국립국어원 2019-01-55

발 간 등 록 번 호
11-1371028-000803-01

의미역 분석 말뭉치 구축

사업 책임자
임 성 모

국립국어원

제 출 문

국립국어원장 귀하

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

■ 사업 기간: 2019년 07월 ~ 2020년 01월

2020년 01월 30일

사업 책임자: 임성모(주식회사 마인즈랩)

사업 수행자 주식회사 마인즈랩
연세대학교 산학협력단
주식회사 딥네추럴

사업 책임자 임성모

사업 참여자 서상원, 이석준, 이원문, 박영선, 송혜원,
윤서영, 임병현, 이하영, 이예준, 김한샘,
유현경, 김재훈, 이공주, 김유섭, 류법모,
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, 박광현, 이영훈,
홍승연, 박성식, 신영진, 강일민, 박요한,
정혜지, 박석주, 오세은, 황석주, 강동찬,
이종현, 최형준, 김담린, 김보은, 김홍진,
오신혁, 박상원, 박정수, 허민강, 박연호,
정동호, 최진혁, 김예진, 이규덕, 임선민

〈사업 수행자〉 주식회사 마인즈랩 · 주식회사 딥네추럴 · 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)
	임병현(주식회사 마인즈랩)
	이하영(주식회사 마인즈랩)
	이예준(주식회사 마인즈랩)
	김한샘(연세대학교)
	유현경(연세대학교)
	김재훈(한국해양대학교)
	이공주(충남대학교)
	김유섭(한림대학교)
	류법모(부산외국어대학교)

사업 참여자	김학수(강원대학교)
	신서인(한림대학교)
	나승훈(전북대학교)
	봉미경(연세대학교)
	김선혜(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)

사업 참여자	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)
	박광현(전북대학교)
	이영훈(전북대학교)

사업 참여자	홍승연(전북대학교)
	박성식(강원대학교)
	신영진(한국해양대학교)
	강일민(충남대학교)
	박요한(충남대학교)
	정혜지(충남대학교)
	박석주(한림대학교)
	정영석(한림대학교)
	오세은(한림대학교)
	황석주(한림대학교)
	강동찬(전북대학교)
	이종현(전북대학교)
	최형준(전북대학교)
	김담린(강원대학교)
	김보은(강원대학교)
	김홍진(강원대학교)
	오신혁(강원대학교)
	박상원(주식회사 딥네추럴)

사업 참여자	박정수(주식회사 딥네추럴)
	허민강(주식회사 딥네추럴)
	박연호(주식회사 딥네추럴)
	정동호(주식회사 딥네추럴)
	최진혁(주식회사 딥네추럴)
	김예진(주식회사 딥네추럴)
	이규덕(주식회사 딥네추럴)
	임선민(주식회사 딥네추럴)

<국문 초록>

의미역 분석 말뭉치 구축

본 사업의 목적은 인공지능 산업 발전을 위한 대규모 우리말 자원 수요에 따른 의미역 분석 말뭉치를 구축하고 이 과정에서 요구되는 의미역 분석 지침을 수립하는 것이다.

본 사업은 의미역 분석 말뭉치 지침 수립과 의미역 분석 말뭉치 구축이라는 두 부분으로 구성된다.

○ 의미역 분석 말뭉치 지침 수립

기존의 한국어 의미역 분석 말뭉치는 구축 주체에 따라 각기 다른 의미역 주석 표지와 프레임셋을 적용하여 구축되었다. 국가 차원의 의미역 분석 말뭉치를 구축하기 위한 지침은 그간 축적된 한국어 의미역 분석 말뭉치 관련 자원을 활용하여 주석할 방법론을 포함하는 것이 바람직하다. 본 사업에서는 한국전자통신연구원(ETRI)의 한국어 의존의미역 주석가이드라인을 기반으로 의존의미역 분석 말뭉치 구축을 위한 실용적인 지침을 개발하였다. 이와 함께 기존 의존의미역 분석 지침에서 부족하다고 판단되는 내용과 예시를 보완하였다.

○ 의미역 분석 말뭉치(200만 어절) 구축

의미역 분석 말뭉치 구축은 다음과 같은 절차로 진행하였다.

의미역 분석 지침 수립 > 수작업 검수 도구 최적화 > 작업자 교육 > 자동 의미역 분석 > 작업자 분석(1차 검수) > 조장과 공동연구원 검수(2차 검수) > 딥러닝(Deep Learning) 기반 정확도와 일관성 검증 > 검수 완료 > 최종 결과물 산출

의미역 분석 지침에 따라 국립국어원에서 제공한 200만 어절 규모의 신문 기사 텍스트로 구성된 문어 말뭉치를 대상으로 의미역 분석 말뭉치를 구축하였다. 이 말뭉치의 동사와 형용사에 대하여 명사구 또는 절을 단위로 의미역 분석 정보를 부착하였으며 최종 결과물은 JSON 형식으로 제출하였다.

작업의 편의성과 일관성을 확보하기 위하여 구축 과정에서 복수의 자동 의미역 분석

기를 활용하여 국립국어원에서 제공한 신문 기사 말뭉치를 문장 단위로 분할하여 자동 의미역 분석을 진행하였다. 이 자동 분석 결과에 대하여 작업자들이 수작업으로 전수 검토하였다. 다수의 자동 분석 결과가 전원 일치하는 문장은 최종 검수자가 검수하고, 전원 일치하지 않는 문장에 대해서는 두 차례의 검수를 시행하여 구축을 완료하였다. 검수 담당 작업자는 총 4개 조로 편성되었으며 각 조의 담당 공동연구원과 조장이 1차 검수 결과물을 검수함으로써 2차 검수를 시행하였다. 검수 완료 후에 연산을 통한 후처리와 파일 형식 변환을 완료하여 최종 산출물을 제출하였다.

본 사업에서는 자동 의미역 분석을 위하여 ETRI, 강원대학교, 전북대학교의 의미역 분석기를 활용하여 1차 자동 분석을 진행하였고, 다수의 기관에서 자동으로 분석한 결과를 교차 검증 후 통합함으로써 신뢰도 향상을 도모하였다. 자동 의미역 분석의 과정에서는 딥러닝을 통한 일관성 검증이 수반되었다.

주요어: 의미역, 의미역 분석, 말뭉치, 의미역 분석 말뭉치

차 례

제1장 서론

제2장 의미역 분석 말뭉치의 구성 및 구축 절차

차 례

제3장 의미역 분석 말뭉치 구축 지침 수립

차 례

제4장 결론

제 1 장

서 론

1. 사업의 목적

○ 의미역 분석 말뭉치 지침 수립
○ 의미역 분석 말뭉치(200만 어절) 구축

본 사업의 목적은 인공지능 산업 발전을 위한 고품질 우리말 자원 수요 증대와 국어 자원의 활용도 및 가치 제고의 필요성을 바탕으로 4차 산업혁명 대비 대규모 말뭉치 구축을 통해 양적 · 질적 수준을 확보한 의미역 분석 말뭉치를 구축하는 데 있다.

한국어 처리 분야의 자연어 이해 영역은 기존의 형태 분석, 어휘 의미 분석, 구문 분석을 넘어 질의응답과 정보 검색 등에 활용할 의미역 분석, 생략 복원, 상호 참조 해결 등의 과제로 확장되고 있다. 통사적 분석을 넘어서 문장 단위의 통사의미 구조를 분석하는 의미역 분석은 활용도가 높아 자동 의미역 분석 과제의 완성도를 높이기 위한 평가 자원으로서의 의미역 분석 말뭉치의 필요성이 증대되고 있다.

기존의 말뭉치 구축 과제에서는 의미역 분석 말뭉치를 본격적으로 구축한 사례가 많지 않기 때문에 본 사업에서는 한국전자통신연구원(ETRI)의 ‘한국어 의존의미역 주석 가이드라인’ 분석 지침과 관련 분야의 선행 분석 지침을 비교 연구하여 의미역 분석 말뭉치 구축 활성화를 위한 기반을 마련하고 선행 연구의 결과로 구축된 의미역 관련 언어 자원을 활용할 방법을 모색하였다. 이와 같은 지침 수립과 의미역 정보 통합을 통해 문어 200만 어절 규모의 의미역 분석 말뭉치를 구축하여 JSON 형식의 최종 산출물을 도출하였다.

2. 사업의 범위

본 사업의 범위는 크게 의미역 분석 말뭉치 분석 지침 수립과 이를 바탕으로 의미역 분석 말뭉치를 구축하는 것이다.

2.1. 의미역 분석 말뭉치 구축 지침 수립

본 사업에서는 ‘한국어 의존의미역 주석가이드라인(한국전자통신연구원)’, 울산대 의미역 주석 말뭉치 주석 체계 등 관련 기관 및 학계의 지침과 주석 체계를 검토하고 이를 보완하여 의미역 분석 말뭉치 구축 지침을 수립하였다.

기존의 의미역 분석 말뭉치는 구축 주체에 따라 필수역과 부가역을 구분하는지, 의미역 분석 대상을 논항에만 한정하는지 부사, 부정소 등도 분석하는지, 각각의 의미역 구

분 기준과 의미역 표지 목록 설정 면에서 각기 다른 원칙으로 구축되었다. 대규모의 의미역 분석 말뭉치를 구축하기 위해서는 실제 의미역 분석 말뭉치에 기반한 상세 주석 지침 항목의 제시가 필요하다. 본 사업에서는 샘플 구축 작업을 통해 실제 의미역 분석 시 발생하는 문제를 발굴하고 이를 처리하기 위한 실용적인 지침을 개발하였다. 이 과정에서 기존 의존의미역 분석 지침에서 부족하다고 판단되는 내용과 예시를 보완한다.

2.2. 의미역 분석 말뭉치 구축

본 사업에서 수립한 의미역 분석 지침에 따라 문어 말뭉치 200만 어절을 대상으로 의미역 분석 말뭉치를 구축하였다.

분석 대상 말뭉치는 국립국어원에서 제공하였으며, 신문 기사 텍스트로 구성되어 있다. 의미역 분석은 문장 단위로 분리하여 수행하였다. 한국어 프롭뱅크, 울산대 유프롭뱅크, 우리말샘 등의 자원을 기반으로 서술어별 프레임셋 통합 자원을 구축하고 이를 기반으로 의미역의 범위를 표시하여 분석 정보를 부착하였다.

작업의 편의성과 일관성을 확보하기 위하여 구축 과정에서 복수의 자동 의미역 분석기를 활용하여 자동 의미역 분석을 진행하였으며, 이 자동 분석 결과에 대하여 작업자들이 수작업으로 전수 검토하였다. 다수의 자동 분석 결과가 전원 일치하는 문장은 최종 검수자가 검수하고, 전원 일치하지 않는 문장에 대해서는 두 차례의 검수를 시행하여 구축을 완료하였다. 검수 담당 작업자는 총 4개 조로 편성되었으며 각 조의 담당 공동연구원과 조장이 1차 검수 결과물을 검수함으로써 2차 검수를 시행하였다. 검수 완료 후에 연산을 통한 후처리와 파일 형식 변환을 완료하여 최종 산출물을 제출하였다.

본 사업에서는 자동 의미역 분석을 위하여 ETRI, 강원대학교, 전북대학교의 의미역 분석기를 활용하여 1차 자동 분석을 진행하였고, 다수의 기관에서 자동으로 분석한 결과를 교차 검증 후 통합함으로써 신뢰도 향상을 도모하였다. 자동 의미역 분석의 과정에서는 딥러닝을 통한 일관성 검증을 수반하였고, 작업자의 수작업 검수에 사용되는 작업 환경도 본 사업을 위한 최적화 과정을 거쳤다.

제 2 장

의미역 분석
말뭉치의 구성 및
구축 절차

1. 의미역 분석 말뭉치의 구성

이 사업에서 구축한 의미역 분석 말뭉치는 200만 어절 규모의 문어 말뭉치이다. 의미역 분석 대상 말뭉치는 국립국어원에서 제공하였으며, 이는 <2018년 국어 말뭉치 연구 및 구축> 사업에서 구축한 말뭉치의 일부이다. 기반 말뭉치의 텍스트 장르는 모두 신문 기사이며, 텍스트 주제는 다음과 같이 구성되었다.

	경제	과학	국제	기획	문화	사람들	사회	스포츠	오피니언	정치	지역	합
어절(천)	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> 기 구축된 신문 기사 말뭉치 주제별 구성

2. 의미역 분석 말뭉치 구축 절차

의미역 분석 말뭉치는 다음 절차에 따라 진행하였다.

의미역 분석 지침 수립 > 수작업 검수 도구 최적화 > 작업자 교육 > 자동 의미역 분석 > 작업자 분석(1차 검수) > 조장과 공동연구원 검수(2차 검수) > 딥러닝 기반 정확도와 일관성 검증 > 검수 완료

2.1. 의미역 분석 지침 수립

의미역 분석 말뭉치를 구축하기 위한 첫 단계로 의미역 분석 지침을 수립하였다. 기존의 의미역 분석 말뭉치는 ETRI Exobrain 의미역 인식 말뭉치(8,531 어절)[1], 한림대학교 의미역 말뭉치(8,080 문장), Korean PropBank 말뭉치(10,093 문장) 등 다양한 주체가 각기 다른 지침과 원시 말뭉치를 기반으로 구축한 것이다. 본 사업에서는 기구축 한국어 자원보다 많은 양의 의미역 분석 말뭉치를 일관성 있게 구축하기 위하여 기존의 의미역 분석 지침인 한국전자통신연구원(ETRI)에서 사용하는 《한국어 의존의미역 주석 가이드라인》을 기반으로 하여 지침을 수립하되 국어학 연구자와 말뭉치 자료를 활용하는 사업자 등의 의견을 반영하여 의미역 분석 말뭉치의 질과 그 활용도를 높이고자

1) 한국전자통신연구원(이하 ETRI)에서 한국어 분석 및 질의응답 기술을 개발하기 위한 과학기술정보통신부 소프트웨어 분야 R&D인 엑소브레인(Exobrain) 과제를 통해 구축한 자료이다. 제시한 말뭉치 규모는 '임수종・권민정・김준수・김현기(2015), ExoBrain을 위한 한국어 의미역 가이드라인 및 말뭉치, 제27회 한글 및 한국어 정보처리 학술대회 논문집' 기준임.

하였다.

ETRI 《한국어 의존의미역 주석가이드라인》과 의미역 분석 전반에서의 수정과 보완이 필요한 사항 중 주요 쟁점 사항은 다음과 같다.

○ 신문 텍스트의 특성
- ▸ 신문 헤드라인은 명사로 종결되거나 성분 생략이 많아 자동 의미역 분석이 정확하지 않음.
- ▸ 인용문이 많고 인용된 문장이 문법적·의미적 오류를 범하는 경우가 있음.

○ 서술성 명사의 의미역 분석 문제
- ▸ 서술성 명사의 의미역 분석이 누락됨.
- ▸ 서술성 명사에 대한 의미역 정보가 구축된 바 없음.

○ 의미역의 범위와 세분화 문제
- ▸ 행동주와 경험주, 피동주와 대상역의 구분이 없고 비교역 등이 설정되어 있지 않아 동일 의미역이 포괄하는 의미역의 범위가 매우 넓음.
- ▸ 특히 부가역의 경우 하나의 성분이 여러 의미역으로 해석될 가능성이 있어 일관성 유지에 어려움이 있음.

○ 조사의 다의성
- ▸ 조사(‘에’, ‘에서’ 등)의 다의성을 고려하지 않은 자동 분석 오류가 빈발함.
- ▸ 다의성을 가지는 조사에 대한 수작업 검수가 요구됨.

○ ‘에서’ 주어 분석
- ▸ 행동주, 위치역 두 가지 분석 결과가 모두 나타남.
- ▸ 형태가 같으나 의미역이 다른 경우의 구분 기준과 예시 보완 필요.

○ 관계 관형절 문제
- ▸ 피수식 명사의 의미역 분석이 부정확한 문제
- ▸ 의미역에 동일 서술어를 포함한 문제

○ 복문의 의미역 분석
- ▸ 내포절 앞에 있는 주절의 성분을 의미역으로 분석하지 못하는 오류
- ▸ 절 경계를 넘어 의미역 분석이 이루어지는 경우
- ▸ 인용절에 대한 의미역 분석 문제

○ 사동주 주석 문제
- ▸ ‘-게 하다’에 의한 사동문의 경우 의미역 분석에 대한 지침 부재함.

○ 의사 보조용언 분석 문제
- ▸ 서법 등의 기능을 하는 의사 보조용언의 경우 의미역 할당의 주체가 될 수 없음.

○ 격 중출 문제
- ▸ 주격 조사, 목적격 조사 등이 중출된 경우의 의미역 분석 문제

○ 기반 말뭉치의 오류

- 원 말뭉치의 오타, 띄어쓰기 등 입력 오류가 의미역 자동 분석에 영향을 주는 경우가 있음.
- 원 말뭉치의 오류를 국립국어원에 보고할 수 있는 시스템이 필요함.

위의 주요 쟁점 사항을 포함하여 본 단계에서는 기존의 지침(ETRI)을 수정・보완하여 분석의 타당성과 신뢰성을 높이고자 하였다. 지침의 수정과 보완은 실제 작업자 교육과 자동 분석 이후에도 지속하였다. 실제 검수 작업 진행 중에 보고된 오류와 문제점을 검토하여 지속적으로 세부 지침을 수정하고 보완하였으며, 이 과정에서 자문회의를 개최하여 의견을 수렴하고 국어원과의 협의를 진행하였다. 최종 분석 지침은 보고서 3장에 수록하였다.

2.2. 수작업 검수 도구 최적화

의미역 분석 지침 수립의 다음 단계는 의미역의 자동화된 분석 결과를 확인하여 분석 오류를 수정할 수 있도록 지원하는 수작업 검수 도구를 마련하는 것이다.

본 사업에서는 딥네추럴 에이아이(DeepNatural AI) 대중 참여 생산 구조에서 이미 개발되어 운영 중이던 의미역 분석 작업 도구를 최종 지침에 적합하도록 최적화하였다. 이 도구는 웹 기반 도구로, 윈도우즈(Windows), 맥오에스(Mac OS), 리눅스(Linux) 등의 운영체제 종류에 상관없이 컴퓨터에 설치된 크롬 브라우저(Chrome Browser) 상에서 구동된다. 다수의 작업자들이 이 웹 기반 도구에 동시 접속하여 병렬적으로 수작업 검수를 수행하였고, 이 도구를 활용함으로써 검수 분과 내에서 작업을 분배하고 통합할 때에 발생 가능한 오류들을 방지하여 효율적 검수와 관리를 가능하도록 하였다. 또한 작업 결과의 수집도 검수 도구 내에서 암호화되어 진행됨에 따라 한층 강화된 보안성을 확보하였다.

또한 의미역 분석 말뭉치를 구축하기 위해서는 기반이 되는 의미역 정보의 구축이 먼저 완수되어야 하는데 본 사업에서는 Korean PropBank[2]의 의미역 정보, ETRI Frameset의[3] 의미역 정보, U-PropBank[4] 의미역 정보 등을 Korean PropBank의 체계에 따라 일관성 있게 변환한 결과를 수작업 검수 도구에서 검색할 수 있게 하여 의미역 분석의 질을 높였다.

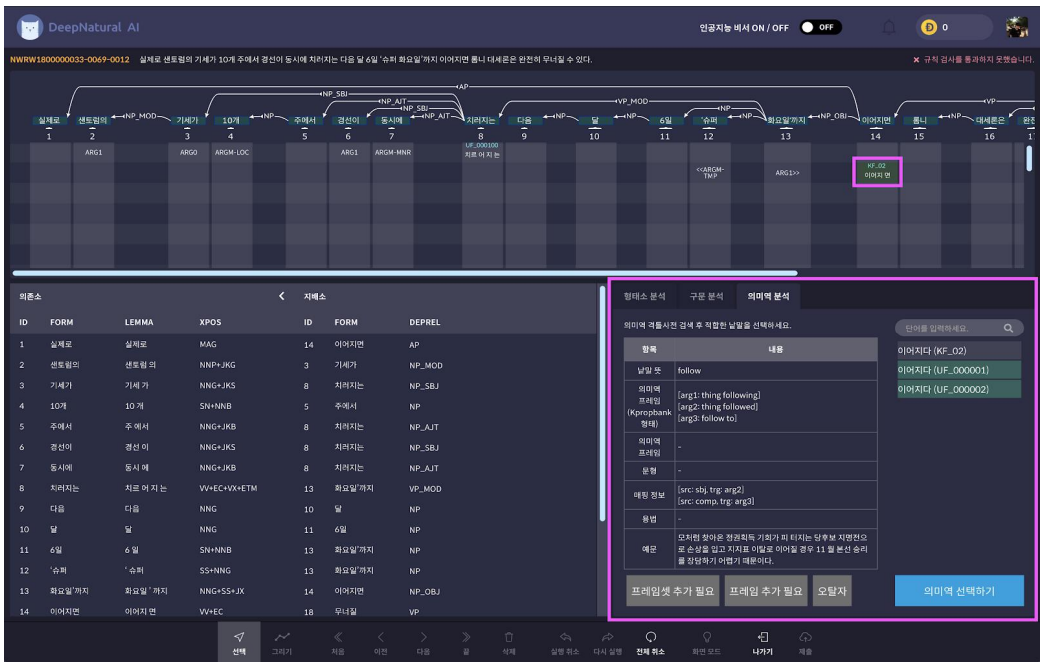
체계적인 검수를 위하여 작업자의 역할에 따라 도구 내 권한을 달리 적용하여 관리하였다. 검수 조원들은 1차 검수자로서의 권한을, 각 조의 조장들은 2차 검수와 진행 현황을 확인할 수 있도록 관리자 권한을 가지도록 설정하여 검수 및 관리의 효율성을 높였다.

---

2) Korean PropBank(Proposition Bank)는 펜실베니아 대학에서 의미역 정보를 기반으로 구축한 한국어 말뭉치이다.
3) ETRI Frameset은 한국전자통신연구원에서 의미역 정보를 기반으로 구축한 한국어 말뭉치이다.
4) U-PropBank는 울산대학교에서 의미역 정보를 기반으로 구축한 한국어 말뭉치이다.

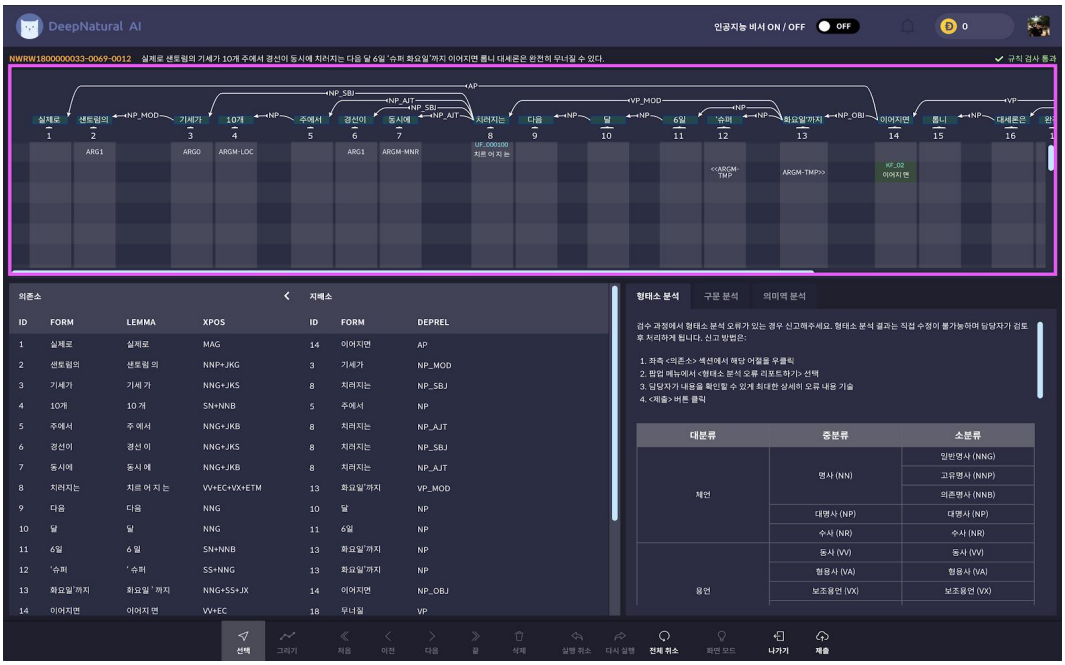
<그림 1> 의미역 분석 검수 도구 초기 화면

1차 검수에 참여하는 조원들은 원시 문장에 대한 자동 의미역 분석 결과를 확인하면서 발견한 분석 오류를 수정하고, 2차 검수 과정에 참여하는 조장들은 1차 검수 결과를 점검하였다.



