국립국어원 2019-01-53

발 간 등 록 번 호
11-1371028-000775-01

상호 참조 해결 말뭉치 구축

사업 책임자
곽용진

국립국어원

제 출 문

국립국어원장 귀하

국립국어원과 체결한 연구용역 계약에 따라 ‘상호 참조 해결 말뭉치 구축’에 관한 연구 보고서를 작성하여 제출합니다.

■ 사업기간: 2019년 7월 15일 ~ 2020년 1월 15일

2020년 1월 15일

사업 책임자: 곽용진(이르테크)

사업 수행기관 ㈜이르테크, 연세대학교

사업 책임자 곽용진

사업 참여자 이순미, 최지선
이석재, 윤영민, 최지명
장호림, 이선덕, 홍은기
한문성, 김상선, 서보원
김영환, 박재은

<사업 수행 기관>	㈜이르테크, 연세대학교 산학협력단

사업 책임자	곽용진(이르테크, PM)
사업 참여자	이순미(이르테크, 사업 관리)
	최지선(이르테크, 지침수립)
	이석재(연세대학교, 구축관 리)
	윤영민(연세대학교, 지침 수립)
	최지명(연세대학교, 구축 관리)
	장호림(이르테크, 품질 보증)
	이선덕(이르테크, 품질 보증)
	홍은기(이르테크, 품질 보증)
	한문성(이르테크, 기술 지원)
	김상선(이르테크, 기술 지원)
	서보원(이르테크, 기술 지원)
	김영환(이르테크, 기술 지원)
	박재은(이르테크, 기술 지원)
연구 보조원	강혜린(연세대학교, 말뭉치 구축)
	김유경(연세대학교, 말뭉치 구축)

	김향수(연세대학교, 말뭉치 구축)
	임은아(연세대학교, 말뭉치 구축)
	전휘목(연세대학교, 말뭉치 구축)
	김영상 (연세대학교, 말뭉치 구축)
	최소영(연세대학교, 말뭉치 구축)
	곽유석(성균관대학교, 말뭉치 구축)
	김민정(성균관대학교, 말뭉치 구축)
	김정원(성균관대학교, 말뭉치 구축)
	김정윤(성균관대학교, 말뭉치 구축)
	김종희(성균관대학교, 말뭉치 구축)
	김지연(성균관대학교, 말뭉치 구축)
	서혜진(성균관대학교, 말뭉치 구축)
	이나래(성균관대학교, 말뭉치 구축)

차 례

제1장 사업 개요

제2장 사업 수행 내용

차 례

제3장 향후 계획

제 1 장

사업 개요

1. 사업일반

1.1. 사업명

상호 참조 해결 말뭉치 구축 사업

1.2. 사업 수행 기간

2019년 7월 15일 ~ 2020년 1월 15일

2. 사업 목적

4차 산업혁명으로 인한 대규모, 고품질 우리말 자원의 수요 증대에 따라, 기초 말뭉치의 양적, 질적 부족에 따른 기반 기술을 개발하고, 인공지능 기술 수준 지체를 해소하기 위해, 국어 자원의 활용도와 가치를 높일 수 있도록 민간에서 활용 가능한 국가 공공재로서의 말뭉치를 확대 구축한다.

3. 사업 범위

본 사업은 개체명, 구문분석 등의 다른 분석 말뭉치 구축 사업과는 다르게, 기존 말뭉치 구축 지침의 표준이 존재하지 않고 해외에서의 연구는 활발하나 국내 연구 및 구축이 거의 이루어지지 않은 분야의 분석 말뭉치로 단순한 양적, 질적 기준이 아닌 앞으로 한국어 분석 말뭉치 연구, 활용의 시금석이 되는 사업으로 말뭉치의 품질, 활용, 구축 지침의 표준을 제시하는 사업이다.

ㅇ 최초의 표준 지침 수립

- '한국정보통신기술협회(TTA)' 등 기타 관련 분야 분석 표지 및 분석 지침을 검토하고, 기존 외국어 연구 사례와 한국어 특성에 대한 지침을 비교한다.
- 기존 지침의 문제점 분석 및 보완책을 제시하여 상호 참조 해결 말뭉치 지침을 수립하며, 세부 지침, 파일명 부여방식, 표지 부착 방식, 형식 등은 주관기관과 협의한다.

ㅇ 지침에 따른 말뭉치 가공

- 원시말뭉치 300만(문어 200만, 구어 200만)어절을 대상으로 문단(문서) 단위로 원어절에 맞추어 상호 참조 해결 결과를 구축하며, 문단(문서) 내 각 명사에 대한 상호 참조 해결 정보를 부착한다.

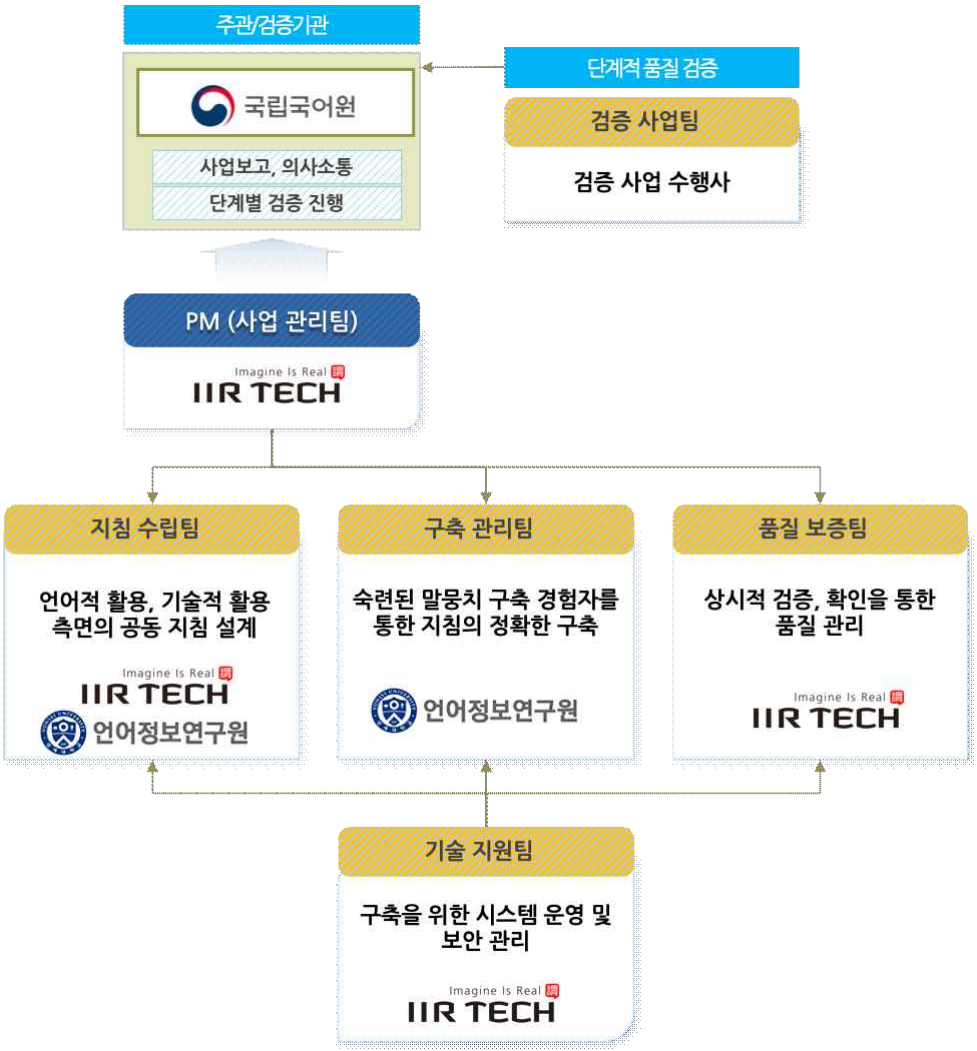
ㅇ **말뭉치의 단계적 품질 검증**

- 오류율 5% 이내 달성을 위해서 검증 사업자와 단계적으로 품질 점검 및 실행 결과를 반영한다.
- 품질 검증 사업자와 지침 표준화 및 검증 데이터 포맷을 표준화한다.

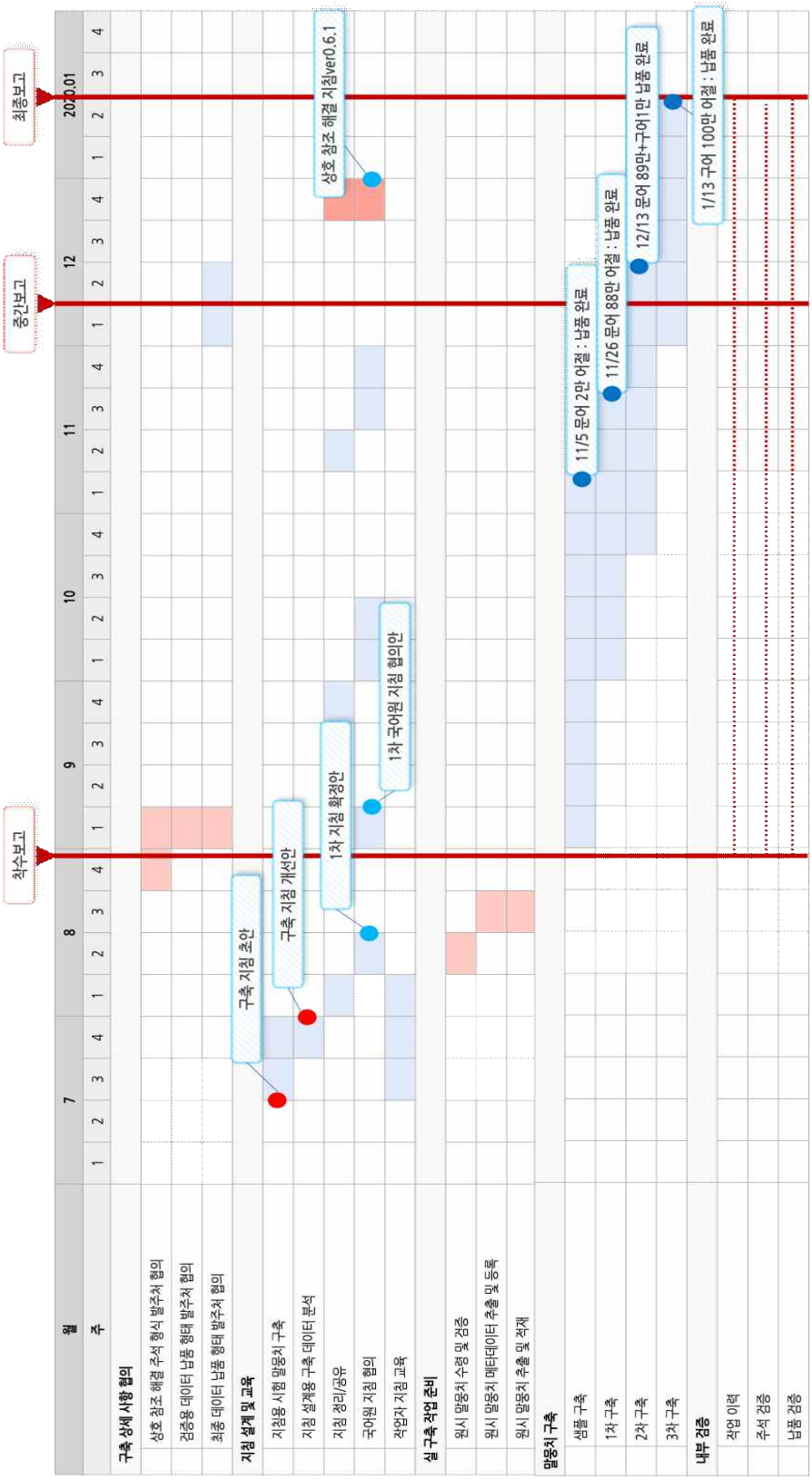
ㅇ **말뭉치 활용성 검증 및 납품**

- 단순히 데이터 자체의 품질만을 확보하여 납품하는 것이 아닌 4차 산업에 활용 가능한 범위에 대한 검증을 실시하기 위하여 산업에서의 활용을 위한 포맷의 결과를 납품한다.

4. 사업 수행



5. 사업 추진 경과


착수보고
중간보고
최종보고
구축 지침 초안
구축 지침 개선안
1차 지침 확정안
1차 국어원 지침 협의안
상호 참조 해결 지침ver0.6.1
11/5 문어 2만 어절 : 납품 완료
11/26 문어 88만 어절 : 납품 완료
12/13 문어 89만+구어1만 납품 완료
1/13 구어 100만 어절 : 납품 완료

제 2 장

사업 수행 내용

1. 수행 환경 구성

1.1. 원시 데이터

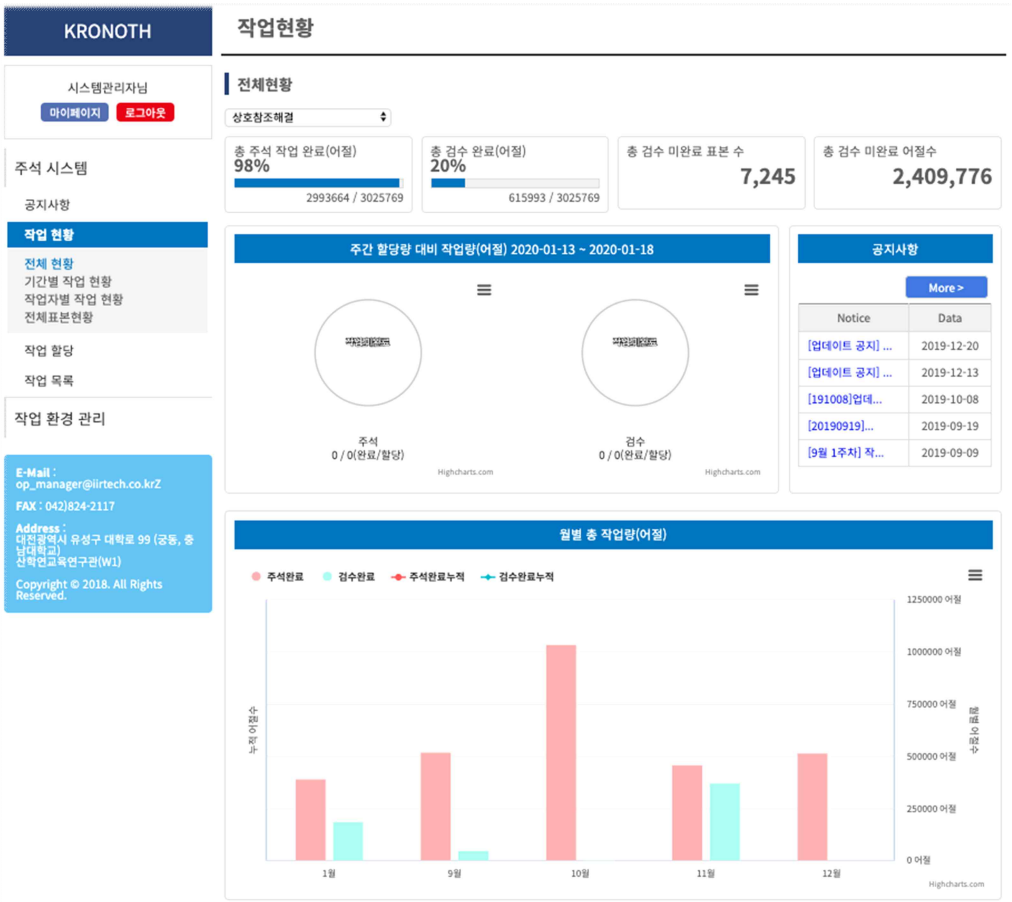
사업 수행 범위에 따라서 국립국어원에서 제공한 원시 말뭉치 약 300만 어절(문어 2,019,322어절, 구어 1,006,447어절)을 수령했으며, 이후 지침 초안을 위한 샘플 데이터 구축하였다.

1.2. 수행 환경 구성

1.2.1. 데이터 구축 시스템 준비 현황

활용 System은 ㈜이르테크 내부 시스템 'KRONOTH Annotation System'이다. 이 System은 승인된 인가자만이 대상 System에 접근하여 제어 및 구축 작업을 할 수 있도록 설계되었으며, 상호 참조 해결 구축 작업에 대한 실시간 현황 관리가 가능하도록 만든 Web System이다.

<그림 1> 작업 현황 관리 - KRONOTH Annotation System



1.2.2. 수행 조직 및 인력 구성

조 직	이 름	담당 업무
사업 관리팀	곽용진	•프로젝트 관리 및 통제 (PM) 및 지침 연구, 설계, 협의
	이순미	•의사 소통 지원 및 행정 지원 (대관업무 및 비용 관리 등)
지침 수립팀	곽용진	•데이터 구축 지침, 표지 부착 방식 등 구축 지침 연구, 설계 •국립국어원과 구축 지침 협의 및 수립
	최지선	•지침 연구 자료 조사 및 수집 데이터 및 샘플 데이터 결과 분석
	윤영민	•구축 지침 연구, 설계 및 지침 운용 현황 확인, 개선안 수립
구축 관리팀	이석재	•구축 관리 수행관리 및 통제
	윤영민	•상호 참조 해결 말뭉치 검수 작업 관리, 작업 배분 및 진척률 관리
	최지명	•상호 참조 해결 말뭉치 검수 작업 관리, 작업 배분 및 진척률 관리
	강혜린	•상호 참조 해결 정보 검수 작업
	김유경	•상호 참조 해결 정보 검수 작업
	김향수	•상호 참조 해결 정보 태깅 작업
	임은아	•상호 참조 해결 정보 태깅 작업
	전휘목	•상호 참조 해결 정보 태깅 작업
	김영상	•상호 참조 해결 정보 검수 작업
	최소영	•상호 참조 해결 정보 태깅 작업
	곽유석	•상호 참조 해결 정보 검수 작업
	김민정	•상호 참조 해결 정보 검수 작업
	김정원	•상호 참조 해결 정보 태깅 작업
	김정윤	•상호 참조 해결 정보 태깅 작업
	김종희	•상호 참조 해결 정보 태깅 작업
	김지연	•상호 참조 해결 정보 태깅 작업
	서혜진	•상호 참조 해결 정보 검수 작업
	이나래	•상호 참조 해결 정보 태깅 작업
품질 보증팀	장호림	•구축 데이터 형식 오류 검사, 규격 검사, 검증 결과 관리
	이선덕	•구축 데이터 형식 오류 검사, 규격 검사, 검증 결과 관리
	홍은기	•산출물 및 검증 결과 문서화 작업
기술 지원팀	한문성	•시스템 아키텍쳐 구성 및 분석 말뭉치 기술 자문
	김상선	•본문 주석 등 부가 필요 주석 개발 지원
	서보원	•사업 수행, 관리, 및 데이터 추적 관리 지원
	김영환	•시스템 설치 및 환경 구성, 보안 관리
	박재은	•납품 데이터 변환 생성, 납품 데이터 변환 추출기 개발

1.3. 초기 지침 및 작업자 교육

TTA(ETRI 상호참조해결 태깅 가이드라인 v3.1) 지침에 따라 작업도구를 수정 개발하고, 작업자들을 대상으로 지침 교육 및 해당 지침에 따른 사용 교육을 실시하였다.

