, ‘B가’는 SBJ, ‘B를’은 OBJ, ‘B에게’는 AJT로 처리한다.

[예시] 그가 철수를 집에 가게 하였다.		
그가	→ NP_SBJ	가게
철수를	→ NP_OBJ	가게

집에	→ NP_AJT	가게
가게	→ VP	하였다.
하였다.	→ VP	ROOT
[예시] 그가 철수에게 집에 가게 하였다.		
그가	→ NP_SBJ	가게
철수에게	→ NP_AJT	가게
[예시] 그가 철수가 집에 가게 하였다.		
그가	→ NP_SBJ	가게
철수가	→ NP_SBJ	가게

5.5. 여러 개의 문장 처리

- 여러 개의 문장이 분할되지 않은 상태로 제시되어 있는 경우, 각 문장의 서술어가 순차적으로 의존하도록 처리한다.

[예시] “철수는 집에 갔어. 영희는 모르겠어. 민지는 집에 있겠지.” 라고 말했다.		
갔어.	→ VP	모르겠어.
모르겠어.	→ VP	있겠지.” 라고
있겠지.” 라고	→ VP_CMP	말했다.

5.6. 명사-부사 통용어 또는 체언 수식 부사의 처리

- 명사-부사 통용어 중 뒤에 조사가 붙지 않고 서술어에 의존하는 경우(‘오늘’ 등) 또는 본래 부사이지만 체언을 수식하는 용법으로 쓰이는 경우에는(‘가장’, ‘아주’, ‘바로’ 등) 이들의 품사적 지위를 기준으로 하여 ‘AP’로 처리한다.

[예시] 친구네 집은 우리 집 바로 뒤에 있어.		
친구네	→ NP	집은
집은	→ NP_SBJ	있어.
우리	→ NP	집
집	→ NP	뒤에
바로	→ AP	뒤에
뒤에	→ NP_AJT	있어.
있어.	→ VP	ROOT

[예시] 오늘(→ AP 해야) 해야 할 일을 다음 날로 미루어서는 안 된다.		
--	--	--

5.7. 술어 생략 세부 유형 가이드라인

- 내포문의 서술어가 생략되었다고 판단되는 경우 내포문의 표층에 나타난 마지막 어절에 선행하는 성분들을 의존하게 하고 마지막 어절을 모문의 서술어에 의존하게 한다.
- 내포문의 서술어가 생략된 경우 형태 기준(조사)으로 분석한다.

[예시] 영희가 철수를 조건으로 (하여) 매장을 임시로 설치하였다.		
영희가	→ NP_SBJ	조건으로
철수를	→ NP_OBJ	조건으로
조건으로	→ NP_AJT	설치하였다.
[예시] 나는 여섯 살, 이름은 철수예요.		
나는	→ NP_SBJ	살,

여섯	→ DP	살,
살,	→ NP	철수예요.
이름은	→ NP_SBJ	철수예요.
철수예요.	→ VNP	ROOT

→ 계사 ‘이다’가 생략되는 경우 형태를 기준으로 ‘살,’을 NP로 주석하여 후행절 서술어에 연결되도록 한다.

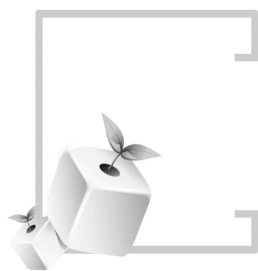
- 모문의 서술어가 생략된 경우 형태 기준(조사)으로 분석한다.

[예시] 과거 하계 올림픽의 정식 종목으로 맞는 것은?(NP_SBJ)

[예시] 밥을.(NP_OBJ)

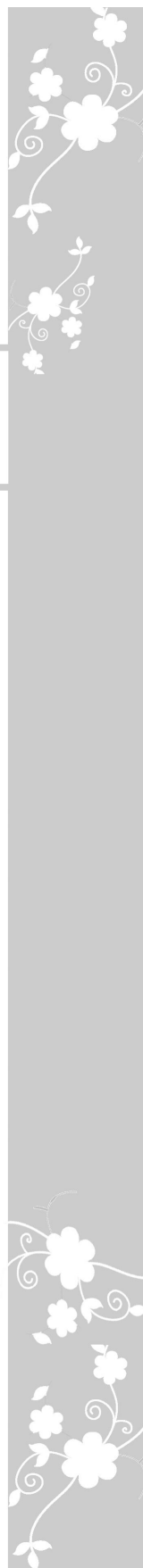
[예시] 철수가(NP_SBJ) 밥을.(NP_OBJ)

철수가	→ NP_SBJ	밥을.
밥을.	→ NP_OBJ	ROOT



제 4 장

결 론



본 사업은 인공 지능 산업 발전을 위한 대규모 우리말 자원 수요를 위한 구문 분석 말뭉치 구축을 목적으로 한다. 또한 이 과정에서 요구되는 구문 분석 말뭉치 지침을 수립하는 것 또한 본 과제의 목적에 포함된다.