<그림 2> 선택한 서술어에 대한 의미역 정보 검색 결과

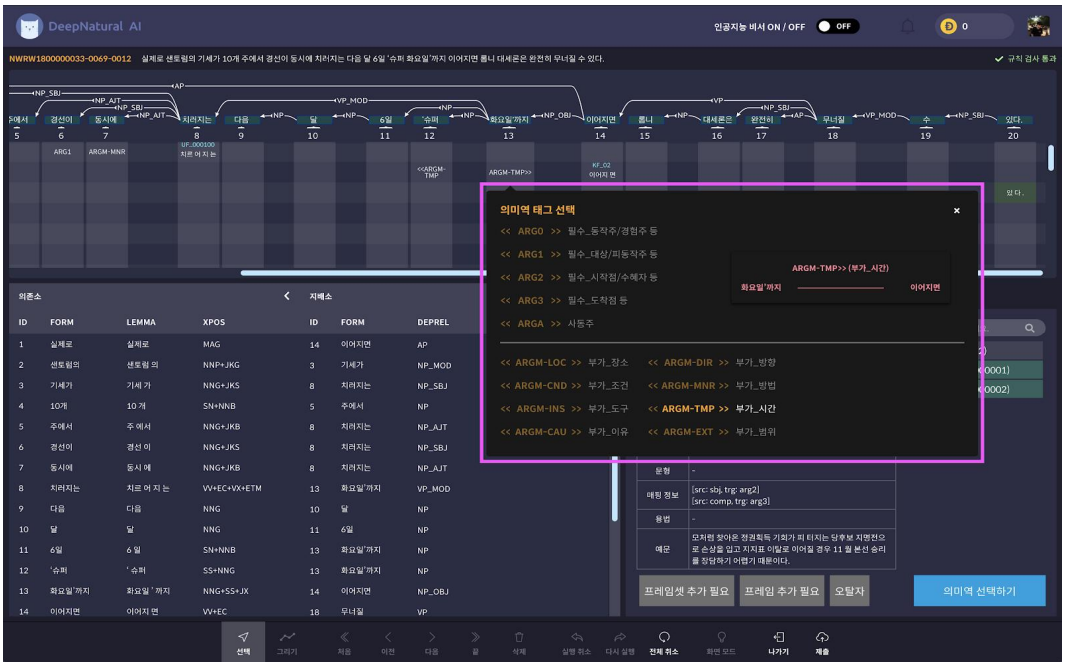
검수자는 우선 서술어에 올바른 의미 번호(Sense ID)가 부여되어 있는지 확인한다. 검수 도구에서 자동 의미역 분석 결과로 탐지된 서술어를 누르면 해당 서술어에 대한 의미역 정보 검색 결과를 <그림 2>와 같이 보여 준다. 검수 도구에 연동된 의미역 정보는 Korean PropBank, ETRI Frameset, U-PropBank와 우리말샘 사전 일부[5]를 포함하고 있다. 자동 분석된 서술어 의미 번호에 오류가 있다면 검수자는 의미역 정보를 검색하여 올바른 의미 번호로 수정할 수 있다. 해당 서술어가 의미역 정보에 없거나 오탈자가 있다면 하단에 위치한 '프레임셋 추가 필요', '프레임 추가 필요', '오탈자' 버튼으로 오류 코드를 서술어에 부착하여 보고할 수 있다. 만약 자동 분석 결과에서 특정 어절이 서술어로 잘못 등록되어 있다면 해당 어절을 두 번 눌러 삭제할 수 있으며, 반대로 자동 분석 결과에서 서술어가 누락되어 있다면 [+] 단추를 눌러 해당 어절을 서술어로 등록할 수 있다.



<그림 3> 의미역 분석 결과를 확인하는 구문 구조도와 의미역 분석표

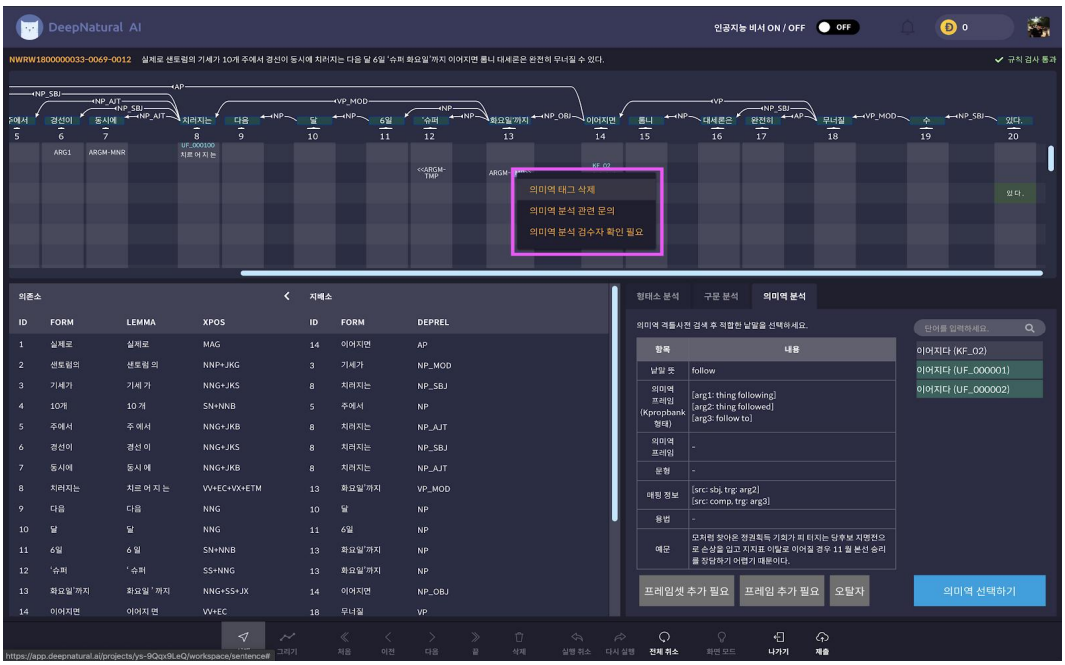
서술어에 대한 의미 번호가 올바르게 부착되었다면 다음으로는 의미역 주석이 올바른지 확인한다. 검수 도구에서 의미역 주석을 확인하고 수정하는 기능은 <그림 3>에 표시된 상단 영역에서 제공하고 있으며, 구문 구조도의 아래에 보이는 의미역 분석표는 문장에 포함된 서술어의 개수와 문장을 이루는 어절 개수에 따라 행과 열을 구성한다. 검수자들은 의미역 분석표에서 서술어(행)와 명사구(열)가 교차하는 칸에 표시된 의미역 주석을 확인하고 올바르지 않다면 주석 추가, 수정, 삭제 기능을 통해 오류를 수정한다.

5) 우리말샘(https://opendict.korean.go.kr/main)은 국립국어원에서 제공하는 개방형 국어사전으로, 본 사업에서는 [2019. 11. 15.] 기준으로 73개 표제어에 대한 격틀 정보를 참고하였다.



<그림 4> 의미역 주석 선택 기능

서술어 행과 명사구 열이 교차하는 칸을 선택하면 <그림 4>와 같이 의미역 주석을 선택할 수 있는 창이 보이고, 특정 의미역 주석 단추를 눌러서 해당 어절에 부착할 수 있다. 연속된 어절에 대한 의미역 주석을 위하여 단독 주석(ARG*), 시작 주석(<<ARG*), 종료 주석(ARG*>>)을 지원한다.



<그림 5> 의미역 주석 오른쪽 누르기 기능 - 주석 삭제, 문의하기, 조장 확인 요청

의미역 분석표에서 부착된 주석에 오른쪽 단추를 누르면 <그림 5>와 같은 창이 나타나며 '의미역 주석 삭제', '의미역 분석 관련 문의', '의미역 분석 검수자 확인 필요' 기능을 사용할 수 있다. 수작업 검수 조원들은 검수 과정에서 질문이 필요하면 '의미역 분석 관련 문의'를 눌러 질문 또는 토론을 위한 문제 제기를 할 수 있다. 이 기능을 통해 새로운 논제를 등록하면 수작업 검수 분과 전원에게 알림이 전송되어, 어떤 문장에 대한 질문인지 확인 가능하며, 이에 대한 답변 또는 논의를 이어갈 수 있다. 이러한 소통 방식을 통하여 수십 명의 검수 분과 구성원이 효율적으로 지침에 대해 문의하고 논의하면서 체계적인 검수 과정을 추진할 수 있다.

<그림 6> 수작업 검수 분과장에게 전달되는 문의 내용과 답변 작성 예

<그림 7> 규칙 기반 유효성 검사

검수 도구는 본 사업의 지침에 최적화하여 적용되었다. 또한 <그림 7>을 통해 제시하였듯이 규칙 기반 유효성 검사 기능을 지원하게 설계되었다. 의미역 분석 결과의 수정이 이루어질 때 마다 실시간으로 유효성 검사가 실행되고 검수자에게 그 결과가 전송된다. 유효성 검사의 통과 유무에 따라 검수 도구의 '제출' 버튼이 활성화, 비활성화되며, 오류에 대한 검수자의 신속한 대응을 통해 분석 결과의 신뢰도를 높일 수 있었다.

<그림 8> 검수 진행 상황 확인 기능 활용

딥네추럴 인공지능 구조에서는 진행 상황 확인 기능이 있어서, 검수 조장과 운영진이 진행 현황 확인 기능을 활용하여 진도를 관리하였다. 검수자별 문장 할당 개수, 검수자별 문장 검수 완료 개수 등을 실시간으로 확인할 수 있어서 작업의 효율적 관리가 가능하였다. 이러한 통계 도구의 활용으로 관리 효율성이 증가함에 따라 검수 분과는 의미역 분석 말뭉치 구축에 더 집중할 수 있게 되었다.

2.3. 작업자 교육

의미역 분석 말뭉치 구축의 세 번째 단계는 작업자 교육 단계이다. 실제 의미역 분석 작업자를 대상으로 의미역 분석 지침, 말뭉치 자료의 이해, 수작업 검수 도구 사용법에 대한 교육을 실시하였다.

의미역 분석 지침 숙지뿐 아니라 실제 말뭉치를 대상으로 직접 검수 작업에 참여해 보는 것이 중요하므로 작업 교육 중에 작업 시연과 실습을 실시하였다. 또한 실제 분석 작업을 중심으로 작업 워크숍을 진행하였으며 이를 통해 발견한 지침상의 수정·보완 사항을 반영하여 의미역 분석 지침을 더욱 세밀하게 보완하였다. 또한 의미역 분석 말뭉치의 질적 향상을 위하여 작업자의 빈발 오류 등을 바탕으로 사업 기간 중 재교육을 지속적으로 실시하였다. 아래는 본 사업 중 진행된 교육 일정이다.

교육 회차	날짜	대상	교육자	비고
1차	8/25-26	연세대 작업자	김선혜, 이찬영, 박혜진	
2차	10/25	연세대 작업자	김선혜, 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 조별 회의, ‘슬랙(Slack)’ 과[6] 수작업 검수 도구를 통한 문의, 카카오톡 그룹채팅방 등을 통한 문의 등 다양한 채널로 작업자들의 문의 사항이나 오류 보고를 수집하여 회신하였으며, 각 조의 조장은 각 조원의 분석 내용을 주기적으로 검수하여 작업자별 검수 결과를 제공하였다.

6) ‘슬랙(Slack)’은 기업에서 여러 구성원이 효율적으로 업무를 할 수 있도록 하는 서비스이다.

2.4. 자동 의미역 분석

의미역 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 엑소브레인의 언어분석기(WiseNLU), 전북대학교 딥러닝 기반 자동 의미역 분석 모형, 강원대학교 딥러닝 기반 의미역 분석기를 사용하고 분석 결과를 종합하는 방식을 이용하여 자동 의미역 분석의 정확도를 높이고자 하였다.

2.4.1. 엑소브레인 언어분석기(WiseNLU)

WiseNLU는 엑소브레인 사업을 통해 ETRI에서 개발한 언어 분석 엔진이다. 의미역 분석을 위한 선행 작업으로 형태소 분석과 개체명 분석, 어휘 의미 분석, 구문 분석이 진행된다.

WiseNLU의 형태소 분석기와 개체명 인식기는 국립국어원의 세종 말뭉치 자료와 ETRI의 15개 대분류, 146개 세부 분류 자료를 이용하여 SSVM(Structural Support Vector Machine)[7] 기반 음절 단위 분류 기술, 경계와 대분류 인식 기술, 세부 분류 기술로 만들어진 엔진들이다. 전처리 사전, 후처리 패턴 기반 사전으로 추가적인 성능의 개선이 이루어졌다. 어휘 의미 분석은 고빈도 의미 기반과 공기 정보 기반으로 진행하였으며, 이 과정에서 818만 어절의 동음이의어, 377만 어절의 다의어 주석 말뭉치를 사용하였다.

구문 구조 분석과 의미역 분석은 한국어 문법에 기반하여 문장 구조를 분석하였으며, SSVM(Structural Support Vector Machine) 기반이다. <그림 9>는 WiseNLU의 성능 지표를 나타낸 것으로 본 사업에서는 2만 어절에 대한 샘플 말뭉치를 WiseNLU로 분석하여 우선 제출하였고, 이후 최종 규모인 200만 어절의 언어 분석 결과를 제출하였다.

7) 복잡하고 구조화된 결과를 데이터 마이닝 기법 및 인공지능에 쓰이는 분류 알고리즘을 이용하여 예측하는 방법론

<그림 9> WiseNLU 성능 지표

고품질 의미역 분석 말뭉치를 구축하기 위하여 본 사업에서는 기계 학습 기반의 자동 의미역 분석 결과를 활용한다. 검수자는 자동 의미역 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 향상시키기 위해 5가지 형태소 분석기(ETRI, 강원대, 전북대, 충남대, 한국해양대), 4가지 구문 분석기(ETRI, 강원대, 전북대, 충남대), 3가지 의미역 분석기(ETRI, 전북대, 강원대)의 자동 분석 결과를 활용하였다. 분석 결과 통합 연산 방식은 여러 기관의 자동 분석 결과를 비교하여 가장 신뢰도가 높은 분석 결과를 도출하는 형태로 하나의 최종 분석 결과를 생성한다.

통합 연산 방식은 다수 분석 결과 신뢰 규칙에 기반을 두고 있으며 판단이 어려운 경우에는 ETRI 분석기의 결과에 높은 가중치를 두고 있다. 검수자들은 통합 연산 방식을 통해 산출된 하나의 자동 분석 결과뿐만 아니라 각각의 분석 엔진 결과도 참조하면서 수작업 검수를 진행했다. 최적화한 검수 도구는 이러한 자동 의미역 분석 결과를 한눈에 확인하고, 분석 결과에 포함된 오류를 쉽게 수정하는 과정의 효율을 극대화하였다.

2.4.2. 전북대학교 딥러닝 기반 자동 의미역 분석 모형

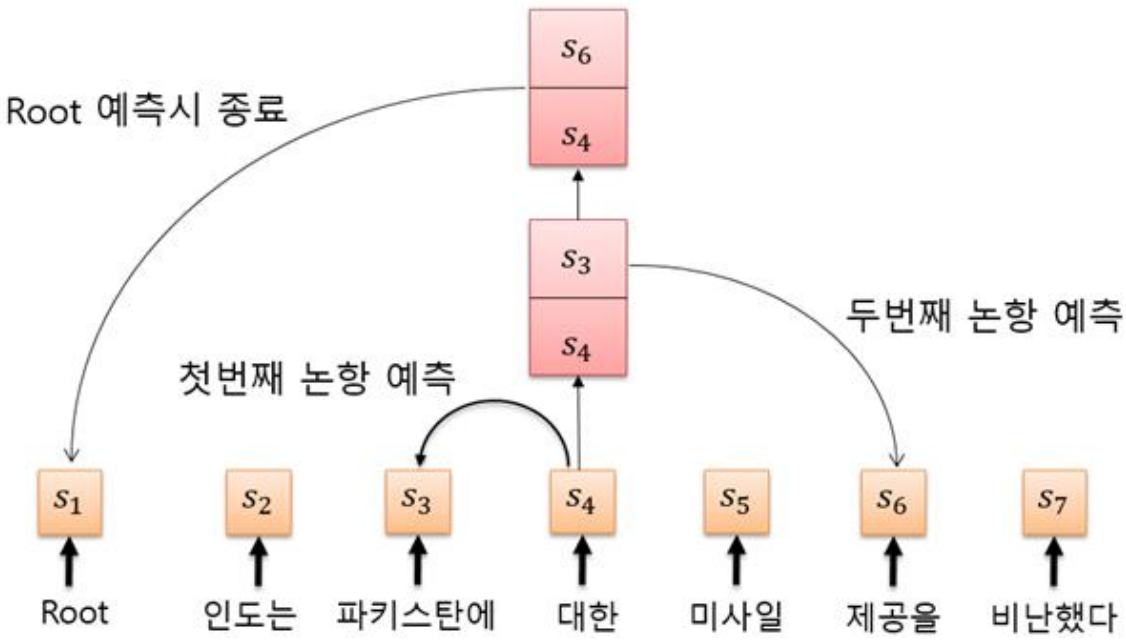
의미역은 서술어에 의해 기술되는 행동이나 상태에 대한 명사구의 의미 역할을 뜻하며, 의미역이 부여된 각 명사구를 논항이라고 한다. 의미역 분석은 서술어 인식 및 분류 단계와 논항 인식 및 분류 단계로 이루어져 있다. 이러한 의미역 분석을 수행하는

작업자는 자동 의미역 분석 결과를 토대로 오류를 수정하게 되고 자동 의미역 분석 결과가 정확할수록 처리해야 하는 작업량이 줄어들기 때문에 높은 성능을 가지는 자동 의미역 분석 모델이 요구된다. 이를 위해 다양한 모델의 연구를 진행하였고 가장 높은 정확성을 가지는 모델을 자동 의미역 분석 모델로 사용하였다.

2.4.2.1. Attention 기반 방법: Pred2arg, Arg2pred

Attention[8] 기반 방법은 여러 번의 Attention을 수행하여 논항 또는 술어를 연속해서 예측하는 모형이다. 기존의 적재-신호 네트워크가 가진 병렬성 문제를 해결하여 병렬적으로 예측이 가능하다.

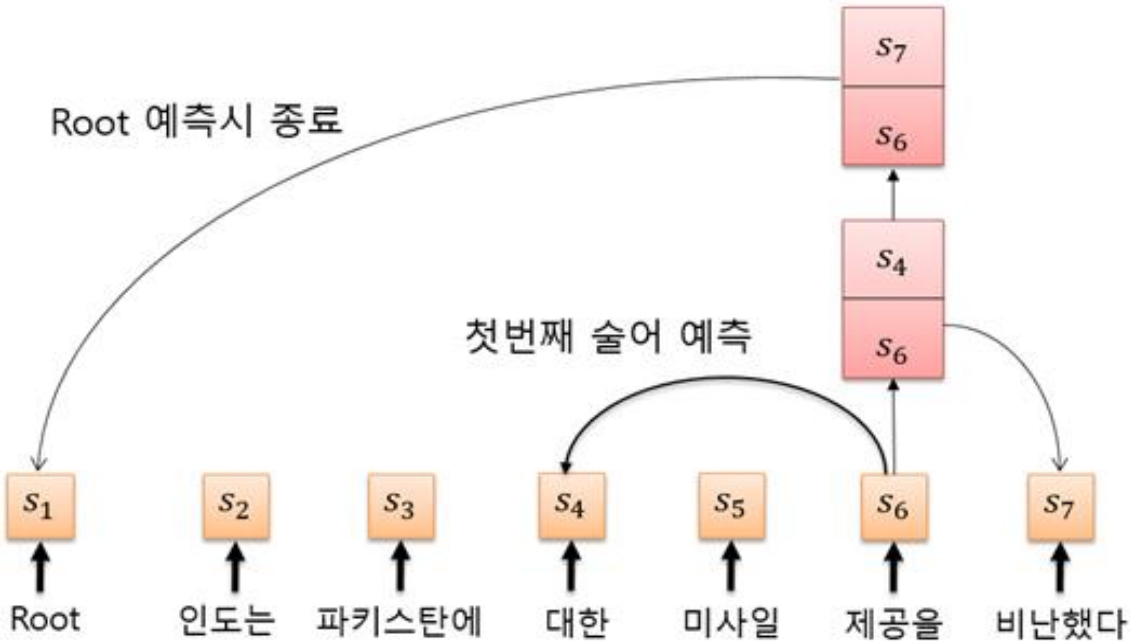
Pred2arg 모형은 주어진 술어에 대한 논항을 예측하는 모형으로 모든 술어에서 병렬적으로 논항을 예측한다. 문장 속에서 모든 단어가 술어라고 가정하고 술어에 대한 논항을 예측한다. 술어에 대한 논항이 없을 경우에는 논항 예측을 종료하고, 논항이 있을 경우에는 다음 논항 예측을 위해 현재 위치에 대한 표상과 예측한 논항 표상을 결합하여 새로운 표상을 얻어 다음 논항 예측을 위한 표상으로 사용한다. 모형은 술어에 대한 모든 논항을 찾을 때까지 수행된다. 아래 그림은 Pred2arg모형의 수행 과정을 보여주고 있는데 “대한[9]”이라는 술어가 있을 때 논항은 “파키스탄에”, “제공을”임을 알 수 있다. 논항의 예측은 문장에 먼저 등장한 순으로 예측이 이루어진다.



<그림 10> Pred2arg 의미역 결정 모델

8) 딥러닝 모델이 특정 벡터에 주목하게 만들어 모델의 성능을 높이는 기법
9) 그림 10 예문과 같은 맥락에서의 ‘대하다’는 의미역 분석에서 배제하는 대상으로 지침에 기술되어 있으며 시스템의 이해를 위한 단순 예시임.

Arg2pred 모형은 Pred2arg 모형과 반대로 주어진 논항에 대한 술어를 예측하는 모형으로, 모든 논항에서 병렬적으로 술어를 예측하는 모형이다. 아래 그림은 Arg2pred 모형의 수행 과정을 보여주고 있다. “제공을”이라는 논항과 관련된 술어는 “대한”, “비난했다”이고 문장에서 등장한 순서대로 “대한”, “비난했다” 순으로 예측을 진행한다.



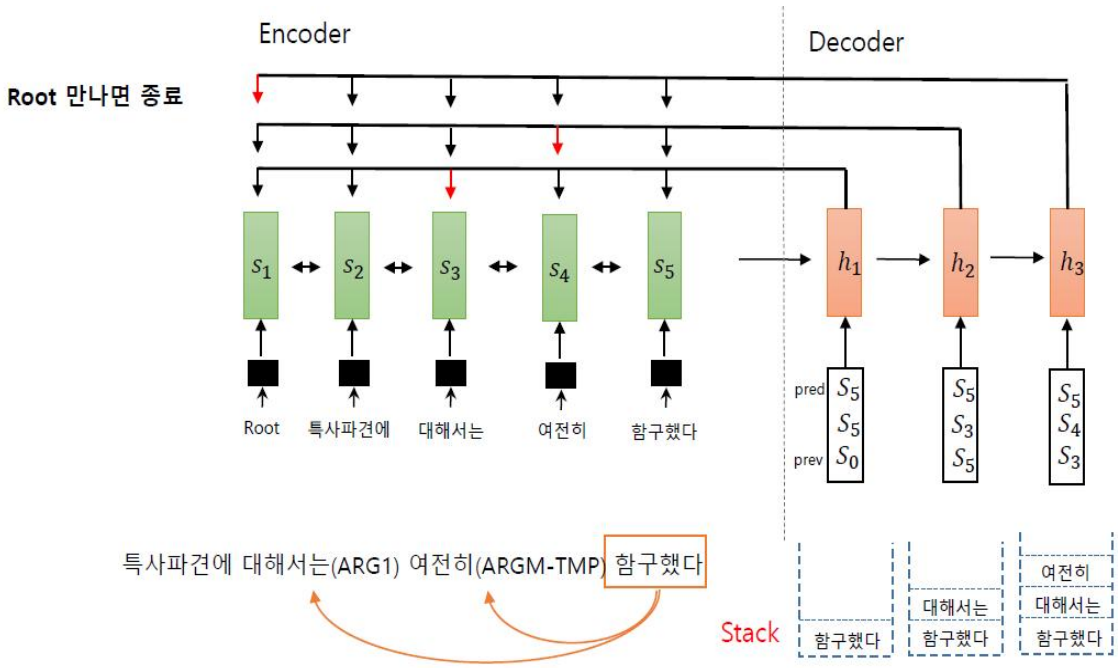
<그림 11> Arg2pred 의미역 결정 모형

2.4.2.2. 적재-신호 네트워크 기반 방법

적재-신호 네트워크 기반 방법은 RoBERTa[10]와 적재-신호 네트워크를 이용한 한국어 의미역 결정의 논문[11]에서 제안한 방법으로, 기존의 의존 구문 분석 모형인 적재-신호 네트워크를 이용하여 의미역 결정 문제를 해결하고자 하는 모형이다. 의존 구문 분석을 위한 적재-신호 네트워크 모형은 하향식으로 지배소에서 의존소를 예측하여 의존 구조도를 구성하는 모형이다. 위 모형이 가지는 의존 구조도를 의미역 결정 모형에 적용하기 위해 부모를 하나만 가지는 구조도를 구성하였다. 구조도의 최상위 요소는 술어가 되고 자식들은 술어와 관계된 논항들이 된다. “특사 파견에 대해서는 여전히 함구했다”라는 문장에서 “함구했다”는 술어, 논항들은 “대해서는”, “여전히”이다. 위 문장에서 구조도를 구성하면 “함구했다” → “대해서는” → “여전히”와 같이 부모가 하나인 구조도이고 논항 선택은 문장에서 등장한 순서대로 선택하였다. 복호기의 적재는 술어인 “함구했다”가 초기 상태로 되어 있고 자식인 “대해서는”을 예측

10) A robustly optimized bert pretraining approach
11) 홍승연, 나승훈 , 신종훈, 김영길, "RoBERTa와 스택-포인터 네트워크를 이용한 한국어 의미역 결정", 한국 정보과학회 동계 학술발표논문집, 2019.12

한다. 자식 예측을 위해 입력 표상과 복호기 표상에 어텐션 기법 중 하나인 Biaffine Attention을 사용하여 자식의 위치를 얻었다. 적재에는 예측된 논항들이 입력되며 술어를 예측하면 예측이 종료된다. 복호기의 입력으로는 현재 적재의 최상층 정보와 이전 적재의 최상층 정보를 활용한다.