작업도구 수정 개발 후 수령된 원시 말뭉치 대상의 테스트용 데이터 구축하였으며, 이를 분석하여 수립된 지침에 따라 추가적으로 수정 개발된 도구의 사용법을 기준으로 사용자 지침/시스템 사용 교육을 실시하였다.

지침 수립 및 변경 후에는 1주일 이내, 변경된 지침에 맞게 작업도구의 변경을 지원하고, 작업자 지침 교육 및 이후 시스템 사용자 교육을 진행하였다.

<그림 2> 초기 지침 수립 및 작업자 교육 순환도

1.4. 데이터 납품, 검증 정책

샘플 및 1, 2, 최종 납품 데이터 주석 형식 및 File Format 사전 정의를 실시하고, 원시 데이터 File Format 및 검증, 납품 형식, 기준에 따른 System Customizing 하였다.

1차 납품, 검증을 위한 구축 작업 전 System Customizing을 위해 본문 주석, File 형식, 기준을 정립하였다.

2. 지침

2.1. 지침 수립 계획

상호 참조 해결 말뭉치 구축 지침 수립을 위해 관련 선행 사례 분석을 통해 기존 지침 및 구축 공정의 문제점, 기구축 도구의 보완점을 도출하고 검토하고 도출한 문제점과 보완 사항을 지침 설계에 반영하는 계획을 수립하였다. ETRI의 '상호 참조 해결 태깅 가이드라인v3.1'을 기준 지침으로 삼아 지침의 형식 보완 및 지침 내용 추가를 통하여 지침을 수립하고 반복적인 개선 작업을 통해 지침을 개선하는 세부 계획을 만들어 수행하였다. 선행 사례를 크게 세가지로 분류하여 이슈 사항을 도출한 결과는 아래와 같다.

○ 기존 지침 분석

기존 상호 참조 해결 말뭉치 구축 지침은 상호 참조 태깅 대상인 멘션(mention)[1] 판단과 상호 참조 태깅 범위 판단이 어렵다는 문제가 있다. 구축 지침은 결국 사람이 보고 활용하는 것이기 때문에 용어에 대한 정의나 상호참조 태깅 예시 및 사례를 보완하여 구축자가 다양한 예시를 통해 다양한 텍스트 문맥 안에서 멘션 추출과 상호참조 해결 링크를 판단할 수 있도록 해야 한다.

○ 기 구축 말뭉치 분석

4차 산업 혁명에 활용 가능한 언어자원으로서 구축 말뭉치는 다양한 형식과 포맷을 지원해야 한다.

○ 구축 도구 분석

기 말뭉치 구축 도구는 전처리 및 자동화 기능의 균형성이 부족한 문제가 있으며 작업 관리적 기능이 부족하다는 문제가 있다. 또한 상호 참조에 대한 시각화 표현이 부족하다. 또한 구축 도구는 지침에 따라 구축 할 수 있도록 지침의 반영이 가능한 도구여야 한다.

2.2. 지침 수립 방안

기존 지침의 문제점 분석과 보완점을 도출한 후 수립할 지침에 반영하기 위해서 기존 지침의 내용은 유지하되 형식과 표현방법을 수정 보완하였다. 사례 기반의 지침 형식을 통하여 작업자가 쉽게 이해하고 태깅 작업에 도입이 가능하도록 중점을 두고 지

1) 이때 멘션(mention)이란 상호참조해결의 대상이 되는 모든 명사구(즉, 명사, 명사구 등)를 의미하는데, 이에 대한 자세한 내용은 '부록'에 수록된 '상호참조해결 구축 지침'을 참고하기 바란다.

침을 수립하였다. ETRI의 ‘상호 참조 해결 태깅 가이드라인v3.1’을 기준으로 다음과 같은 수정, 보완을 통해 지침을 개선하였다.

○ 구축 지침 형식 수정, 보완

○ 구축 지침 내용 보완

○ 구축 데이터 형식 수립

2.2.1. 선행 연구

2.2.1.1. 국내 선행 연구

○ 엑소브레인 언어분석 말뭉치

현재 상호 참조 해결 말뭉치에 대한 국내 선행 연구는 한국전자통신연구원(ETRI)에서 진행한 엑소브레인 프로젝트가 있으며 해당 과제에서는 한국어 분석 및 질의 응답 기술을 개발하기 위한 언어분석 말뭉치를 구축 하였다. 형태소분석, 다의어 어휘의미분석, 세분류 개체명인식, 의존구문분석, 의미역인식, 상호참조해결 기술의 태깅 가이드와 자연어 질의응답을 위한 질문/정답 포맷의 뉴스기사 및 위키백과 대상 총 2,593문장 33,131어절의 태깅 말뭉치로 이루어져있다.

(1) 말뭉치 구성

엑소브레인 말뭉치의 유형별 구성 및 텍스트장르는 다음과 같으며 상호참조해결 말뭉치는 유형 3에 해당한다.

- 유형1: 위키백과QA 분야 대상 형태소분석, 세분류 개체명인식, 의존구문분석, 의미역인식(문서: 439, 문장: 725, 어절: 8,520)
- 유형2: 위키백과QA 분야 대상 형태소분석, 세분류 개체명인식, 다의어 어휘의미분석, 의존구문분석, 의미역인식, 상호참조해결(문서: 645, 문장: 1,086, 어절: 12,835)
- 유형3: 위키백과 문서 대상 형태소분석, 세분류 개체명인식, 의존구문분석, 의미역인식(문서: 19, 문장: 461, 어절: 6,490)
- 유형4: 뉴스 문서 대상 형태소분석, 세분류 개체명인식, 의존구문분석, 의미역인식(문서: 12, 문장: 321, 어절: 5,286)

(2) 말뭉치 형식

엑소브레인 언어분석 말뭉치의 형식은 UTF-8 인코딩된 JSON 및 TEXT 파일 포맷을 따른다. 상호참조해결의 입출력은 한 문서에 대한 N-Doc 구조체이며 JSON을 이용한다. N_Doc 구조체에서 상호참조해결의 각 엔티티의 결과는 “entity”안에 포함된다.

<표 1> 엔티티 N_Doc 데이터 구조

```
{ “entity” : [
        {“id” : entity`s id, “type” : NE label, “number” : {singular | plural},
         “gender” : {male | female}, “person” : {1 | 2 | 3}, “animacy” : {human | thing| time | place},
         “mention” : [{“id” : mention`s id, “text” : mention`s text, “sent_id” : sentence`s id, “start_eid” : begin id of mention(NP), “end_eid”: end id of mention(NP), “ne_id” : id of NE label}, … ]},

        {“id” : 0, “type” : “PS_NAME”, “number” : “singular”,
         “gender” : “male”, “person” : “3”, “animacy” : “human”,
         “mention” : [
                 {“id” : 48, “text” : “토니 블레어 총리”, “sent_id” : 5, “start_eid” : 4, “end_eid” : 6,     “ne_id” : 3},
                 {“id” : 49, “text” : “토니 블레어”, “sent_id” : 5, “start_eid” : 4, “end_eid” : 5, “ne_id” : 3}, {“id” : 55, “text” : “그”, “sent_id” : 6,“start_eid” : 0, “end_eid” : 0, “ne_id” : 3},{“id” : 81, “text” : “블레어”, “sent_id” : 9,“start_eid” : 5, “end_eid” : 5, “ne_id” : 3}]}, { … },
        ] }
```

○ 엑소브레인 언어분석 말뭉치 상호참조 해결 태깅 가이드라인

엑소브레인 언어분석 말뭉치의 각 층위별 태깅 가이드 라인 중 ‘상호참조 해결 태깅가이드라인 v3.1’이 있으며 상호참조해결의 개요와 태깅도구 사용법, 상호참조해결 태깅 가이드가 기술되어 있다. 상호참조해결 태깅 가이드는 크게 멘션 탐지 태깅 규칙과 추출된 멘션의 상호참조해결 태깅 규칙으로 이루어져 있다.

2.2.1.2. 해외 선행 연구

○ 일본

일본은 ‘co-reference’ 및 이에 대한 국내의 ‘상호참조’에 대하여 ‘동일 지시(同一指示)’ 또는 ‘동일 지시 관계’라고 표현하고 있다. 한국에서와 같이 대대적인 상호참조 해결 말뭉치 구축 및 결과물은 아직 소개된 바 없다.

다만, 일본에서의 ‘상호참조’ 관련 학계 연구와 보고는 1980년 초반부터 본격적으로 확인되는 가운데 일본어와 영어를 생성문법(generative grammar)의 관점에서 비교 대조하려는 이론적 접근이었다고 할 수 있다.

본 과제와 밀접한 연관성을 가진 것으로 보이는 연구를 조사한 바에 의하면 현재 일본어 내에서의 상호참조 연구는 첫째, 명사와 명사 또는 대명사 간의 지시 관계, 둘째, 명사구(NP)가 가지고 있는 ‘지표성’과 서술 관계, 셋째, 이른바 ‘~이다’ 구문(copula文) 내에서의 명사 및 명사구의 지시 체계, 넷째, 상호참조 관계를 기술하기 위한 Annotation 모델 디자인 등과 같이 구분 지을 수 있다. 특히 상호참조 관계를 기술하기 위한 Annotation 모델 디자인 등의 경우, 문어에서의 정보 추출은 ‘MUC-7’[2], 구어에서의 정보 추출은 ‘MATE’[3] 툴을 활용한 사례가 있었으며, 일본 언어처리학회에서는 이 두 가지 툴을 비교한 상호참조 분석 모델이 제안되기도 하였다. 아래에 몇 가지 사례를 소개하겠다.

(1) 명사와 명사 또는 대명사 간의 지시 관계(中村真衣佳, 2017)[4]

이 연구는 일본어의 전형적인 명사 수식구 구조인 「NのN」이 가지는 다의성이 ‘の’에 의해 결정된다고 한 기존 연구에 더하여 명사의 성질과 화용론적 표상의 영향도 함께 작용한다고 밝힌 것이다. 특히, 논증의 방법으로서 「NのN」 구조상에 나타나는 반전 현상(反転現象)을 들었는데, 반전 현상은 이른바 「青のボールペン」과 「ボールペンの青」, 「みじん切りの玉ねぎ」와 「玉ねぎのみじん切り」와 같이 「の」 전후의 명사를 치환하여도 지시 대상이 동일한 일본어의 한 현상을 가리킨다.

또한 반전 후의 「N2のN1」형식은 반전 표현이라고 불린다. 실제 언어 운용상에서 반전이 되기 전과 지시 대상이 동일한 반전 표현이 우선적으로 사용되는 경우가 있지만, 왜 반전 표현이 사용되는지, 지시 대상은 왜 동일한지에 대해서는 명백한 설명이 없는 가운

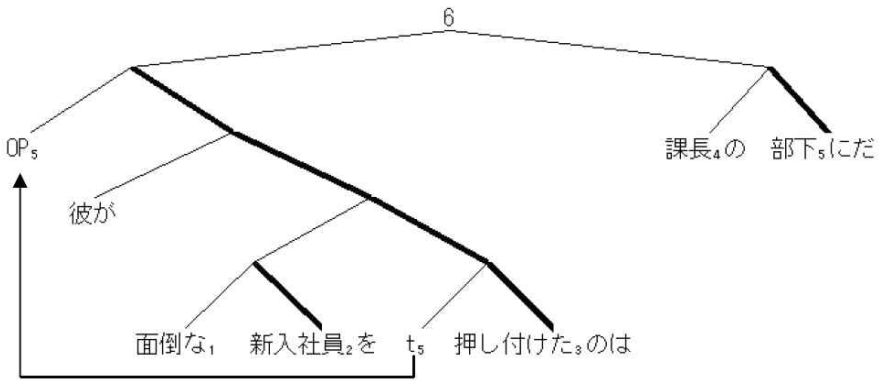
2) https://catalog.ldc.upenn.edu/LDC2001T02
3) https://www.ims.uni-stuttgart.de/suche/?q=MATE
4) 中村真衣佳(2017) 同一指示と解釈される「N1のN2」と「N2のN1」: 反転表現「N2のN1」の焦点化の要因, 北海道大学大学院文学研究科研究論集 (17), 169-183

데 이 연구에서는 그 이유를 멘션으로 잡을 수 있는 명사 간의 속성과 멘션 간의 동시 지시성이 작용하는가의 여부로 들었다.

이 연구는 반전 표현의 유형을 [材質N1＋の＋モノN2] 형식, [色彩N1＋の＋モノN2] 형식, [サイズ・方法N1＋の＋モノN2] 형식으로 구분하였으며, 최장 NP 범위 규정 및 멘션 간 동일 지시성의 설명에 있어서도 에트리의 지침과 함께 좋은 참조가 되었다.

(2) 명사구(NP)가 가지고 있는 '지표성'과 서술 관계(池田則之, 2011)[5]

이 연구는 멘션 간 상호참조의 판단 및 방향성(양방향성, 단방향성)의 근거를 확립하는 데 적지 않은 참고가 되었다. 이 논문에서는 NP가 가지고 있는 지시 지표(referential index)와 동일한 지표를 가지고 있는 '순서'에 포함된 개체가 해당 NP에 의해 지시 혹은 상호참조로서 용인되는 개체라고 정의하였다.



(3) '~이다' 구문 내에서의 명사 및 명사구의 지시 체계(東郷雄二, 2004)[6]

이 연구는 언어학에서의 'copula문'에 대한 특수성을 고찰한 것으로 본 과제에 있어서 특히 한국어의 주격보어 구문에 대한 상호참조 처리에 기초적인 참조가 된 것이다. 이 연구에서는 이른바 「A is B」 구조의 'A' 와 'B' 모두 명사구인 경우를 협의의 'copula문'으로, 주격보어인 'B'가 형용사 혹은 그와 같은 속성을 보이는 명사구인 경우를 '준(準)copula문'으로 규정하였다.

① copula문: 「A is B」 → 'A', 'B' 모두 명사구

a. John is the private secretary of the president.

b. Jean est le fiancé de Suzanne. "Jean is Suzanne's fiancé."

c. 山田太郎は娘の家庭教師だ。

5) 池田則之(2011) 同一指示解釈と叙述関係, 九州大学言語学論集(32), 31-52

6) 東郷雄二(2004) 名詞句の指示とユピュラ文の意味機能, 科学研究費補助金 基盤研究(C) 課題番号 14510569 平成14年度～16年度 研究成果報告書, 1-65

② 준copula문: 「A is B」 → ‘A’는 명사구, ‘B’는 형용사 또는 형용사적구

a. A cat is four-legged.

b. Mon père est professeur de latin. "My father is (a) professor of Latin."

c. あなたの考え方は独特だ。

이 연구에서는 copula문의 일반형을 「A is B」로 표기함으로써 단순하게 보면 그 의미는 「A = B」이겠으나 실제로 그렇지만은 않다는 점, copula문은 「A is B」라고 하는 하나의 형식으로 대표형을 삼을 수는 있겠으나 결코 하나의 구문이 아닌 그 의미 기능에 의해 몇 가지 하위분류가 필요하다는 점, 만일 「A is B」가 수학적인 동일성을 나타내는 것이라면 A와 B의 순서를 치환하여 「B is A」라고 해도 좋겠지만, 실제로는 그렇지 않다는 점을 명확하게 근거하였는데 의의가 적지 않다고 할 수 있다.

(4) 상호참조 관계를 기술하기 위한 Annotation 모델 디자인(吉田悦子・谷村緑, 2008)[7)]

이 연구는 기본적으로 언어학적 정보를 활용하여 영어 담화에서 이루어지는 상호참조 관계를 규명하기 위한 것으로 일본어의 언어적 현상에 그대로 적용하기에는 무리가 있다. 다만, 명사구, 확장 명사구, 대명사를 대상으로 Annotation 모델인 MUC-7과 MATE의 태그 체계를 비교하고 상호참조 관계 시 선행사의 개념 보다 지시 대상의 개념을 중시하는 방법이 일치율면에서 유용한 결과가 제시되었음을 입증하였다.

또한, 이 연구는 상호참조 관계를 충실하게 반영하기 위해서는 선행사를 규정하는 것이 아닌 지시 대상을 컨텍스트 내부에서의 변화에 따라 복원할 수 있는 일관된 기술(記述) 체계를 확립해야 한다는 점을 중요하게 지적하고 있다.

(5) 인식 동사 구문에 쓰인 명사와 명사 간의 지시성과 지시성의 불투명성(今田水穂, 2010)[8)]

본 연구는 일본어에서 흔히 ‘~라고 생각된다(と思う)’ 류와 같은 ‘인식(認識) 동사 구문’에 사용된 명사 간의 지시성에 대한 일치성 및 불일치성에 대한 판단 근거를 고찰한 것이다. 과제 초기 지침 수립 시 많은 도움이 된 연구로서 멘션으로 잡은 명사(구) 간 상호참조 판단이 작업자마다 차이가 날 수 있다는 점과 이는 주절과 종속절 간의 현실 영

7) 吉田悦子・谷村緑(2008) 同一指示係を記述するためのアノテションモデルの討-MUC-7とMATEを比較して-, 言語處理學會第14回年次大會發表論文, 440-443

8) 今田水穂(2010) 認識動詞構文と間スペース同定, 言語学論叢 オンライン版第3号(通巻29号), 33-44

역과 속성 영역 판단에 기인하고 있다는 것에 대한 하나의 근거가 되었다.