본 사업의 범위는 크게 두 부분으로 나눌 수 있다. 첫째는 구문 분석 말뭉치 지침 수립으로, 한국정보통신기술협회(TTA) 등 관련 분야의 구문 분석 지침을 참고하여 비교·연구하였으며 한국전자통신연구원(ETRI)의 한국어 의존 구문 분석 가이드라인 등 관련 분야의 지침 등을 바탕으로 기존 지침의 문제점을 수정·보완한 의존 구문 분석 지침을 수립하였다. 둘째는 구문 분석 말뭉치(200만 어절) 구축으로, 구문 분석 말뭉치 구축 지침을 바탕으로 말뭉치를 구축하였다.

○ 구문 분석 말뭉치 지침 수립

구문 분석 지침을 수립하기 위하여 ‘한국정보통신기술협회(TTA)’ 등 관련 분야의 구문 분석 지침을 참고하여 비교·연구하였으며 ‘한국어 구문 가이드라인(한국전자정보통신원)’ 등 관련 분야 분석 지침을 검토하여 기존 지침의 문제점을 분석, 보완된 지침을 제시하였다.

기존의 의존 구문 분석 말뭉치는 주로 <21세기 세종계획>에서 구축된 구 구조 기반 구문 분석 말뭉치를 의존 구문 분석 말뭉치로 변환하는 것에 초점을 맞추어 개발되었으며 이에 따라 의존 구문 분석 말뭉치를 구축하기 위한 지침은 그 양적·질적 측면에서 부족하였다고 할 수 있다. 본 사업에서는 한국정보통신기술협회의 의존 구문 분석 가이드라인을 기반으로 하여 의존 구문 분석 말뭉치를 구축하기 위한 실용적인 지침을 개발하고 기존 의존 구문 분석 지침에서 부족하다고 판단되는 내용과 예시를 보완하였다.

○ 구문 분석 말뭉치(200만 어절) 구축

구문 분석 말뭉치의 구축 절차는 다음과 같다.

구문 분석 지침 수립 > 수작업 검수 도구 커스터마이징 > 작업자 교육 > 자동 구문 분석 > 작업자 분석(1차 검수) > 팀장 및 교수진 검수(2차 검수) > 딥 러닝 기반 구문 분석 말뭉치 검증 > 전문가 집단 심층 면접 > 최종 결과물 산출

의존 구문 분석 지침을 바탕으로 200만 어절 규모의 구문 분석 말뭉치를 구축하였다. 말뭉치는 문어 200만 어절 규모로 신문 기사 텍스트로 구성되어 있다.

본 사업에서는 국립국어원에서 제공한 문어 말뭉치를 대상으로 문장 단위 구문 분석을 진행하였다. 또한 각 어절별로 구문 분석 정보를 부착하였으며 제이슨(JSON) 형식의 최종 산출물을 제출하였다.

구축 과정에서 다수의 자동 구문 분석기를 활용하여 작업의 편의성과 일관성을 확보하고자 하였으며 자동으로 분석된 결과를 작업자들이 수작업으로 전수 검토하도록 하였다. 상세한 과정은 다음과 같다. 먼저 국립국어원에서 제공한 신문 기사 말뭉치를 문장 단위를 기준으로 자동 구문 분석을 시행한다. 이때 복수의 자동 분석기 결과가 전체 일치하는 문장에 대해서는 최종 검수자의 검수만으로 검수가 완료되며 자동 분석 결과가 전체 일치하지 않는 문장에 대해서는 1차 검수와 2차 검수를 거쳐 검수가 완료된다. 작업자는 담당 교수, 팀장을 포함하여 총 4개 조로 편성되었으며 각 조의 담당 교수 및 팀장이 나머지 작업자의 1차 검수 결과물을 검수하는 방식으로 2차 검수를 실시하였다. 2차 검수 후에는 알고리즘을 통한 후처리 및 파일 형식 변환을 통해 최종 산출물을 제출하였다.