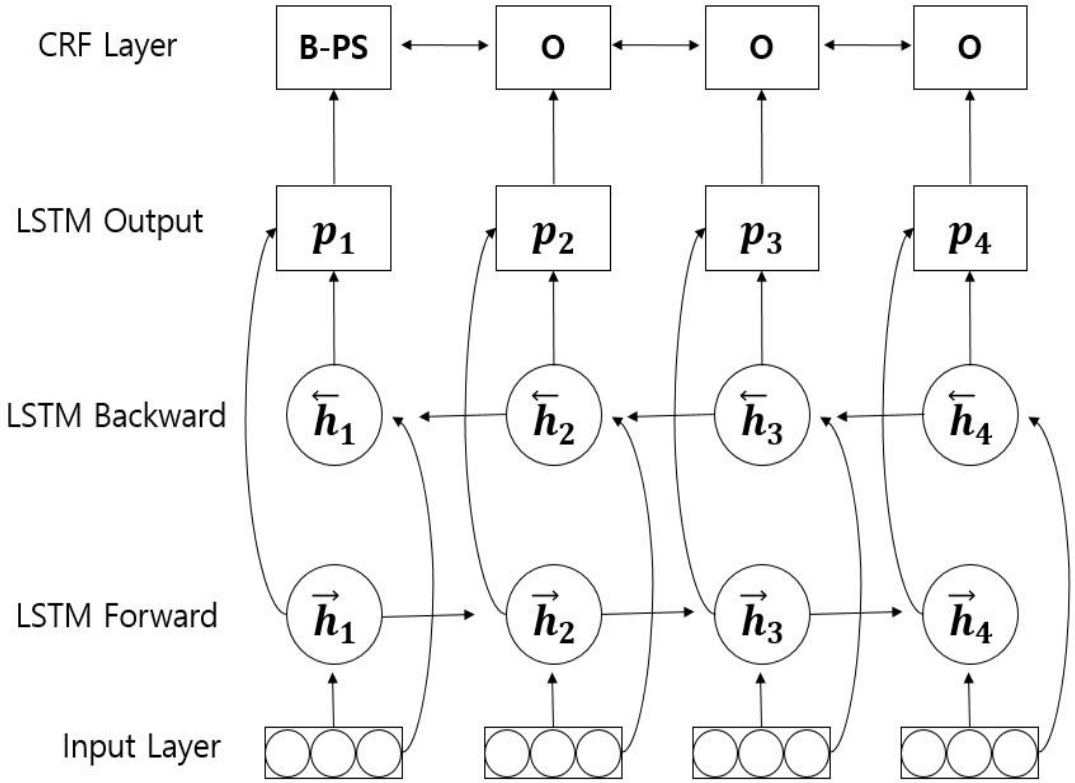


<그림 12> 스택-포인터 네트워크를 이용한 한국어 의미역 결정 구조

2.4.2.3. Bi-LSTM CRF[12] 기반 방법

Bi-LSTM CRF 기반 방법은 품사 주석, 개체명 인식, 의미역 결정과 같은 순차 주석 문제를 해결하는 데 주로 사용되는 모형이다. 순차 주석 문제는 현재의 의미역 표지 정보를 결정하기 위해 이전의 의미역 표지 정보가 큰 영향을 미친다. 기존의 Bi-LSTM은 출력층의 마디들 간의 의존성이 모형화되어 있지 않아 이전의 의미역 표지 정보를 사용하지 않는 문제를 가지고 있다. 이를 해결하기 위해 CRF(conditional random field)를 이용하여 출력층의 마디 간의 의존성을 반영할 수 있도록 CRF층을 추가한 모형이 Bi-LSTM CRF 모형이다. 아래 그림에서 볼 수 있듯이 기존의 LSTM Output층까지 존재하는 Bi-LSTM 모형에 CRF층을 추가하여 얻어진 의미역 표지 간 의존성이 반영되도록 한다.

12) Bi-Long Short Term Memory network Conditional Random Field



<그림 13> Bi-LSTM CRF 의미역 결정 모형

2.4.2.4. RoBERTa와 RoBERTa 기반 의미역 주석 결과

BERT[13]는 대용량 말뭉치를 이용하여 학습한 트랜스포머(transformer) 기반 언어 모형이다. 그리고 RoBERTa는 BERT를 개선한 언어 모형인데 BERT의 문장 예측 부분을 제거하고 역동적 마스킹을 도입하였다. 한국어 적용을 위하여 입력은 형태소-주석 단위로 하였으며 단어장에 없는 경우에는 BPE(Byte Pair Encoding) 단위로 토큰화하였다. 한국어 대용량 말뭉치로 사용한 것은 위키피디아 코퍼스이다.

의미역 분석 모형은 사전 학습된 RoBERTa의 마지막 층위값을 이용하여 각 어절의 마지막 형태소의 출력값을 다른 정보들과 결합하여 RoBERTa를 적용하는 방식으로 이루어졌다.

13) Bidirectional Encoder Representations from Transformer

원문
철수는 학교에 갔다
형태소 분석
철수/nnp 는/jx, 학교/nng 에/jkb, 갔/vv~EP 다.ef
의미역 결정 모델에서의 입력
[CLS] 나/np, 는/jx 학, 교, 에/jkb 갔/vv~ep 다.ef [SEP]

<그림 14> RoBERTa의 입력 예시

2.4.2.5. 최종 성능

실험 자료로 Korean PropBank의 뉴스와이어(Newswire) 학습 말뭉치 문장을 사용하였고 나무 구조의 말뭉치를 변환하는 도중 말뭉치에 오류가 있는 경우는 제외하였다. 전체 23,059문장 중 19,602문장은 학습 자료, 1,152문장은 개발 자료, 2,305문장은 평가 자료로 하여 학습과 평가를 진행하였다. 표에서 보여 주는 성능은 서술어 인식과 분류, 논항 인식과 분류 중 논항 인식과 분류 성능을 나타낸다. 성능 지표로 정확률(Precision)과 재현율(Recall)의 조화평균값인 F1 값을 사용하였다.

모델	개발 기준	평가 기준
RoBERTa + Bi-LSTM CRF	86.07%	85.00%
RoBERTa + StackPtrNet	85.61%	84.64%
RoBERTa + Pred2Arg	85.57%	84.70%
RoBERTa + Arg2Pred	85.65%	84.34%

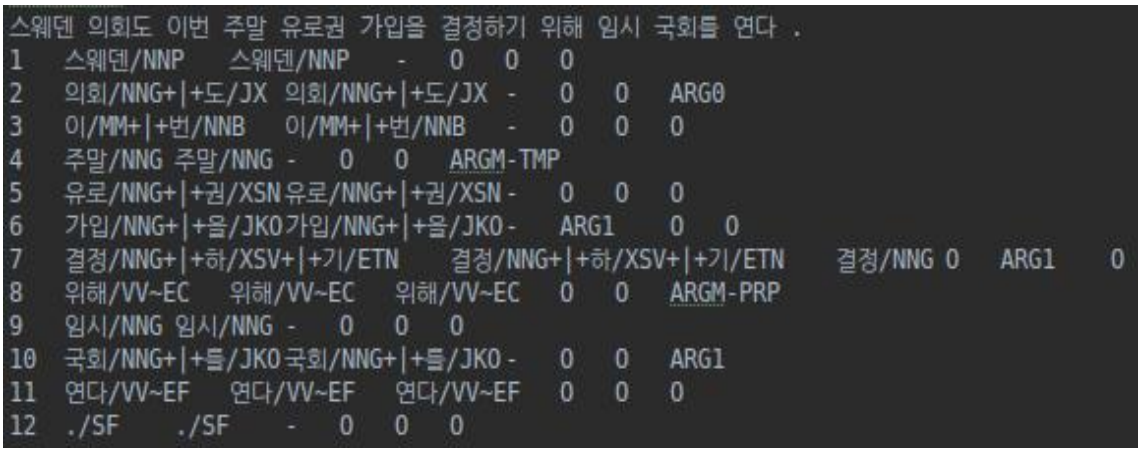
<표 3> 모형별 최종 성능

현재 가장 높은 성능을 보이고 있는 모형은 RoBERTa를 적용한 Bi-LSTM CRF 모형으로, F1 85.00%의 성능을 보이고 있다.

2.4.2.6. 출력 예제

문장이 주어졌을 때 형태소 분석을 실시하여 어절 단위의 형태소 분석 결과를 얻은 후에, 이를 모형에 적용하여 술어와 논항들을 결정한다. 결정된 술어를 표시하기 위해 “-” 대신에 술어를 출력하여 해당 어절이 술어임을 표시하였다. 아래 그림에서 술어는 3개임을 알 수 있고 순서는 “결정”, “위하”, “열”로 각 어절마다 논항 결정

시에 술어 순서에 맞추어 “O”, “O”, “ARG0”과 같이 술어와 논항과의 관계를 표시한다.



<그림 15> 출력 예제

2.4.3. 강원대학교 의미역 분석기

의미역 분석 말뭉치 구축 단계에서는 딥러닝(Deep Learning) 기반의 의미역 분석 도구를 사용하여 자동으로 의미역이 부착된 말뭉치를 우선 구축한다. 자동 의미역 분석 말뭉치는 한국어 의미역 분석에 대한 사전 교육을 받은 작업자들의 검수를 통해 더 정확한 의미역 말뭉치로 정제된다. 말뭉치 검수 작업의 효율과 정확도를 높이기 위해서는 자동 의미역 분석 도구의 높은 정확도가 요구된다. 의미역 분석에 있어서 구문 분석 정보는 정확도 향상과 직결되는 중요 자질이다. 따라서 본 사업에서는 구문 분석 정보를 자동으로 습득하여 의미역 분석에 활용하는 구문 의미역 통합 분석 도구(강원대 구문 의미역 분석기)를 사용하였다.

강원대 구문 의미역 분석기는 의존 관계와 의미역 분석 말뭉치인 U-PropBank 말뭉치(약 13만 문장 규모)를 훈련에 사용한 딥러닝 기반 구문과 의미역 통합 분석기이다. 구문 분석기와 의미역 분석기가 연결되어 있기 때문에 구문 분석 정보를 의미역 분석기의 자질로 직접 사용하며 보다 정확한 의미역 분석이 가능하다. 의미역 분석기는 서술어 인식 단계와 논항 인식과 분류 단계를 거쳐 진행된다.

서술어 인식 단계에서는 언어 분석에 특화된 딥러닝 기법인 순환 신경망(Recurrent Neural Network, RNN)을 사용해 순차적으로 각 어절이 서술어에 해당하는지를 판단한다. 서술어 정보가 사전에 주어져야 했던 기존의 다른 분석기들과는 다르게 서술어 인식을 자체적으로 수행하기 때문에 더 효율적으로 자동 의미역 분석 말뭉치를 구축할 수 있다. 논항 인식과 분류 단계에서는 순환 신경망을 병렬로 설계하여 문장에 존재하는 모든 서술어에 대한 각 논항 열을 한 번에 출력한다. 문장에 존재하는 모든 서술어에 대한 논항 열을 독립적으로 분석하는 것이 아니라 병렬 처리를 통해 한 번에 분석

하기 때문에 대용량의 말뭉치도 비교적 빠른 속도로 처리가 가능하다는 장점이 있다. 결과적으로 의미역 분석기는 U-PropBank 말뭉치에 대해서 서술어 인식 단계에서는 99%의 F1 score를 보였으며 논항 인식과 분류 단계에서는 73%의 F1 score를 보였다.

<그림 16>은 의미역 분석기의 결과를 보인 것이다. 결과 양식은 의미역 분석에서 보편적으로 사용되는 양식인 코넬(conll) 양식을 따른다. 첫 줄은 분석 대상 문장이며 두 번째 줄부터 각 어절에 대한 의미역 분석 결과다. 어절에 대한 의미역 분석 결과는 탭 단위로 구분되며 순서대로 어절 순차 번호, 어절 어휘, 어절의 서술어 포함 유무 그리고 각 서술어에 대한 논항 분류이다.

```
국회가 2007년 남북 정상회담 대화록과 관련된 자료 일체에 대한 제출을 2일 의결한 데 이어 3일 국가기록원에 자료 제출 요구서를 보냈다.
1  국회가  -  O  O  ARG0  O  O
2  2007년  -  O  O  O  O  O
3  남북  -  O  O  O  O  O
4  정상회담  -  O  O  O  O  O
5  대화록과  -  ARG2  O  O  O  O
6  관련된  관련된  O  O  O  O  O
7  자료  -  O  O  O  O  O
8  일체에  -  ARG1  ARG2  O  O  O
9  대한  대한  O  O  O  O  O
10  제출을  -  O  ARG1  ARG1  O  O
11  2일  -  O  O  ARGM-TMP  O  O
12  의결한  의결한  O  O  O  O  O
13  데  -  O  O  O  ARG2  O
14  이어  이어  O  O  O  O  ARGM-CAU
15  3일  -  O  O  O  O  ARGM-TMP
16  국가기록원에  -  O  O  O  O  ARG2
17  자료  -  O  O  O  O  O
18  제출  -  O  O  O  O  O
19  요구서를  -  O  O  O  O  ARG1
20  보냈다.  보냈다.  O  O  O  O  O
```

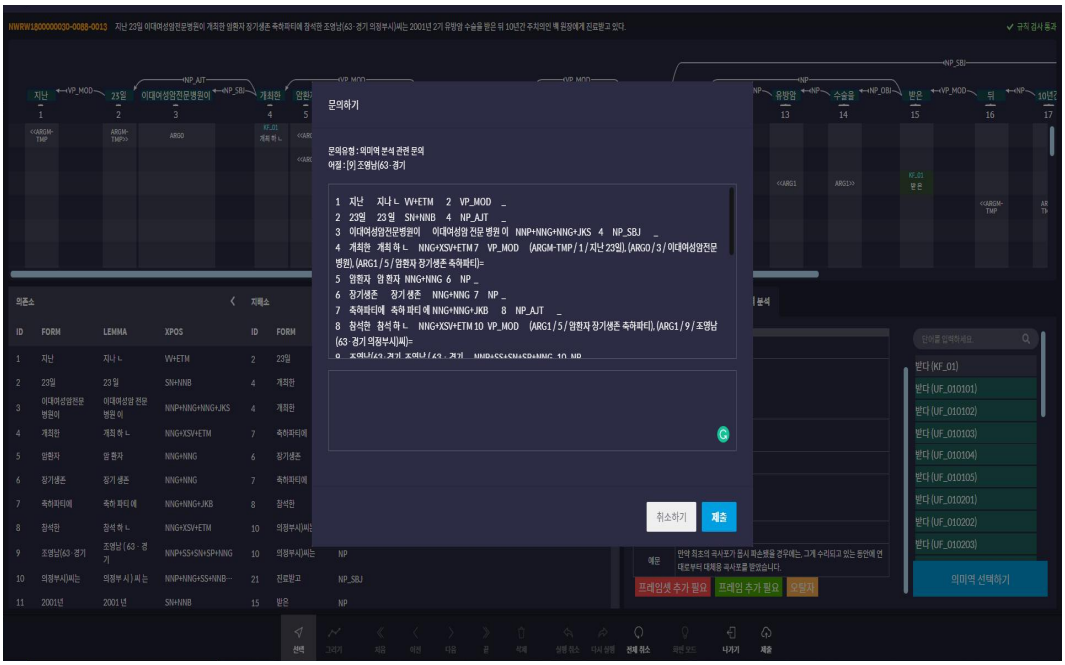
<그림 16> 의미역 분석 결과 예시

2.5. 작업자 분석(1차 검수)

자동 의미역 분석은 일관의 측면에서 분석 결과가 신뢰 가능하나, 100%의 정확도를 기대하기 엷다. 따라서 전문가의 수작업 검수 단계를 거쳐 이를 보완하고자 하였다. 본 사업에서는 자동 의미역 분석 결과를 수작업으로 전수 검수함으로써 의미역 분석 말뭉치의 질적 향상을 제고하였다.

자동 의미역 분석 결과는 웹 기반 작업 도구를 통하여 작업자들에게 문장 단위로 할당하였으며, 작업자들은 이 도구에서 본인에게 할당된 문장의 의미역 분석 결과를 직접 검수하여 수정하는 방식으로 검수 단계를 진행하였다. 자동 분석 결과에 문제가 있는 경우(의미역 정보 적용이 애매한 경우 등) 작업 도구를 활용하여 보고하였다. 의미역 분석 오류 보고 화면은 다음과 같다.

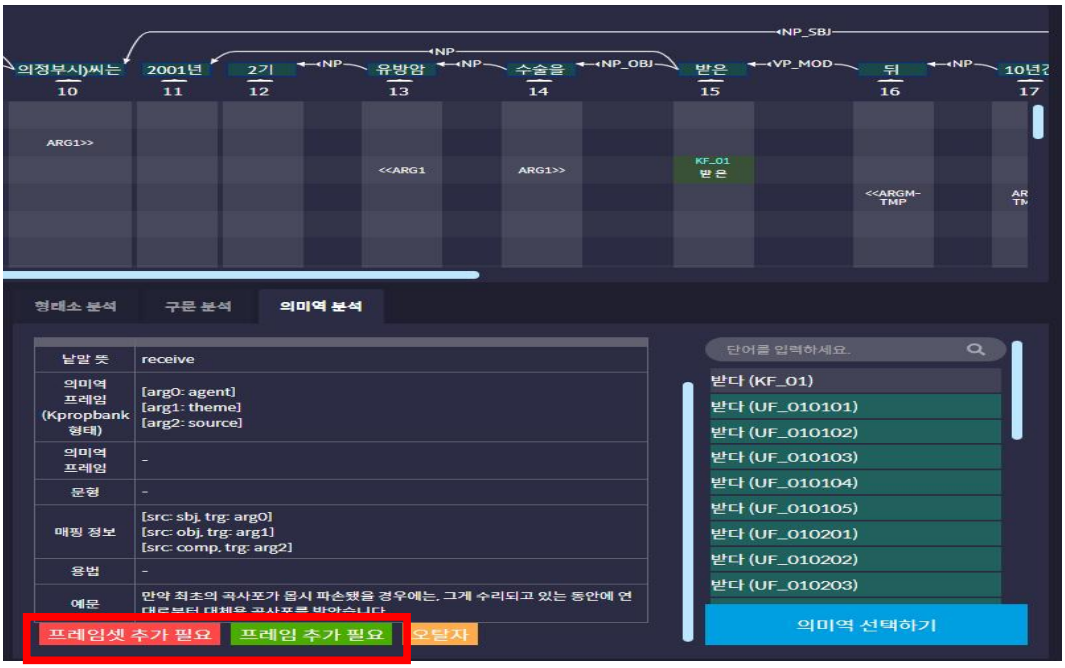
<그림 17> 의미역 분석 문의 화면 1



<그림 18> 의미역 분석 문의 화면 2

작업자들은 총 4개의 조로 편성하고, 각 조를 담당 공동연구원 1명, 조장 1명, 작업자 7~8명으로 구성하였다. 조장을 포함한 작업자들은 작업 도구를 통하여 각 문장에 대한 의미역 자동 분석 결과를 검수하며 특히 Korean PropBank 의미역 정보, ETRI Frameset 의미역 정보, U-PropBank 의미역 정보를 순차적으로 적용하여 이를 바탕으로 의미역 분석을 진행하였다. 이 과정에서 기존 의미역 정보에 없는 용언이나 해당 의미가 없는 의미역 정보가 나타나는 경우에는 보고할 수 있도록 하였으며 이는 조장

급 검수자를 통해 보완되었다. 아래는 작업 도구에서 의미역 정보 관련 보고를 접수하는 화면이다.



<그림 19> 의미역 정보 추가 보고 화면

또한 작업 도구의 오류나 작업의 편의를 위하여 필요한 개선 사항이 있으면 작업자들과 조장이 보고하여 도구에 반영될 수 있도록 하였다. 의미역 분석 작업 중에 발견되는 형태소 분석 오류, 원시 말뭉치상의 오류 등도 검수 도구 내에서 수집하여 국립국어원에 보고하였다.

2.6. 조장과 공동연구원 검수(2차 검수)

조장과 조별 담당 공동연구원은 각 조원의 검수 결과를 수시로 검수하였다. 이때 각 조장과 담당 공동연구원의 전체 작업물 검수는 표본 추출을 통하여 시행하였다. 검수 결과를 바탕으로 담당 조원에게 개별 피드백을 하고 이를 통하여 의미역 분석 말뭉치의 질과 일관성 향상을 도모하였다. 이때 각 분과에서 자주 보이는 오류 유형은 조장들이 수합하고 공유하여 정리한 후에 재교육을 진행하였다.

1차 검수 과정에서 수집된 질문이나 분석 상의 어려움은 지속적으로 지침에 반영하여 미침을 보완하였고, 지침의 보완 사항은 각 조원의 개별 교육을 통하여 작업에 반영되도록 하였다. 또한 기계 처리를 통해 일관성 확보가 가능한 유형은 수집하여 자동 의미역 분석과 후처리에 반영함으로써 분석의 일관성 확보 및 검수 환경 개선에 도움이 되도록 하였다.

또한 1차 검수 과정에서는 의미역 정보가 없어 분석하지 못한 문장을 2차 검수 과정

에서 처리하였으며 이때 새로 추가한 의미역 정보를 작업 도구에 반영할 수 있도록 주기적으로 작업 도구를 개선하였다.

2.7. 딥러닝 기반 정확도 및 일관성 검증

본 사업에서는 자동 의미역 분석기를 통하여 분석이 이루어진 후 전수 수작업 검수를 거친 말뭉치를 딥러닝 기반의 의미역 말뭉치 검증 모형을 통해 정확도와 일관성을 확보하였다.

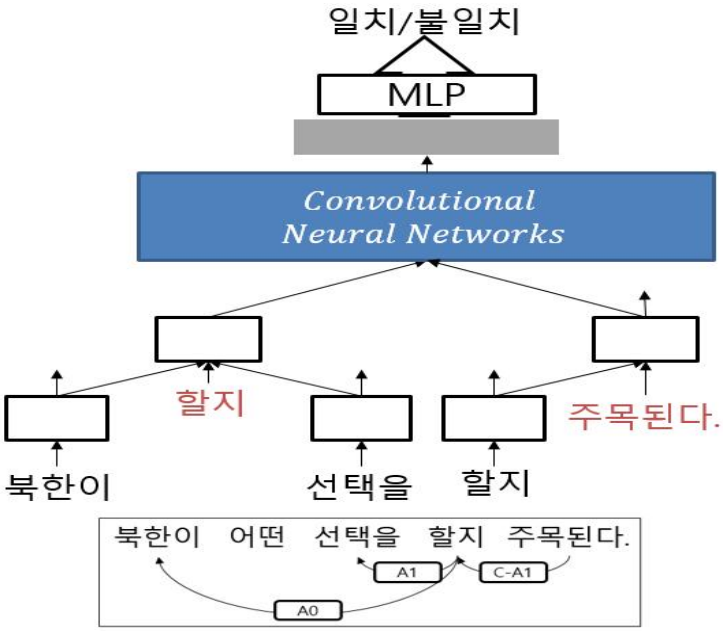
2.7.1. 전북대학교 딥러닝 기반 의미역 말뭉치 검증 모형

의미역 말뭉치 검증의 목표는 수작업 구축 자료나 자동 주석된 결과에 대해서 문장별 또는 서술어-논항 단위의 주석 품질을 예측하도록 하는 딥러닝 기반의 의미역 주석 품질 검증/교정기를 개발하는 데에 있다. 이를 위한 딥러닝 기반 의미역 말뭉치 검증 방법으로 제안하는 것은 신경망 모형에 기반한 검증 모형과 베이지안 모형[14] 불확실성에 기반한 검증 모형의 두 가지이다. 그래프 뉴럴 모형 기반 검증 모형은 뉴럴 모형을, 베이지안 모형 불확실성 기반 검증 모형은 MC Dropout 등을 이용한다. 그리고 모형의 결과가 어느 정도의 불확실성을 보이는지에 대한 이른바 '신뢰도'는 반복적인 표준 추출을 통한 결과의 일관성으로 측정한다.

2.7.1.1. 문장 단위 검증 모형

Tree LSTM의 문장 단위의 검증은 문장 내에 주어진 서술어에 대해 논항이 전부 올바르게 찾았는지에 대한 여부를 판별하는 검증 모형으로, 의미역 말뭉치 검증은 Tree LSTM에 기반한 검증 모형을 제안한다. 제안하는 방식은 아래의 그림과 같다.

14) 한 환경에 분포되어 있는 여러 변수들의 상태를 관찰하고, true 혹은 false 값으로 나눠 확률을 계산하는 모형

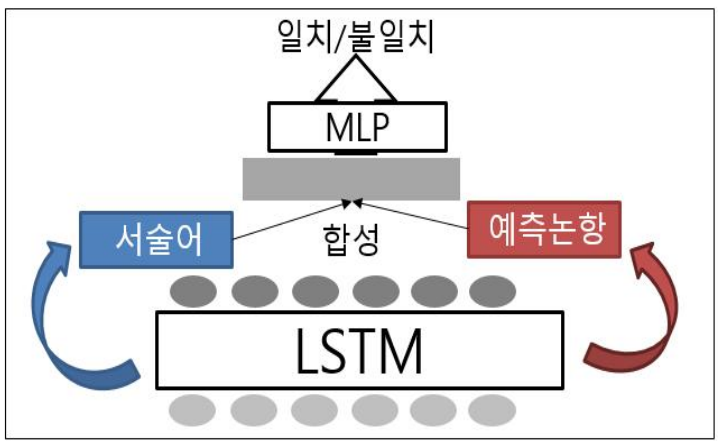


<그림 20> Tree LSTM

먼저, 문장 내의 각 서술어에 대해서 서술어를 부모, 논항들을 자식들로 하는 구조도로 구성한 후 Tree LSTM을 통해 입력하여 서술어-논항 표상을 얻고, 얻어진 문장 내의 모든 서술어-논항 표상에 대해서 CNNs(Convolutional Neural Networks)을 이용하여 단일 벡터로 합성하는 과정을 거친다. 얻어진 단일 벡터를 MLP를 통해 출력층으로 한 후 출력층에서 판별한다.

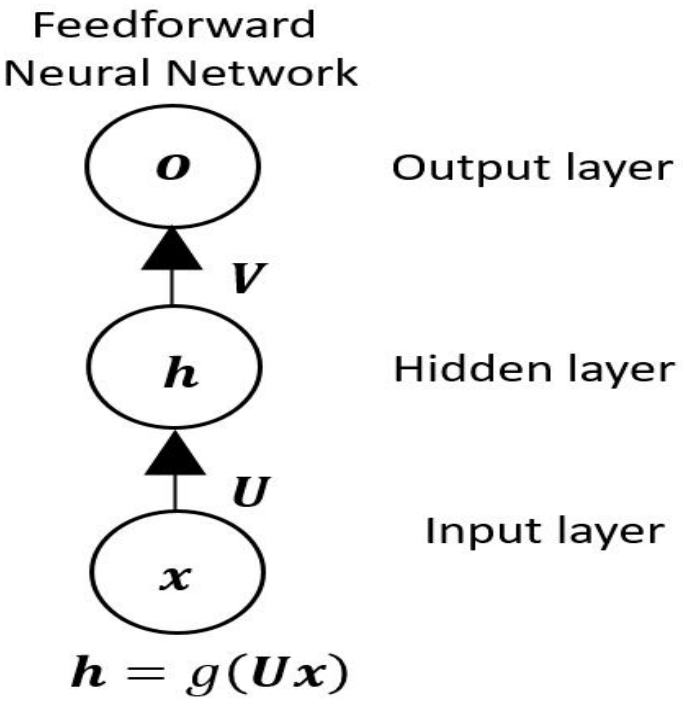
2.7.1.2. 서술어-논항 단위 검증 모형

$$\bullet SFU(x,y)$$
$$\tilde{x} = relu(W_r[x;y;x \circ y;x-y;])$$
$$g = \sigma(W_g[x;y;x \circ y;x-y;])$$
$$o = g \circ \tilde{x} + (1-g) \circ x$$



<그림 21> 서술어-논항 단위 검증 모델

FNN(Feed forward Neural Networks)은 가장 기본적인 인공 신경망으로 다음 그림과 같이 d차원의 입력 벡터 x와 학습 가능한 가중치 U에 대해 행렬 연산을 수행한다. 행렬 연산 후 활성함수(activation function) g를 적용하게 되는데 이 활성 함수는 비선형 함수로 가중치가 적용된 벡터에 대해 비선형 변환을 적용하여 은닉벡터 h를 얻는다. FNN은닉벡터에 대해 한 번 더 가중치 V 행렬 연산을 수행하여 출력 결과 o를 얻는 구조를 가진 신경망이다.