즉, “타로는 [[UFO:a]가 아닌 것]을 [UFO:a′]라고 생각했다.”라는 인식 동사 구문은 자연스럽지만, “타로는 [[UFO:a]가 아닌 것]의 [UFO:a′]라고 생각했다.”는 부자연스러운데, 이 차이는 전자의 “[[UFO:a]가 아닌 것]을”이 주절, 후자의 “[[UFO:a]가 아닌 것]의” 종속절의 요소인 것에 기인하며 이 경우 a와 a′ 사이에 ‘현실 공간의 요소’와 ‘신념 공간의 대응물’이라는 속성의 판단을 고민하게 됨을 상기시키고 있다.

○ 서구

서구권에서는 ‘상호참조 해결(coreference resolution)’에 대하여 대용어 해결 혹은 조응 표현 해결(anaphora resolution)의 하나로서 실제 세계에 존재하는 동일 지시 대상(the same referent in the real world)을 가지는 대용어(anaphor 또는 referring expression)와 선행자(antecedent)를 찾아 링크해 주는 것을 말한다는 Mitkov의 정의에 따르고 있다.

대표적으로는 CSTNews corpus annotation, OntoNote coreference, 위키피디아 상호참조 주석 등을 들 수 있는 가운데 본 과제의 수행을 위해 참조한 서구권의 선행 연구를 정리하면 다음과 같다.

(1) MUC-7(Message Understanding Conference)

1997년도에 개최된 MUC-7(Message Understanding Conference)은 정보추출을 위한 North American News Text Corpora를 대상으로 상호참조해결 태스크를 진행하였다. 정보 추출을 목적으로 한 MUC-7은 개체명만을 대상으로 상호참조 태깅을 하였으며 동사와 관련된 관계는 무시된다.

① 표기법

MUC-7의 상호참조해결 주석 표기는 SGML태그이며 참조 표현 및 선행어는 다음과 같이 태그가 지정된다.

<COREF ID="100">Lawson Mardon Group Ltd.</COREF> said <COREF ID="101" TYPE="IDENT" REF="100">it</COREF> ...

위 주석에서 대명사 it은 Lawson Mardon Group Ltd.와 상호참조 태깅이된다. TYPE 속성은 대명사와 선행어와의 관계를 나타내며 “IDENT”와 같은 하나의 관계에만 주석을 한다. ID와 REF속성은 상호참조 링킹을 표현한다. ID는 마크업된 문자열에 고유한 값으로 할당되며 REF는 해당 ID를 사용하여 상호참조 링크를 나타낸다.

② 주석 대상

MUC-7의 상호참조 해결 주석의 대상은 명사, 명사구, 대명사이며 이를 마커블(Markables)이라고 한다. 날짜 및 숫자표현은 명사구로 간주된다. 이름과 개체명 또한 모두 마커블로 간주된다.

(2) ACE(Doddington et al., 2004)

상호참조 해결 시스템 공개 성능 평가에 사용된 1백만 어절 규모의 대규모 영어 말뭉치이다. LDC(Linguistic Data Consortium)에서 구축한 Event에 대한 상호참조 말뭉치로 위키피디아처럼 협업적 공개 구축방식을 활용했다. 이로 인해 엔터티의 부분집합이 매우 제약적이고 일관성이 부족하다는 평가가 있다.

(3) OntoNote coreference(Pradhan et al., 2012)

이 연구는 앞서 소개한 MUC-7나 ACE 등의 제한된 범위에 국한한 연구들이 있었기에 가능한 것으로 2005년부터 BBN Technologies, 콜로라도 대학, 펜실베니아 대학 등이 협력하여 진행되었다. 상호참조 외에 구문 및 의미적 구조를 포함한 여러 층위의 주석을 통합하고 주석의 주 대상은 명사구이나 MUC나 ACE와는 달리 일부 유형의 개체에 국한하지 않고 범용적인 상호참조 주석을 실시(개체 및 특정 사건)하고 있으며 copula verb로 연결되는 주어와 보어는 주석의 대상으로 삼지 않거나(보어는 속성(attributives)으로 보아 다른 층위의 주석에 의해 커버된다고 판단) 동격(appositives) 표현은 특수한 유형의 상호참조 대상으로 보아 별도로 주석하고 있다.

자연언어 처리용 기계 학습을 위한 말뭉치 제작을 목적으로 진행된 가운데 뉴스 기사, 전화 대화, 웹 블로그, 방송 등의 다양한 공개 말뭉치를 기반으로 이루어져 있으며, 영어, 중국어, 아랍어 등도 포함된다. OntoNotes의 결과물은 상호참조 뿐만 아니라, 개체명, 참조해소, WSD(Word Sence Disambiguation),구문 분석 등에 대한 정보도 담고 있고 있으며 1.3백만 어절 규모(2012년 결과물)이다. 현재까지는 MOU시리즈와 함께 가장 널리 사용되고 있는 상호참조 해결 말뭉치이다.

① 표기법

CoNLL에 따른 표기[9]와 MOU에서 사용한 SGML을 함께 지원하며, JSON 형식을 지원하는 공개 자료도 있다.

9) https://github.com/ontonotes/conll-formatted-ontonotes-5.0

<표 2> Ontonotes 5.0 : 통합 분석 결과에서의 상호참조 정보

```
Leaves:
-------
0   Nicaraguan
       coref: IDENT        000-69 0-3    Nicaraguan President Daniel
                                         Ortega
       name:  NORP              0-0      Nicaraguan

8   the
       coref: IDENT        000-75 8-9     the weekend
       name:  DATE               8-9     the weekend

11  his
       coref: IDENT        000-69 11-11   his

26  the
       coref: IDENT        000-71 26-28   the Contra rebels
27  Contra
       coref: IDENT        000-70 27-27   Contra
       name:  ORG               27-27     Contra
```

<표 3> Ontonotes 5.0 : 자료 유형별 결과에서의 상호참조 정보

```
      Chain 000-71 (IDENT)
                  0.26-28    the Contra rebels
                  1.13-14    the Contras
                  2.19-25    the rebels seeking *PRO* to topple him
                  2.29-30    the Contras
                  2.37-37    they
                  2.44-44    their
                  7.29-30    the Contras
                  8.16-17    the rebels
                  11.12-13   the Contras
                  19.19-20   the Contras
                  19.34-35   the Contras
                  20.18-19   the Contras
                  20.25-25   themselves
                  26.11-12   the Contras
                  27.6-6     they
```

<표 4> Ontonotes 5.0 : 전체 자료 보기를 위한 상호참조 데이터

```
<DOC DOCNO="bc/cnn/00/cnn_0003@0003@cnn@bc@en@on">
<TEXT PARTNO="000">
..
..
<COREF ID="m_5" TYPE="IDENT" SPEAKER="Linda_Hamilton">I</COREF> mean
<COREF ID="m_5" TYPE="IDENT" SPEAKER="Linda_Hamilton">I</COREF> have *-
1 to tell you that when <COREF ID="m_5" TYPE="IDENT"
SPEAKER="Linda_Hamilton">I</COREF> married <COREF ID="122"
TYPE="IDENT"><COREF ID="123" TYPE="APPOS" SUBTYPE="ATTRIB"><COREF
ID="m_5" TYPE="IDENT" SPEAKER="Linda_Hamilton">my</COREF> first
husband</COREF> <COREF ID="123" TYPE="APPOS"
SUBTYPE="HEAD">Bruce</COREF></COREF> *T*-2 <COREF ID="m_5" TYPE="IDENT"
SPEAKER="Linda_Hamilton">I</COREF> <COREF ID="m_5" TYPE="IDENT"
SPEAKER="Linda_Hamilton">I</COREF> <COREF ID="138"
TYPE="IDENT">went</COREF> into hiding for the first year /.
<COREF ID="m_5" TYPE="IDENT" SPEAKER="Linda_Hamilton">I</COREF> just
started *-1 reading books /.
```

② 주석 체계

Ontonotes는 Noun phrase, Possessives, Premodifiers, Verbs를 멘션의 대상으로 하며, 크게 Identical과 Appositives 2개의 상호참조 연결 유형(Co-reference link types)을 구하고 각각에 대한 세부 유형의 이슈와 9개의 개별 이슈에 대한 사례를 제시하고 있다.

특이한 점은 IDENT와 APPOS 2개의 상호참조 연결 유형이 주석 단계에 함께 포함된다는 점이다. 아래와 같이 'the Contra rebels'와 'the Contras'가 INDENT 관계임이 함께 주석되어 있다. 주석 체계에 나타난 보다 하위의 유형들에 대해서까지는 공개되어 있지 않다. 이러한 주석은 상호참조 관계에 대한 검증 뿐만 아니라 보다 세밀한 기계학습의 가능성을 보여준다.

<표 5> Ontonotes 영어 상호참조 가이드라인 7.0의 체계

Link Types	Sub Types	
IDENT (Identical)	Pronouns & Demonstrative	
	Generic Mentions	
	Pre-Modifiers	
	Nested Mentions	Head-Sharing NPs
		Proper Names
	Copular Structes	
	Determining which entity to add	
	Small clauses	
	Temportal expressions	
APPOS (appositives)	Marking appositive heads	
	Linking appositive spans	
Special Issues	Organization and Members	
	Gender and Number	
	Idenfinite uses of proper nouns	
	GPEs and governments	
	Quantifying Expressions	Quantifiers
		Partitives
		Linking quantifying exp
	Possessive extents	
	Formulaic mentions	
	Sentence fragments	
	Metonyms	

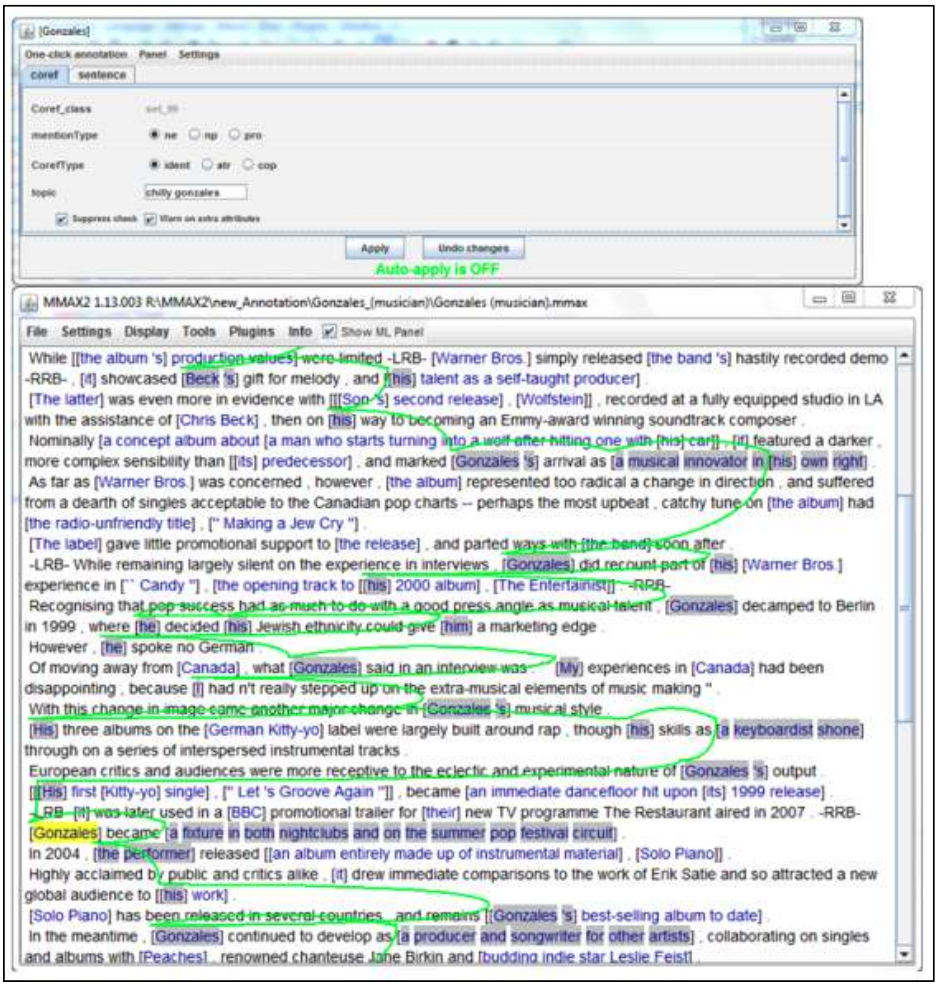
(3) 위키피디아 상호참조 주석 (WikiCoref: An English Coreference-annotated Corpus of Wikipedia Articles 2016)

LDC에서 위키피디아 문서만을 대상으로 진행한 상호참조 말뭉치이다. 상호참조 말뭉치가 규모에 비해 텍스트의 길이와 주제가 다양한 문제를 해소하기 위해 진행되었으며, 위키피디아 문서의 주제와 텍스트 길이를 감안하여 작성되었다. 자동화된 전처리와 상호참조 해결 프로그램 및 수작업 편집도구를 활용하였다.

특징으로는 위키피디아 영문판 기사 일부를 샘플링 후 상호참조(30개 기사 약 6만 단어), 모호성 제거와 작업자간 일치도 개선을 위해 OntoNotes 지침을 일부 수정 후 사용, 최장 멘션의 선정에서 전체 멘션을 하나의 단위로 취급(OntoNotes와 달리 내포 멘션을 표시하지 않는다. [Zsa Zsa, who slap a security guard] vs. [[Zsa Zsa], who slap a security guard]]), 동사(구) 제외 및 명사(구)만을 대상으로 함, 주석된 멘션(mention)에는 3개의 속성을 부여: 멘션 유형, 상호참조 유형, 해당 Freebase 토픽(Freebase DB 내의 엔티티 정보 부착), Identical 유형과 Attributive 유형은 서로 다른 기능을 하므로 상호참조 주석은 Identical 유형과 Attributive 유형(예: 동격과 주격보어)으로 나누어 함, 2명의 주석자의 주석 결과를 일부 샘플링하여 대조 평가하여 멘션 추출

에서는 0.78 (Kappa)의 일치도를, 상호참조에서는 83.3%의 성공률을 보였다.[10)]

<그림 3> WifiCoref 편집-검증 도구 : MMAX2



(4) PreCo(Hong Chen et al., 2018)

PreCo는 Yitu Tech라는 중국계 인공지능 기업에서 구축된 결과물이며, 1,250만 어절의 대규모 자료 구축을 위한 방법론을 제안했다. 웹 상에서 3만8천 여개의 유치원생 영어평가 자료를 수집, 정제하여 구축하였다.

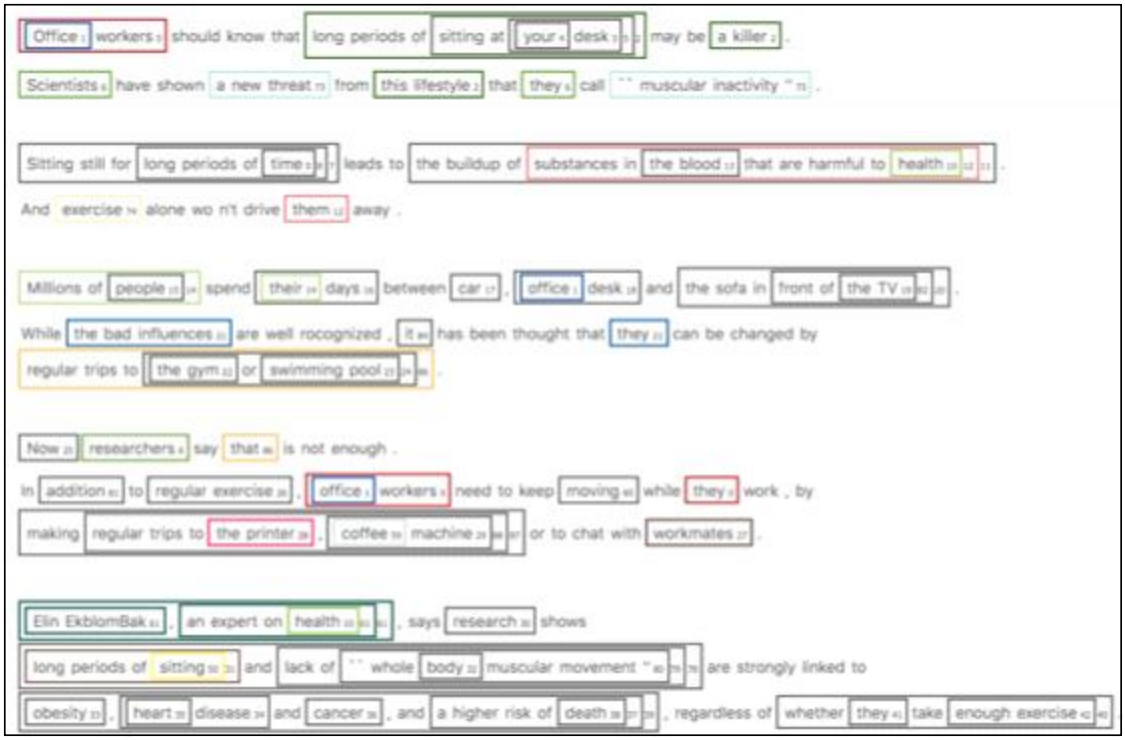
MOU, ACE, Ontonotes가 매우 정밀한 반면에, 데이터 규모가 작아 실용적인 기술 개발에 있어서 OOV(Out-Of-Vocabulary)가 발생하는 문제를 해결하고, 도메인 제약을 극복하는 데 중점을 두었다.

구축과정에서 3인씩 2단계의 교차검증을 적용하여 100%의 일치율을 제시하고 있으나, 초회 평가값은 78.8정도를 제시하고 있다. 주석체계는 Ontonotes를 참고하여 비교제시

10) 이는 MUC F1 공식에 의한 수치임.