본 사업에서는 자동 구문 분석을 위하여 한국전자통신연구원, 강원대학교, 전북대학교, 충남대학교의 의존 구문 분석기를 활용하여 자동 분석을 진행하였고, 다수의 기관에서 자동으로 분석한 결과를 교차 검증 및 통합하여 신뢰도를 높이하고자 하였다. 자동 구문 분석의 과정에서는 딥 러닝을 통한 일관성 검증이 수반되었으며 작업자의 1차 검수와 2차 검수에 사용되는 작업 환경 또한 본 사업 과정에서 보완되었다.

<Abstract>

Building a Korean parsed corpus

The purpose of this project is to build a parsed corpus to meet the demand for large-scale Korean language resources in the AI industry. In addition, this project also aims to set up annotation guidelines for compiling a dependency-parsed corpus in Korean.

This project has two main parts. The first is establishing dependency parsing guidelines for a Korean corpus. We compared and analyzed several guidelines of related fields such as “Dependency tag sets and dependency relation establishment methods for constructing dependency tagged corpora” of the National Telecommunications Technology Association (TTA), which we corrected and improved to maintain the consistency across the parsed corpus. The second is constructing a parsed corpus of Korean. The corpus contains two million words which were selected from the national corpus of written Korean constructed in 2018 by the National Institute of Korean Language.

○ Establishing a guideline for a dependency-parsed corpus

To establish dependency parsing guidelines, we compared and reviewed parsing guidelines such as the ‘Telecommunications Technology Association (TTA)’. We present an improved version of guidelines based on the problem analysis.

Most dependency-parsing guidelines focus on conversion of an old constituency-based parse tree constructed in <21st Century Sejong Plan> into a new dependency-parsed corpus. Therefore, the guidelines for constructing

dependency-parsed corpus were insufficient in terms of quantity and quality. In this project, based on the Korea Telecommunications Technology Association (TTA) dependency parsing guideline, we suggest more practical and detailed guide for constructing a dependency-parsed corpus with contents and examples complementing what was insufficient in the original version.

○ Constructing a dependency-parsed corpus (two million words)

The construction procedure is as follows.

Establishing parsing guidelines > Customizing manual inspection tools > Researcher training > Automatic parsing > Manual analysis (1st inspection) > Team leader and faculty inspection (2nd inspection) > Deep learning-based parsed corpus validation > Producing the final result

Based on the dependency parsing guidelines, a dependency-parsed corpus containing two million words was compiled. The corpus is composed of texts from newspaper articles.

In this project, sentence-by-sentence parsing was conducted for the written corpus provided by the National Institute of Korean Language. Also, parsing information was annotated to each word and the final output was submitted in JSON format.

At the initial stage of the construction, multiple automatic parsers were used for the convenience and the consistency of the work. Then, researchers manually reviewed the results. The details of this process are as follows. First, we divided the given raw corpus into sentences and conducted the automatic parsing in a sentence level. When multiple automatic parsers give

the same results, the results were submitted after a final inspector's check. When automatic parsers generate different results, different researchers review the sentence in the first and second inspection, and the final inspector would review the sentence. The researchers were organized into four groups, including the professor in charge and team leaders in each team. Researchers in a group would review sentences first, and the professors and team leaders would review the results. The final results were submitted after the second inspection followed by post-processing and file format conversion.

The automatic analysis was performed using dependency analyzers of ETRI, Kangwon National University, Jeonbuk National University, and Chungnam National University. We used different analyzers from different organizations to improve the credibility of our project by cross-analyzing and integrating the results. In the process of automatic parsing, consistency was verified by deep learning method.

Keywords: Syntax structure, parsing, dependency parsing, corpus, dependency-parsed corpus

Project Director: Yim Seongmo(MindsLab)