본 연구에서는 연결형 FNN을 통해 하나의 벡터로 합성하게 되는데 지배소에 대한 표상 h와 의존소에 대한 표상 m을 연결한 후 FNN을 통해 비선형 변환을 수행하여 단일벡터로 합성한다.

SFU(Semantic Fusion Unit)는 두 벡터를 합성하는 함수[15]로 두 벡터 x,y에 대해 출력 벡터 o를 얻는 연산($o=fusion(x,y)$)으로 다음 수식과 같이 얻어진다.

$\tilde{x}=\tanh(W_r[x;y;x\circ y;x-y])$

$g=\sigma(W_g[x;y;x\circ y;x-y])$

$o=g\circ\tilde{x}+(1-g)\circ x$

여기서 σ는 sigmoid 함수로서 이를 적용하여 얻어진 g는 업데이트한 결과인 $\tilde{x}$을 기존의 x에 얼마나 업데이트할지를 결정하는 게이트의 역할을 수행한다. 본 연구에서는

15) Minghao Hu, et al. IJCAI '18 제안 참조.

지배소에 대한 표상 h와 의존소에 대한 표상 m을 $fusion(h,m)$을 통해 하나의 합성된 출력 벡터 o를 얻는다.

본 연구에서는 MC Dropout(Monte Calro Dropout)을 적용하여 모형이 불확실성을 가지게 되어 표본 추출을 통해 신뢰도를 측정하는 방식으로 베이지안 모형을 도입하였다. 드롭아웃은 모델의 과적합(overfitting)을 방지하기 위해 일정 비율의 뉴런들의 값을 0으로 바꾸어 값이 전달되지 않도록 하는 학습하는 기술이다. MC Dropout은 드롭아웃을 평가 단계에 적용하여 표본을 추출하는 방법으로 사후 확률을 근사 추정하여 모델의 불확실성을 예측하는 기술이다. 이는 학습 모델을 변경시키지 않고 적용할 수 있다는 장점을 가지고 있다. 본 연구에서는 학습된 구문 분석 모델에 대해서 평가 단계에서 일정 비율의 드롭아웃을 적용하여 n개의 표본을 생성한 후 얼마나 동일한 결과가 나오는지에 대한 신뢰도를 50%, 60%, 70%, 80%, 90% 단위로 측정하여 그중 최대한 많은 문장을 포함하면서 일정 수준의 품질을 보장할 수 있는 신뢰도를 선택한다.

2.7.2. 해양대학교 의미역 말뭉치 검증 모형

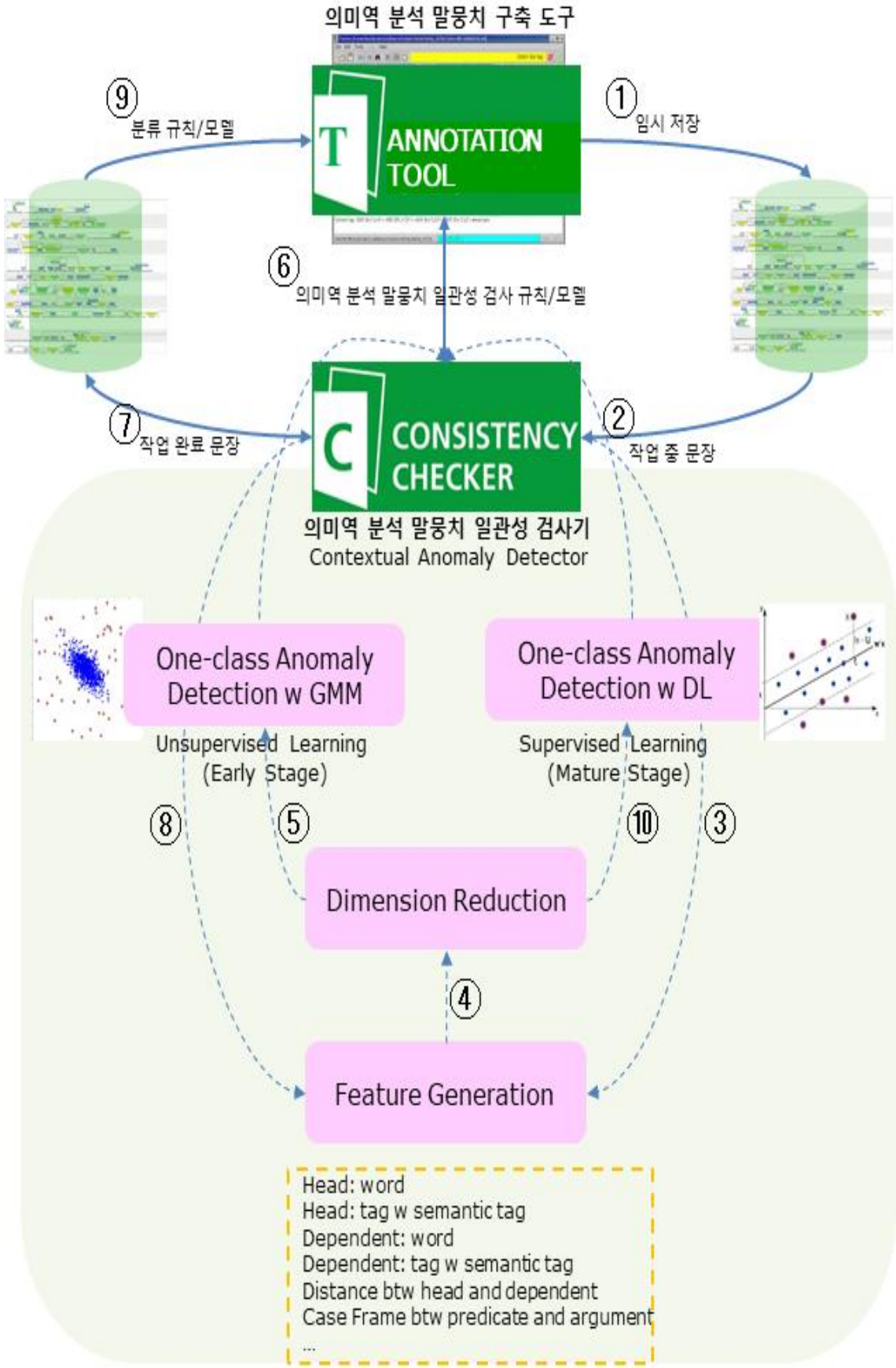
의미역 말뭉치 검증의 목적은 첫째, 말뭉치의 신뢰성과 일관성 확보, 둘째, 말뭉치 오류의 최소화에 있다. <그림 22>는 의미역 말뭉치 검증 시스템의 구조를 보인 것이다. 검증은 총 일곱 단계이다.

2.7.2.1. 수동 말뭉치 구축

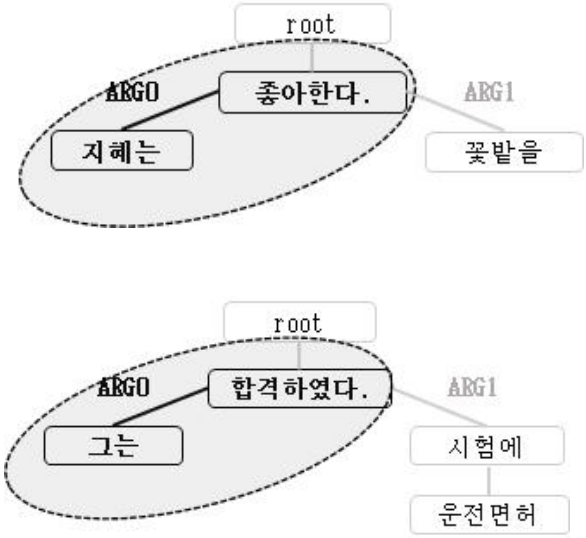
초기 학습 시에는 말뭉치의 신뢰성을 담보하고자 전문가가 직접 의미역 분석 말뭉치를 구축한다. 구축은 구축 도구를 통해 이루어지지만, 분류 모형에 의한 의미역 오류 후보를 제시하므로 초기 학습 이후 말뭉치 구축의 생산력 향상을 도모할 수 있다.

2.7.2.2. 자질 생성

의미역 분석 말뭉치의 신뢰성 확보에는 경험적 지식에 근거한 자질 선택과 생성으로 도모한다. 자질의 추가는 평가 자료 집합에 대한 추론 이후 일관성 검사기에서 나타나는 의미역 오류 분석으로 이루어진다. 본 연구에서는 <그림 23>에서 보인 의미역 구조도를 바탕으로 <그림 24>에서와 같은 자질 추출을 통하여 하나의 덩어리 벡터를 생성하고 이 벡터는 검증 시스템의 입력으로 활용된다.



<그림 22> 의미역 말뭉치 검증 시스템의 구조도



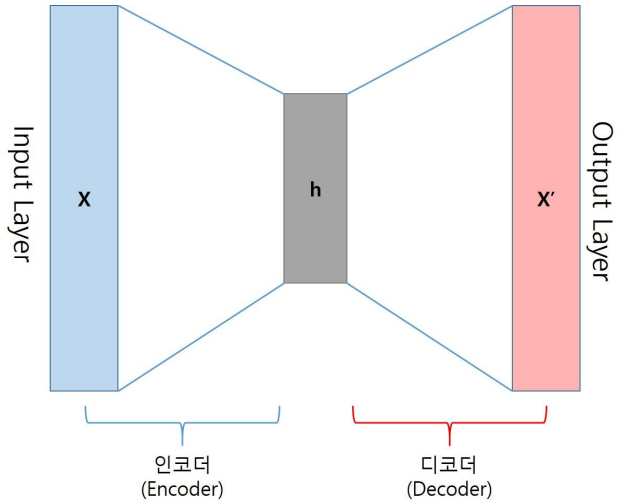
<그림 23> 의미역 트리의 예

<그림 24> 의미역 검증 시스템의 입력 자질 집합

<그림 23>과 같이 의미역 자질의 문맥 표상에서는 지배소의 단어($Word_{head}$), 지배소의 주석(Tag_{head}), 의존소의 단어($Word_{dep}$), 의존소의 주석(Tag_{dep}), 그리고 지배소와 의존소의 거리($Dis_{head|dep}$) 총 5개를 포함한다. 각 단어는 200차원의 크기를 가지고 주석과 거리는 30차원의 크기를 가진다. 결과적으로 의미역에서의 문맥 표상의 크기는 총 490이다. 이와 같은 방법으로 의미역 말뭉치 전체에서 같은 의미역을 가지는 구조도에 대해서 자질 벡터를 추출하며 자료로 사용한다.

2.7.2.3. 자질 축소

일반적으로 군집화의 복잡도는 자질 벡터의 수, 군집의 수, 자질벡터의 크기, 반복횟수(dp)의 해이다. 그러므로 자질 벡터의 차원이 크면 오류 후보 탐지에도 적지 않은 시간이 소요되기 마련이다. 이에 대한 해결 방안으로는 자질 벡터의 크기를 축소하는 것을 들 수 있다. 이 연구에서는 자기 부호화기를 이용한 차원 축소를 수행한다. <그림 25>는 이에 대한 구조를 보인 것이다.

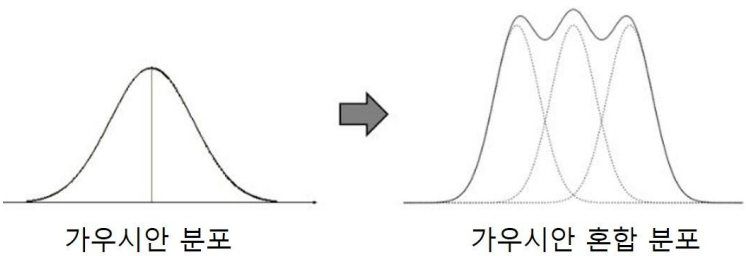


<그림 25> 자질 축소를 위한 자기 부호화기

<그림 25>에서 자기 부호화기는 입력층의 크기를 490, 은닉층의 크기를 100, 출력층의 크기를 입력층과 동일한 490으로 설정하여 학습시켜 사용한다. 이렇게 미리 학습한 자기 부호화기의 부호화 부분에서 실시간으로 문맥 표상의 차원 크기를 100으로 축소하여 사용한다.

2.7.2.4. 가우시안 혼합 모형(Gaussian Mixture Model, GMM)

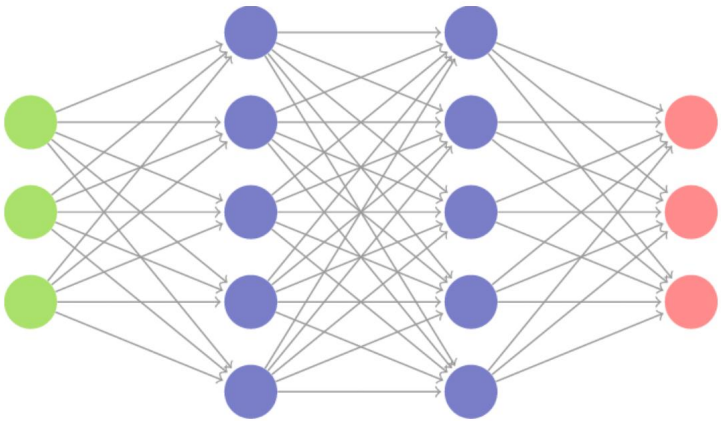
가우시안 혼합 모형은 <그림 26>에서 보인 바와 같이 여러 개의 가우시안 분포를 하나로 결합한 모형이다. 말뭉치의 양이 충분하지 않을 때 사용하는 효과적인 모형으로 군집화 연산의 대표적 방식 가운데 하나라고 할 수 있다. 군집화 연산 방식은 비지도 학습(unsupervised learning)으로 일관성 검사기에서는 하나의 군집만 사용한다. 이는 하나의 표지에 대한 하나의 모형을 대응하고 이렇게 대응된 각 모형의 입력 자질이 가우시안 혼합 분포에서 얼마나 벗어났는지의 여부로 오류 후보를 출력한다.



<그림 26> 가우시안 혼합 분포에 대한 설명

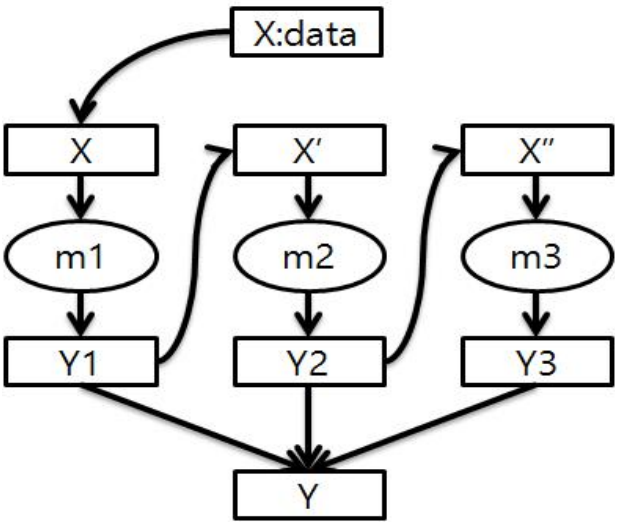
2.7.2.5. 심층 학습 분류 모형

이 연구에서는 지지 벡터(support vector)를 이용하여 오류를 추정하는 방법이 고려되었으나 지지 벡터보다 더욱 좋은 성능으로 알려진 심층 학습 모형을 적용한다. <그림 27>은 심층 학습 모형을 이용한 의미역 분류의 예를 제시한 것이다.



<그림 27> 심층 학습을 이용한 의미역 분류 모형

오류 검증에 하나의 분류 모형만을 적용하는 것은 성능과 품질면에서 만족스러운 결과를 기대하기 어렵기 때문에 이 연구에서는 앙상블 모형을 적용하고자 한다. 이에 본 연구에서는 XGBoost[16] 모형을 채택하였으며 구조는 <그림 28>과 같이 나타낼 수 있다.



<그림 28> 오류 검증을 위한 XGBoost 모형

그러나 전체를 학습 말뭉치로 사용하므로 말뭉치 전체의 오류를 검증하는 데에는 효

16) 병렬 처리가 가능한 기계 학습 기법

율성을 확보하기 어려운 것이 사실이다. 이에 대한 효과적인 해결 방안으로 본 연구에서는 '교차 검증' 방식을 사용하였다. 이 방식은 말뭉치 일부를 학습 데이터로 삼고 그 나머지에 대한 오류를 검증하는 것이다.

구체적으로는 교차 검증 1을 통해 A와 B를 학습하여 다음 C의 오류를 검증하고 교차 검증 2를 통해서는 A와 C를 학습하여 B의 오류를 검증하며 교차 검증 3을 통해서는 B와 C의 학습, A의 검증과 같이 반복함으로써 전체 말뭉치에 대한 오류 검증을 수행하는 것이다.

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

<그림 29> 교차 검증을 도입한 오류 검증 예

2.7.2.6. 오류의 최소화

앞선 과정을 통해 충분한 말뭉치가 확보되었다면, 말뭉치 학습을 통해 분석 말뭉치 오류의 최소화를 도모하여야 한다. 이는 양질의 말뭉치를 축적함으로써 구축자 또는 검수자가 직접 검수하는 분량을 최소화하여 능률과 생산성의 향상을 모색하는 데 있어서도 상당히 중요한 작업이다. 이에 본 연구에서는 GMM(Gaussian mixture model)을 이용한 비지도 학습 보다는 심층 학습을 이용한 지도 학습 방법을 채택하고자 한다.

2.7.2.7. 결과의 분석

의미역 분석 말뭉치 검증에서는 오류의 수정이 유기적인 프로세스로 이루어진다. <그림 30>과 함께 보면 우선 오류로 판단될 경우 첫 번째 라인과 두 번째 라인에 sent_id와 원문이 입력되고 이하 각각 오류로 나타난 의미역 id, 단어 및 구문 분석, 주석 의미역, 정답으로 판단되어 제시된 의미역의 순서로 입력이 된다. 이때 각각의 의미역 우측으로는 의미역 판단 확률이 제시되며, 이 확률은 모두 더하면 1이 된다.

sent_id = NWRW1800000022-0318-0014
sent_id = 1990년에는 열 집 중 한 집이 채 안되던(9.0%) 1인가구가 2000년에는 일곱 집 중 한 집꼴(15.5%)로, 이제는 다섯 집 중 한 집꼴(20.2%)로 19년 만에 2배
26‘늘어난 늘어나 ㄴ (ARGM-TMP / 이제는 다섯 집 중 한 집꼴(20.2%) / 16 / 0/193) (ARGM-CND / 이제는 다섯 집 중 한 집꼴(20.2%) / 16 / 0.274)

<그림 30> 의미역 말뭉치 검증 결과 예시

2.8. 자문 및 전문가 집단 심층 면접

자연어 처리, 전산 언어학, 국어학 등 관련 분야 전문가와의 자문회의 및 연구진의 심층 면접을 통해 지침과 말뭉치 검토 결과에 대해 토론하고 의미역 분석 말뭉치 보완 방향에 대해 논의하였다.

2.8.1. 전문가 자문

1) 자문위원 명단

■ 남기심(연세대, 국어학)
■ 이기용(고려대, 전산언어학)
■ 옥철영(울산대, 자연언어처리)

2) 자문위원 의견
■ 남기심(연세대, 국어학)

I. 심리용언 구문의 의미역 태깅

1. “나는 지금 사과가 먹고 싶다.”의 ‘사과가’는 subject로 보고, “나는 지금 사과를 먹고 싶다.”의 ‘사과를’은 object로 보아야 할 텐데 이때 의미역은 동일하게 처리해야 할 것이다.

2. “나는 사과가 좋다”의 ‘나는’의 의미역이 ARG1라면 “나는 사과를 좋아한다.”의 “나는”도 ARG1로 할 수 있겠지만 “나는 지금부터 (열심히) 사과를 좋아하겠습니다.”의 ‘나는’의 의미역을 ARG0로 변별해야 할지 결정이 필요하다.

II. 중주어 구문의 의미역 태깅

“저 산이 소나무가 많다.”의 “저 산이”의 의미역은 ARG1인지 ARGM-LOC인지 결정이 쉽지 않다.
“이 붓이 글씨가 잘 써진다.”의 ‘이 붓이’의 의미역은 ARGM-INS가 될 수도 있다.

III. 소절(小節) 구문의 태깅

"그가 김춘식을 사위로 삼았다."의 '사위로'의 의미역은? '김춘식'과 '사위' 사이의 주, 술 관계를 고려할 필요가 있다..

■ 이기용(고려대, 전산언어학)

(1) 작업지침서가 인코딩 작업을 정확히 수행할 수 있도록 적절한 예시와 함께 명료하고 체계적으로 잘 작성되었음.

(2) 단, 이 작업지침서의 기본 틀에 이론적이면서 실제적인 심각한 문제가 있다고 고려되므로 관련 협력 기관과 충분한 논의가 있기를 강력히 제안함.

문제점: 논항 구조(argument structure)와 논항들의 의미역(semantic role) 또는 담화적 기능을 혼동하고 있음. 다음 것은 문법 기능 또는 담화 기능을 논항구조와 의미역에 혼합시키고 있음.

예:

- 보조 서술 (ARGM-PRD)
- 목적 (ARGM-PRP)
- 담화 연결 (ARGM-DIS)
- 부사적 어구 (ARGM-ADV)
- 부정 (ARGM-NEG)

제안:

가. Table 1을 보면, 첫째 칸에 의미역 표지로 ARG0, ARG1 등이 있고 이에 대한 정의가 둘째 칸에 있는데. 첫째 칸의 제목은 "논항 구조", 둘째 칸의 제목은 "의미역 표지"로 할 것을 제안.
그러면 Table 2의 [arg0:행동주] 또는 영어로 [arg0:agent]는 논항 구조와 그것의 의미역으로 해석됨.

나. 4.3 부가의미역도 일관성 있게 ARGM-LOC을 [argm:장소] 또는 [argm:loc] 등으로 수정하는 방법을 고려할 것.

다. 표지문자를 라틴자, 소문자로 할 것도 논의. *lowerCamel case*로 통일하면. 대문자를 피하는 것이 추세인 것 같습니다. hyphen (-) 같은 것도 피하고요.
예: ARGM-TMP 대신에 argM:tmp 또는 argMtmp, ARGM-LOC 대신에 argMloc 또는 argM:loc, 또는 단순히 argm:loc ISO-TimeML, ISO-Space도 lowerCamel을 받아들여서, <EVENT XML:ID="e2" .../>를 <event xml:id="e2" .../>로 바꾸었음.

맺음: 이 작업은 문장의 논항 구조(argument structure)와 논항들의 의미역 표지(semantic role labeling, SRL)를 동시에 처리하려고 하고 있음. 좋은 방법이라 생각됨. 단, 이런 점을 분명히 밝혔으면 합니다. 그래서 과제 제목/ 건명도도 "**논항 구조 및 의미역 분석 말뭉치 구축**"으로 바꾸면 좋겠습니다.

■ 옥철영(울산대, 자연언어처리)

<한국어 의미역 태깅의 제문제>

I. 사전의 문제

1. 한국어 의미역 주석에 한국어사전의 정보를 활용하는 것이 가장 바람직하나 현존하는 사전의 문형 정보가 의미역 주석에 최적화되어 있지 않다.

2. 표준국어대사전을 기준으로 살펴볼 때 문형 정보가 용례의 문형을 반영하는 데에 충실하여 일반적인 문형과 거리가 있는데 용례마저도 모든 용언에 제시되어 있는 것이 아니며 충실성을 판단하기 힘들다.

II. 문장 구조의 문제

1. 합성어가 띄어서 써서 두 개 이상의 용언으로 분리되어 있을 때 처리가 쉽지 않다.
2. 미등재어나 조사가 생략된 동사구의 경우 처리가 쉽지 않다.
3. 피동문, 사동문 등의 처리가 쉽지 않다.

III. 의존 관계의 문제

1. 본용언, 보조 용언의 연쇄가 일반적인 의존 관계와 달라 주의를 요한다.
2. 의사보조용언 구성의 의존 관계 설정과 의미역 주석에 유의해야 한다.

3) 주요 내용 및 과제 반영 여부

■ 심리 용언 구문의 주석 [수용]
- 통사 층위의 문장 성분과 의미역을 변별해야 한다.

■ 중주어 구문의 주석 [수용]
- 문장 성분이 주어로 나타난다고 하더라도 의미역은 도구로 주석할 수 있다.

■ 소절 포함 문장의 주석 [미수용 - 서술어 용언의 프레임셋 기준으로 주석]
- 문장 내에 소절에 해당하는 구성이 있을 때에 이를 의미역 주석에 고려해야 한다.

■ 논항 구조와 의미역, 담화 기능 등의 층위 변별
[미수용 - 본 과제에서 채택한 다국어에 적용된 주석 체계의 기본 틀 유지]

• 문법 기능, 담화 기능을 논항구조와 의미역에 혼합시키는 문제를 해결해야 한다.

■ 의미역 기준 정보의 부족 [수용]

① 한국어 의미역 주석에 한국어사전의 정보를 활용하는 것이 가장 바람직하다.

② 미등재어 등 기존 사전의 정보 부족을 보완해야 한다.

■ 의미역 기준 정보의 부족 [수용]

• 보조용언, 의사보조용언의 의미역 주석에 대한 의미역 주석에 유의해야 한다.