하고 있다.

<그림 4> PreCo : 상호참조 구축결과 시각화 자료



(5) CSTNews corpus annotation (The Coreference Annotationof the CSTNews Corpus 2017)

가장 최근에 보고된 상호참조 해결 말뭉치로 50개의 주제 그룹으로 나뉘는 총 140개의 포르투갈어 뉴스 기사를 대상으로 상호참조 주석을 진행하였다. 이 텍스트들은 전 단계에서 여러 층위의 언어적 주석을 부착하여 공동 작업(collective task)을 전제로 하며 3~4명으로 이루어진 개별 팀을 5개 구성하여 각 팀별로 동일한 텍스트에 대한 주석을 진행하였다. 각 주석자는 약 한 달 반 동안 11개 내지 13개의 기사를 처리하는 작업량((1일 1개 정도)이었으며, 각 팀별 주석 일치도는 전체 평균 일치도 kappa = 0.54로 나타났다. 또한, 정확한 상호참조 집합을 구성하기 위해 배경 지식과 도메인 지식이 필요한 경우에는 인터넷 등의 참고 자료 이용할 수 있으며 상호참조 유형을 나누어 우선 명사구끼리 주석한 후 다음 단계로 명사구와 대명사간 주석을 하고 있다.

2.2.2. 구축 지침 내용

개선한 구축 지침은 기존의 '어절' 단위 태깅에서 '형태' 단위 태깅으로 태깅의 단위가

바뀌었으며 기존 지침에 없던 지침이 추가되었다. 또한 한국어에 없는 문법적 용어 등을 한국어에 맞는 용어로 변경하였다. 변경된 지침의 상세 내용은 다음과 같다.

<표 6> 지침 변경 내용

변경 내용	변경 상세
태깅 단위	태깅 단위 '어절'에서 '형태'단위로 변경
용어 변경	'한정사구'를 '관형사구'로 변경
	'대등 접속사'를 '대등 접속'으로 변경
	'소유격'을 '관형격'으로 변경
지침 추가	2.2.12 유의어 및 일반화/구체적 표현 지침 추가
	2.2.13 일반대상을 지칭하는지 구체적인 특정 대상을 지칭하는지 판단이 어려울 경우 지침 추가

(1) 태깅 단위 변경

개선한 구축 지침에서는 상호 참조 해결 태깅을 위한 멘션 탐지 규칙에서 멘션의 태깅을 '형태' 단위로 정의하였다. 멘션이 명사(구)에 기반한다는 대전제 하에 멘션을 추출하는 데 작업자의 직관과 일치하며 불필요한 조사가 삽입되지 않아 데이터의 무결성을 보장한다.

(2) 용어 변경

기존 지침에서는 영어에서 쓰이는 문법적 용어를 그대로 사용하여 한국어와는 맞지 않는 문제가 있었다. 한국어 특성에 맞는 언어자원을 구축하는 지침으로서 한국어에 해당하는 문법적 용어와 설명으로 대체하였다. 영어에서 쓰이는 '한정사구(Determiner Phrase)'에 대응하는 용어인 '관형사구'로 변경하였다. 또한 한국어의 접속사에는 없는 '대등 접속사'라는 표현을 지우고 '대등 접속'이라 변경하였다. 여기서 '대등 접속'이란 대등적 연결어미 '~고', 접속조사 '와/과', '나' 등을 포함하는 명사구를 가리킨다. 마지막으로 '소유격'이라는 표현 또한 한국어에는 없는 용어로 다음은 기존 지침에서 용어를 변경한 지침이다.

- 2.1.11 지시관형사를 포함한 관형사구와 대명사는 수식어를 제외한다.
- 2.2.11 명사(구)와 관형사구

(3) 지침 추가

지침 수립 과정 중 구축 결과물 분석을 통한 지속적인 지침 개선을 통해 새로운 지침의 필요성을 도출하여 기존 지침에 없던 내용을 추가하였다. 구축 사례를 통해 세부적인 지침이 필요한 경우와 태깅 대상에 관한 세부적인 판단을 요하는 경우에 대한 지침을 추가하였다. 특히 구축 과정 중 가장 이슈가 되었던 태깅의 대상과 범위에 관련한 지침을 추가하여 보다 더 명확한 태깅 대상 정의가 가능한 지침이 되도록 하였다. 태깅의 대상이 혼동 되는 경우는 주로 유의어 및 일반적인 표현과 문서 내에서 일반대상을 지칭하는지 구체적인 특정 대상을 지칭하는지 판단이 어려운 경우이다. 따라서 위 두가지 사항을 신규 지침으로 정의하였다. 다음은 추가된 지침 항목들이다.

<표 7> 지침 추가 사항

2.1.16 숫자, 날짜 및 수량 표현(지침 추가 사항)

▶ 날짜, 금액, 수치 등 숫자 표현들을 멘션 추출 대상에 포함한다.

예1) 지시하는 날짜, 금액, 수치가 같은 수량 표현의 경우
<예문> 2006년 보다 200만원 하락, …2006년보다 200만원 정도 싼 1500만원 후반대로 결정될 것으로 보인다.
<보기> [2006년](O), [하락](O), [200만원](O), [2006년](O), [200만원](O), [200만원 정도](0),[1500만원](O), [1500만원 후반대](0)
<오류>
<상호참조 태깅>{[2006년], [2006년]}(O), {[200만원], [200만원]}(O)
<예외> 문서 내에 구체적 정보가 없이 대상 또는 시간 명사만 있을 경우 상호참조 태깅을 하지 않는다.
<예문> 네 오늘은 또 어떤 모범 밥상을…오늘은 또 어떤 요리가 저희를 기다리고…
<보기> [오늘](O), [모범 밥상](O), [오늘](O), [요리](O), [저희](O),
<오류>
<상호참조 태깅> {∅}(예문 내 멘션들 중 상호참조관계 없음)

예2) 고정된 숫자값의 경우
<예문> 판교 중대형 분양가 3.3㎡당 1500만원 대, 대우건설 등은 3.3㎡당 분양가를…신청했지만…(후략)표본명: NWRW1800000021-0001(판교 아파트)
<보기> [판교](O), [중대형 분양가](O), [3.3㎡당](O), [1500만원 대](O), [대우건설](0), [대우건설 등](O), [3.3㎡당](O),[3.3㎡당 분양가](O)
<오류>
<상호참조 태깅> {[3.3㎡당], [3.3㎡당]}(O)

예3) 숫자 표현과 문자로 된표현이 같은 것을 지시하는 경우
<예문> 하지만 당초 2008년이던 이전 시기가 차일피일 미뤄지면서 빈집은 절반을 웃돌고 있다…2004년 7월 기지 이전을 2008년 말까지 끝내기로…(후략)
표본명: NWRW1800000021-0004(동두천)
<보기> [당초2008년이던 이전 시기](0), [당초 2008년](0), [빈집](0), [절반](0),[2004년 7월](0), [기지](0), [기지 이전](0), [2008년 말](0)
<오류>

<상호참조 태깅> {[당초 2008년], [당초 2008년이던 이전 시기]}(0)

2.1.17 지리 정보(지침 추가 사항)

▶ 지리 정보 전체를 하나의 멘션으로 잡고 그 내부의 개별 지리명을 각각멘션으로 처리한다.

예1)

<예문> 중국 저장성 동부 타이저우시 해안에 있는 다천다오에서

<보기> [중국 저장성 동부 타이저우시 해안에 있는 다천다오](O), [중국 저장성 동부 타이저우시해안](0),[중국 저장성 동부 타이저우시](O), [중국 저장성 동부](O), [중국 저장성](O), [중국](O)

<오류>

<상호참조 태깅> {∅}(예문 내 멘션들 중 상호참조 관계 없음)

예2)

<예문> 서울 종로구 신문로 1가 197금호아시아나 신사옥

<보기> [서울 종로구 신문로 1가 197금호아시아나 신사옥](O), [서울 종로구 신문로 1가 197 금호아시아나](O), [서울 종로구 신문로 1가 197](O), [서울 종로구 신문로 1가](O), [서울 종로구 신문로](O),[서울 종로구](O), [서울](O)

<오류>

1.1.1.동일 어구 반복 상호참조해결(지침 추가 사항)

▶ 동일한 어구로 동일한 대상을 지칭하는 경우 가장 명시적인 상호참조 현상이다.

예1)

<예문> 폴크스바겐은 2018년까지 미국 시장에서의 판매량을 3배로 늘리기로 하고20년 만에 처음으로 미국 공장에 10억 달러를 투자하는 한편 미국 시장을 겨냥한 새로운 모델들을 개발하고 있다…세계 최대 자동차 시장인 미국 시장은 2000년대 초반만 해도 매년 1700만 대가 팔렸으나 지난해에는 1300만 대로 시장이 크게 축소됐다.'

<보기> [폴크스바겐](O), [2018년](O), [미국](O), [미국 시장](O), [미국시장에서의 판매량](O),[3배](O),[20년](0), [20년 만](O), [미국](O), [미국 공장](O), [10억 달러](O), [미국](O), [미국 시장](O), [미국 시장을 겨냥한 새로운 모델들](O), [세계](O), [최대](O), [세계 최대 자동차](O), [세계 최대 자동차 시장](O), [세계 최대 자동차 시장인 미국 시장](O), [미국](O), [2000년대](O),[2000년대 초반](O), [매년](O), [1700만 대](O), [지난해](O), [1300만 대](O), [시장](O)

<오류>

<상호참조 태깅>{[미국], [미국], [미국], [미국]}(O), {[미국시장], [미국 시장], [세계 최대 자동차 시장], [세계 최대 자동차 시장인 미국 시장], [시장]}(O)

1.1.2.줄임말 및 약어(지침 추가 사항)

▶ 동일한 대상을 지칭하는 명사구와 줄임말 혹은 약어 사이에도 상호참조 현상이 나타난다.

예1)

<예문> 도널드 트럼프 미국 대통령은…(후략)…트럼프 대통령은…(후략)(자체 생성 예시문)

<보기> [도널드 트럼프](O), [미국](O), [도널드 트럼프 미국 대통령](O), [트럼프](O), [트럼프 대통령](O)

<오류> [미국 대통령은](X)à2.1.4. 중심어중복 제거 규칙 적용
<상호참조 태깅> {[도널드 트럼프], [도널드 트럼프 미국 대통령], [트럼프], [트럼프 대통령]}(O)

1.1.3.생략된 명사구(지침 추가 사항)

▶ 상호참조 관계의 멘션 중 하나가 생략된 명사구로 나타날 경우에도 상호참조해결처리한다.

예1)
<예문> '노근리 59주년'인권평화캠프 열린다…6.25전쟁 초기에 발생한 '노근리 사건'의 현장인…노근리 사건 발생 59주기를 맞아…(후략)
<보기> [노근리](O), [노근리 59주년](O), ['노근리 59주년'인권평화캠프](O), [6.25전쟁 초기에 발생한 '노근리 사건'의 현장] [6.25 전쟁](O),[6.25 전쟁 초기](O), [노근리 사건 발생 59주기](O), [노근리 사건 발생](0), [노근리 사건](0)
<오류>
<상호참조 태깅> {[노근리], [노근리 사건],[노근리 사건]}(O)

1.1.4.기관/단체와 소속자/소속물의 관계(지침 추가 사항)

▶ 특정기관/단체와 소속자 혹은 소속물 사이는 상호참조 대상이 아니다. 단, 대명사 표현으로 인해 동일성이 존재할 경우 상호참조해결한다.

예1)
<예문> 삼성전자는…삼성직원들은…삼성관계자는 "저희는…했습니다"라고 대답했다.
(기존 지침에서 발췌 후변형)
<보기> [삼성전자](O), [삼성직원들](O), [삼성관계자](O), [저희](O)
<오류> {[삼성전자], [삼성직원들]}(X)
<상호참조 태깅> {[삼성전자], [저희]}(O)

1.1.5.명사(구)와 관형사구(지침 추가 사항)

▶ 명사(구)와 '지시관형사(이/그/저)+ 명사'또는 '이런/저런+명사'패턴 사이에 상호참조가 일어나기도 한다.

예1) 이/저/그+명사의 예
<예문> 제10호 태풍 크로사가…, 이 태풍은…, 크로사는…(기존 지침에서 발췌 후변형)
<보기> [제10호 태풍 크로사](O),[이 태풍](O), [크로사](O)
<상호참조 태깅> {[제10호 태풍 크로사],[이 태풍], [크로사]}(O)

예2) 이런/저런+명사의 예
<예문> 00일 새벽 0시 규모 9의 지진이…, 이런 규모의 지진은…
(기존 지침에서 발췌 후 변형)
<보기> [00일 새벽 0시 규모 9의 지진](0), [00일 새벽 0시](O), [규모 9](O), [이런 규모의 지진](0), [이런 규모](O)
<상호참조 태깅> {[규모 9], [이런 규모]}(0),{[00일 새벽 0시 규모 9의 지진], [이런 규모의 지진]}(O)

1.1.6.유의어 및 일반화/구체적 표현

▶ 다음과 같은 유형들은 상호참조 대상으로 본다.

예1) 유의어관계
<예문> 칠레에서 시위가 이어지는 가운데… 칠레의 소요 사태는…(후략)
<보기> [칠레](0), [시위](0),[칠레](0), [칠레의 소요 사태](0)

<상호참조 태깅> {[칠레], [칠레]}, {[시위], [칠레의 소요 사태]}

<예문> 지방자치단체장이 독단적으로 지방의료원을 폐업하지 못하도록 하는 이른바 '진주의료원법'…(증략)…'지방의료원의 설립 및 운영에 관한 법'개정안을 6일 법안심사 소위원회로 넘겼으나…(증략)…일부 새누리당 의원들은 '지방자치권 침해'를 이유로 개정안 처리에 반대하고 있는 것으로 전해졌다.
<보기> [지방자체단체장](0), [지방의료원](0), [진주의료원법](0), [지방의료원의 설립 및 운영에 관한 법](0), [지방의료원](0), [지방의료원의 설립 및 운영](0), ['지방의료원의 설립 및 운영에 관한 법'개정안](0),[6일](0), [법안심사 소위원회](0), [새누리당](0), [일부 새누리당 의원들](0), [지방자치권](0), [지방자치권 침해] (0), [개정안] (0), [개정안 처리](0)
<상호참조 태깅> {[진주의료원법], ['지방의료원의 설립 및 운영에 관한 법'개정안], [개정안]}(0), {[지방의료원], [지방의료원]}(0)

예2) 일반화및 구체화
<예문> 교통 사고 희생자 수는…(중략)… 2016년부터 지난 해 동안 교통 사고로 사망한 사람들의 수는…(중략)…금년들어서도 이미 154명이 교통 사고로 사망하면서 가장 희생자가 많았던 지난 해 같은 기간의 146명을 넘어섰다.
<보기> [교통 사고](0),[교통 사고 희생자](0), [교통 사고 희생자 수](0),[2016년](0), [지난 해](0), [지난 해 동안](0), [교통 사고](0), [교통 사고로 사망한 사람들](0), [교통 사고로 사망한 사람들의 수](0), [금년](0), [154 명](0),
[교통 사고](0), [희생자](0),[가장 희생자가 많았던 지난 해](0), [가장 희생자가 많았던 지난 해 같은 기간](0), [146 명](0)
<상호참조 태깅> {[교통 사고 희생자], [교통 사고로 사망한 사람들], [희생자]}(0)

1.1.7.일반대상을 지칭하는지 구체적인 특정 대상을 지칭하는지 판단이 어려운 경우

▶ 일반 대상을 가리키는지 특정한 대상을 가리키는지 판단이 어려운 경우에는다음과 같은 기준들에 따라 처리한다.
- 이어진문장들에서 동일한 어구의 형태로 계속 반복적으로 나타나는가?
- 이어진문장들에서 줄임말, 유의어 등을 사용한 동일 대상 지칭 변용 표현들이 나타나는가?
- 텍스트의내용과 밀접히 관련된 핵심적인 명사구인가?

예1)
<예문> 세법 바뀌었는데 내 연금저축 어떡하나 세금 환급액…세법 개정으로 연금저축 연말정산이…연금저축 가입으로 인해 돌려받는 세금이 절반 이하로 쪼그라들기 때문이다. 반면 …연금저축 가입에 따른 세금 환급액이 더 늘어난다.
<보기> [세법], [내 연금저축], [세금 환급액], [세금], [세법], [세법 개정], [연금저축], [연금저축 연말정산], [연금저축], [연금저축 가입],[연금저축으로 인해 돌려받는 세금], [절반 이하], [연금저축], [연금저축 가입], [연금저축 가입에 따른 세금 환급액]
<상호참조 태깅> {[세법], [세법]}(0),{[연금저축], [연금저축], [연금저축]}(0), {[세금 환급액], [연금저축 가입에 따른 세금 환급액], [연금저축으로 인해 돌려받는 세금]}(O)

또한 구어의 경우 텍스트 특성상 기존 문어를 대상으로 한 지침으로는 구축이 어려운 경우가 있기 때문에 구어에 관한 지침을 추가 하였다. 구어에 대한 상호 참조 주석의 지침은 문어와 동일하게 적용하되 동일한 지침이라도 구어의 표현, 문맥적 특성상 그 양상이 문어와 현저히 다른 경우에는 해당 지침에 사례를 추가하고 사례번호에 '구어자료의 경우'로 명시하였다. 다음은 구어 지침의 표기 형식이다.

<표 8> 구어 지침 사례 표기 형식

2.1.2. 명사구에서 가장 의미를 갖는 명사는 중심어(head)이며, 멘션은 중심어를 기반으로 해당 명사구에 존재하는 수식어까지 포함해야 한다. 예1) ….. 예4) 담화 표지서 지시사의 수식어 포함(구어의 경우) <예문> 이 쪼끄만 내장 있는 데 쌉쓰름한 맛이 있어요 이릏게. <보기> [쪼끄만 내장](X), [이 쪼그만 내장](O),

(4) 기타

(ㄱ) 기존 지침의 예외 사항 적용

기존 지침 '2.1.4 중복되는 위치의 중심어가 두 개 이상의 멘션을 추출할 경우, 바운더리가 더 큰 멘션만 선택한다.(중심어 중복 제거 규칙)'은 '사람 이름과 직함의 예'에서 직함이 중심어가 될 경우 [이름+직함]으로 멘션을 추출하며 내포 된 '이름'또한 멘션으로 별도 추출하는 예를 보여준다.