<부록 1> 제이슨(JSON) 형식의 기본 구조

1 수준	2 수준	3 수준	4 수준	5 수준	타입	설명
id					string	* 원시 말뭉치 파일 ID 혹은 작업세트 파일 ID * 고유 ID로 중복이 없어야 함
metadata					object	* 파일의 메타 정보
	title				string	* 파일 제목
	author				string	* 작성자, 게시자
	publisher				string	* 출판사, 신문사
	year				string	* 출판년도
	note				string	* 부가 설명. 샘플링 방식 등 기타 정보
document					array(object)	* 문서 정보
	id				string	* 문서 ID * '원시말뭉치파일 ID.파일 내 문서 순서'로 구성
	metadata				object	* 문서의 메타 정보
		title			string	* 문서 제목
		author			string	* 작성자, 게시자
		publisher			string	* 출판사, 신문사
		url			string	* URL 주소 (웹 말뭉치)
		date			string	* 작성일시, 게시일시, 크롤링 일시
		category			string	* 분류. 분류 단계는 '〉'로 구분. 예) '신문 > 전국 종합지'
		annotation_level			array(string)	* 분석 층위 (복수 나열) * 원시, 형태 분석, 개체명 분석, 어휘 의미 분석, 상호참조 해결, 무형 대용어 복원, 구문 분석, 의미역 분석
		note			string	* 부가 설명. 구어 사용 맥락 정보, 샘플링 방식 등 기타 정보
	sentence				array(object)	* 문장
		id			string	* 문장 ID. * '문서 ID.문서 내 문장 순서'로 구성. 문서 내 문장 순서는 1부터 시작
		form			string	* 문장 정보
		word			array(object)	* 어절 정보
			id		number	* 어절 ID. 문장 내 순서로 1부터

						시작
			form		string	* 어절
			begin		number	* 어절의 문장 내 시작 위치 (UTF-8 문자 위치로 0부터 시작)
			end		number	* 어절의 문장 내 끝 위치 (UTF-8 문자 위치로 0부터 시작)
		DP			array (object)	* 구문 분석 정보
			word_id		number	* 어절 ID
			word_form		string	* 어절
			head		number	* 어절의 지배소 정보(어절 ID). 어절이 최상위 지배소면 -1
			label		string	* 구문 태그
			dependent		array (number)	* 어절의 피지배소 정보 (피지배소들의 어절 ID 숫자 배열)

<부록 2> 제이슨(JSON) 형식의 예시

```
{
  "id": "NXRW1802000000",
  "metadata": {
    "title": "동아일보, 조선일보, 한겨레 2009~2017년 기사",
    "author": "동아일보, 조선일보, 한겨레",
    "publisher": "동아일보사, 조선일보사, 한겨레",
    "year": "2009~2017",
    "note": "부분 추출 - 임의 추출"
  },
  "document": [
    {
      "id": "NWRW1800000026-0008",
      "metadata": {
        "title": "동아일보, 조선일보, 한겨레 2009~2017년 기사",
        "author": "동아일보, 조선일보, 한겨레",
        "publisher": "동아일보사, 조선일보사, 한겨레",
        "url": "https://www.korean.go.kr/",
        "date": "20100107",
        "category": "신문 > 국제",
        "annotation_level": ["형태 분석", "어휘 의미 분석", "구문 분석"],
        "note": ""
      },
      "sentence": [
        {
          "id": "NWRW1800000021-0205.19",
          "form": "세계유산위원회는 27일 등재 결정문에서 왕릉 주변 개발 완충 지역 내 개발의 가이드라인을 만들라고 권고했다.",
          "DP": [
            { "word_id": 1, "word_form": "세계유산위원회는", "head": 13, "label": "NP_SBJ", "dependent": [] },
            { "word_id": 2, "word_form": "27일", "head": 3, "label": "NP", "dependent": [] },
            { "word_id": 3, "word_form": "등재", "head": 4, "label": "NP", "dependent": [2] },
            { "word_id": 4, "word_form": "결정문에서", "head": 13, "label": "NP_AJT", "dependent": [3] },
            { "word_id": 5, "word_form": "왕릉", "head": 6, "label": "NP", "dependent": [] },
            { "word_id": 6, "word_form": "주변", "head": 7, "label": "NP", "dependent": [5] },
            { "word_id": 7, "word_form": "개발", "head": 8, "label": "NP", "dependent": [6] },
            { "word_id": 8, "word_form": "완충", "head": 9, "label": "NP", "dependent": [7] },
            { "word_id": 9, "word_form": "지역", "head": 10, "label": "NP", "dependent": [8] },
            { "word_id": 10, "word_form": "내", "head": 11, "label": "NP", "dependent": [9] },
            { "word_id": 11, "word_form": "개발의", "head": 12, "label": "NP_MOD", "dependent": [10] },
            { "word_id": 12, "word_form": "가이드라인을", "head": 13, "label": "NP_OBJ", "dependent": [11] },
            { "word_id": 13, "word_form": "만들라고", "head": 14, "label": "VP_CMP", "dependent": [1, 4, 12] },
            { "word_id": 14, "word_form": "권고했다.", "head": -1, "label": "VP", "dependent": [13] }
          ]
        }
      ]
    }
  ]
}
```