2.8.2 집단 심층 면접

1) 집단 심층 면접 참석 전문가 의견

■ 임수종(한국전자통신연구원, 의미역 분석 포함 자연언어처리)

1. 부가역 검증 대상 여부
  - PropBank의 numbered role(필수역)은 프레임이라는 기준이 있기 때문에 정답 유무가 객관적이지만 부가역의 경우 기준을 세우는 것이 굉장히 어려운 문제임
  - 부가역에 대해 검증을 할 경우 정답이 하나라고 가정을 하면 실제로는 오류가 아닌 경우를 오류로 판정할 수 있으니, 부가역에 대해 검증을 할 경우는 부가역의 이러한 특성을 반영하여 유연성을 갖아야 함

2. 사동주 주석
  - ~게 하다 와 같이 사동의 의미를 내포하는 경우라 해도 형태소 분석 결과 VX(보조용언) 으로 판정이 된 경우는 AUX로 태깅
  - 실제적인 의미를 파악한다는 관점으로 보면 형태소 분석 결과가 VX라고 되어 있더라도 사동의 의미를 파악하여 이를 의미 구조로 기술하는 것은 필요하다고 판단이 되지만, 이는 일반적으로 인정되는 '의미역 인식' 혹은 '의미역 결정'의 범위를 벗어나는 문제임
  - 좀더 복잡다단한 의미적인 분석 문제는 의미역 인식을 포함하는 semantic parsing 개념으로 접근이 필요함. 아래에 추가적으로 의견을 드리는 문제도, 실제로 제기된 문제가 틀렸다는 것이 아니라, **의미역 인식에서 다루는 범위를 벗어나는 문제**들이 상다수 포함되어 있음.

3. 붙여쓰기로 처리된 본용언+ 보조용언 구성의 AUX
  - 사동주 주석과 마찬가지로 형태소 분석 결과가 VX인 경우에는 AUX로 태깅하는 것이 지침임
  - ETRI에서 작업을 할 때는 초기에는 원문을 모두 오탈자가 없도록 수정하여 붙여쓰기로 인한 오류를 원천적으로 배제하였고, 후에 작업을 할 때는 오탈자를 포함하여 비문은 태깅 대상에서 제외하는 방향으로 지침을 정하였음
  - 그러나, 국립국어원 과제의 경우 문장을 선별하지 않는 것으로 알고 있으니 이러한 **비문 등에 대해 지침을 수립**하는 것도 필요하다고 판단됨.

4. 의사 보조 용언의 처리

- 있다, 없다, 같다, 하다, 되다 과 같은 용언은 원칙적으로는 VV/VA인 경우에는 predicate으로 태깅을 해야 하지만, 정보성이 없을 경우에는 이를 대상으로 하지 않고 '의사' 보조용언으로 간주함
- 현재는 '정보성이 없다'는 개념적인 지침만 존재하지만, 객관적이고 일관성이 유지되도록 지침을 수립하는 것이 필요함.

5. 중복역에 대한 의미역 주석
- 의미역 태깅에 있어서도 필수역(numbered role)은 1개의 predicate에 대해 중복으로 발생하지 않는 것이 원칙이지만, 구문분석 결과 이중 주어, 이중 목적어 구문인 경우에는 예외적으로 중복역을 허용함
- 구문적으로 중복역이더라도 의미를 따져서 이를 적당한 부가역으로 태깅할 수는 있겠지만, 이럴 경우 작업자의 주관이 개입되기 때문에 구문분석 결과 이중 주어/목적어에 해당하는 경우 의미역도 중복역을 허용하는 것이 오류를 줄 일 수 있는 방향으로 판단됨.

6. 의존명사 '간'이 포함된 명사구 처리
- 의존 명사 '간'이 포함되었다는 문제로 국한되는 것이 아니라 predicate에 대한 논항의 경계인식(span)의 문제로 실질적으로는 의미역 본질적인 문제와는 상관이 없지만, 지식으로써 전체 의미역 구조(perfect proposition)의 관점에서 논항에 대한 지식을 온전히 갖게 위한 보조 수단으로, 현재는 구문분석 결과를 이용하여 chunk를 구성한다는 관점으로 봐야 하며 의존 명사 '간의 경우로 국한하자면, 명사구의 일부이기 때문에 이를 분리하지 않고 같이 묶어야 함.

7. 어절 내 연속 동사구 처리
- 분리된 어절과 다르게 같은 어절내의 연속 동사구는 전체 동사가 사전에 등재된 경우 이를 채택하면 되지만, 그렇지 않은 경우에는 주요 의미를 갖고 앞 쪽을 선택하여 predicate으로 채택함.
- 어절이 분리된 연속 동사구의 경우는 작업자도 판단이 어렵지만, 후행 용언에 대해 직접적인 구문 관계를 파악하기 어렵기 때문에 기계학습 기반의 의미역 인식 엔진에서도 난이도가 높은 유형에 해당함.

8. ARGM-ADV, EXT, NEG 처리
- 위의 부가 의미역은 기준을 세우기 명확하지 않고 정답이 있기보다는 선택의 문제인 경우가 많아서 ETRI에서는 이를 목록화하여 처리하고 있음
- ADV 의 경우에는 목록에 있는 것만 태깅을 하고 있으며, EXT, NEG 의 경우는 기계학습 엔진이 목록에 있는 것을 인식하지 못 하는 경우에 후처리 개념으로 태깅하고 있음.
- 관련하여 목록은 정리하여 작업자가 일관되게 태깅해야 함.

9. 의미역 주석 시 스팬 설정
- ETRI의 의미역 주석 스팬을 설정은 질의-응답 시스템 활용 시 적절한 응답을 추출하기 위한 것임.
- ETRI의 의미역 주석 지침의 '의존 구문분석 결과 기반 구/절 인식 가이드라인'은 의미역 주석을 위해 통사적 정보를 참고하기 위한 용도임.
- 통사적 정보에만 의존한 스팬 정보는 실제 응용 시스템에 활용도가 떨어짐.
- ETRI의 경우 질의-응답을 전제로 하여 질의에 대해 자연스러운 응답으로 제시할 수 있는 의미역

정보의 스팬을 따로 주석하고 있음.

■ 홍문표(성균관대, 전산언어학)

현재 구문 분석 지침에서는 기본적으로 선행절의 서술어에 의존하는 것으로 분석하게 되어 있는데, 이는 의미역 분석 지침 부분과 일치하지 않는 측면이 있다. 의미역 분석 지침에서는 후행절의 서술어의 논항으로 주어를 분석하게 되어 있기 때문이다. 이러한 문장에서는 문장 맨앞의 주어를 후행절의 서술어에 의존하는 것으로 분석하고 선행절의 주어는 생략된 것으로 보아 차후에 주어 복원을 할 수 있게 처리하는 것에 대한 논의가 필요해 보인다.

한편 용언으로 된 인용절의 마지막 어절과 인용동사의 활용형이 하나의 어절로 줄어든 꼴을 분석할 때 어떤 용언의 논항인지 표시할 길이 없다. 예를 들어 '느껴진다"면서'는 '느껴진다고 하면서'가 줄어든 꼴로 보고 있는데(구문 분석 지침 3.2.1.3.), 의미역 분석을 할 때 '느껴진다'의 논항과 '하면서'의 논항을 구분하여 표시할 수 없다. '경기 뒤 "~ 피곤함도 느껴진다"면서'와 같은 문장에서 '피곤함도'는 '느껴진다'의 논항(ARG1)이고 '경기 뒤'는 '하면서'의 부가어(ARGM-TMP)인데 현재로서는 이들을 한 줄에 나란히 표시할 수밖에 없다.

■ 유혜원(단국대, 통사론/구문 분석)

1. 의미역 분석 관련 논의 사항

1.1. 사동주 분석

연구팀은 '-게 하다'의 사동이 여타의 보조용언 구성과 달리 '-게 하'에 의해 사동주 의미역이 할당되는 특수성을 반영하기 위하여, 본용언과 보조용언의 합이 전체 문장에 의미역을 할당하는 것으로 처리하여 접미사에 의한 사동사 의미역 주석과의 일관성을 확보하였다. 이는 '-게 하' 구성의 특수성을 가장 합리적으로 고려한 처리 방식이라 판단된다. '-게 하' 구성이 사동주라는 의미역을 생성하는 역할을 하고 있으나, 이를 장형 사동으로 사동의 한 형태로 보는 것이 국어학계 전반의 합의라는 점을 반영한 처리 방식이라 할 수 있다. 국어원이 제기한 방안은 장형 사동 문장을 복문으로 이해한 방식으로, 학교문법을 비롯한 공적 영역의 문법관에도 맞지 않을 뿐만 아니라, 국어학계 전반의 문법 지식에도 매우 벗어난 것이라 볼 수 있다. 기존의 학설과 맞지 않다고 하더라도 그것이 활용도가 높다든지 응용 분야에 이점을 갖는다든지 하는 근거를 찾을 수 있다면 이러한 안을 고려해 볼 수 있겠으나, 실용적 관점에서도 이러한 처리 방식은 매우 부적절하다. 만약 이 문장을 복문으로 처리할 경우 구문 코퍼스에서 이를 사동으로 인지할 수 있는 방법이 현실적으로 마련되어 있지 않고 여타의 부사절을 포함한 복문 구성과 구별되지 않아, 이 구성의 특수성을 기계적으로 파악할 방법이 없다는 현실적인 문제가 발생한다. 더구나 '아들에게(ARG0) 청소를(ARG1) 하다'라는 내포문이 성립하지 않으며, 가능하다 하더라도 여격으로 실현된 '아들에게'에 AGR0를 할당하는 것은 기존 지침에 제시된 의미역 할당 방식에 위배되는 문제가 발생한다. 따라서 이 문제와 관련하여 연구팀에서 제안한 본용언과 보조용언의 합이 전체 문장에 의미역을 할당하는 방식이 가장 적합하다는 결론에 이를 수 있다. 아울러 '-게 하다'에 집중하여 '나는(ARGA) 아들이(ARG0) 청소를(ARG1) [하게 하였다].', '아들이(ARG0) 청소를(ARG1) [하게 하였다].'로 처리하는 것이 사동문 의미역 할당의 주체를 밝힌다는 이론적 측면에서나 작업의 효율성과 활용 가능성에서 더 나은 방안이 된다고 판단된다.

1.2. 의미역을 할당 받은 논항의 조사 삭제 여부

국어학 연구에서 조사는 논항 할당과 의미역 파악에 중요한 정보를 포함하고 있다. 의미역을 할당

받은 논항의 조사를 삭제하여 제시하는 방안은 이 사업에서 얻을 수 있는 매우 중요한 언어적 정보를 임의적으로 누락하게 된다는 문제가 있다. 회의 과정에서 한국전자통신연구원의 자문 결과에 따라 모든 의미역의 조사를 떼야 하는 이유가 어플리케이션 구축의 편의성과 관련되어 있음을 인지하였다. 그러나 이러한 접근은 두 가지 측면에서 문제가 된다. 첫 번째는 이 사업의 목표와 관련된 부분이다. 국립국어원에서 구축하는 코퍼스는 국어 연구의 공신력 있는 기초 자료로서의 의미를 가진다. 대규모 코퍼스를 통해 다양한 의미역의 논항이 어떤 형태로 실현되는지는 국어학 연구에서 매우 중요한 의미를 갖는다. 격조사가 실현될 수 있는 논항은 조사가 생략되기도 하고 보조사가 오기도 하고 조사 복합형으로 실현되기도 한다. 이러한 다양한 출현 환경은 화자나 필자가 자신의 의도에 따라 선택할 수 있는 선택지가 다양하다는 것을 의미하고, 국어학 연구가 심층적으로 이루어지기 위해서는 이러한 분포가 면밀히 분석되어야 한다. 그런데 조사가 삭제된 상태에서 논항이 제시되는 정보는 연구를 위한 기초자료로서 그 가치가 매우 감소되는 문제를 유발하게 된다. 많은 인력과 자원을 투입하여 구축하는 기초 코퍼스가 이러한 정보를 충분히 담지 못한다면, 국가사업으로서의 위상에도 문제가 있으리라 생각한다.

두 번째 문제는 이 사업에서 구축된 코퍼스를 활용하여 다양한 어플리케이션을 만들 수 있는데 조사를 삭제하여 제공하는 방식은 현재 수준에서 사용하기 편리한 방식이라는 점에서 한계가 있다. 질의응답 시스템을 비롯한 현재의 방식은 문장 생성이 매우 제한적 수준에서 이루어지고 있기 때문에 조사가 없는 방식이 선호될 수 있으나, 미래의 기술이 다양한 실현형의 기능을 고려한 방식으로 진화한다면 조사를 삭제한 코퍼스를 이용할 수 있는 범위가 제한적이라는 문제가 발생한다. 기술 발전 속도를 고려한다면 이러한 기초 코퍼스는 좀 더 미래지향적으로 설계되고 구축되는 것이 바람직하다. 조사를 삭제하지 않은 형태로 제공되는 지금의 방식이라고 하더라도 한국어 조사는 폐쇄부류여서 해당 형태에서 조사를 기계적으로 삭제하는 것이 어려운 기술은 아니다. 따라서 이러한 작업은 어플리케이션을 개발하는 쪽에서 자신들의 의도에 맞게 가공하는 것이 더 현실적 방안이 될 수 있다.

결론적으로 이 사업에서 추구하는 코퍼스 구축의 사업 목표와 현재와 미래를 아우르는 활용 가능성을 생각한다면, 조사를 삭제하는 것은 재고의 여지가 있다고 판단된다. 특히 조사 정보가 중요한 일부 부가역에 대한 목록과 지침을 마련한다는 것은 국어의 다양한 현상을 매우 제한적으로 파악하고 있다는 인식을 반영하는 것으로, 대규모 언어 자원을 다루고 처리하는 사업의 기본 전제로 삼기에 부적절하다고 판단된다.

1.3. 의사 보조용언

의사 보조용언의 처리 방식은 연구팀의 최종 지침이 타당한 것으로 파악된다. ‘우리 학생도 인턴 자리를 구할 수 있다.’라는 의미역 할당의 주체를 ‘구하다’와 ‘있다’ 둘로 처리하는 방식은 이 문장을 복문으로 파악한다는 점에서 우리의 언어직관과 맞지 않을 뿐만 아니라, 코퍼스의 활용 측면에서도 매우 큰 문제가 된다. 간혹 ‘~ 구할 수가 있다’라고 주격조사가 실현되는 형이 가능하지 않은 것은 아니나 현실 언어에서 이는 매우 빈도가 낮을 뿐만 아니라, 국어 문법의 관점에서 이 ‘가’ 형을 여타의 주격조사와 같은 형으로 파악할 수 있는가 하는 이론적 논쟁도 제기되어 있는 상태이다. 국어 문법 연구에서 ‘가’가 주격조사로 실현되기는 하나 보조사적 쓰임을 갖는 기능을 갖기도 한다는 점에서 분명히 전형적인 주격으로 기능한다고 보기 어려운 측면이 있다. 따라서 현대국어 코퍼스를 구축한다는 사업의 취지를 생각해 본다면, 문법화의 정도를 객관적으로 판단하기 어렵고, 이러한 관점을 코퍼스 구축에 고려하기는 힘들다고 판단된다. 따라서 ‘수 있다’와 같은 구성을 의사 보조용언으로 처리하는 현재의 지침이 국어학적 활도이나 응용적 활용도가 높은 코퍼스를 구축하는 최적의 방안이 될 수 있다.

1.4. 격 중출

이 사업의 대 원칙은 표층 형태에 충실하게 주석을 한다는 것이다. 이러한 관점에서 격 중출은 실현형을 중심으로 의미역을 중복적으로 부여하는 것이 적절하다. 만약 **'내가 사과를(AGR1) 반을(AGRM_EXT) 먹었어'**와 같이 처리할 경우 작업의 일관성을 지키기 어려울 뿐만 아니라 '반을'이 '먹었어'의 대격 논항이 아니라고 볼 근거가 없기 때문에 적절하지 않다고 판단된다. 국어학 연구에서 '내가 사과를 반을 먹었어'라는 문장과 관련 있는 문장을 거론되는 것은 '내가 사과의 반을 먹었어'이고 이러한 환경에서 '반을'은 명백히 대격인 대상역이다. 이러한 맥락에서 '반을(AGRM_EXT)'과 같이 처리하는 방식은 국어문법 연구의 특정한 관점을 반영하는 방식이 되기 때문에 사업의 목표와 맞지 않는 부분이 있는 것 같다.

주격 중출이나 목적격 중출의 경우 동일한 현상처럼 보인다고 하더라도 서술어의 의미적 특성에 따라 논항의 세부역을 달리 부과해야 한다는 다양한 견해가 있고, 어떤 의미역이 각각 부과되어야 하는지 학자마다 의견이 분분하다. 따라서 이러한 상황을 코퍼스 구축에 선택적으로 반영하는 것은 문제가 있다고 판단된다. 즉 이 분야의 연구는 매우 심층적인 국어 연구의 영역이고, 이 코퍼스가 이러한 연구의 기초자료라고 한다면, 표층적 실현형에 근거하여 중복역으로 주석하는 것이 이 분야의 연구자들이 코퍼스를 효율적으로 활용할 수 있게 하는 최적의 방안이 될 것이다.

1.5. 의존명사 '간'이 포함된 명사구의 처리

'여야 간 이견'은 하나의 명사구로 처리하는 것이 바람직하다. 다음은 표준국어대사전의 '간'에 대한 기술 내용이다.

1. 한 대상에서 다른 대상까지의 사이. 서울과 부산 간 야간열차.
2. (일부 명사 뒤에 쓰여) '관계5'의 뜻을 나타내는 말. 부모와 자식 간에도 예의를 지켜야 한다.
3. ('-고 -고 간에', '-거나 -거나 간에', '-든지 -든지 간에' 구성으로 쓰여) 앞에 나열된 말 가운데 어느 쪽인지를 가리지 않는다는 뜻을 나타내는 말. 공부를 하든지 운동을 하든지 간에 열심히만 해라.

위의 내용을 살펴보면 '여야 간 이견'은 1에 해당하는 것으로 엄밀히 말하면, '여야 간의 이견'에 해당하는 것으로 분석하는 것이 의미적으로 타당하다. 그런데 '여야 간'에 LOC을 설정하는 것은 2, 3의 쓰임을 혼동한 결과라 이해된다. 따라서 '여야 간 이견'을 하나의 명사구로 처리하는 것은 표준국어대사전의 사전 기술 내용과 부합할 뿐만 아니라, 구문 의미역 분석에서 하나의 단위에 해당하는 것을 하나의 논항으로 파악하는 합리적 방안이 될 수 있다.

1.6. 인용문의 처리

'"~한다"며'와 '~한다고'를 구문 분석이나 의미역 분석에서 달리 처리하는 것이 적절한지 검토가 필요한 것으로 보인다.

1.7. 괄호나 기호 등으로 붙어서 나타난 어절의 처리

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

2) 주요 내용 요약 및 과제 반영 여부

■ 상세 분석 필요

① 하위 표지 설정 문제 [미수용 - 지침 기본 원칙 준수]

• 관형사절의 피수식어에 대한 의미역 표지를 변별할 필요가 있음.

② 어절 내부 요소 분석 문제 [미수용 - 지침 기본 원칙 준수]

• 한 어절에 둘 이상의 서술어가 포함될 때 분석 누락 문제

예) '느껴진다"면서'와 같은 어절을 의미역 분석을 할 때 '느껴진다'의 논항과 '하면서'의 논항을 구분하여 표시할 수 없다. '경기 뒤 "~ 피곤함도 느껴진다"면서'와 같은 문장에서 '피곤함도'는 '느껴진다'의 논항(ARG1)이고 '경기 뒤'는 '하면서'의 부가어(ARGM-TMP)인데 현재로서는 이들을 한 줄에 나란히 표시할 수밖에 없다.

■ 사동주 분석의 방식 [수용]

장형 사동 문장을 복문으로 이해하여 분석하는 방식은 학교문법을 비롯한 공적 영역의 문법관에도 맞지 않을 뿐만 아니라, 국어학계 전반의 문법 지식에도 매우 벗어난 것이라 볼 수 있다. 복문으로 처리할 경우 구문 말뭉치에서 이를 사동으로 인지할 수 있는 방법이 현실적으로 마련되어 있지 않고 여타의 부사절을 포함한 복문 구성과 구별되지 않아, 이 구성의 특수성을 기계적으로 파악할 방법이 없다는 현실적인 문제가 발생한다. 본용언과 보조용언의 합이 전체 문장에 의미역을 할당하는 방식이 가장 적합하다.

■ 의미역 할당받은 논항의 조사 포함 여부 [미수용 - 응용 시스템 활용 형식 고려]

조사는 논항 할당과 의미역 파악에 중요한 정보를 포함하고 있다. 의미역을 할당 받은 논항의 조사를 삭제하여 제시하는 방안은 이 사업에서 얻을 수 있는 매우 중요한 언어적 정보를 임의적으로 누락하게 된다는 문제가 있다. 격조사가 실현될 수 있는 논항은 조사가 생략되기도 하고 보조사가 오기도 하고 조사 복합형으로 실현되기도 한다. 이러한 다양한 출현 환경은 화자나 필자가 자신의 의도에 따라 선택할 수 있는 선택지가 다양하다는 것을 의미한다. 미래의 기술이 다양한 실현형의 기능을 고려한 방식으로 진화한다면 조사를 삭제한 말뭉치를 이용할 수 있는 범위가 제한적이라는 문제가 발생한다. 조사를 기계적으로 삭제하는 것이 어려운 기술은 아니므로 어플리케이션을 개발하는 쪽에서 자신들의 의도에 맞게 후가공하는 것이 더 현실적 방안이 될 수 있다.

■ 격 중출 처리 [수용]

① 국어학의 관점

대원칙인 표층 형태에 충실하게 주석을 한다는 관점에서 격 중출은 실현형을 중심으

로 의미역을 중복적으로 부여하는 것이 적절하다. 주격 중출이나 목적격 중출의 경우 동일한 현상처럼 보인다고 하더라도 서술어의 의미적 특성에 따라 논항의 세부역을 달리 부과할 수 있는 의견이 있고, 어떤 의미역이 각각 부과되어야 하는지 학자마다 의견이 분분하다. 따라서 이러한 상황을 말뭉치 구축에 선택적으로 반영하는 것은 문제가 있다. 기초자료로서 표층적 실현형에 근거하여 중복역으로 주석한 연구자들이 말뭉치를 효율적으로 활용할 수 있게 하는 것이 합리적이다.

② 언어 처리의 관점

ETRI에서는 의미역 주석에 있어 필수역(numbered role)은 1개의 서술어에 대해 중복으로 발생하지 않는 것이 원칙이지만, 구문 분석 결과 이중 주어, 이중 목적어 구문인 경우에는 예외적으로 중복역을 허용한다. 구문적으로 중복역이더라도 의미를 따져서 이를 적당한 부가역으로 주석할 수는 있겠지만, 이럴 경우 작업자의 주관이 개입되기 때문에 구문 분석 결과 이중 주어, 이중 목적어에 해당하는 경우 의미역도 중복역을 허용하는 것이 오류를 줄 일 수 있는 방향이다.

■ 의사 보조 용언 처리 [수용]

문법화의 정도를 객관적으로 판단하기 어렵고 의사 보조 용언을 구성하는 조사의 기능이 보조사적 쓰임을 갖는다는 점에서 복문으로 파악하여 처리하는 것은 말뭉치 활용 측면에서 문제가 된다. 예를 들어 ‘우리 학생도 인턴 자리를 구할 수 있다.’라는 의미역 할당의 주체를 ‘구하다’와 ‘있다’ 둘로 처리하는 방식은 이 문장을 복문으로 파악한다는 점에서 우리의 언어직관과 맞지 않고, 간혹 ‘~ 구할 수가 있다’라고 주격조사가 실현되는 형이 가능하지 않은 것은 아니나 현실 언어에서 이는 매우 빈도가 낮을 뿐만 아니라, 국어 문법의 관점에서 이 ‘가’ 형을 여타의 주격조사와 같은 형으로 파악할 수 있는가 하는 이론적 논쟁도 제기되어 있는 상태이다. 따라서 ‘수 있다’와 같은 구성을 의사 보조용언으로 처리하는 현재의 지침이 국어학적 활용도나 응용적 활용도가 높은 말뭉치를 구축하는 최적의 방안이 될 수 있다.

■ 부가역 주석과 검증 [수용]

① 부가역 주석의 기준

- 부가 의미역은 기준을 세우기 명확하지 않고 정답이 있기보다는 선택의 문제인 경우가 많아서 ETRI에서는 ARGM-ADV를 비롯한 부가 의미역을 목록화하여 처리하고 있으며 구축 분과에 제공 가능하므로 구축 분과에서도 ETRI 기준으로 주석
- EXT, NEG의 경우는 ETRI는 후처리 개념으로 주석하고 있으며 구축 분과에서도 일관성을 위해 후처리하여 주석

② 부가역의 검증

- PropBank의 필수 의미역은 의미역 정보라는 기준이 있기 때문에 정답 유무가 객관

적이지만 부가역의 경우 기준을 세우는 것이 굉장히 어려운 문제임
- 부가역에 대해 검증을 할 경우 정답이 하나라고 가정을 하면 실제로는 오류가 아닌 경우를 오류로 판정할 수 있으니, 부가역에 대해 검증을 할 경우는 부가역의 이러한 특성을 반영하여 유연성을 가져야 함

■ 의미역 스팬의 설정 [수용 - 필수역 중심으로 명확한 범위에 대해 부착]

의미역 주석에서 스팬을 설정하는 것은 질의-응답 시스템 활용 시 적절한 응답을 추출하기 위한 것이다. 통사적 정보를 참고할 수는 있지만 통사적 정보에만 의존한 스팬 정보는 실제 응용 시스템에 활용도가 떨어지며 ETRI의 경우 질의-응답을 전제로 하여 스팬을 따로 주석하고 있다.

2.9. 최종 결과물 산출

의미역 분석 말뭉치 구축의 최종 단계는 2차 검수를 마친 의미역 분석 말뭉치를 제이슨(JSON) 형식으로 변환하여 최종 결과물을 산출하는 것이다. 이때 2차 검수 과정에서 수집된 후처리가 필요한 자료를 수정하는 과정이 포함된다. 제이슨 형식의 기본 구조와 예시는 부록에서 제시하였다.