<표 9> 지침 2.1.4

2.1.4. 중복되는 위치의 중심어가 두 개 이상의 멘션을 추출할 경우, 바운더리가 더 큰 멘션만 선택한다.(중심어 중복 제거 규칙) 예1) ….. 예2) 사람 이름과 직함의 예: 직함이 중심어가 될 경우 [이름+직함]으로 멘션을 추출한다. <예문> 아베 신조 일본 총리(구축지침 대조정리 문서에서 발췌) <보기> [아베 신조 일본 총리](O), [아베 신조](O), [일본](O), <오류> [일본 총리](X) <상호참조 태깅> {[아베 신조], [아베 신조 일본 총리]}(O) <예외> <예문> 아베 신조는 10년 째 일본 총리이다. <상호참조 태깅> {[아베 신조], [10년 째 일본 총리]}(O)

위 지침은 ‘예시2)’에서 수식어구가 결합한 ‘이름+직함’의 분석 예시는 나와 있지 않다. 다음은 수식어구가 결합한 ‘이름+직함’문장에서는 내포 멘션인 이름을 추출할 때 수식어구를 포함하지 않는 예를 보여준다. 이 예시는 최장NP 추출 규칙인 지침 ‘2.1.2 명사구에서 가장 의미를 갖는 명사는 중심어(head)이며, 멘션은 중심어를 기반으로 해당 명사구에 존재하는 수식어까지 포함해야 한다.’에 예외인 결과를 보여준다.

<예문> 지난해 FA컵에서 첫 우승컵을 들어올렸던 황선홍 감독은 "이제야 선수들과 진짜로 하나가 되는 법을 알게 됐다"고 했다.
<상호참조 태깅> {[지난해 FA컵에서 첫 우승컵을 들어올렸던 황선홍 감독], [황선홍]}

위 예문에서 ‘황선홍’은 [이름+직함]의 경우 ‘이름’을 별도로 추출한 결과이며 앞의 수식어구 ‘지난해 FA컵에서 첫 우승컵을 들어올렸던’을 함께 추출하지 않는다.

(ㄴ) 기존 지침의 세부 사항 적용

또한 ‘사람 이름과 직함의 예’지침에서 이름대신 ‘성씨’가 등장한 경우 기존 지침과 동일하게 ‘성씨’또한 별도의 멘션으로 추출하였다. 다음은 ‘성씨+직함’의 경우 상호참조 태깅 예시이며 성씨인 ‘박’만 추출한 결과를 보여준다.

<예문> 새누리당 박상은 의원에게 정치자금법 위반으로 구속영장을 청구하는 방언을 검토중이다. 검찰은 박 의원의 혐의를 해운업체를 압수수색하는 과정에서 포착했다.
<상호참조 태깅> {[새누리당 박상은 의원], [박상은], [박], [박 의원]} (O)

3. 데이터 구축 수행 도구 활용

3.1. System 설치 및 구성

Web System 으로 AWS(Amazon Web Service) 클라우드 서비스에 설치하여 안정적인 Cloud 환경의 시스템을 운영하였으며, 데이터 보관의 시간적, 공간적 제약 없는 작업 환경을 구성하였다.

3.2. 자료 보안 및 외부 인력 접근 제어

회원 가입 형태가 아닌 컨소시엄 내 수행 인력에 한해서만 접근 사용자 권한이 발급되므로, 미승인(외부 인력) 사용자의 접근은 원천적으로 불가하다. 외부 Hacking 등에 따른 원시/가공 데이터 접근에 대한 Cloud Service(AWS) 차원의 침입 방지 보안 서비스가 기본적으로 제공되어 자료 보안에 도움이 된다.

작업 권한 부여 시 승인된 도구 사용 권한은 작업된 데이터에 대한 외부 접근을 차단하므로, 납품 시 최고 관리자(사업책임자)외에는 데이터 Export가 불가능하다.

<그림 5> 작업 권한자 접근 화면/기능 - KRONOTH Annotation System

3.3. 구축 도구 활용

3.3.1. 원시 데이터 검사

원시 데이터 수령 이후, 구축 도구를 활용하여, File Open Check와 File Size Check를 실시하여, 수령 파일 자체의 무결성을 확인하고, 단위 파일 당 시스템 허용 용량을 검사하였으며, 주석 부착 여부 Check로 원시 데이터의 메타 정보 등 작업을 위한 최소한의 데이터 부착 상태를 점검하였다.

또한 Tag 무결성 Check를 통하여, 부착된 Tag가 정상적으로 부착된 상태인지 주석 형식을 검사하고, Tag 기준 Check로, 부착된 Tag가 사업 검증, 납품 기준에 맞게 부착되었는지 여부를 알아보는 주석 Policy 검사를 실시하였다.

예외 문자, 문자 코드 Check를 하여, 파일 내 포함된 Text가 정상적인지를 확인하는 내용 형식 검사를 실시하였으며, ETL 실행을 통하여, 원시 데이터 file Check를 진행하였다. 이 검사로 File이 정상적인 File일 경우 File 내용의 Extract 및 System Transfer, Load 작업을 수행하였다.

3.3.2. 승인된 사용자 시스템 사용 등록

시스템 사용시 보안을 위하여, 채용(소속)된 구축 작업자 비밀유지서약서, 보안각서 등의 인적 보안서를 수령하였으며, 채용(소속)된 구축 작업자 ID/Password를 수령 및 등록하여, 등록된 System 사용자에게만 작업 권한을 부여하였다.

또한 등록된 작업자에게만 System 접속 URL을 제공하며, 등록 승인된 작업자만 진입이 허용된다. System 최초 로그인 시 작업자는 개인별 Password를 필수적으로 변경해야만 진입이 가능하다. 이때 최고 관리자는 작업자별 작업 권한에 따라 System 접근 제어 실행 권한을 다시 부여하며, 작업자에게는 승인된 작업 화면만 활성화되고, 접근도 가능하다.

3.3.3. 등록된 원시 말뭉치 데이터의 분배

등록된 원시 말뭉치는 최고 관리자 권한으로 System 접근 시에만 작업 배분이 가능하며, 일별 수행 가능한 할당량 및 주간/월간 공정량을 확인하여 최고 책임자가 작업을 할당할 수 있다. 최고 관리자가 할당한 작업은 Tag 부착 작업과 검수 작업으로 구분되며, 관리자는 할당된 작업들의 실시간 상태 확인이 가능하다.

3.3.4. 할당된 작업 수행

작업자가 System에 접근할 때, 첫번째로 보이는 페이지는 공지사항 페이지로, 진입 시 변경된 지침 등의 공지사항을 확인하고 작업을 시작하게 된다.

작업자는 현재 할당된 작업 및 할당된 작업 중 미수행, 미완료(임시저장 상태)된 작업 현황을 확인하여, 선택한 작업의 작업 창으로 이동하고 작업을 수행한다. 작업 중 임시 저장을 할 수 있으며, 작업 완료 시 작업 내용을 저장할 수 있다. 작업을 완료하여 저장한 표본은 다시 작업할 수 없다.

작업 창 진입 시 원문 데이터 및 기본적인 형태소 분석 정보 등을 제공하여 작업의 편의성을 높였으며, 보류 기능과 검토 요청 기능을 보완하여, 모호한 기준의 작업일 경우 지침을 재확인하거나 상위 검수자들에게 검토를 요청하도록 하여 작업의 정확도에 기여하도록 하였다. 작업자가 작업 완료 후 저장한 결과 데이터는 격리되어 보존된다.

3.3.5. 수행 현황 확인, 관리

수행 현황을 확인하고 관리하는 것은 최고 관리자 권한으로 System 접근할 때만 가능하다. 최고 관리자는 현재 수행 완료된 작업의 현황을 확인하고, 공정률 확인할 수 있으며, 현재 수행 중인 작업 및 남은 작업 현황을 확인할 수 있다.

기간별, 수행 작업자별 수행 현황도 확인이 가능하다. 이런 과정을 거쳐 완료된 작업 대상의 데이터를 Export 및 납품(1차, 2차, 최종)하는 것도 최고 관리자이다.

3.3.6. 작업 결과 데이터 관리

작업 결과 데이터는 최고 관리자 권한 외에는 접근(열람, 조회, 추출 등)이 불가능한 보안 권한 체계로 구성되어 있다. 최고 관리자는 사업 책임자 외 1명으로 구성하여 데이터의 유출 등 인적 유출 가능성을 최소화하여 구성(사업 책임자 부재 시 작업 배정 등을 위해 1명 더 배치)되었다.

민간 클라우드(Cloud) 서비스를 통한 시스템 구성으로 외부 침입 불가하며, 최고 관리자 등 사업 참여인력은 보안 서약 등을 제출해야만, 시스템에 접근 가능한 사용자 계정을 개설하여 준다.

데이터 납품을 위한 구축 결과 데이터 Export는 System Log를 통해 이력 관리되며, 이에 대한 모든 접근자는 추적이 가능하다.

4. 말뭉치 구축 및 납품

4.1. 말뭉치 구축

4.1.1. 말뭉치 구축 절차

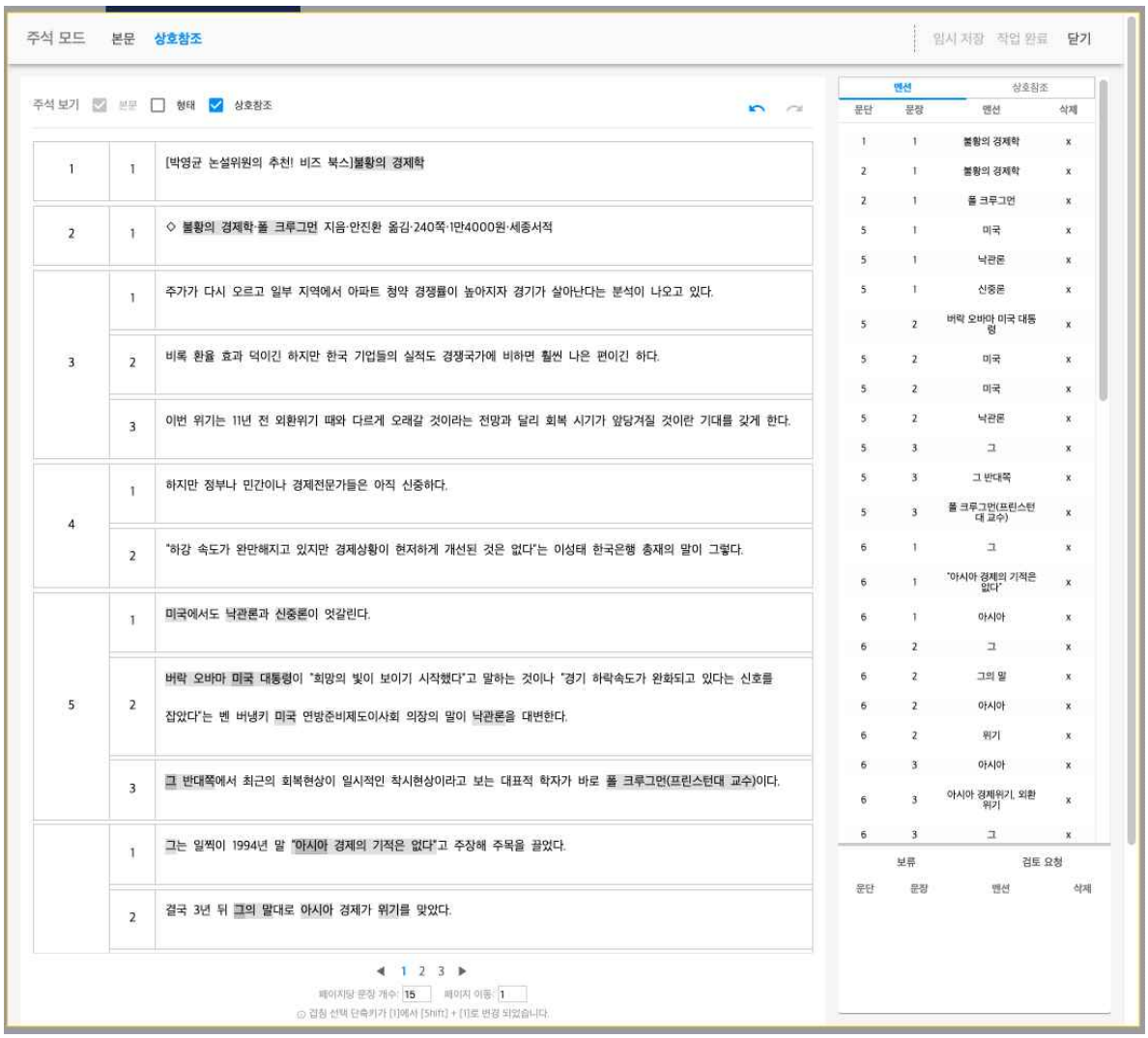
말뭉치의 구축을 위해, 원본 데이터 형태 분석, 개체명 분석 및 문장, 문단 등 본문 주석을 자동으로 처리하고, 상호참조 대상 멘션 후보를 추출한다. 이와 같은 전처리 작업을 통해 보다 정확하고, 효율적인 가공 작업을 지원하고, 작업 중 작업 보류 및 검토 요청을 통해 부정확한 가공을 사전 배제하며, 작업 완료 시 자동 검사를 통해 참조 관계가 없는 멘션(싱글톤)이나 누락된 멘션이 없는지, 작업 내용을 재확인 후 작업을 완료하는 절차로 진행한다.

<그림 6> 말뭉치 가공 절차도

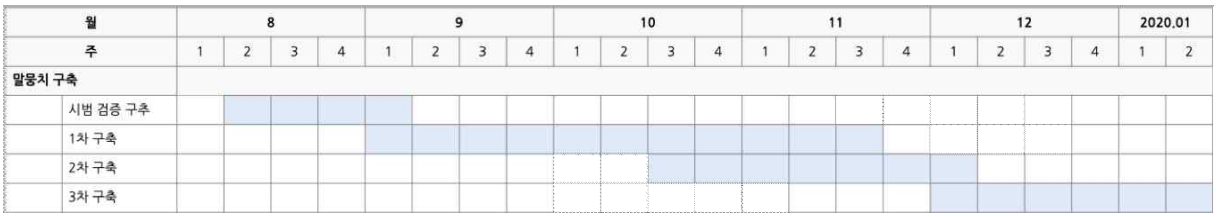
4.1.2. 자동 전처리를 통한 구축 소요 단축

상호 참조 해결 정보 부착 작업은 많은 자원이 소요될 뿐만 아니라 고품질의 작업 결과를 위해서는 고도의 집중을 필요로 한다. 따라서 구축 도구에서 본문 자동 전처리와 형태소 분석 및 개체명 분석을 통해 자동 멘션 추출 후보를 제시하고 작업자가 상호 참조 해결 작업에 집중할 수 있게 하였다.

<그림 7> 구축 도구의 전처리 후 멘션 후보 시각화



4.1.3. 말뭉치 구축 기간



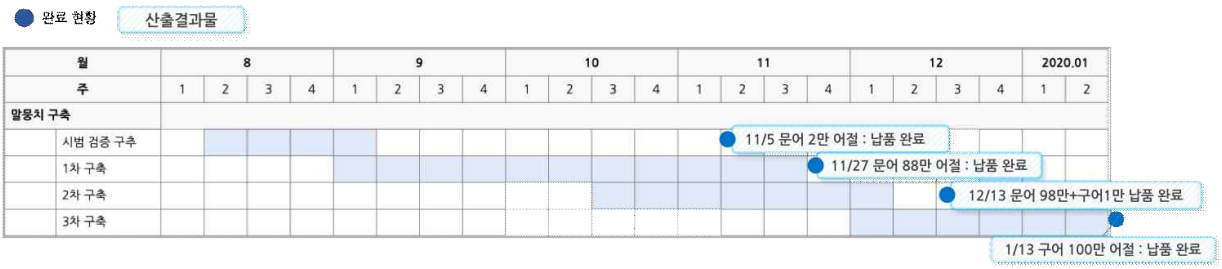
- 시범 검증용 샘플 구축 기간은 2019년 2019년 8월 19일부터 2019년 9월 6일까지 샘플 표본 2만여 어절(표본 수 73개)을 구축하였다.

- 1차 구축 기간은 2019년 9년 9일부터 2019년 11월 22일까지이며, 시범 검증 표본을 포함하여, 문어 1,036,168어절(표본수 3,870개)을 구축하였다.

2차 구축 기간은 2019년 10월14일부터 2019년 12월 6일까지이며, 문어 983,154어절(표본수 3,395개) 구어 10,743어절(표본수 5건)을 구축하였다.

3차 구축 기간은 2019년 11월 8일부터 2020년 1월 3일까지 이며, 구어 1,012,410어절(표본수 423개)을 구축하였다.

4.2. 말뭉치 납품



4.2.1. 납품 어절 및 표본 수

- 결과물로 납품된 총 어절 수는 3,025,769어절(표본수 7,688개)이며, 이 가운데 문어는 2,019,322어절(표본수 7,265개), 구어는 1,006,447(표본수 423개)이다.
- 납품은 총 3차로 이루어졌으며, 1차 납품 2019년 11월 27일, 2차 납품 2019년 12월 13일, 3차 납품 2020년 1월 13일이다.
- 1차 납품 시 문어 시범 검증 표본을 포함한 문어 1,036,168어절(표본 3,870개), 2차 납품 시 어 983,154어절(표본수 3,395개) 구어 10,743어절(표본수 5건), 3차 납품 시 구어 시범 검증 표본을 포함한 전체 1,006,447어절(표본수 423개)를 결과물로 납품하였다.