<부록 3> 원시 말뭉치 엑스엠엘(XML) 형식의 예시

```
<?xml version="1.0" encoding="utf-8"?>
<SJML>
<header>
  <fileInfo>
    <fileId>NXRW1802000000</fileId>
    <annoLevel>원시</annoLevel>
    <sampling>부분 추출 - 임의 추출</sampling>
    <class>신문</class>
  </fileInfo>
  <sourceInfo>
    <title>동아일보, 조선일보, 한겨레 2009~2017년 기사</title>
    <author>동아일보, 조선일보, 한겨레</author>
    <publisher>동아일보사, 조선일보사, 한겨레</publisher>
    <year>2009~2017</year>
  </sourceInfo>
</header>
<text id="NWRW1800000021-0205" date="20090629" subclass="정치">
  <p>
    <s>...</s>
    <s>...</s>
  </p>
  <p>
    <s>세계유산위원회는 27일 등재 결정문에서 왕릉 주변 개발 완충 지역 내 개발의 가이드라인을 만들라고 권고했
다.</s>
  </p>
  <byline>윤완준 기자 zeitung@donga.com</byline>
</text>
</SJML>
```

사업 책임자	임성모(주식회사 마인즈랩)
사업 참여자	서상원(주식회사 마인즈랩)
	이석준(주식회사 마인즈랩)
	이원문(주식회사 마인즈랩)
	박영선(주식회사 마인즈랩)
	송혜원(주식회사 마인즈랩)
	윤서영(주식회사 마인즈랩)
	박지원(주식회사 마인즈랩)
	이예준(주식회사 마인즈랩)
	김한샘(연세대학교)
	유현경(연세대학교)
	김재훈(한국해양대학교)
	이공주(충남대학교)
	김유섭(한림대학교)
	류법모(부산외국어대학교)
	김학수(강원대학교)
	신서인(한림대학교)
	나승훈(전북대학교)
	봉미경(연세대학교)
	김선희(연세대학교)
	김수경(연세대학교)
	이찬영(연세대학교)
	박혜진(연세대학교)
	장연지(연세대학교)
	신아영(연세대학교)
	정주연(연세대학교)
	정진경(연세대학교)
	강혜린(연세대학교)

김교연(연세대학교)
김상민(연세대학교)
김지영(연세대학교)
정해운(연세대학교)
천성호(연세대학교)
박서윤(연세대학교)
박진현(한림대학교)
이수현(한림대학교)
이한범(한림대학교)
전상호(한림대학교)
김유미(한림대학교)
이지원(한림대학교)
김재균(한국해양대학교)
남궁영(한국해양대학교)
윤호(한국해양대학교)
최민석(한국해양대학교)
최용석(충남대학교)
박천용(충남대학교)
오병두(한림대학교)
허탁성(한림대학교)
민진우(전북대학교)
이영훈(전북대학교)
홍승연(전북대학교)
박성식(강원대학교)
신영진(한국해양대학교)
강일민(충남대학교)
박요한(충남대학교)
정혜지(충남대학교)
정영석(한림대학교)
오세은(한림대학교)

황석주(한림대학교)

강동찬(전북대학교)

이종현(전북대학교)

최형준(전북대학교)

김담린(강원대학교)

김보은(강원대학교)

김홍진(강원대학교)

오신혁(강원대학교)

박상원(주식회사 답네츄럴)

박정수(주식회사 답네츄럴)

허민강(주식회사 답네츄럴)

박연호(주식회사 답네츄럴)

정동호(주식회사 답네츄럴)

최진혁(주식회사 답네츄럴)

김예진(주식회사 답네츄럴)

이규덕(주식회사 답네츄럴)

임선민(주식회사 답네츄럴)

담당 연구원

이승재(국립국어원 언어정보과장)

서셋별(국립국어원 언어정보과 학예연구사)

발행인: 국립국어원장

발행처: 국립국어원

서울시 강서구 금남화로 154

전화 02-2669-9775, 전송 02-2669-9727

인쇄일: 2020년 2월 6일

발행일: 2020년 2월 6일

인 쇄: 비즈카피

※ 이 책은 국립국어원의 용역비로 수행한 ‘구문 분석 말뭉치 구축’ 사업의 결과물을 발간한 것입니다.