또한 본 사업의 최종 납품 후 국어원의 회신을 반영하여 보수 기간 동안 수정 작업이 진행될 예정이다.

제 3 장

의미역 분석 말뭉치 구축 지침 수립

1. 지침 수립 과정

의미역 분석 말뭉치 구축 지침은 기본적으로 ‘ETRI 한국어 의존의미역 주석가이드라인’[17]을 기반으로 하되, Korean PropBank와 ETRI Frameset, 그리고 U-PropBank, 우리말샘 기반 프레임셋의 네 가지 의미역 정보를 적용하는 방식과 실제 작업에서 나타날 수 있는 문제점들을 해결하기 위한 주의사항을 기술하였다.

기존 프레임셋에 기술된 정보들이 실제 말뭉치에 등장하는 모든 서술어들을 포괄하기에는 부족하고 의미역 표지와 해당 언어 표현의 대응이 쉽지 않기 때문에 실제로 원시 말뭉치를 대상으로 하여 의미역 주석 작업을 하는 데 있어서 결정이 어려운 사항들에 대한 세부 지침을 제시할 필요가 있다. 이에 본 지침에서는 ‘ETRI 한국어 의존의미역 주석가이드라인’ 을 기반으로 하되, 일부 내용을 수정하고 항목별 설명을 상세화하며 다양한 사례를 포함함으로써 실제 구축 과정에서의 정확성과 일관성을 기하고자 하였다. 이 장에서는 구축 지침 각 장의 핵심적인 기술 내용과 추가·보완된 사항을 제시하여 지침이 수립된 과정을 보인다.

1.1. 의미역 정의와 주석 원칙

본 사업에서는 Korean PropBank와 ETRI Frameset, U-PropBank, 우리말샘의 네 가지 의미역 정보를 순차적으로 적용하여 각 서술어에 대한 의미역을 할당한다. 의미역은 크게 ‘필수 의미역’과 ‘부가 의미역’으로 나눌 수 있다. 의미역 주석 표지는 구문 분석 결과를 고려하여 유형과 범위를 설정하도록 한다.

서술성 명사나 계사 ‘이다’ 구문의 경우에는 문장 내에서 서술어로서 기능을 하지만 적절한 의미역을 할당하기 어렵다. 본 사업에서는 논항 구조를 가지는 서술어에 대한 의미역 할당에 초점을 맞추고, 이들 두 유형에 대해서는 주석하지 않도록 한다.

1.2. 의미역 주석 작업 순서

본 사업의 주석 작업에서는 Korean PropBank, ETRI Frameset, U-PropBank, 우리말샘의 네 가지 의미역 프레임셋을 순차적으로 적용하여 의미역 할당을 한다. 이들 의미역 정보는 그 규모에 있어서 큰 차이를 보일 뿐만 아니라 우선순위를 가지고 있기 때문에, 적용 순서가 명시적으로 제시되어 실제 작업에 적용될 필요가 있다.

기존 의미역 프레임셋이 매우 많은 수의 서술어에 대한 논항 구조를 가지고 있기는 하지만, 실제 원시 말뭉치에서는 어디에도 포함되지 않는 서술어가 출현하는 경우가 있다. 그리고 이들 중에서도 의미역 정보를 추가해야 하는 경우와 단순한 오류로 보아

17) 한국전자통신연구원(ETRI)에서 2017.08.25.일에 발행한 지침으로, 엑소브레인 언어 분석 말뭉치 구축을 위한 지침이다.

주석할 필요가 없는 경우 등으로 나뉘기 때문에, 각각의 유형에 따라 서로 달리 주석 작업을 수행하도록 한다.

1.3. 의미역 정보

지침에서는 우선 '의미역 프레임셋'에 대한 정의 및 설명을 제시하고, 주석의 기준이 되는 프레임셋을 소개하였다. 특히 K-propbank와 ETRI frameset은 동일한 의미역 표지 체계를 공유하고 있지만 U-propbank의 의미역 표지 체계는 이와 상이하므로, 이들 간의 차이점을 구체적으로 제시하였다.

각각의 프레임셋에 대해서는 실제 작업 창에서 제시되는 모습을 제시하고, 각 영역에서 해당 서술어의 의미역 프레임을 어떻게 확인할 수 있는지, 그리고 문장 내 성분에 대해 적절한 의미역을 어떻게 태깅할 수 있는지에 대한 방법을 제시하였다.

1.4. 의미역 주석 표지

필수의미역과 부가의미역에 대한 구체적인 설명을 제시하였다. 'ETRI 의존의미역 태깅 가이드라인'에서는 필수의미역과 부가의미역에 대한 설명이 소략하게 제시되어 있고 구체적인 의미역들이 어떠한 표지에 대응되는지에 대한 설명이 부족하기 때문에 이를 보완하였다.

우선 필수의미역의 경우, 실제 문장에서 여러 번 등장하는 '경험주(experiencer)' 및 '대상(theme)' 등에 대한 의미역 표지 대응 정보를 보완할 필요가 있으며, 또한 구체적인 예시를 들어 문장 내의 의미적 속성에 따른 의미역 표지 대응 관계를 제시해 주어야 한다. 특히 필수역의 경우 특정한 의미역이 절대적으로 특정 표지에 대응되는 경우도 있지만, 프레임셋 정보에 따라 술어별로 숫자가 적은 의미역부터 출발하여 차례로 의미역이 할당되는 것이 기본이기 때문에 주의를 요한다. 이에 대한 개념 정리 및 실제 예시를 제시하였다.

부가의미역의 경우 'ETRI 의존의미역 태깅가이드라인'에서는 의미역 표지와 정의만을 제시하고 있는데, 각 의미역에 대한 정의와 설명, 예시 등을 구체적으로 제시하여 주석자가 주석 시에 적용할 수 있도록 하였다.

1.5. 주석 표지-의미역 정보 주석 가이드라인 예시

의미역 주석의 기준이 되는 네 가지 프레임셋을 기반으로 하여 서술어의 의미역 태깅을 검토하는 작업이 구체적으로 어떻게 이루어지는지에 대한 가이드라인을 예시를 들어 단계별로 제시하였다.

1.6. 주석 표지-의미역 정보 주석 주의사항

‘ETRI 의존의미역 태깅가이드라인’에서는 ‘조사 관련 태깅 원칙’, ‘문장부호 관련된 태깅 원칙’, ‘한국어 어휘별 태깅 원칙’, ‘Predicate 배제 리스트 및 태깅 원칙’, ‘상호참조 관계에 대한 태깅 원칙’ 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 6장에서는 작업 중 확인되는 부분들에 대한 태깅 원칙을 지속적으로 추가해 나가고 이를 작업자들에게 꾸준히 교육함으로써 의미역 태깅 검토 작업이 일관적으로 이루어질 수 있도록 한다.

2. 의미역 분석 말뭉치 구축 지침

이 장에서는 본 사업에서 의미역 분석 말뭉치를 구축하기 위하여 적용한 지침을 보이고자 한다. 지침의 구성은 다음과 같다.

1) 의미역 정의 및 태깅 원칙
2) 작업 순서 및 방법
3) 의미역 프레임셋
4) 의미역 태그셋
5) 태그셋-프레임셋 태깅 가이드라인
6) 태그셋-프레임셋 태깅 주의사항

1) 의미역 정의 및 태깅 원칙

술어(동사나 형용사)는 문장 완성을 위해 필수적인 성분인 논항(주어, 목적어 등)을 요구한다. 이를 ‘필수 논항’이라고 하는데 이러한 필수 논항은 술어와 통사적 관계를 맺을 뿐만 아니라 특정한 의미적 관계(주어-서술어가 가리키는 동작의 행위주, 목적어-서술어가 가리키는 동작의 대상)를 맺게 되는데 이것을 ‘필수의미역’이라 부른다.

또한 문장에는 술어가 필수적으로 요구하는 성분 외에도 부가적인 성분(부사어 등)들이 존재한다. 이를 ‘부가어’라고 하며 이 부사어들 역시 술어와 특정한 의미적 관계를 맺게 되는데 이것을 ‘부가의미역’이라고 부른다.

의미역 정의 및 태깅을 위하여 Proposition Bank(Korean Proposition Bank, 약칭 KPB)

의 의미역 구분 기준을 바탕으로 아래의 구축 원칙을 따른다.

- 파싱되어 나온 결과를 바탕으로 수정한다.
- 프레임셋 기준으로 필수역을 중심으로 의미역을 태깅한다.
- 의미적 관계를 살필 때 절 단위를 넘어가지 않고 같은 절 내에서의 관계만 살피며 주로 명사구를 중점적으로 검토한다.
- FrameSet의 ARG0, ARG1, ARG2, ARG3에 속하는 의미역에 해당되는 것만 '필수의미역'으로 표시
- KPB의 FrameSet의 의미역 할당을 따르되 KPB에 없다면 ETRI, ETRI에 없다면 UPB의 의미역 할당을 따라 의미역을 태깅한다. 이때, 참고한 Frameset에 대한 정보를 남긴다.
- FrameSet에서 '필수의미역'에 속하지 않는 것은 모두 '부가의미역'에 해당한다.
- 계사 '이다' 문에 대해서는 의미역을 태깅하지 않는다.

- 단, 계사 '이다' 문을 논항으로 가진 서술어에 대해서는 해당 계사문을 논항으로 태깅하도록 한다.

[예시] 앞에 놓인 것은 평범한 음식들이었다.
→ 서술어 '평범한'의 ARG1으로 '음식들이었다.'를 태깅함.

- 필수 논항 중 여러 어절이 하나의 의미 단위(청킹 단위)를 이룰 수 있는 경우, 구문 분석에서 서술어에 딸린 문장 성분으로 잡히는 명사구의 수식 어구를 포함한 최대 범위를 논항으로 설정한다. 첫번째 어절에 해당 의미역 태그를, 마지막 어절에 '>>>' 태그를 부착함으로써 범위를 나타낸다. (자세한 내용은 6장 참고)
- 필수논항은 구문분석 결과에 연결되어 있지 않더라도 문장에 나타나 있으면 반드시 태깅하며, 부가 논항은 구문분석 결과에 의존한다.
- 논항이 되는 SBJ, OBJ, CMP, AJT 구 별로 별도의 독립적인 구로 인식한다.
- 부가역의 경우 의미적으로 연관관계가 있으면 무조건 논항으로 채택하기보다는 구문 관계 등을 고려하여 명확한 경우를 중심으로 태깅한다.
- 부가 논항은 구문 분석 결과에 의존하므로 서술어에 직접 연결되는 어절에 의미역을 부착한다.

[예시] 자동차 매연과 무분별한 공장 가동 **때문에(ARGM-CAU)** 오염되고 있다.

- 부가역 중 명사구의 경우에는 필수역과 마찬가지로 '>>>' 태그를 사용하여 범위를 나타낸다.

[예시] 갑자기 비가 **[오니까는](ARGM-CAU)** 사람들이 건물 안으로 들어갔지.
시험 감독을 들어가기 **[위해서는](ARGM-PRP)** 30분 일찍 일어나야 해.

• 서술성 명사나 용언이 아닌 성분으로 끝나는 헤드라인의 경우에는 태깅하지 않는다. (단, 구문 분석의 경우에는 태깅한다는 점에 주의)

2) 의미역 태깅 작업 순서

① KPROPBANK(예: KF_01)에 프레임셋이 있는 용언이 서술어라면 이 프레임셋에 따라 필수역에 대해 넘버링을 한다. ETRI 프레임셋, UPropbank 프레임셋, 우리말샘 프레임셋에 있는 경우도 마찬가지이다.(예: EF_01, UF_02, ,US_03). 우선 순위는 KF, EF, UF, US로 적용한다. 선택된 용언에 해당하는 태깅 작업 시의 의미번호는 JSON 변환 시 부록 4와 같은 체계로 변환되어 정보가 부여된다.

② KF, EF, UF, US로 지정되어 있지 않은 경우 999로 정보가 주어져 있는데, 프레임셋을 검색해서 지정할 수 있는 경우 KF, EF, UF, US 중 하나로 주석하고, 프레임셋이 기술되어 있지 않은 경우 아래와 같이 주석하여 지체하지 않고 다음 분석으로 넘어간다. 777~999는 리뷰 단계에서 일괄 처리한다.

- 777: 의미역 검색창에서 프레임셋을 검색해도 해당 어형이 존재하지 않아 추가 기술이 필요한 경우

> [예시] '팔로잉하다, 마이너하다' 등의 어휘 미기술

- 888: 해당 어형은 있으나 해당 의미나 프레임이 없는 경우

> [예시] 춤추다 - {A0_X:행동주}로만 프레임이 기술되어 있음

- 999: 오탈자가 포함된 경우 주석하지 않고 기존의 999로 표지를 유지한다.

> [예시] 계다

③ 파서의 분석 결과 프레임셋에서 필수역으로 처리한 것이 부가역으로 처리되어 나온 경우에는 해당 부가역을 필수역으로 다시 넘버링한다.

④ 반대로 프레임셋에 포함되지 않은 성분이 필수역으로 처리되어 있다면 이 결과를 삭제한다.

<그림 31> 의미역 분석 화면 예시

3) 의미역 프레임셋

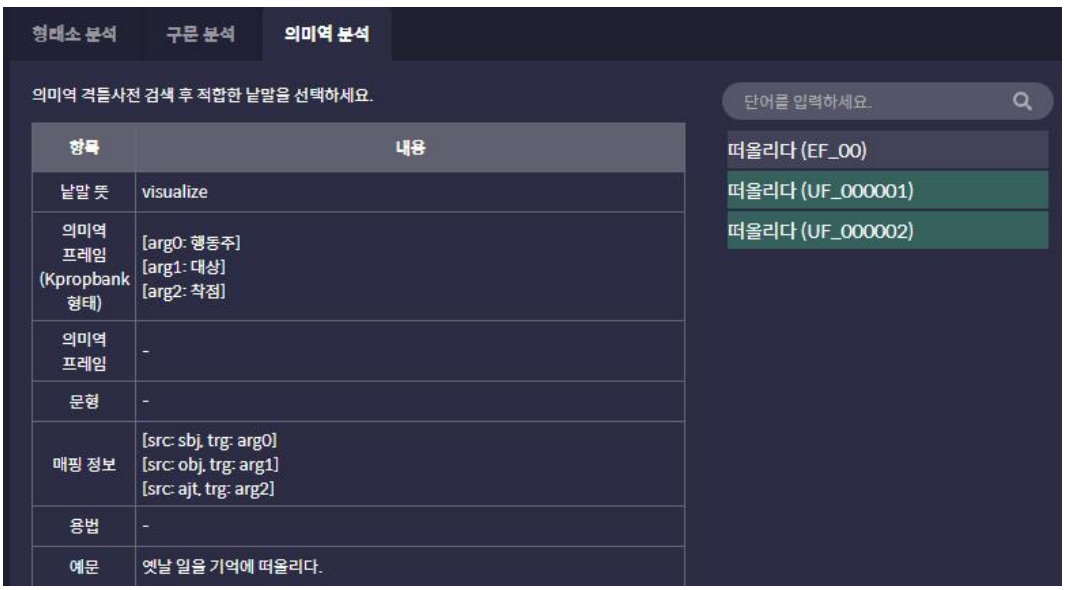
가. 개요

의미역 프레임셋이란, 문장에서 서술어가 되는 각 용언들의 필수 논항(용언이 필수적으로 요구하는 성분)과 그 필수 논항이 요구하는 의미역 정보를 표시한 틀을 말한다. 의미역 태깅 작업은 문장 서술어가 되는 용언을 중심으로 그 필수 논항에 의미역 정보를 표시하는 것이 주가 되므로 의미역 프레임셋 기반 작업이라고 할 수 있다.

우리의 작업에서 사용되는 의미역 프레임셋은 'K-Propbank'의 프레임셋을 기본으로 하지만 KPB의 용언의 수가 적어 실제 작업에서 어려움이 있으므로 KPB에 누락된 용언에 대해서는 ETRI 프레임셋, UPropbank 프레임셋, 우리말샘 프레임셋 등을 순서대로 적용하여 주석한다.

KPB와 ETRI의 태그셋은 유사한 방식이며 UPB의 태그셋은 상당히 큰 차이를 갖는데 툴에서는 KPB의 태그셋을 바탕으로 매핑한 프레임셋이 제시된다. KPB의 태그셋 중 필수역에 대한 태그셋을 보이면 아래와 같다[18].

의미역 표지	정의
ARG0	서술어의 동작주, 행위자(agent), 경험주(experiencer)
ARG1	서술어의 피동작주(patient), 대상(theme) 등
ARG2	시작점(starting point), 수혜자(benefactive) 등
ARG3	도착점(ending point) 등



<그림 32> 의미역 프레임셋 확인 및 검색 화면 예시

18) 각 태그셋에 대한 자세한 내용은 '4. 의미역 태그셋' 참조.

나. K-Propbank

K-Propbank는 약 2,700여 개의 서술어를 대상으로 하여 의미역 번호(태그셋) 및 내용(롤셋, role set)을 제시해 놓은 것이다. K-Propbank는 우리의 작업 툴에서 다음과 같이 제시된다.

서술어: 추진 (KF_01)	
낱말 뜻	drive forward
의미역 프레임 (Kpropbank 형태)	**[arg0: agent]** **[arg1: thing driven forward]**
매핑 정보	**[src: sbj, trg: arg0] [src: obj, trg: arg1]**
예문	추진하다: 금감위는 서민금융기관의 대형화를 유도하고 영업구역 및 업무 영역을 확대해 경쟁력 있는 중소형 금융기관으로 육성하는 한편 적기시정 조치 및 경영관리제도를 통한 경영개선 또는 정리를 지속 추진키로 했다.

여기에서 주목할 부분은 '의미역 프레임'이다. K-Propbank에는 'arg0, arg1'과 같이 의미역 태그셋과 함께 각각의 태그셋에 대응되는 롤셋이 제시되어 있다. 우리는 여기에서 적절한 롤셋이 그에 맞는 태그셋에 대응되어 있는지를 확인할 수 있다.

[예시] **국정개혁 과제 전반을(ARG1)** 보다 속도감 있게 추진하겠다.

■ 위 문장의 분석 결과를 확인할 때, 'agent'에 해당하는 논항은 표면적으로 드러나 있지 않으므로 따로 태깅을 할 필요가 없다. 또한 'thing driven forward'에 해당하는 '국정개혁 과제 전반'이 ARG1로 태깅되어 있으므로, 위 문장은 적절하게 의미역 태깅이 이루어진 것으로 판단할 수 있다.

다. ETRI frameset

ETRI frameset은 약 300개의 서술어를 대상으로 하여 의미역 번호(태그셋) 및 내용(롤셋)을 제시해 놓은 것이다. ETRI frameset은 우리의 작업 툴에서 다음과 같이 제시된다.

서술어: 건드리 (EF_00)	
낱말 뜻	touch
의미역 프레임 (Kpropbank 형태)	**[arg0: 행동주] [arg1: 대상]**
매핑 정보	**[src: sbj, trg: arg0] [src: obj, trg: arg1]**
예문	건드리다: 남의 물건을 함부로 건드리지 말아라.

여기에서도 ‘의미역 프레임’ 부분을 참고하여 태깅 작업을 진행하면 된다. ‘arg0, arg1’과 같이 의미역 태그셋이 제시되어 있다는 점에서 K-Propbank와 체계가 같고 실제 문장에서 적절한 태그셋이 부여되어 있는지를 확인하는 데 활용할 수 있다. 한편 롤셋이 영어로 되어 있고 각 논항의 구체적인 의미가 기술되어 있는 K-Propbank와는 달리, ETRI frameset의 롤셋은 한국어로 되어 있고 특정 의미역 표지가 부여되어 있다. 이를 통해 ETRI frameset에 포함되어 있는 서술어들은 보다 용이하게 태그셋-롤셋 대응 관계를 확인할 수 있다.

[예시] **책을(ARG1)** 살짝 건드리기만 해도 떨어질 것 같다.

■ 위 문장에서 ‘행동주’는 표면적으로 드러나 있지 않고, 프레임셋에서 ‘대상’에 해당하는 ‘책’이 ARG1로 태깅되어 있다. 따라서 위 문장 역시 적절하게 의미역 태깅이 이루어진 것으로 판단할 수 있다.

라. U-Propbank

U-Propbank는 약 90,000여 개의 서술어를 대상으로 하여 알파벳 형식의 의미역 태그셋 및 내용(롤셋)을 제시해 놓은 것이다. 따라서 KPB나 ETRI에 없는 용언들을 대개 UPB에서 확인할 수 있다.[19] U-Propbank는 우리의 작업 툴에서 다음과 같이 제시된다.

19) KPB나 ETRI에서 제시되는 용언도 대부분 UPB에 존재하는데 프레임셋에 차이가 있을 수 있다. 그러나 이 경우 우선 순위에 따라 KF, EF, UF, US의 순으로 적용한다.

서술어: 흔들거리다 (UF_000101)	
낱말 뜻	이리저리 자꾸 흔들리다. 또는 그렇게 되게 하다.
의미역 프레임 (Kpropbank 형태)	**{A1_X:대상} {A0_X:행동주 A1_Y:대상-을/를}**
의미역 프레임	**{X:대상} {X:행동주 Y:대상-을/를}**
문형	◁(…을)▷
예문	긴장해서 다리가 흔들거렸다. 바람에 등잔불이 흔들거렸다. 지진으로 땅이며 벽이며 모두 흔들거렸다. 배가 좌우로 흔들거렸다. 몸을 흔들거리며 걸었다. 그는 요람을 흔들거리며 우는 아이를 달랬다.

U-Propbank의 프레임 제시 방식은 K-Propbank 및 ETRI frameset과 적지 않은 차이가 존재하나 최종 결과물의 일관성을 위해 ‘X, Y, Z’ 3항 체계로 기술된 UF가 넘버링된 A0~A3의 4항 체계로 변환된 프레임셋이 주석 도구에 제시되어 있다. UPB에서 제시하는 의미역은 총 22개이다.

행동주, 피동주, 대상, 경험주, 도구, 처소, 착점, 기점, 방향, 경로, 수혜자, 자극, 원인, 자격, 비교기준, 동반주, 목적, 재료, 방법, 정도, 내용, 시간

UPB는 필수역에 대해서만 의미역을 할당하는 체계이므로 이 점에 유의해야 한다. 즉, UPB의 ‘시간’이나 ‘비교기준’, ‘방법’ 등도 프레임셋에 나타난다면 모두 필수역이므로 KPB 기준으로 ARGN으로 넘버링을 해야 한다. 이때 넘버링은 주어 부분만 ARG0, ARG1 여부를 결정한 뒤 나머지 성분은 프레임셋에서 제시하는 매핑된 것이 대부분이다. 각 태그셋의 정의에 맞도록 하는 것을 원칙으로 하되 주로 ARG(N+1)으로 넘버링하고 있다는 점을 알아 둔다.

[UPB에서 ARG0과 ARG1을 받을 수 있는 의미역]

(1) **행동주, 경험주 = ARG0**

UPB에서 ARG0을 받을 수 있는 의미역은 주어에 해당하는 의미역들 중 주로 ‘행동주’ 및 ‘경험주’이다다. 행동주는 주어가 의도성을 가지고 행위의 주체가 되는 의미역이다. 주로 ‘이/가’, ‘께서’, ‘에서’, 보조사 ‘은/는’이 붙는 성분들이다.

경험주는 인지하거나 지각하거나 어떤 감정을 느끼는 주체를 말한다. 행동주와 같은 주어 성분이지만 직접적인 행위를 하지 않는다는 차이가 있다. 형용사의 주어로 많이 나타나며 주로 ‘은/는’ 조사와 함께 쓰인다.

‘죄송하다, 무섭다’ 등과 같이 경험주이지만 ARG1 표지를 부착해야 하는 경우도 있으므로 주의해야 한다.

[예시] **내가[행동주]** 저녁을 만들겠다.
선생님께서[행동주] 주번에게 칠판을 닦으라고 하셨다.
정부에서[행동주] 일본측에 항의하였다.
영희는[행동주] 학교에서 아이들을 가르친다.
나는[경험주] 영희가 좋다.

---

(2) **피동주, 대상 = ARG1**

UPB의 의미역 중 ARG1을 받을 수 있는 대표적인 의미역으로는 피동주, 대상이 있다. 피동주는 피동문에 나타나는 주어로 사건에 주체적으로 참여하는 행동주가 아니라 영향을 입는 의미역이다. 대상은 서술어의 행위에 의해 옮겨지거나 묘사되는 등 서술어의 행위나 사건에 영향을 받는 논항을 의미한다.(한국어에서는 주로 을/를 조사와 함께 나타나는 목적어가 대표적이다.)

[예시] **도둑이[피동주]** 경찰에게 잡혔다.
내 말이[피동주] 엉뚱한 의미로 오역되었다.
내가 **그를[대상]** 때렸다.
정부에서 **이번 일을[대상]** 추진하였다.
그는(ARG0) 요람을(ARG1) 흔들거리며 우는 아이를 달랬다.

■ 먼저 U-Propbank에서 제시되어 있는 두 개의 격틀 중에서 해당 문장이 어느 격틀에 대응되는지를 판단한다. 위 문장의 경우 두 번째 격틀인 {X:행동주 Y:대상-을/를}에 대응된다. 다음으로는 각각의 의미역에 대해 적절한 태그셋을 부여하면 되는데, ‘그’는 행동주로, ‘요람’은 대상으로 파악하여 각각 ARG0, ARG1이 적절하게 부여되어 있음을 확인할 수 있다.

[UPB에서 ARG2, ARG3를 받을 수 있는 의미역]

각 용언의 격틀에 따라 같은 의미역이라고 할지라도 ARG2로 분석될 수도 ARG3로 분석될 수도 있다는 점에 유의한다. 수혜자는 주로 ARG2를 받는 것이 일반적이나 논항이 3개인 세 자리 서술어의 경우 다른 의미역이 A0~A2까지 기술되어야 하는 경우 ARG3로 분석되기도 한다.

주로 ‘에’, ‘에게’, ‘로’, ‘에서’, ‘와/과’ 등의 조사와 함께 쓰인다. 다음은 대표적인 예를 보인 것이다.