<표 10> 말뭉치 유형별 납품 현황

자료유형	차수	어절	표본수(개)	구축기간	납품일
문어	1차(시범 포함)	1,036,168	3,870	19.09.09~19.11.22	19.11.27
	2차	983,154	3,395	19.10.14~19.12.06	19.12.13
	문어 총계	2,019,322	7,265		
구어	3차	1,006,447	423	19.12.01~20.01.03	20.01.13
	구어 총계	1,006,447	423		
총계		3,025,769	7,688		

4.2.2. 정제 일정

상호 참조 해결 말뭉치의 정제 기간은 국립국어원과의 협의 사항에 따라 문어 1차 2020년 1월 9일, 문어 2차 2020년 1월 15일까지이며, 국립국어원의 구어 지침 변경에 따라 구어 정제는 협의에 의한다.

5. 검증 및 산출물 보고

5.1. 내부 검증

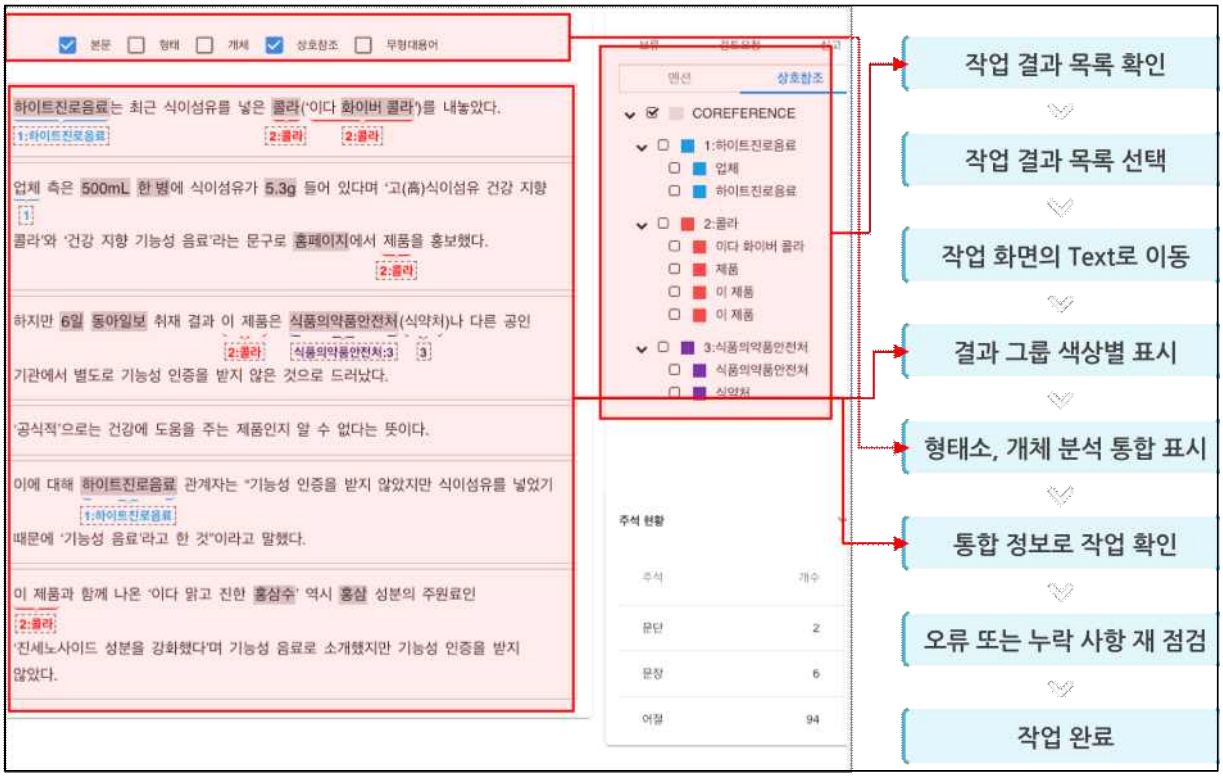
<표 11 >말뭉치 검증 단계

단계	검증 내용	검증 수행
1단계	작업자 검증	▪ 작업 결과 목록을 통한 작업 추적 ▪ 시각화를 통한 직관적 확인 ▪ 형태, 개체 분석 통합 확인
2단계	기계적 검증	▪ 작업 보류 유무 확인 ▪ 단일 개체 유무 확인 ▪ 누락 대상 여부 확인
3단계	절차적 검증	▪ 작업 -> 검수 -> 확인 및 반려 프로세스 ▪ 단계별 역할 부여, 작업 진행 및 단계 종료시 접근 격리 ▪ 검수 및 반려를 통한 재작업 프로세스
4단계	관리적 검증	▪ 보류 이력 현황 확인 ▪ 검토 요청 등록 현황 확인 ▪ 처리 현황 및 사례 추적 확인
5단계	외부 검증	▪ 작업 결과 목록을 통한 작업 추적 ▪ 시각화를 통한 직관적 확인 ▪ 형태, 개체 분석 통합 확인

5.1.1. 작업자 검증

작업 도구는 작업자의 직관이 반영된 도구로 시각화를 통해 작업 내역의 쉽고 빠른 추적이 가능하도록 하였다. 작업창에서 작업 내용 클릭시 해당 요소 하이라이팅 및 불확실한 작업 요소에 대해서 보류 및 검토 기능을 통해 부정확한 가공을 사전 배제하도록 하였다.

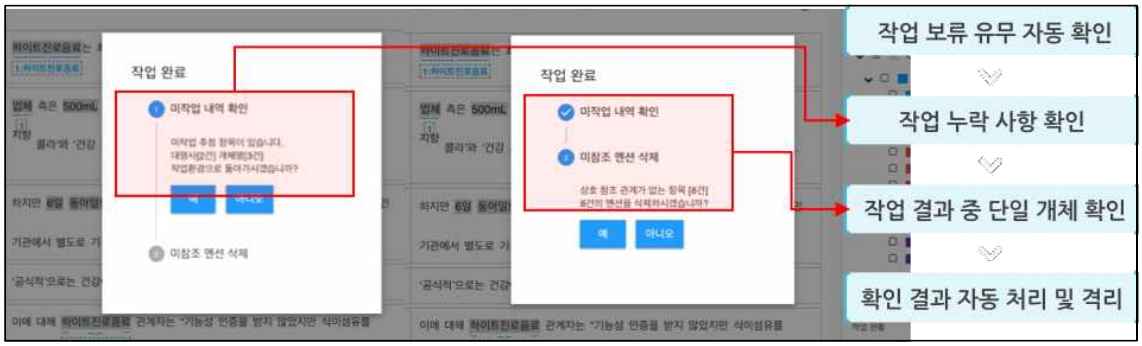
<그림 8> 작업자 작업 및 검증 화면



5.1.2. 기계적 검증

분석 말뭉치 구축을 위해 부착하는 주석 정보는 텍스트에 따라 주석의 양이 과해지면 작업 내용에 대한 추적이 어려워진다. 따라서 효율적인 구축 작업을 위해서 작업한 대상의 위치를 쉽게 추적 가능하도록 하는 작업과 동시에, 작업 완료 시 텍스트 내 미작업 대상에 대한 자동 검수를 통해 분석 대상 누락을 빠르게 점검하여, 구축 작업의 효율성을 높이는 동시에 작업 품질을 높이는 역할을 하였다.

<그림 9> 단일 개체(싱글톤) 및 누락에 대한 검증



5.1.3. 절차적 검증

검수 단계의 검증은 작업자 결과물에 대해 지침 준수 여부와 작업 결과물 사이의 불일치를 검출하는 검증으로써 단순 오류 검출 뿐만아니라 다수의 작업자 간 지침 해석에 다른 불일치 대상에 대한 처리 방침을 최종 결정하여 교정작업이 이루어지게 된다. 이 과정에서 데이터 변경, 작업 이력에 대한 통계 관리로 지침과 말뭉치의 품질을 개선하였다.

작업자는 작업 과정 중 지침이 모호하거나 적용하기 어려운 작업을 검수자에게 검토요청을 하고, 검수자는 작업자의 검토 요청 데이터에 주석을 부착하여 저장한다. 이 단계에서 완료된 데이터를 격리 저장하고 단계별로 역할을 부여하여 철저한 검증이 이루어지도록 하였다.

5.1.4. 관리적 검증

이 단계에서 검증에서 관리자는 작업자의 작업 내용과 현황을 확인할 수 있으며, 동일 시간 내 작업자별 수행량 확인, 작업자 별 검토 요청 확인, 검수자별 수정 현황 확인, 작업자와 검수자 작업 보류 이력 사항 확인 등이 가능하다. 이런 과정을 통해 작업자 작업의 품질을 관리하고, 작업자의 수행 성과를 관리하며, 작업 결과를 확인하여 지침 보완 사항을 검토할 수 있으며, 작업의 반려 및 지침 개선 시 적용할 수 있도록 하였다.

5.1.5. 활용성 검증

구축 도구 시스템에 저장된 데이터에 대하여 국어원과 협의하고, 데이터의 활용 및 평가 검증을 통해 객관적인 활용 의견을 수합하고, 해당 의견에 대한 지침의 검토와 구축된 말뭉치에 대한 검토를 통해 지침 개선이나 말뭉치 재작업 등의 순환 개선을 실시하였다.

<그림 10> 의견 검토 및 순환 개선 절차도

5.2. 외부 검증

상호 참조 해결 말뭉치 시범 검증 대상 73개의 표본으로 실시한 내용 및 오류 검증은 현재 같은 구문 분석 결과가 없기 때문에 작업자간 최장 명사구(NP) 설정에 대한 기준이 모호하다.

5.2.1. 내용 오류 검증

1) 내용 오류 검증 내용 및 방법

내용 오류 검증은 형식 오류 검증 통과한 구문 분석 말뭉치에 대하여, 국어원이 정답으로 제시한 정답 세트와의 일치도를 검사하는 것으로, 동일한 멘션을 추출하였는가 하는 동지시 관계 개체 정의(동일 mention 지정)검사와 추출한 멘션들을 같은 상호참조 그룹으로 묶었는가 하는 동지시 관계 개체 군집 검사(mention chain check)를 검사한다.

<표 12> 내용 검증 검사 항목

검사 항목	검사 내용
동지시 관계 개체 정의 (동일 mention 검사)	• 동지시 관계 개체의 범위가 겹칠 경우, 동일 개체로 인정 • 말뭉치 구축 주체 마다 개체 범위의 기준이 되는 구문 분석 결과가 다르기 때문에 동지시 관계 일치도를 높이기 위하여 관대한 기준을 적용 • 동지시 관계 개체를 설정할 때 지침상 최장 명사구(NP)를 기준으로 삼음
동지시 관계 개체 군집 검사 (mention chain check)	• 동지시 관계 개체가 형성한 군집의 일치를 검사하기 위해 문서 내에 존재하는 선행 동지시 관계가 일치할 경우 일치한 것으로 판단(MUC변형) • 동지시 관계 개체 군집이 일부만 일치해도 부분 점수 가능

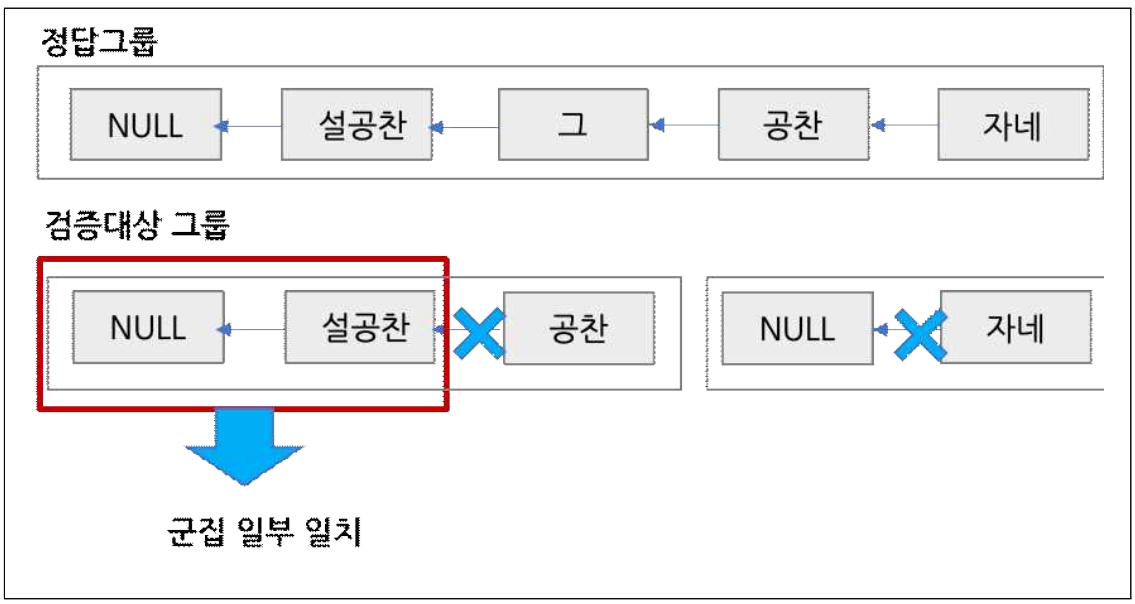
2) 내용 오류 검사항목 예시

- 일치도 검사항목(1) - 동지시 관계 개체 정의 (동일 mention검사)

구축 주체	동지시 관계 주석 예시
국립국어원	• 1997년 한글소설본이 발견되어 '홍길동전'을 밀어내고 최초의 한글소설 자리를 꿰찬 '설공찬전' • 조선 중종조에 이미 금서(禁書)가 된 작품 • 이 고전소설
검증 사업단	• 1997년 한글소설본이 발견되어 '홍길동전'을 밀어내고 최초의 한글소설 자리를 꿰찬 '설공찬전' • 조선 중종조에 이미 금서(禁書)가 된 작품 • 이 고전소설
구축 사업단	• 1997년 한글소설본 • 최초의 한글소설 자리를 꿰찬 '설공찬전' • 이 고전소설

위 군집 내에서 완정히 일치한 동지시 관계는 '이 고전소설' 밖에 없으며 최장 동지시 관계 내에 또 다른 동지시 관계가 포함되어 있다.

- 일치도 검사항목(2)-동지시 관계 개체 군집 검사(mention chain check)



5.3. 산출물

5.3.1. 1차 결과물 납품

납품 규모 : 문어 1,036,168어절(표본수 3,870개)

납품 시기 : 2019년 11월 27일

5.3.2. 2차 결과물 납품

납품 규모 : 문어 983,154어절(표본수 3,395개)

납품 시기 : 2019년 12월 13일

5.3.3. 3차 결과물 납품

납품 규모 : 1,006,447어절(표본수 423개)

납품 시기 : 2020년 1월 13일

5.3.4. 산출물 납품 형태

상호 참조 해결 말뭉치(300만 어절), 중간 산출물, 상호 참조 해결 말뭉치 지침을 저장 매체(USB 등, 3개)에 담아 제출하며, 사업 수행 과정 및 결과 요약 보고서 형태의 최종 보고서(인쇄본 30부)를 납품 완료 시(2020년 1월 15일)에 제출한다.

5.4. 사업 보고

착수 보고 2019년 8월 30일

중간 보고 2019년 12월 3일

최종 보고 2020년 1월 14일

제 3 장

향후 계획

1. 개선 방향

본 사업의 경우 4차 산업 및 언어적 연구 모두 만족하는 말뭉치를 구축하여야 하나, 각 분야의 복원 방법, 복원 대상, 범위 측면에서 각 층위별 서로 다른 이해관계를 가지고 있어 이를 모두 만족하고, 실현 가능한 지침 수립을 고려해야 하므로, 지침에 관한 이슈들이 많았다. 특히, 상호 참조 해결 층위는 기존의 연구가 미비하며, 고품질의 데이터 구축을 위해서는 지침의 설계와 수립이 선행되어야 한다. 또한 상호 참조 해결 말뭉치의 지침이 완결을 위해서는 상호 참조 해결 대상의 정의를 명확하게 하기 위한 구문분석 및 개체명에 대한 확실한 정의는 필수불가결한 요소라고 보인다.

이번 사업은 상호 참조 해결 말뭉치 구축에 대한 기본 지침 초안의 수립과 이를 바탕으로 한 상호 참조 해결 데이터 구축에 중점을 두었다. 이 사업을 통하여, 기본 지침이 수립되었으므로, 향후 이 지침을 보완해 가는 방향으로 나아가야 할 것이며, 구문분석 및 개체명 분석과 연동하여, 상호 참조 해결의 대상과 범위에 대한 보완점을 찾는 방향으로 이끌어 가야할 것이다. 이에 주석 체계의 보완, 무형대용어 복원 말뭉치 등 다른 분석 말뭉치와의 연계, 구어 자료의 처리의 관점에서 개선 방향을 제시해 본다.

1) 주석 체계의 보완

본 사업에서 구축된 상호참조 해결 말뭉치의 주석체계는 표지없이 멘션들의 집합으로 형성되었다. Ontonotes 등의 영어권 자료들이 참조 관계 표지를 부착하는 것과는 대조적이다. 이러한 참조 관계 표지의 문제는 구축 기간 동안 지속적으로 논의되었던 멘션들 간의 의미와 지시 관계에 대한 단방향적인 참조와 양방향적인 참조 관계의 논의와도 연관된다.

참조 관계에 대한 표지의 생략은 산업적 활용을 위한 대량의 주석에는 효과적이나, 상호참조의 언어학적, 인지적 정보 표현에는 부족한 면이 있다. 상호참조 관계에 있는 멘션들간의 관계들을 세분화하고 표현하되, 그 규모를 차등하게 유지하는 등의 보다 다각화된 정보 제공이 필요하다.

WikiCoref(Ghaddar and Langlais, 2016), PreCo(Hong Chen et al., 2018)과 같은 최근 영어권의 자료의 보이는 바와 같이, 상호 참조 해결 말뭉치는 비교적 작은 규모이며 여전히 개선과 발전 방안을 모색하고 있는 상황이다. 그러므로, 보다 다각화된 주석 체계와 정보 수준을 제공함으로써 산업적인 데이터 분석 처리뿐만 아니라 언어학적인 연구에도 활용의 폭을 넓힐 수 있다.

2) 다른 층위의 분석 말뭉치와의 연계

Ontonote 프로젝트에서 확인할 수 있는 바와 같이, 상호참조는 구문분석, 개체명, 어휘의미 분석과 함께 제공되는 경향이 있다. 언어학적으로는 (2014)에서 정리한 바와 같이, 상호참조는 넓은 의미에서 대용 관계와 함께 논의되었으며 의미, 통사, 화용의 측면에서 골고루 조명되어 왔다.

본 사업과 함께 진행된 국어빅데이터 구축에는 구문분석, 개체명, 의미역, 어휘의미 분석이 모두 포함되므로 자연스럽게 Ontonote와 같이 여러 층위의 주석 데이터를 통한 연계가 가능하다. 다만, 모든 어절에 분석 결과가 반영되는 형태, 구문분석과 달리 개체명, 의미역, 어휘의미의 경우 1:1로 대응되지 못하는 부분이 있다. 또한, 무형 대용어의 경우에는 선행어와 서술어와의 관계를 표현하고 있어서 통합적인 연계 분석이 쉽지 않다.

이러한 점들을 고려하여, 통합된 여러 층위의 분석 결과를 효과적으로 연계하고 조작할 수 있는 자료 구조와 처리 도구의 개발이 필요하다.

3) 구어 자료의 처리

구어 자료는 '21세기 세종계획'에서도 구축되어 활용된 바가 있으나, 금번 사업과 같이 상호참조나 무형대용어, 어휘의미 분석과 같은 수준의 주석이 적용된 대규모 자료는 없었다. 또한 전산처리의 관점에서도 구어 자료의 본격적인 분석과 활용은 최근에야 확대되고 있는 실정이다.

기본적으로 구어 자료는 문어체와 구어체라는 자료 자체의 차이 뿐만 아니라, 메타데이터와 본문 주석(문장, 단락 등의 단위적 정보)과 전사 주석(발화현상에 대한 주석) 등 자료의 표현과 의미가 문어 자료와 전혀 다르다.

본 사업에서는 구어 자료의 원시 말뭉치 상태를 최대한 유지하고 문어자료와 데이터 처리의 일관성이 유지되도록 다루었다. 이러한 방식은 문어 말뭉치와의 일관된 자료 처리라는 점에서 이점이 있는 반면에, 구어 자료의 고유한 특성이 부각될 수 없다는 한계를 동시에 갖게 된다. 그러므로, 향후의 개선에서는 구어 자료의 특성을 고려한 자료 구조 및 데이터 처리의 제공과 함께 문어 자료와의 호환성까지 고려한 처리 방안이 필요하다.

2. 기대 효과

본 사업을 통해 7개의 분석 말뭉치 층위 중 상호 참조 해결 층위의 기본 지침이 수립되었으며, 이 지침을 바탕으로 한 상호 참조 해결 말뭉치의 시범 데이터 및 말뭉치 데이터 구축이 완료되어, 통합 검증 및 다층위 말뭉치 호환성 검증을 진행한다.

상호 참조 해결 말뭉치 구축을 위해, 기본적인 범위, 단위, 대상, 방법을 명확히 제시하고, 작업자가 일관되게 판단 할 수 있도록 올바른 사례, 지침 개선에 반영될 수 있는 여러 예외 상황에 대한 검토와 시험을 진행하였으며, 이러한 지침에 따라 일관되게 작업할 수 있는 도구를 활용하여 상호 참조 해결 말뭉치를 구축함으로써, 4차 산업 및 언어적 연구에서 즉각적인 전산처리가 가능한 말뭉치 구축에 이바지 하고자 하였다.

상호 참조 해결 말뭉치는 여타의 말뭉치와 달리 영어권에서도 150만 어절 이상 규모의 말뭉치가 구축된 사례가 거의 없다. 영어권과 달리 참조 관계 유형에 대한 주석이 되어 있지 않다고 하더라도 대규모 자료라 할 만하다. 또한 기존에 공개된 국내 상호참조 말뭉치와는 표본 텍스트의 중복성이 없고 구축지침을 상당 부분 공유하였기 때문에 확대와 보완이 가능할 것으로 기대한다.

이 사업을 통해 4차 산업 혁명을 위한 우리나라 인공지능 기술 수준의 지체 및 기초 말뭉치의 양적 질적 부족에 따른 갈증을 해소하고, 민간에서 활용 가능한 공공재로서의 대규모, 고품질 우리말 자원의 구축과 이에 대한 활용이 활발해지기를 기대한다.

참고문헌

ETRI(2017), 상호 참조 해결 태깅 가이드라인 v.3.1

ETRI(2017), 엑소브레인 언어분석 통합말뭉치 자료 구조 v.2.3

BBN TECHNOLOGIES(2006), Co-reference Guidelines for English OntoNote

김광희(2014), 대용 표현 연구의 이론과 논점, 한국어 통사론의 현상과 이론

이선희(2014), 말뭉치에 기반한 국어 대용어 연구 : 재귀사 '자기', '이, 그, 저' 지시 표현, 영형태를 중심으로

Hong Chen et al(2018), PreCo: A Large-scale Dataset in Preschool Vocabulary for Coreference Resolution

Abbas Ghaddar, Philippe Langlais(2016), WikiCoref: An English Coreference-annotated Corpus of Wikipedia Articles

Thiago Alexandre Salgueiro Pardo et al(2017), The Coreference Annotationof the CSTN ews Corpus

Ulrich Schäfer, Christian Spurk, Jörg Steffen(2012), A Fully Coreference-annota ted Corpus of Scholarly Papers from the ACL Anthology

Arrick Lanfranchi, Kevin Crooks, Mariah Hamang(2013), Clinical Coreference A nnotation Guidelines (with excerpts from ODIE guidelines and modified for SHA RPn/THYME)

中村真衣佳(2017), 同一指示と解釈される「N1のN2」と「N2のN1」：反転表現「N2のN 1」の焦点化の要因, 北海道大学大学院文学研究科研究論集 (17), 169-183

池田則之(2011), 同一指示解釈と叙述関係, 九州大学言語学論集(32), 31-52

東郷雄二(2004) 名詞句の指示とコピュラ文の意味機能, 科学研究費補助金 基盤研究(C) 課題番号 14510569 平成14年度∽16年度 研究成果報告書, 1-65

吉田悦子 · 谷村緑(2008) 同一指示係を記述するためのアノテションモデルの討-MUC-7と MATEを比較して-, 言語處理學會第14回年次大會發表論文, 440-443

今田水穂(2010) 認識動詞構文と間スペース同定, 言語学論叢 オンライン版第3号(通巻29 号), 33-44

LDC[웹사이트], (2020.03.12.), URL:https://catalog.ldc.upenn.edu/LDC2001T02

Universität Stuttgart[웹사이트], (2020.03.12.), URL:https://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/matetools/
CoNLL formatted[웹사이트], (2020.03.12.) URL:https://github.com/ontonotes/conll-formatted-ontonotes-5.0https://github.com/ontonotes/conll-formatted-ontonotes-5.0

부 록

상호 참조 해결
말뭉치 구축 지침

개정이력

버전	작성일	변경내용	작성자	승인자
0.1	2019-06-08	최초작성	곽용진	곽용진
0.2	2019-06-13	1.4. 지침 구성 추가	곽용진	곽용진
0.3	2019-08-23	2. 주석 대상 마커블 지침 추가	최지선	최지선
0.4.0	2019-08-26	1) 지침 작성 구성 체계 변경에 따른 문서 전체갱신 2) 0.4 완료시까지 0.4.1의 하위버전에 의한 갱신	곽용진	곽용진
0.5	2019-09-11	ETRI 상호참조태깅가이드라인3.1 기반 목차 수정	곽용진	곽용진
0.5.1	2019-09-17	예시 사례/오류/예외 정리 및 사례 추가	홍은기	곽용진
0.5.2	2019-09-19	국어원 요청 지침 추가, 목차 정리	윤영민	곽용진
0.5.3	2019-09-25	국어원 검토/수정(사례의 정오표시 확정)	서근화	곽용진
0.5.4	2019-09-26	각 지침 사례에 대한 검토 및 수정	홍은기	곽용진
0.5.5	2019-10-08	태깅 단위 변경(어절->형태)	홍은기	곽용진
0.5.6	2019-10-18	지침 내용 수정 1.1 상호참조 해결의 정의와 태깅 단방향성 정의 2.1.6 대등접속사로 연결된 멘션들, 2.1.8 명사들이 반점, 온점으로 묶인 경우 2.1.9 지정사가 포함된 경우 용어검토 2.1.11 한정사구 용어 검토	최지선	곽용진
0.5.7	2019-10-25	중심어 중복 제거 규칙 지침 검토에 따른 2.2.9 생략된 명사구 예문 수정	최지선	곽용진
0.5.8	2019-11-11	2.2.12, 유의어 및 일반화/구체적 표현 지침 추가 2.2.13 일반 대상을 지칭하는지 구체적인 특정 대상을 지칭하는지 판단이 어려운 경우 추가 2.1.8 접속사(A and B)형태단위 분석 내용 지침 추가	최지선	곽용진
0.5.9	2019-11-15	관형사 수식 명사구 , 한정사구 용어 통일	최지선	곽용진
0.6.	2019-12-13	2.1.2, 2.1.9, 2.2.3, 2.1.16 구어 지침 사례 추가	최지선	곽용진
0.6.1	2020-01-06	1.4 구어 지침 추가	곽용진	곽용진

* 모든 용어, 정의, 개념은 참조 문헌/자료/논거(각주)를 제시해야 한다.
** 모든 지침은 1개 이상의 사례를 반드시 제시해야 한다.

1. 개요

1.1. 상호참조해결의 정의와 태깅

상호참조해결의 태깅은 선행사와현재 등장한 대용어 간의 관계를 연결해주는 것이다.(ETRI v.3.1. p.4, 1.상초참조해결 태깅)

상호참조해결(Coreference resolution)은 임의의 개체(entity)에 대하여 다른 표현으로 사용되는 단어들을 찾아, 서로 같은 개체로 연결해주는 자연어처리 문제이다. 하나의 개체를 다른 단어로 표현하는 경우는 별명, 약어, 대명사, 관형사구[1] 등이 있으며, 이들 간의 참조 관계를 올바르게 찾아낼 수 있으면 담화나 문서 내에서 언급하는 대상에 대한 정보를 일관성 있게 유지할 수 있고, 정확하게 전달할 수 있다. 따라서 상호참조해결은 문서에서 등장하는 개체를 이해하는 데 매우 중요한 역할을 하며, 질의 응답, 문서요약, 기계 번역, 정보 추출 등에 응용될 수 있다. 이와 같은 상호참조관계의 선행사와 대용어를 연결하는 것은 상호참조해결의 태깅이라 한다.

예를 들어 “오바마는 작년 봄에 한국에 방문하였다.”, “그 이후에 미국 대통령은 말레이시아를 방문하기로 예정되어 있었다.”, “그는 현재 백악관에서집무 중에 있다.”와 같은 세 문장에서 ‘오바마, 미국 대통령, 그’가 서로 상호참조관계이며, 이들을 {오바마, 미국 대통령, 그}와 같이 서로 연결해 놓은 것이 상호참조해결 태깅이다.

> 국립국어원은 상호참조 분석 범위를 ① world 내, ② document 내 두 가지 방법으로 모두 가능하다고 본다. ①의 분석 범위로 하는 것이 ‘일반언어학적 관점’에서 더 적합하나, ‘기계 처리 관점’에서는 ②의 분석 범위로 하는 것이 기계 성능 및 학습/활용, 분석의 일관성 측면에서 더 장점이 있다. 또한 이번 말뭉치 사업은 ‘산업 활용’이 주목적이므로 국립국어원 상호 참조 해결 말뭉치 구축 지침서에서는 상호 참조 분석 범위를 ②document(문서) 내로 한정하여 a) 기계/시스템 처리 관점에서의 단방향성, 양방향성 개념(‘엔티티’ 구조와 개념), b) 많은 분석 대상, c) 분석의 일관성을 갖춘다.
>
> ‘단방향성’은 상호참조 분석 범위를 문서(document) 내로 한정하는 것을 전제로 한다. 동일 문서 내에서 같은 대상을 지시하는 다른 형태의 두 멘션이 있을 때, 한 멘션이 다른 멘션을 지시할 수 있는 경우를 말한다. ‘단방향성’의 방향은 선행자가 후행자를 지시(A→B)할 수도 있고, 역으로 후행자가 선행자를 지시(A←B)할 수도 있다.

1.2. 용어의 정의

멘션(Mention)은 상호참조해결의 대상이 되는 모든 명사구(즉, 명사, 명사구 등)를 의미한다. 멘션에서 해당 구의 실질적인 의미를 나타내는 단어를 중심어(head)라 하며,멘션은 중심어를 중심으로 이를 수식하는 수식어를 포함한다. 엔티티는 동일한 멘션의 집합으로, 상호참조해결의 결과이다. 선행 멘션(antecedent)과 현재 등장한 멘션간의 참조를 해결하면 하나의 엔티티

1) 이때의 ‘관형사구’는 영어의 ‘한정사구(Determiner Phrase)’에 대응하는 개념이다.

로 포함된다.

멘션 탐지 단계는 의존 구문 트리에서 등장하는 모든 명사구를 멘션으로 잡는다. 문서 내의 멘션을 탐지하기 위해 다음과 같은 규칙을 적용한다.

- 멘션은 기본적으로 형태 단위로 처리한다.
- 수식 정보를 포함한 멘션 생성(즉, 명사구)[2)]
- 개체명의 원자성[3)]
- 중심어의 중복 처리[4)]
- 대명사 분류

표 1. 멘션 탐지를 수행한 예

입력문서

프랑스의 르노 자동차 그룹은 한국 삼성자동차 인수를 공식 제의할 것이다.

- Step 1: 문장에서 명사 단어 추출

[프랑스], [르노], [자동차], [그룹], [한국], [삼성자동차], [인수]

- Step 2: 중심어에 대한 수식어 확장

[프랑스], [르노], [자동차], [르노 자동차], [한국], [삼성자동차], [한국 삼성자동차], [인수], [한국 삼성자동차 인수]

- Step3: 각 멘션에 대한 개체명 확인 (개체명 원자성)

i. Marking NE

[프랑스]: LCP_COUNTRY, [르노], [르노 자동차]: OGG_BUSINESS, [그룹], [르노 자동 차 그룹], [프랑스의 르노 자동차 그룹], [한국]: LCP_COUNTRY, [삼성자동차]: OGG_ BUSINESS, [인수], [한국 삼성자동차 인수]

ii. Result of this step

[프랑스], [르노 자동차], [그룹], [르노 자동차 그룹], [프랑스의 르노 자동차 그룹], [한국], [삼성자동차], [인수], [한국 삼성자동차 인수]

- Step 4: 중복되는 중심어 제거 (긴 멘션을 사용)

[프랑스], [르노 자동차], [프랑스의 르노 자동차 그룹], [한국], [삼성자동차], [한국 삼성자 동차 인수]

- 멘션탐지결과

$[[\text{프랑스}]^{1}_{1}$의 $[\text{르노 자동차}]^{2}_{2}$ 그룹$]^{0}_{0}$은 $[[\text{한국}]^{4}_{4}$ $[\text{삼성자동차}]^{3}_{3}$ 인수$]^{5}_{5}$를 공식 제의할 것이다.

2) 의존 구문 트리는 단어마다 수식 정보(즉, mod)를 가지기 때문에 이것을 이용하여 수식 정보를 포함한 완벽한 명사구를 생성한다. 즉, 각 단어에 포함 된 mod 인덱스를 따라가 수식어를 포함한 하나의 멘션을 만드는데, 멘션의 중심어는 항상 명사류(즉,NP 가 포함된 태그 정보)로 정의한다. 만약, 멘션의 중심어가 동사류(즉, VP가 포함된 태그 정보)이면 멘션 탐지 정의에 의하여 해당 단어는 멘션으로 추출하지 않는다.

3) 개체명은 멘션에서 의미가 있는 단어의 최소 단위로 본다. 그러므로 상호참조해결을 진행할 때 개체명을 멘션으로 또는 멘션의 자질로서 참조하기 때문에 추출하는 멘션이 개체명보다 작은 단위가 있어서는 안 된다. 따라서 멘션 추출 시 '중심어 중복 처리'의 상황과 둘 중 하나가 개체명이라면, 개체명을 멘션으로 하고 그보다 작은 멘션은 제거한다.

예를 들어, $[[\text{르노}]^{0}_{1}$ 자동차$]^{0}_{0}$ ⇒ $[\text{르노 자동차}]^{0}_{0}$와 같은 결과를 가진다. 즉, $[\text{르노}]^{0}_{1}$ 는 개체명이 없지만, $[\text{르노 자동차}]^{0}_{0}$는 'OGG_BUSINESS'라는 개체명을 가진다. 따라서 $[\text{르노}]^{0}_{1}$를 $[\text{르노 자동차}]^{0}_{0}$에 포함시켜 제거한다.

4) 멘션 추출 시 중심어와 중심어의 위치가 같은두 개 이상의 명사구가 추출될 경우, 더 큰 멘션을 선택하며 작은 멘션은 제거 한다. 크기가 같은 멘션은 중복 처리(즉,하나만 남긴다)를 한다.

예를 들어, [오바마 $[\text{대통령}]^{0}_{1}]^{0}_{0}$à $[\text{오바마 대통령}]^{0}_{0}$과 같이 두 개의 멘션 $[\text{오바마 대통령}]^{0}_{0}$,$[\text{대통령}]^{0}_{1}$은 서로 중심어와 그 위치가 같다. 따라서 멘션이 더 큰 $[\text{오바마 대통령}]^{0}_{0}$을 추출하고 멘션의 크기가 작은 $[\text{대통령}]^{0}_{1}$을 제거한다.