(1) ARG2 : 기점, 도구, 방향, 비교 기준, 동반자, 수혜자, 내용, 처소, 자격 등

[예시] 기점: 기차는(ARG0) **서울역에서(ARG2)** 출발했다.
처소: 나는(ARG0) **지사에(ARG2)** 근무하고 있다.
처소: 이 물품은(ARG1) 인근 **지역에서(ARG2)** 강취한 것이다.
수혜자: 그는(ARG0) **내게(ARG2)** 자신이 사는 곳이 어디인지를(ARG1) 가르쳐 주었다.
원인: 산이(ARG1) **나무로(ARG2)** 빽빽하다.
동반자: 우리는(ARG0) 이 일을(ARG1) **그와(ARG2)** 합의했다.

(2) ARG3 : 착점, 도구, 방법, 원인, 재료, 내용, 동반자, 처소, 정도 등

[예시] 방향: 아이가(ARG1) **깊은 산속으로(ARG3)** 유괴범에게(ARG0) 끌려갔다.
착점: 저는(ARG0) 이제 부산에서(ARG2) **서울로(ARG3)** 갑니다.

4) 의미역 태그셋

가. 개요

의미역 태그셋은 크게 '필수의미역'과 '부가의미역'의 두 가지로 구분된다. 필수의미역은 총 4개로 구성되어 있는데, 'ARG0', 'ARG1', 'ARG2', 'ARG3'과 같이 'ARG' 옆에 숫자가 붙어 있는 형식을 띤다. 한편 부가의미역은 'ARGM'으로 시작하며 'ARGM-LOC', 'ARGM-CND'와 같은 형식을 띤다. 태깅 작업은 필수의미역을 중심으로 이루어지며 넘버링은 프레임셋에 의존하여 태깅한다. 이 과제의 의미역 주석은 완전히 어휘 개별 프레임셋에 의존하는 작업이므로 아래 내용을 참고하되 반드시 프레임셋의 정보를 확인하여 작업해야 한다.

나. 필수의미역

필수의미역은 기본적으로 술어의 의미별로 정의된다. 의미역의 내용과 표지의 대응 양상을 보이면 아래와 같다.

의미역 표지	정의
ARG0	서술어의 동작주, 행위자(agent), 경험주(experiencer)
ARG1	서술어의 피동작주(patient), 대상(theme) 등
ARG2	시작점(starting point), 수혜자(benefactive) 등
ARG3	도착점(ending point) 등

필수의미역의 경우 의미역 표지와 의미역의 내용이 원칙적으로 대응되지만, 의미역 표지가 숫자로 이루어져 있는 만큼 이들 간에는 상대적인 순서가 존재한다.

1) 주어가 명확한 동작주 또는 행위자인 경우에는 ARG0로 태깅한다.

[예시] **닭이(ARG0)** 모이를 먹었다.
그는(ARG0) 문 쪽으로 쏜살같이 달렸다.

2) 행위의 주체로 판단되더라도 명확한 동작주 또는 행위자가 아닌 경우에는 ARG0이 아닌 ARG1로 태깅하는 것에 주의한다. 피동주 역시 ARG1로 태깅한다.

[예시] **따뜻한 바람이(ARG1)** 불었다.
동생은(ARG1) 모든 것이 불만스러운 말투이다.
도둑이(ARG1) 경찰에게 잡혔다.

3) 첫 번째 논항이 명확한 동작주 또는 행위자가 아닌 경우에는 모두 ARG1로 태깅한다.

[예시] 말을 하고 싶어서 **입이(ARG1)** 근질거린다.
작은 돌멩이와 나뭇가지들이(ARG1) 발에 챈다.

4) 의미역 태그셋의 정의를 바탕으로 태깅할 때 보통 주로 주어에 1씩 더한 값으로(ARG1, ARG2, ARG3) 프레임셋이 기술된 경우가 많지만 꼭 그렇지 않다는 점에 유의하여야 한다. 특히 주어가 ARG0로 잡히는 경우 ARG1이 나타나지 않는 경우가 많으며 한국어의 특성상 ARG2가 ARG1에 선행하기도 한다.

■ 넘버링이 순서대로 태깅되는 예

[예시] 우리는(ARG0) 돌밭을(ARG1) 옥토로(ARG2) 가꾸었다.
그는(ARG1) 그림에도(ARG2) 강하다는 소문이 있다.
그는(ARG0) 아들을(ARG1) 잘 타일렀다.

■ 넘버링이 중간에 숫자가 누락되는 예

[예시] 그는(ARG0) 운전면허 시험에(ARG2) 합격하였다.
그는(ARG0) 미국으로(ARG2) 망명하였다.
철수는(ARG0) 영희와(ARG2) 연애하고 있다.
그는(ARG0) 아들에게((ARG2) 이제 그만 두라고(ARG3) 타일렀다.

다. 부가의미역

부가의미역은 한국어 문장의 부사적 기능을 수행하는 어휘에 해당되며, 장소, 방향, 조건, 방법, 시간, 범위, 목적 등의 의미를 나타내는 논항이 대부분으로, 필수의미역과 다르게 아래표에서와 같이 각각의 정의된 의미 역할을 갖는 어휘에 할당된다. '부가'라는 말에서도 알 수 있듯이, 모든 용언이 필수적으로 요구하는 것이 아니므로, 의미역이 쓰일 수도 있고, 쓰이지 않을 수도 있다. 그러나 사전에 격틀 정보로 제시되어 있는 필수 성분이라면 수작업 검수 과정에서 필수역으로 분석해야 한다.

의미역 표지	정의
ARGM-LOC	장소 (locatives)
ARGM-DIR	방향 (directional)
ARGM-CND	조건 (condition)
ARGM-MNR	방법 (manner)
ARGM-INS	도구 (instrument)
ARGM-TMP	시간 (temporal)
ARGM-CAU	이유/원인 (cause)
ARGM-EXT	범위 (extent)
ARGM-PRD	보조 서술 (secondary predication)
ARGM-PRP	목적 (purpose clauses)
ARGM-DIS	담화 연결 (discourse)
ARGM-ADV	부사적 어구 (adverbials)
ARGM-NEG	부정 (negation)

① ARGM-LOC (장소)

- FrameSet의 ARG-SRC(시작점)과 ARG_GOA(도착점)에 해당하지 않아야 한다.
- 사건이 발생하는 상황적 공간을 가리키는 처소 논항을 ARGM-LOC로 분석한다.
- 동사의 의미에 이동성이 없고, '-에서/에' 조사와 함께 쓰이는 경우 ARGM-LOC로 분석한다.
- FrameSet 상에 ARG-N으로 정의되지 않고, 명확한 지명이나 장소를 뜻하는 경우를 M-LOC로 분석한다.

> [예시] 여름에는 값싼 집에 머물며 **해변에서(ARGM-LOC)** 휴양을 하기도 했다.

② ARGM-DIR (방향)

- 동사의 의미가 이동성을 가질 때, 방향격조사 '-로/으로'와 함께 나타나는 논항을 ARGM-DIR로 분석한다.
- '오른쪽, 왼쪽, 위쪽, 아래쪽, 앞으로, 뒤로, 동서남북' 등에 해당하는 논항을 ARGM-DIR로 분석한다.

- 의미적으로 ‘방향’과 ‘도착점’이 혼동되는 경우가 있는데, 이러한 경우 ‘에’ 논항일 경우에는 도착점, ‘로’ 논항일 경우에는 방향으로 분석한다.

[예시] 달이 **서쪽으로(ARGM-DIR)** 기울었다.
버스가 터미널을 떠나 **서울로(ARGM-DIR)** 출발했다.
철이는 금새 학교에(ARG3) / **학교로(ARGM-DIR)** 달려갔다.

---

③ ARGM-CND (조건)

- 인물이나 사물의 자격이나 술어 발생 조건을 가리키는 논항을 ARGM-CND로 분석한다.
- ‘-중에’, ‘-가운데에’에 해당하는 성분이나 ‘~ 측면에서’, ‘~ 면에서’ 등의 성분도 ARGM-CND로 분석한다.

[예시] 모인 열 사람 **가운데(ARGM-CND)** 아홉 사람은 차를 가지고 있다.
에메랄드는 가격 **면에서(ARGM-CND)** 다이아몬드를 능가한다.

- 어떤 참여자가 사태에 참여하는 자격, 지위, 신분 등을 나타내는 논항을 ARGM-CND로 분석한다.
- 주로 ‘~로’, ‘~로서’ 등으로 나타난다. 이때 ‘-로(서)’가 이끄는 성분이 해당 용언의 필수 논항일 경우에는 필수의미역으로 넘버링해야 한다는 것에 주의한다. 예컨대 아래에서 ‘유명하다’, ‘선출하다’ 등은 각각 의미역이 [arg1: thing famous][**arg2: famous for**], [arg0: voters][arg1: candidate][**arg2: role, position**]로 기술되어 있으므로 ‘로’가 이끄는 성분을 필수역으로 처리해야 한다.

[예시] 그는 **회장으로서(ARGM-CND)** 기업을 이끌고 있다.
우리는 **검찰로서(ARGM-CND)** 이러한 입장을 취할 수밖에 없다.
그는 **과학자로**(ARG2) 유명하다.
철수는 **회장으로**(ARG2) 선출되었다.

- ‘~로서의’가 명사를 수식하는 경우에는 서술어가 취하는 성분이 아니므로 의미역을 태깅하지 않는다.

[예시] 저는 *교수로서의 책임*을 통감하고 자리에서 물러나는 바입니다.
이러한 이론은 *표현 수단으로서의 언어*에 중점을 두고 있다.

---

④ ARGM-MNR (방법)

- 술어를 수행하는 방법에 대한 성분(명사절, 부사, 부사절)을 ARGM-MNR로 분석한다.
- ‘늦게, 빠르게’ 등을 ARGM-MNR로 분석하지 않도록 주의한다.

- ‘방망이로, 택시로, 금으로’ 등 구체명사에 ‘로’가 붙어서 술어를 수행하는 방법이나 수단(교통 수단 포함)을 나타내는 경우에는 도구(ARGM-INS)로 분석하고, 그 외 구체명사가 아닌 경우에는 모두 방법으로 분석한다.

[예시] 그는 **큰 소리로(ARGM-MNR)** 떠들었다.
그들은 **엄숙한 태도로(ARGM-MNR)** 김 박사를 맞았다.

※ 방법과 도구 중 어느 것으로 할지 혼동이 되거나 양쪽 다 성립하는 경우에는 파서의 결과를 존중하여 주석한다.

⑤ ARGM-INS (도구)

- 술어를 행할 때 사용하는 구체적인 사물로서의 도구에 대한 논항을 M-INS로 분석한다.
- 술어를 수행하는 방법인 ARGM-MNR보다 구체적인 ‘사물’이 있는 논항, 즉 ‘물리적 도구’를 나타내는 논항을 ARGM-INS로 분석한다. 주로 ‘~로’ 성분으로 나타난다.
- 특히 ‘~을 이용하다’를 대입하여, 문장이 어색하지 않을 경우, ARGM-INS로 분석한다.
- cf. “철이 엄마와 철이는 수화로 소통한다.“와 같은 문장의 “수화“는 “이용하다“를 대입하여 어색하지 않더라도 “언어 소통 방법“으로 이해하여 ARGM-MNR로 태깅한다.

[예시] 철수는 **숟가락으로(ARGM-INS)** 밥을 먹었다.
철이는 **색연필로(ARGM-INS)** 색을 칠했다.

- 어떤 물리적인 이동이 발생할 때 시작점과 도착점 사이에 경유하는 장소(‘~로’로 나타남) 또한 ARGM-INS로 분석한다. 이러한 성분은 ‘~을 이용하다’의 대입이 가능하고 ‘구체명사+로’로 나타나는 성분이라는 점에서 위에서 설명한 ‘도구’에 해당한다. 다만 용언의 필수 논항일 경우에는 필수의미역으로 넘버링해야 한다는 것에 주의한다. 예컨대 아래에서 ‘다니다’는 [arg0: entity moving] [arg1: place, path]로 프레임셋이 기술되어 있으므로 필수역으로 처리해야 한다.

[예시] 철수는 **창문으로(ARGM-INS)** 집에 들어갔다.
철수는 이 **길로(ARGM-INS)** 학교에 간다.
늦은 시간에는 **대로로(ARG1)** 다녀라.

- cf. 주의해야 할 것은, ‘~를 통해’, ‘~를 거쳐’ 등과 같이 복문 구조로 나타나는 경우 ‘통하다’, ‘거치다’의 대상(ARG1)로 분석하여야 한다.

[예시] 이 기차는 서울에서 **대전을(ARG1)** *거쳐* 부산으로 간다.
나는 **수로를(ARG1)** *통해* 터널 밖으로 빠져나갔다.

⑥ ARGM-TMP (시간)

- 술어(행위)가 발생한 시간을 지시하는 논항을 ARGM-TMP로 분석한다.
- 명확한 날짜, 시기, 시대를 나타내는 경우, 논항을 ARGM-TMP로 분석한다.
- 단, '-부터 -까지'와 같이 기간을 나타내는 경우, FrameSet과 상관없이 시작점과 도착점으로 구분하여 분석한다.
- '-ㄴ 뒤/후' 등도 M-TMP로 분석한다.

[예시] 진달래는 **이른 봄에(ARGM-TMP)** 핀다.
해산 후(ARGM-TMP) 농민들은 집강소를 설치하였다.

⑦ ARGM-CAU (이유)

- 술어가 발생한 이유에 해당하는 논항을 'ARGM-CAU'로 분석한다.
- '~때문에', '~덕분에', '~로 인하여' 등을 넣었을 때, 문장의 의미가 통하는 경우 ARGM-CAU로 분석. 다만 용언의 필수 논항일 경우에는 필수의미역으로 넘버링해야 한다는 것에 주의한다. 예컨대 아래에서 '고생하다'는 [arg1: patient][arg2: reason]로 프레임셋이 기술되어 있으므로 필수역으로 처리해야 한다.

[예시] 지난 밤 **강풍으로(ARGM-CAU)** 가로수가 넘어졌다.
이번 겨울에는 **감기로(ARG2)** 고생했다.

- 다만 구문 분석에서 서술어에 직접 연결되는 어절에 의미역을 부착하여야 하므로 '~ 때문에', '~ 덕분에', '~ 탓에' 등의 성분에서는 '때문에', '덕분에', '탓에'에 ARGM-CAU를 태깅한다.

[예시] 한강이 하수 **때문에(ARGM-CAU)** 오염되고 있다.
네 **덕분에(ARGM-CAU)** 시험에 합격했어.

⑧ ARGM-EXT (범위)

- 크기 또는 높이 등의 수치와 정도를 의미하는 논항이다.
- 다만 시간, 장소 등의 의미가 명확한 '-에서 -까지' 등의 범위는 ARGM-EXT를 할당하지 않는다.
- 또한 시간에 해당하는 명사구가 나타나는 경우 '시간'으로 분석한다.

[예시] 철수는 운동장을 **10km(ARGM-EXT)** 뛰었다.
1년 사이 나는 키가 **7cm나(ARGM-EXT)** 자랐다.
오늘 증시는 다우 2.90%, 나스닥 3.47%의 **3% 안팎으로(ARGM-EXT)** 폭락했다.
cf. 철수는 세 시간 **동안(ARGM-TMP)** 공부했다.

⑨ ARGM-PRD (보조서술)

- 시간, 장소, 조건, 방법, 원인, 범위 등에 해당되지 않으나 술어의 상태를 보조적으로 수식하는 의미를 갖는 논항에 해당한다.
- 대상과 같은 의미이거나 대상의 상태를 나타내면서 술어를 수식하는 논항이다.
- 주로 '~로서', '~로' 조사와 결합 빈도가 높고, NP_AJT가 주로 이에 해당한다.
- '말자로', '최초로' 등 대상이 술어에 대해 행해진 순서를 나타내는 논항이다.

[예시] 현지에서 생산되지 않는 물품을 **공납으로(ARGM-PRD)** 부과한다.
석회암 지대에서 깔때기 **모양으로(ARGM-PRD)** 파인 웅덩이가(ARG1) 생겼다.
삼남 삼녀 가운데(ARGM-CND) **말자로(ARGM-PRD)** 태어났다.

⑩ ARGM-PRP (목적)

- 술어의 주체가 목표를 가리키는 논항을 ARGM-PRP 로 분석한다.
- 행위(술어)의 주체가 하려는 바, 목표하는 바를 지시하는 표현을 ARGM-PRP 로 분석한다.
- 행위의 의도가 분명히 드러나는 논항을 ARGM-PRP 로 분석한다.
- '~고자, ~러, ~기 위해, ~를 위해'와 같은 연결어미 또는 명사형 어미와의 결합형이 주로 해당된다.

[예시] 주나라의 '백이'와 '숙제'는 절개를 **지키고자(ARGM-PRP)** 수양산에 거처했다.
회의에 늦지 않기 **위해(ARGM-PRP)** 30분 일찍 일어났다.

⑪ ARGM-DIS (담화연결)

- '그러나', '그리고', '즉' 등의 문장접속부사를 ARGM-DIS 로 분석한다.
- 문장을 연결하는 접속부사 '그-, 하-' 계열의 '그러나, 그러니까, 하지만, 한데, 더욱이, 게다가' 등과 단어를 접속하는 '곧, 즉, 또, 또한' 등도 ARGM-DIS 로 태깅한다.

[예시] **하지만(ARGM-DIS)**, 여기서 동(東), 서(西)는 중국과 유럽을 뜻한다.

⑫ ARGM-ADV (부사적 어구)

- ‘마치’, ‘물론’, ‘역시’와 같이 부사적 어구에 해당하는 어휘를 선정하여, ARGM-ADV 로 분석한다.

> [예시] 산의 능선이 **마치(ARGM-ADV)** 닭벼슬을 쓴 용의 형상을 닮았다.
> **특히(ARGM-ADV)**, 한강 유역을 장악함으로써 삼국 경쟁의 주도권을 쥐었다.

⑬ ARGM-NEG (부정)

- 술어에 대해 부정의 의미를 가지는 논항을 ARGM-NEG로 분석한다.
- ‘~지 않다, 없다’ 등의 어휘를 분석한다.

> [예시] 산은 불에 타지 않았다.(ARGM-NEG)
> 그 일에 대해서는 알 수 없었다.(ARGM-NEG)

⑭ ARGM-AUX (보조용언)

- 일반적으로 보조용언은 한국어에서는 형태소 품사 태깅에서 결정이 되는 경우가 많기 때문에, 이에 대한 다른 기준은 설정하지 않는다.

> [예시] 이론이 세상에 널리 알려지게 되었다.(ARGM-AUX)

5) 태그셋-프레임셋 태깅 가이드라인 예시

가. 필수역 태깅 가이드라인

3장에서 제시한 프레임셋에 따라 문장 내 서술어의 의미역 태깅을 검토한다. 이때 ‘K-Propbank > ETRI frameset > U-Propbank > 우리말샘 기반 프레임셋’ 순으로 우선순위를 가진다. 즉 어떤 서술어가 K-Propbank에 포함되어 있으면 해당 서술어의 의미역은 K-Propbank에 제시되어 있는 방식대로 태깅을 하고, K-Propbank에 포함되어 있지 않으면 ETRI frameset 포함 여부를 확인하여 태깅하는 방식으로 이루어진다.

필수역의 경우 ARG0~ARG3까지의 태그셋만 부여되기 때문에 별도의 롤셋은 부여되지 않는다. 따라서 해당 서술어가 포함되어 있는 프레임셋에서 제시된 격틀에 따라 순서대로 적절한 태그셋을 부여하면 된다. 필수역 태깅 예시를 보이면 다음과 같다.

> [예시] **정부는(ARG0)** 차량 폭발 테러 이후 해당 지역에 **계엄령을(ARG1)** 발동하였다

1) 위 문장의 서술어 ‘발동하다’를 작업 툴에서 검색해 보면 우선 K-Propbank와 ETRI frameset 모두에서 검색되지 않고 U-Propbank에만 포함되어 있음을 확인할 수 있다. 따라서 이 문장에 대한 필수역 태깅은 U-Propbank를 기준으로 해야 한다.

2) U-Propbank에는 ‘발동하다’에 대해 4개의 의미가 제시되어 있는데, 여기에 제시된 ‘낱말 뜻’, ‘프레임’, ‘예문’ 등의 정보를 이용하여 위 문장에서의 ‘발동하다’가 이들 중 어떤 용법으로 쓰인 것인지를 판단한다. 검토 결과 ‘발동하다(UF-000200)’에 대응됨을 알 수 있다.

서술어: 발동하다 (UF_000200)	
낱말 뜻	발동(3). >>공공 기관이 법적 권한을 행사함.
의미역 프레임 (Kpropbank 형태)	**{A0_X:행동주 A1_Y:대상-을/를}**
의미역 프레임	**{ X:행동주 Y:대상-을/를}**
예문	대통령이 올바르게 정치권력을 발동해야 선진 민주 국가가 될 수 있다. 학생들의 시위가 거칠어지자, 정부는 공권력을 발동했다. >>공권력 발동. 국가의 긴급한 상황에서는 국가 긴급권의 발동을 통해 위기를 극복하기도 한다.

3) 해당하는 프레임셋을 찾은 뒤에는 의미역 프레임에 제시되어 있는 대로 의미역 태그셋을 찾고, 해당 태그셋이 적절하게 태깅되어 있는지를 확인한다. 위 문장의 경우 ‘행동주’의 의미 역할을 하는 성분이 ARG0을, ‘대상’의 의미 역할을 하는 성분이 ‘ARG1’을 부여받아야 하고, 이에 따라 태깅이 올바르게 이루어져 있음을 확인할 수 있다.

나. 부가역 태깅 가이드라인

부가역은 상대적인 번호로 부여되어 있는 것이 아니라 각 성분의 의미에 따른 명칭을 가지고 있다.

[예시] **영빈관에서(ARGM-LOC)** 국가유공자 및 유가족 초청 오찬이(ARG1) 열렸다.

■ 위 문장에서는 ‘국가유공자 및 유가족 초청 오찬’이 필수역인 ARG1로 태깅되어 있고, ‘영빈관’이 ARGM-LOC로 태깅되어 출력되었다. ‘ARGM-LOC’은 ‘장소’를 나타내는 부가역인데, 위 문장에서 ‘영빈관’이 서술어 동사가 이루어진 ‘장소’를 의미하는 것이 맞으므로 이 부가역은 적절하게 부여된 것으로 판단할 수 있다.

6) 태그셋-프레임셋 태깅 주의사항

가. 분석 배제 리스트

영어의 전치사구에 해당하는 해당 용언의 프레임 번호(의미)[20]는 **서술어로 채택하지 않고 의미역 정보를 지운다.** 배제된 리스트는 다음과 같다.

- 대하.01, 관하.01, 위하.01, 의하.01, 따르.02, 통하.01, 비하.01, 인하.02, 불구하.01, 비롯하.01, 더불.01, 말미암.01, 그러.01, 이러하.01, 그리하.01, 그러하.01, 어떠하.01, 그렇.01

[예시] 정당은 헌법재판소의 심판에 의하여 해산된다.
전통 문화에 대한 관심.
개발에 따른 공해 문제.
몸살에도 불구하고 출근했다.

배제 대상은 활용형이 제약적이고 의미적으로 영어의 전치사구에 해당하는 경우로 한정한다. 용언마다 의미의 분포가 다르므로 **무조건 삭제하지 않도록 주의한다.** 아래 '대하다'의 의미(표준국어대사전) 중 배제 대상이 되는 것은 '「3」 【…에】 (('대한', '대하여' 꼴로 쓰여)) 대상이나 상대로 삼다.' 만이다. 한편 '관하다'는 단일 의미로 기술되어 있으므로 일괄적으로 분석 표지를 부여하지 않는다.

[예시] 대하다2
「1」【…을】【(…과) …을】 (('…과'가 나타나지 않을 때는 여럿임을 뜻하는 말이 주어로 온다)) 마주 향하여 있다.
- 그는 벽을 대하고 앉아서 명상에 잠겼다.
- 나는 어머니와 얼굴을 대하기가 민망스러워서 자리를 피했다.
- 친구들이 서로 얼굴을 대하고 앉아서 차분하게 이야기를 나눈 지도 꽤 오래되었다.

「2」【…에/에게 -게】【…을 …으로】【…을 -게】 (('…으로'나 '-게' 성분은 '…처럼, -은/을 듯이' 따위의 부사어나 '-이/히' 부사로 대체될 수 있다)) 어떤 태도로 상대하다.
- 그는 누구에게나 친절하게 대한다.
- 그 여자는 특히 잘생긴 남자 사원에게 상냥하게 대한다.
- 낯선 사람을 친구처럼 대하다.

「3」【…에】 (('대한', '대하여' 꼴로 쓰여)) 대상이나 상대로 삼다.
- 전통문화에 대한 관심.
- 강력 사건에 대한 대책.
- 건강에 대하여 묻다.
-「비슷한말」 관하다(關하다)

「4」【…을】 작품 따위를 직접 읽거나 감상하다.

---

20) 리스트에 제시된 어휘의 번호는 가장 먼저 적용하는 KF(Korean PropBank Framenet) 기준임.

- 이 소설을 처음 대하는 독자는 다소 당황하게 될 것이다.

나. ARGA의 문제

KPB의 프레임셋에서 주로 장형 사동 구문에서의 사동주를 ARGA로 표시하는 경우가 많다. 이때는 사동주를 ARGA, 실제 행위자를 ARG0으로 분석하도록 한다.[21] 예는 다음과 같다.

[예시] 나는(ARGA) 아들에게(ARG0) 청소를(ARG1) [하게 하였다].
나는(ARGA) 아들에게(ARG0) 밥을(ARG1) 먹였다.

다. 붙여쓰기로 처리된 '본용언 + 보조용언' 구성

'본용언+보조용언' 구성의 경우 띄어쓰기가 원칙이나, 경우에 따라서는 붙여쓰기로 처리되어 나올 수 있다. 이러한 결합 형식은 사전 미등재어이기 때문에 프레임셋에서도 검색이 되지 않는다. 이 경우에는 본용언을 기준으로 하여 프레임셋 서술어를 설정하고, 그에 따라 의미역 태깅을 한다.

[예시] 나는 너의 마음을 **알고있다**.
→ '알고있다'는 프레임셋에서 검색되지 않음. 이러한 경우 '알다'를 기준으로 검색하고 그에 따라 의미역 태깅을 함.

라. 가운뎃점 또는 괄호에 병기되어 나타나는 서술어 문제

아래 예문에서 '체결·공포된'과 같이 가운뎃점 또는 괄호 등으로 구분되어 있지만 실제 의미적으로는 '체결되고, 공포된'으로 해석되는 경우에, 앞에 나온 용언(여기에서는 '체결되다'만 서술어로 처리하여 의미역을 할당한다.

[예시] 헌법에 의하여 체결·공포된 조약과(ARG1)
→ 서술어 '**체결되다**'의 대상역으로 '조약'을 할당함. 이때 '공포되다'는 서술어로 처리하지 않음.

마. 보조용언 구성의 처리

본용언 뒤에 출현하여 서법 등의 문법적 의미를 나타내는 보조용언과 의사 보조용언의 경우 별도로 처리하지 않고, 일반적인 명사구 및 동사구 분석 기준에 맞춰서 분석

21) 장형 사동 '-게 하다' 구문의 경우 이와 같이 예외적인 조항을 둔 것은 다음과 같은 이유 때문이다. 먼저 일반적인 보조용언의 경우 의미역 할당의 주체가 될 수 없으므로 AUX 표지를 부여하게 된다. 그러나 '-게 하다' 등 일부 보조용언의 경우 전체 문장의 구조와 의미를 바꾸게 되며, 피동 접미사에 의한 사동사와의 일관성을 고려하여 [V-게 하다] 전체를 하나의 사동 서술어로 간주하여 사동주를 ARGA, 실제 행위주를 ARG0, 행위를 ARG1로 처리한다.

한다. 보조용언이나 의사 보조용언은 의미역 할당의 주체가 될 수 없으므로 본용언이 할당하는 의미역을 주석한다.

의사 보조용언 구성에 해당하는 예는 아래와 같다.

i. -ㄹ 수/리(가) 있다/없다

[예시] 김 씨는 어느 기관에 지원해야 전공을 살려 인턴 자리를 **구할 수 있을지 알 수 없었다**.
→ 알다: 김 씨는(ARG0), 어느 기관에 지원해야 전공을 살려 인턴 자리를 구할 수 있을지(ARG1)
→ 구하다: 인턴 자리를(ARG1)

ii. -ㄴ/ㄹ 의존(일반)명사+이다 (의존명사: 것/터/뿐/따름/모양/지경/참/중 & 일반명사: 노릇/예정/길)

[예시] 철수가 밥을 곧 먹을 것이다.
→ 먹다: 철수가(ARG0), 밥을(ARG1)

iii. '-ㄴ/ㄹ 의존(일반)명사+으로

[예시] 정부는 이 프로그램이 적은 비용으로도 어학연수를 **대체할 것으로** 예상했다.
→ 예상하다: 정부는(ARG0), 이 프로그램이 적은 비용으로도 어학연수를 대체할 것으로(ARG1)
→ 대체하다: 이 프로그램이(ARG2) 어학연수를(ARG1)

iv. -ㄹ {만/법/듯}하다

[예시] 그 영화는 청소년들도 볼 만하다.
→ '보다'의 ARG0으로 '청소년들도', ARG1으로 '그 영화'를 할당

v. -는 말이다

[예시] 밥을 먹었단 말이냐?
→ '먹다'의 ARG1로 '밥을'을 할당

vi. -ㄴ/ㄹ 듯(도) 하다

[예시] 따스한 손길로 소리 없이 머리를 어루만져 주시는 듯도 했다.
→ '어루만지다'의 ARG1로 '머리를'을 할당 / '주시다', '듯도 했다'는 각각 보조용언, 의사보조용언이므로 의미역 할당하지 않음.

vii. ㄹ 것 같다

[예시] 비가 곧 올 것 같다.
→ '오다'의 ARG0로 '비가'를 할당

viii. ㄹ 것을(걸) 그랬다

[예시] 동창들한테 협찬금 받아낼 걸 그랬잖다.
→ '받다'의 ARG1로 '협찬금'을, ARG2로 '동창들한테'를 할당

ix. -어서는 안 되다

[예시] 우리는 경계를 늦추어서는 안 된다.
→ '늦추다'의 ARG0로 '우리는'을, ARG1으로 '경계를'을 할당

x. -고 해서

[예시] 시간도 없고 해서 그는 친척집에 들르지 않았다.
→ '없다', '들르다'에 대해서만 의미역 할당

xi. -든지 하다

[예시] 철수는 밥을 먹든지 빵을 먹든지 할 것이다.
→ 두 개의 '먹다'에 대해서만 의미역 할당

xii. -기로 하다

[예시] 그가 나와 같이 가기로 하였다.
→ '가다'의 ARG0로 '그가'를 할당
심의를 거치지 못한 법안은 9일 다시 임시국회를 열어 처리하기로 했다.
→ 처리하다: 심의를 거치지 못한 법안은(ARG1)

xiii. -기{도/만/는} 하다

[예시] 철수는 하루 종일 음악을 듣기만 하였다.
→ '듣다', '하다'에 대한 의미역을 각각 할당

xix. -기 시작하다

[예시] 철수가 밥을 먹기 시작하였다.
→ 먹다: 철수가(ARG0), 밥을(ARG1)

xx. -ㄴ/ㄹ지(도) 모르다

[예시] 그가 집에 갈지도 모른다.
→ '가다'에 대한 의미역을 할당

바. 관형절의 논항 범위 설정과 유형별 처리

- 관형절의 경우 논항 중 하나의 어절에만 태깅하는 것이 아니라, 해당 논항의 범위 전체를 설정한다.
- 논항이 시작되는 어절에는 해당 의미역(ARG0, ARG1 등)의 표지를 부착하고, 논항이 끝나는 어절에 '>>>'을 부착하는 방식으로 나타낸다.
- 기본적으로 관형절 전체와 핵어 명사까지를 모두 범위로 설정하는 것을 원칙으로 한다. 구체적인 유형별 처리 예시는 아래와 같다.

1) 관계관형절

[예시] 영희는 비싼(ARG1) 화장품을(>>>) **좋아한다**.
영희는 가격이(ARG1) 많이 비싼 화장품을(>>>) **좋아한다**.

2) 관형절 + NP의 + 핵어명사

[예시] 영희는 가격이(ARG1) 많이 비싼 화장품의 좋은 향기를(>>>) **좋아한다**.

3) '와/과' 접속 명사구가 포함된 관형절

[예시] 영희는 가격이(ARG1) 많이 비싼 화장품과 예쁜 옷을(>>>) **좋아한다**.

4) 문장으로 환원되지 않는 관형절

[예시] 영희는 어린(ARG1) 시절을(>>>) **회상했다**.
영희는 아기가(ARG1) 자고 있는 동안을(>>>) **활용했다**.

5) '것' 명사절

[예시] 영희는 아기가(ARG1) 잠든 것을(>>>) **몰랐다**.

사. 인용절의 처리

인용절은 길이 및 언어학적 단위와 상관없이 전체를 범위로 설정한다.

[예시] 영희는 "아기가(ARG1) 잠들었어."라고(>>>) **말했다**.
그들은 국가폭력에(ARG1) 무고한 국민이 희생되는 것을 막으려면 공소시효와 관계없이 가해자들을 처벌하는 특별법을 제정해야 한다고(>>>) **주장한다**.
그들은 '음악만(ARG1) 해도 먹고살 수 있고, 영화만 찍어도 먹고살 수 있으면 좋겠다'고(>>>) **말했다**.

발화자가 동일한 인용문 2개가 연속해서 나오는 경우, 마지막 서술어('말했다', '했다' 등)의 의미역은 마지막 인용문만 할당한다.

[예시] 김 장관은 "이번 조치는 역사적 과오를 망각한 것이라고 생각된다"며 "우리(ARG1) 정부도 단호히 대처해 나갈 것"이라고(>>>) **말했다**.
→ '말했다'

아. 다중 문장성분 출현 문장의 처리

주어나 목적어가 두 번 이상 나타난 문장의 경우 각각 문장의 성분으로서 서술어와 연결되어 있다는 점을 고려하여 중복역을 할당한다.

[예시] 코끼리가(ARG1) 코가(ARG1) 길다.
나는(ARG0) 사과를(ARG1) 반을(ARG1) 먹었다.

이중 목적어 구문의 경우, '을/를' 조사 형태가 실현된 경우에 한하여 목적어로 처리하고, 그렇지 않은 경우에는 **최대한 적절한 부가역을** 할당한다.

[예시] 내가(ARG0) 반이나(ARGM-EXT) 먹은 사과
내가(ARG0) 반만(ARGM-EXT) 먹은 사과
내가(ARG0) 반(ARGM-EXT) 먹은 사과

특히 서술어에 따라 대상역이 두 개 이상 출현하는 가운데 그중 하나가 관형절을 내포하고 있는 경우에는 '을/를' 또는 '이/가' 표지에 따라 대상역을 할당하는 것에 주의한다.

[예시] 내가 반을 먹은 사과를 친구에게 주었다.
→ 먹다: 내가(ARG0), 반을(ARG1), **사과를(ARG1)**
→ 주었다: 내가(ARG0), 반을 먹은 사과를(ARG1), 친구에게(ARG2)
중산층 백인 가정 자녀는 **68%가** 부모 세대보다 많이 벌었지만 중산층 흑인 가정 자녀는 **31%만** 부모 세대보다 많이 벌었다.
→ 격조사 표지 출현 여부에 따라 필수역과 부가역을 구분하므로, '68%가'는 첫 번째 '벌다'에 대한 ARG0, '31%만'은 두 번째 벌다'에 대한 ARGM-EXT로 처리

자. 관용 표현의 문제

프레임셋에 관용 표현 전체가 의미역을 할당할 수 있다는 것이 반영되어 있지 않으므로 용언의 프레임셋을 기준으로 처리한다. 이때 중복역이 발생할 수 있다.

[예시] 그 모습에(ARG3) 눈길이(ARG0) 갔다.
선생님이(ARGA) 철수를(ARG0) 비행기를(ARG1) 태웠다.

차. 접속 명사구의 처리

쉼표, '및', '와/과', '간' 등으로 접속된 명사구의 경우 전체를 범위로 설정한다.

[예시] 대학생(ARG0) 및 졸업생들이(>>>) 미국에서 최장 18개월 동안 **체류하며** 어학연수를 할 수 있도록 한 제도.
미국 국무부가 관리하는 스폰서 업체가 연구기관과(ARG1)일자리를 (>>>) **알선해** 준다.
이 기사의 취재에는 김철수(26、서강대(ARG0) 4학년), 송영희(24、서울시립대 3학년) 씨가(>>>) **참여했습니다**.
여야(ARG1) 간 이견이(>>>) **없는** 법안 2건만 이번 임시국회에서 협의한다.

카. 연속 동사구의 처리

띄어쓰기가 이루어지지 않은 연속 동사구의 경우 선행 용언의 프레임셋을 기준으로 분석한다.

[예시] 그는(ARGA) 혼자 식구들을(ARG0) 먹여살렸다.
→ Kpropbank '먹다'의 프레임셋 적용

타. 절 경계를 넘어서는 성분의 처리

절 경계를 넘어서는 성분, 즉 주절의 서술어에 대한 주어 또는 목적어 등이 내포절에 포함되어 있는 경우에는, 해당 문법 관계가 분명치 않은 경우를 제외하고는 최대한 복원하여 논항을 할당한다.

[예시] 유 의원 측은 "수신료 인상안을 여당 간사와 공식적으로 논의한 것은 어제가 처음이었다"며 "여당이 기습적으로 상정한 것은 입법 취지에 어긋난 것"이라고 말했다.
→ 주절의 서술어 '말했다'의 의미상 주어는 선행절의 '유 의원 측'이 됨. 절 경계를 넘어서지만 '말했다'의 ARG0 논항으로 '유 의원 측'을 할당함.
김 장관은 해당 부처와의 협의를 통해 법안을 처리하기로 했다.
→ '김 장관'은 선행절에 포함되어 있고 '처리하기로'는 후행절에 포함되어 있음. 또한 **'처리하다'의 주체가 엄밀한 의미에서 '김 장관'인지 여부를 문맥이 없이 판**

단하기 힘들지만, 해당 부처의 책임자라는 것을 고려하면 의미적으로 주어로 파악할 수 있음. 따라서 '처리하다'의 ARG0로 '김 장관'을 할당함.

제 4 장

결론

이 과제에서는 국립국어원에서 구축한 문어 말뭉치 중 신문 텍스트 200만 어절을 대상으로 문장 단위의 의미역 분석을 진행하였다. 각 절의 서술어를 기준으로 의미역 분석 정보를 부착하였으며 JSON 형식으로 납품하였다. 이 과정에서 의미역 분석의 기준이 되는 술어 의미역 정보 기술 형식을 상대역 방식으로 통일하였고 기존의 의미역 기술 목록 통합으로도 해결되지 않는 서술어의 의미역 정보를 추가로 기술하였다. 또한 한국전자통신연구원(ETRI)의 한국어 의존의미역 주석가이드라인을 기반으로 하여 구체적인 의미역 할당 지침과 의미역 정보 적용 등에 대한 실무 지침을 개발하였다. 의미역 분석 지침을 수립하고 수작업 검수 도구를 최적화한 후에 자동 분석 결과를 수작업으로 검수하였으며 의미역 분석 결과물의 정확도와 일관성을 높이기 위해 공학적 검증 과정을 거쳤다. 지침 개발부터 최종 검수에 이르기까지 한국어 의미역 분석의 전 과정을 수행하면서 얻은 경험을 바탕으로 향후 한국어 의미역 분석 자원의 구축과 활용을 위해 다음과 같이 제언을 하는 것으로 결론을 갈음한다.

현재 국내의 의미역 분석 방식은 개별 용언에 대해 의미역 구조를 순차적 번호로 기술해 놓은 의미역 정보에 의존하여 의미역을 주석하는 것으로 통일되어 있다. 이는 영어권의 프롭뱅크 의미역 체계를 한국어에도 적용한 것으로 의미역 주석 표지가 필수역과 부가역으로 나뉘어 주석된다. 이러한 방식은 주석 대상 말뭉치에 대한 의미역 정보의 점유율이 충분히 확보되고 일관되게 기술되어 양과 질의 측면에서 견고한 언어 자원을 기반으로 주석되었을 때 활용도가 높은 방식이다. 그러나 아직 한국어 의미역 분석을 위한 의미역 정보는 구축 주체마다 다른 목록의 용언에 대해 서로 다른 방식으로 기술되어 있으며 용언의 센스를 구획하는 기준도 상이하다. 이번 의미역 분석 말뭉치 구축 시에 이러한 단점을 보완하기 위해 달리 기술되어 있는 의미역 정보의 형식을 최대 4항의 순차적 번호로 변환하여 통일하였지만 용언의 센스 코딩 기준이 다르고 센스의 의미영역이 겹치는 문제 등을 완전히 해결하는 것은 불가한 작업이었다.

과제 초반에 연구진에서 제안한 대로 불완전한 의미역 정보에 의존한 상대역 분석 방식 대신 절대역 분석 방식으로 전환할 가능성에 대해 검토할 필요가 있다. 상대역 방식은 의미역 정보 미등재어의 처리가 불가하며 통사 · 의미적으로 유사한 어휘의 의미역 정보를 적용한다고 하더라도 유의어 정보가 안정되게 공급되어 자동 분석에 반영되지 않는다면 기존의 의미역 정보에 포함되어 있지 않은 어휘의 의미역 정보 구축이 지속되어야 한다는 한계가 있다. 의미역 분석 대상 말뭉치의 레지스터가 구어나 웹(WEB)이거나 장르가 신문, 잡지 등인 경우에는 고정된 의미역 정보로 분석할 수 있는 문장의 비율이 낮아서 구축의 난도가 높고 활용성이 떨어지는 결과를 가져온다.

최근 AMR(Abstract Meaning Representation)을 비롯한 의미 자원의 구축 방식이 프롭뱅크 기반에서 특정 의미 자원에 의존하지 않는 절대역 방식으로 전환되고 있다는 국제적인 흐름을 고려할 때 한국어 의미역 분석 말뭉치도 절대역 방식을 취할 가능성이 없는지 재고해 보아야 한다. 현재까지 상대역 방식으로 구축된 의미역 분석 말뭉치는

의미역 정보와의 매핑을 통해서 절대역으로 변환이 가능하기 때문에 기존 자원과의 호환성은 고려의 대상으로 삼지 않아도 될 것이다.

더욱 활용도가 높은 의미역 말뭉치를 구축하기 위해서는 정보의 일관성을 확보하기 위해 필수적인 부가역 분석의 세부 지침 및 부가역 적용 표현 목록의 확정이 필요하다. 마지막으로 질의 · 응답 등의 응용 영역에서의 활용을 위해서는 단순히 통사적인 구문 구조에 기계적으로 대응하는 의미역 범위 설정에서 한 단계 나아가 인간의 직관으로 수용할 수 있는 의미역 범위 설정 방안에 대한 연구도 필요하다.

<Abstract>

Building a corpus through the semantic role labelling

The purpose of this project is to build a semantic domain analysis corpus in respons e to South Korea's demand for large-scale linguistic resources with the development o f artificial intelligence. Additionally, this project aims to develop guidelines for analytic al corpus of semantic domains required in this process.

This project consists of two main parts. The first is constructing a semantic analysis corpus based on the comparative research of corpus analysis guidelines in related fiel ds such as the Korea Electronics and Telecommunications Research Institute (ETRI). B ased on the on-site guidelines, we have developed guidelines for semantic role taggin g by revising and improving existing guidelines. The second is building a corpus (2 mil lion words) for semantic analysis based on the corpus annotation procedure for seman tic analysis.

○ Formulating guidelines to build a sematic role annotated Korean corpus

Most previous semantic role annotated Korean corpora were compiled by applying di fferent semantic role annotation markers and framesets under each construction subje ct. We reviewed the methodologies used to annotate such corpora as a guideline for constructing a national-level semantic analysis corpus. In this project, a practical guid eline for building a corpus of dependent semantic role analysis was developed mainly based on the Korea Electronics and Telecommunications Research Institute (ETRI)'s gu idelines for semantic role annotation. We suggest a practical guideline with contents a nd examples complementing what was insufficient in the previous dependency parsing guidelines.

○ Constructing semantic-role tagging corpus (2 million words)

The procedure for a construction of semantic-role tagging is as follows:

Establishing Semantic role tagging guideline > Customizing inspection tools > researc her training> Automatic semantic role tagging analysis > Manual analysis by researche rs (primary inspection)> Supervision and coordinator inspection (secondary inspection)>

Verifying through Deep learning (Deep learning based on base accuracy and consisten cy)> Final result

Based on the guidelines for semantic domain analysis, 2-million-word-scale semantic domain analysis corpus was compiled. The corpus contains texts from newspaper articl es with the size of 2 million words.

In this project, we conducted a semantic domain analysis of the corpus provided by the National Institute of Korean Language and separated it sentence by sentence. The verbs and adjectives' semantic analysis information were attached to the units of no un phrases and clauses, and the final output was submitted in JSON format.

In the establishment process multiple automatic semantic analyzers were used for th e convenience and consistency of work, and the automatically analyzed results were manually reviewed by the researchers.

The detailed process is as follows.

First, the corpus of newspaper articles provided by the National Institute of the Kor ean Language is analyzed automatically in a sentence level. When multiple automatic parsers give the same results, the results were submitted after a final inspector's ch eck. When automatic parsers generate different results, different researchers review t he sentence in the first and second inspection, and the final inspector would review t he sentence.

The researchers were organized into four groups, including the co-researcher in cha rge and the team leader. Researchers would review the results first, and then profess ors and team leaders in each group inspected the results as a second inspection. Afte r the second inspection, the final results were submitted after post-processing using a n algorithm and file format conversion.

In this project, the first automatic analysis was carried out using multiple dependent semantic role analyzers of ETRI, Kangwon National University, Jeonbuk National Unive rsity to improve the reliability by cross-validating and integrating the results.

In the process of automatic parsing, consistency was verified by deep learning meth od, and the methodologies used for the first and second inspections by researchers w ere also complemented during this project.

Key words: semantic, semantic tagging, semantic analysis, semantic annotation, corpus, sematic analysis corpus, the construction of a sematic analysis Korean corpus

Project Director: Yim Seongmo(MindsLab)

<부록 1> 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_le			array(string)	* 분석 층위 (복수 나열) * 원시, 형태 분석, 개체명 분석,

		vel				어휘의미 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)
		SRL			array(object)	* 의미역 분석 정보
			predicate		object	* 서술어 정보
				form	string	* 서술어
				begin	number	* 서술어의 문장 내 시작 위치. (UTF-8 문자 위치로 0부터 시작)
				end	number	* 서술어의 문장 내 끝 위치 (UTF-8 문자 위치로 0부터 시작)
				lemma	string	* 서술어의 표제어 (서술어의 기본형)
				sense_id	number	* 서술어의 의미 번호
			argument		array(object)	* 논항 정보

				form	string	* 논항
				label	string	* 의미역 태그
				begin	number	* 논항의 문장 내 시작 위치. (UTF-8 문자 위치로 0부터 시작)
				end	number	* 논항의 문장 내 끝 위치 (UTF-8 문자 위치로 0부터 시작)

<부록 2> 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": "NWRW1800000021-0205",
            "metadata": {
                "title": "동아일보, 조선일보, 한겨레 2009~2017년 기사",
                "author": "동아일보, 조선일보, 한겨레",
                "publisher": "동아일보사, 조선일보사, 한겨레",
                "url": "https://www.korean.go.kr/",
                "date": "20090629",
                "category": "신문 > 정치",
                "annotation_level": ["형태 분석", "어휘의미 분석", "구문 분석"],
                "note": ""
            },
            "sentence": [
                {
                    "id": "NWRW1800000021-0205.19",
                    "form": "세계유산위원회는 27일 등재 결정문에서 왕릉 주변 개발 완충 지역 내 개발의 가이드라인을 만들라고 권고했다.",
                    "word": [
                        { "id": 1, "form": "세계유산위원회는", "begin": 0, "end": 8 },
                        { "id": 2, "form": "27일", "begin": 9, "end": 12 },
                        { "id": 3, "form": "등재", "begin": 13, "end": 15 },
                        { "id": 4, "form": "결정문에서", "begin": 16, "end": 21 },
                        { "id": 5, "form": "왕릉", "begin": 22, "end": 24 },
                        { "id": 6, "form": "주변", "begin": 25, "end": 27 },
                        { "id": 7, "form": "개발", "begin": 28, "end": 30 },
                        { "id": 8, "form": "완충", "begin": 31, "end": 33 },
                        { "id": 9, "form": "지역", "begin": 34, "end": 36 },
                        { "id": 10, "form": "내", "begin": 37, "end": 38 },
                        { "id": 11, "form": "개발의", "begin": 39, "end": 42 },
                        { "id": 12, "form": "가이드라인을", "begin": 43, "end": 49 },
                        { "id": 13, "form": "만들라고", "begin": 50, "end": 54 },
                        { "id": 14, "form": "권고했다.", "begin": 55, "end": 60 }
                    ],
                    "SRL": [
                        {
                            "predicate": { "form": "만들라고", "begin": 50, "end": 54, "lemma": "만들", "sense_id": 4444401 },
                            "argument": [
                                { "form": "27일 등재 결정문", "label": "ARGM-LOC", "begin": 9, "end": 19 },
                                { "form": "왕릉 주변 개발 완충 지역 내 개발의 가이드라인", "label": "ARG1", "begin": 22, "end": 48 }
                            ]
                        },
                        {
                            "predicate": { "form": "권고했다.", "begin": 55, "end": 60, "lemma": "권고", "sense_id": 4444401 },
                            "argument": [
                                { "form": "세계유산위원회", "label": "ARG0", "begin": 0, "end": 7 },
                                { "form": "27일 등재 결정문", "label": "ARGM-LOC", "begin": 9, "end": 19 }
                            ]
                        }
                    ]
                }
            ]
        }
    ]
}
```

<부록 3> 원시 말뭉치 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 만들라
고 권고했다.</s>
        </p>
        <byline>윤완준 기자 zeitung@donga.com</byline>
    </text>
</SJML>
```

<부록 4> JSON 구조의 술어 의미번호 부여 체계

술어 정보	의미번호(sense_id)	비고
K-propbank	44444**	펜실베니아 대학에서 구축한 한국어 의미역 프레임셋
Etri	55555**	K-propbank의 형식대로 한국전자통신연구원에서 추가 구축한 한국어 의미역 프레임셋
U-propbank	6******	울산대 구축 의미역 프레임셋
우리말샘	***	우리말샘의 격틀 정보를 활용한 의미역 프레임셋
기존 격틀에 어형 없음	777	비슷한 의미와 논항 구조를 가지는 술어를 참조하여 의미역 주석
기존 격틀과 어형은 일치하나 의미가 불일치	888	비슷한 논항 구조를 가지는 술어를 참조하여 의미역 주석
오탈자 포함	999	미주석

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)
	임병현(주식회사 마인즈랩)
	이하영(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)
	김선혜(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)
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)
박광현(전북대학교)
이영훈(전북대학교)
홍승연(전북대학교)
박성식(강원대학교)
신영진(한국해양대학교)
강일민(충남대학교)
박요한(충남대학교)
정혜지(충남대학교)
박석주(한림대학교)
정영석(한림대학교)
오세은(한림대학교)
황석주(한림대학교)
강동찬(전북대학교)
이종현(전북대학교)
최형준(전북대학교)
김담린(강원대학교)

김보은(강원대학교)
김홍진(강원대학교)
오신혁(강원대학교)
박상원(주식회사 딥네추럴)
박정수(주식회사 딥네추럴)
허민강(주식회사 딥네추럴)
박연호(주식회사 딥네추럴)
정동호(주식회사 딥네추럴)
최진혁(주식회사 딥네추럴)
김예진(주식회사 딥네추럴)
이규덕(주식회사 딥네추럴)
임선민(주식회사 딥네추럴)
담당 연구원 이승재(국립국어원 언어정보과장)
이선영(국립국어원 언어정보과 연구원)

발행인: 국립국어원장
발행처: 국립국어원
서울시 강서구 금낭화로 154
전화 02-2669-9775, 전송 02-2669-9757
인쇄일: 2020년 1월 30일
발행일: 2020년 1월 30일
인 쇄: 비즈카피

※ 이 책은 국립국어원의 용역비로 수행한 '의미역 분석 말뭉치 구축' 사업의 결과물을 발간한 것입니다.