1.3. 구축지침의 구성

1.3.1. 예문 제시

상호참조 해결 규칙을 제시할 때는 반드시 예문, 올바른 적용예시와 잘못된 예시 3가지 항목을 함께 제시한다. 이 밖에 규칙의 예외사항이 있다면 그 예시 또한 별도로 제시한다. 예문은 <예문>, 중심어 중복 제거 규칙을 적용한 올바른 모든 멘션은 <보기>, 잘못된 멘션 후보는 <오류>, 상호참조 관계인 멘션은 <상호참조 태깅>, 예외 예시는 <예외>목록에 각각 기록하며, (O), (X) 기호를 이용하여 추가로 정/오답을 표시한다.

1.3.2.기호 정의

예문은 “ ”, 멘션은 [ ], 상호참조관계는 { } 기호를 이용하여 일관되게 표현하며, 정/오답을 알려주는 o,x표시는 ( )기호를 이용하여 표현한다

1.4. 구어 자료의 처리

1.4.1. 주요 문법적 단위의 처리

구어는 문어와 사뭇 다른 문법적 단위와 양상을 보인다. 구어 자료는 문장의 구분이 없고 발화자가 존재하며, 발화상에서 발생하는 물리적 현상(겹침, 끼어듦, 발화 외적 소리(웃음, 박수, 기침 등), 불분명한 발음/발화 등)이 많다. 이러한 구어 자료의 특수성이 반영된 구어 전사 원시 말뭉치는 문어 자료와 다른 문법 구조와 함께 문서 구조를 가지게 된다. 그러나 본 지침에서는 효율적인 자료 처리와 작업자의 직관 활용을 위해 이러한 차이를 최소화하여 처리한다.

문법적 단위라 할 수 있는 문어의 단락(<p>), 문장(<s>), 어절(공백문자)은 각각 구어 자료의 발화자(<u>의 화자정보 속성, 줄바꿈을 가짐), 억양 단위(<u>의 값 데이터), 어절(공백문자)와 대응하는 것으로 본다.

또한, <unclear> 등과 같이 값을 갖는 전사 표지는 주석의 대상이 되는 문자열의 일부로 보고, <vocal desc="laughing"/>와 같이 값을 갖지 않는 전사 표지는 주석 대상에서 제외한다.

1.4.2. 지침 적용 사례의 표현

구어에 대한 상호 참조 주석(태깅)의 지침은 문어와 동일한 지침을 적용한다. 다만, 동일한 지침이라도 구어의 표현, 문맥적 특성상 그 양상이 문어와 현저히 다른 경우에는 해당 지침에 사례를 추가하고 사례번호에 ‘구어자료의 경우’로 명시하여 활용할 수 있게 한다.

> 2.1.2. 명사구에서 가장 의미를 갖는 명사는 중심어(head)이며, 멘션은 중심어를 기반으로 해당 명사구에 존재하는 수식어까지 포함해야 한다.
>
> 예1) …..
>
> 예4) 담화 표지서 지시사의 수식어 포함(구어의 경우)
>
> <예문> 이 쪼끄만 내장 있는 데 쌉쓰름한 맛이 있어요 이릏게.
>
> <보기> [쪼끄만 내장](X), [이 쪼그만 내장](O),

2. 상호참조해결 가이드

2.1. 멘션 탐지 태깅 규칙

2.1.1. 모든 멘션은 명사(구)에 기반하며, 형태 단위로 태깅한다.

<예외> 구어의 경우 불특정 다수를 언급하는 명사(구)는 멘션으로 추출하지 않는다.

예1)

<예문> 미국의 대통령인 오바마는 어제 한국을 방문했다.(etri지침 예시_29쪽)

<보기> [미국의 대통령인 오바마](O), [미국의 대통령](O), [미국](O),[어제](O), [한국](O)

<오류> [미국의 대통령인 오바마는](X), [미국의 대통령인](X), [미국의](X), [한국을](X)

<상호참조 태깅>{[미국의대통령인 오바마], [미국의 대통령]}(O)

예2) 어절 좌우에 기호가 포함되거나 띄어쓰기로 인해 어절이 잘못 연결된 예

<예문> ▽서울시, “영향 없다”느긋=경기지역은 미군기지 이전 연기에 따른 불만의 목소리가 높지만 용산 미군기지가 있는 서울시는 다소 느긋한 표정이다.

<보기> [서울시](O), [경기지역](O), [미군기지](O), [용산](O), [용산 미군기지](O), [용산 미군기지가 있는 서울시](O)

<오류> [▽서울시,](X), [느긋=경기지역은](X)

<상호참조 태깅> {[서울시], [용산 미군기지가 있는 서울시]}(O)

▶ 구어에서 ‘여러분, 저희’와 같이 불특정 다수를 언급하는 경우, 대상이 명확하게 드러나지 않는 경우 멘션으로 추출하지 않는다.

예1)

<예문> 어떤 모범 밥상을 여러분들께 소개해 드릴지 기대가 되는데요… 또 어떤 요리가 저희를 기다리고 있을까요?

<보기> [모범 밥상](O), [요리](O)

<오류>

<상호참조 태깅> {∅}(예문 내 멘션들 중 상호참조 관계 없음)

▶ 일반명사 외의 복합명사, 대명사 등도 멘션으로 정의한다.

예1)

<예문> 노벨 평화상(etri지침 예시_29쪽)

<보기> [노벨](O), [노벨 평화상](O)

<오류>

<상호참조 태깅> {∅}(예문 내 멘션들 중 상호참조 관계 없음)

예2) 대명사 멘션의 예

<예문> 이것은 여우입니다.(자체 생성 예시문)

<보기> [이것](O), [여우](O)

<오류>

<상호참조 태깅>{[이것], [여우]}(O)

예3) 어절이 분리되지 않은 복합명사의 예(어절이 분리되지 않은 복합명사 내부의 각 형태소를 멘션으

로 추출할 수 없다.)

<예문> 미군기지가 옮겨와 지역경제가 활성화할 것으로 믿고 5년 전 이곳에 온 이씨는…고 주장했다.

<보기> [미군기지](O), [지역경제](O), [5년](O),[이곳](O), [미군기지가 옮겨와 지역경제가 활성화할 것으로 믿고 5년 전 이곳에 온 이씨](O)

<오류> [미군](X),[기지가](X), [5년 전](X)

<상호참조 태깅>{∅}(예문 내 멘션들 중 상호참조 관계 없음

예4) 약어형 복합명사

<예문>… 한미 양국에 제출했다. …. 한국 측은 여전히 … 미국 측이 …

<보기> [한미](O), [한미 양국](O),[한국](O), [한국 측](O), [미국](O), [미국 측](O)

<오류> [한](X), [미](X)

<상호참조 태깅> {[한미],[한미 양국]}(O), {[한국],[한국 측]}(O), {[미국],[미국 측]}(O)

2.1.2.명사구에서 가장 의미를 갖는 명사는 중심어(head)이며, 멘션은 중심어를 기반으로 해당 명사구에 존재하는 수식어까지 포함해야 한다.

<예외> etri구축 지침 또는 원시 말뭉치에 오류가 있는 경우에는 태깅하지 않고 '오류'로 분류하여 보고한다.

<예외> 구어의 경우 발화단위를 넘어 멘션 추출하지 않는다.

2.1.2. 명사구에서 가장의미를 갖는 명사는 중심어(head)이며, 멘션은 중심어를 기반으로 해당 명사구에 존재하는 수식어까지 포함해야 한다.

<예외> etri구축 지침 또는 원시 말뭉치에 오류가 있는 경우에는 태깅하지 않고 '오류'로 분류하여 보고한다.

<예외> 구어의 경우 발화단위를 넘어 멘션 추출하지 않는다.

예1)

<예문> 위상기하학의 유명한 문제인 푸앵카레 추측(etri지침 예시_29쪽)

<보기> [위상기하학의 유명한 문제인 푸앵카레 추측](O), [위상기하학의 유명한 문제](O), [위상기하학](O), [푸앵카레](O)

<오류> [유명한 문제인](X), [유명한 문제인 푸앵카레 추측](X)

<상호참조 태깅> {[위상기하학의 유명한 문제], [위상기하학의 유명한 문제인 푸앵카레 추측]}(O)

예2)

<예문> 64일 동안 쉬지 않고 4500km를 달리는 유럽 종단 울트라마라톤 대회에 참가한다.

<보기> [64일](O), [64일 동안](O), [4500km](O), [유럽](O), [유럽 종단](O), [유럽 종단 울트라마라톤] (O), [64일 동안 쉬지 않고 4500km를 달리는 유럽 종단 울트라마라톤 대회](O)

<오류>

<상호참조 태깅> {∅}(예문 내 멘션들 중 상호참조관계 없음)

예3)

<예문> 한화는 … 대우조선의 경영권을 확보하고 … 분할 매입 방안을 제시했다.

(표본명: NWRW1800000021-0019(대우조선-산은), 5번 문장)

<보기> [한화](O), [대우조선](O), [대우조선의 경영권](O), [분할](O), [분할 매입](O), [분할 매입 방안](O)

<오류> [경영권을](X),

<상호참조 태깅> {∅}(예문 내 멘션들 중 상호참조관계 없음)

▶ 구어의 경우 발화 단위를 넘어서서 멘션 추출 하지 않는다. 발화 단위로인해 끊어진 명사는 앞의 수식어를 제외하고 중심어만 멘션 추출하며, 앞의 수식어가 붙은 의미로 여기고 상호참조 태깅한다.

예4)

<예문>

U: 차가운 커피 위에

U:색색의 크림으로

U:그림을 그리는

U:크리마트가 인기라고 하는데요

…(중략)…

U:어~ 크리마트는 일단 라떼아트 처럼

<보기> [차가운 커피](O), [색색의 크림](O), [크리마트](O), [크리마트](O)

<오류> [차가운 커피 위에 색색의 크림으로 그림을 그리는 크리마트](X)

<상호참조 태깅> {[크리마트], [크리마트]}(O)

2.1.3. 한 멘션의 중심어 부분을 제외한 내부 명사(구)도 멘션으로 태깅한다.

예1)

<예문> 위상기하학의 유명한 문제인 푸앵카레 추측(etri지침 예시_30쪽)

<보기> [위상기하학의 유명한 문제인 푸앵카레 추측](O), [위상기하학의 유명한 문제](O), [위상기하학](O), [푸앵카레](O)

<오류> [유명한 문제인](X), [유명한 문제인 푸앵카레 추측](X)

<상호참조 태깅> {[위상기하학의 유명한 문제], [위상기하학의 유명한 문제인 푸앵카레 추측]}(O)

예2)

<예문> PMC는 미군기지 이전사업 전반을 총괄하는 한미 양국의 민간용역업체 컨소시엄으로…제출했다…(중략)…군 당국은 주한미군기지 이전사업 종합관리업체(PMC)가 최근 최종제안서를 제출함에 따라…(후략)

<보기> [PMC](O), [미군기지 이전사업 전반을 총괄하는 한미 양국의 민간용역업체 컨소시엄](O), [미군기지 이전사업](O), [미군기지](O), [주한미군기지](O), [주한미군기지 이전사업](O), [주한미군기지 이전사업 종합관리업체](O), [한미](O), [미군기지 이전사업 전반을 총괄하는 한미 양국](O)

<오류> [민간용역업체 컨소시엄으로](X),

<상호참조 태깅> {[PMC],[미군기지 이전사업 전반을 총괄하는 한미 양국의 민간용역업체 컨소시엄], [주한미군기지 이전사업 종합관리업체]}(O), {[미군기지 이전사업], [주한미군기지 이전사업]}(O), {[미군기지], [주한미군기지]}(O), {[한미], [미군기지 이전사업 전반을 총괄하는 한미 양국]}(O)

▶ 구어에서 명사 혹은 관형사 구실을 하는 담화 표지 등은 최장NP 규칙을 적용하고 최장NP 내부의 명사는 내포 명사구로 잡지 않는다.

예3)

<예문> 무슨 신남돈가 무슨 그런 섬이 있다 그러더라고

<보기> [무슨 신남돈가 무슨 그런 섬](O)

<오류> [무슨 신남돈](X), [무슨 그런 섬](X), [그런 섬](X)

<상호참조 태깅> {∅}(예문 내 멘션들 중 상호참조 관계 없음)

2.1.4. 중복되는 위치의 중심어가 두 개 이상의 멘션을 추출할 경우, 바운더리가 더 큰 멘션만 선택한다.(중심어 중복 제거 규칙)

예1)

<예문> 프랑스의 르노 자동차 그룹은 한국 삼성자동차 인수를 공식 제의할 것이다.

(etri지침 예시_31쪽)

<보기> [프랑스](O), [르노](O),[르노 자동차](O), [프랑스의 르노 자동차 그룹](O),[한국](O), [삼성자동차](O), [한국 삼성자동차 인수](O)

<오류> [그룹](X), [자동차 그룹](X), [르노 자동차 그룹](X), [인수](X), [삼성자동차 인수](X)

<상호참조 태깅> {∅}(예문 내 멘션들 중 상호참조 관계 없음)

예2) 사람 이름과 직함의 예: 직함이 중심어가 될 경우 [이름+직함]으로 멘션을 추출한다.

<예문> 아베 신조 일본 총리(구축지침 대조정리 문서에서 발췌)

<보기> [아베 신조 일본 총리](O), [아베 신조](O), [일본](O),

<오류> [일본 총리](X)

<상호참조 태깅> {[아베 신조], [아베 신조 일본 총리]}(O)

<예외>

<예문> 아베 신조는 10년 째 일본 총리이다.

<상호참조 태깅> {[아베 신조], [10년 째 일본 총리]}(O)

2.1.5. 멘션의 의미 단위 정의

▶ 멘션 의미의 최소단위는 개체명 정보, 개체명이 없는 경우 단어(형태)가 최소단위이다.

예1)

<예문> 아름다운 아메리카 플로리다주(etri지침 예시_31쪽)

<보기> [아메리카](O), [아름다운 아메리카 플로리다주](O)

<오류> [아름다운 아메리카](X)

<상호참조 태깅> {∅}(예문 내 멘션들 중 상호참조 관계 없음)

예2)

<예문> 알프스 산맥(etri지침 예시_31쪽)

<보기> [알프스](O), [알프스 산맥](O)

<오류> [산맥](X)

<상호참조 태깅> {∅}(예문 내 멘션들 중 상호참조 관계 없음)

2.1.6. 대등 접속[5]으로 연결된 멘션들

예1) 수식어는 대등 접속으로 연결된 멘션에서 잡는다.("~~ A, B and (or) C" → [A], [B], [C],[~~ A, B and (or) C]와 같이 추출)

<예문> 크다는 의미의 한국어 '아리'와 한자 수(水)를 결합한다.

(etri 지침 예시 32쪽)

<보기> [크다는 의미의 한국어 '아리'와 한자 수](O), [한국어 '아리'](O), [한자 수](O)

<오류> [크다는 의미의 한국어'아리'](X)

<상호참조 태깅> {∅}(예문 내 멘션들 중 상호참조 관계 없음)

▶ 동사구가 대등 접속으로 연결된 경우 멘션으로 추출하지 않는다.

예1) 파생동사의 어기가 대등접속으로 연결된 예

<예문> 재건 및 증축하면서 이름이 바뀌었다.(etri지침 예시 32쪽)

<보기> [이름](O)

<오류> [재건](X),[증축](X)

<상호참조 태깅> {∅}(예문 내 멘션들 중 상호참조 관계 없음)

2.1.7. 중심어를 수식하는 괄호의 경우

▶ 괄호 안의 정보들은 멘션으로 추출하지 않는다.(예2 참조)

예1)

<예문> 평양 순안 국제공항(平壤順安國際空港, Pyeongyang Sunan International airport)은 평양직할시 중심부이다. (etri 지침 예시 32쪽)

<보기> [평양 순안 국제공항](O), [평양직할시 중심부](O), [평양직할시](O), [평양] (O), [평양 순안](O)

<오류> [平壤順安國際空港,](X),[Pyeongyang Sunan International airport)](X)

<상호참조 태깅>

{[평양 순안 국제공항], [평양직할시 중심부]}(O), {[평양], [평양직할시]}(0)

예2)

5) 여기서 '대등 접속'이란 대등적 연결어미 '-고', 접속조사 '와/과', '나'등을 포함하는 명사구를 가리킨다

<예문> 인수대금(약 6조4000억 원)의 60%를 우선 납부해 대우조선의 경영권을 확보하고…분할 매입 방안을 제시했다.(표본명: NWRW1800000021-0019(대우조선-산은), 5번 문장)

<보기> [인수대금](O),[인수대금(약 6조 4000억 원)의 60%](O), [대우조선](O), [대우조선의 경영권](O), [분할](O), [분할 매입](O), [분할 매입 방안](0)

<오류> [약 6조4000억 원](X)

<상호참조 태깅> {∅}(예문 내 멘션들 중 상호참조 관계 없음)

▶ 멘션 [NP1(괄호정보) + 와/과, 반점 + NP2(괄호정보)]를 추출할 경우 NP1과 NP2의 괄호정보를 포함하여 멘션을 추출한다.

<예문>'무관의 여왕' 디나라 사피나(러시아 1위)와 돌아온 '러시안 뷰티' 마리야 샤라포바(31위)는 2라운드에 진출했다.(NWRW1800000024-0321)

<보기> [무관](O), [무관의 여왕](O), ['무관의 여왕' 디나라 사피나](O), [돌아온 '러시안 뷰티']](O), [러시안](O), [돌아온 '러시안 뷰티' 마리야 샤라포바](O), ['무관의여왕' 디나라 사피나(러시아 1위)와 돌아온 '러시안 뷰티' 마리야 샤라포바(31 위)](O), [2라운드](O)

<오류>

<상호참조 태깅> {[무관의 여왕], ['무관의 여왕' 디나라 사피나]}(O), {[돌아온 '러시안 뷰티'], [돌아온 '러시아 뷰티'마리야 샤라포바]}(O)

▶ 멘션 [수식어 + NP1]를 추출할 경우 NP1을 수식하는 수식어에 결합한 괄호 및 문장부호는 삭제하지 않는다.

<예문>'홀렙(HOLEP)'이라고 부르는 새 수술법은…(후략) (NWRW1800000022-0401)

<보기> ['홀렙(HPLEP)'이라고 부르는 새 수술법](O), [홀렙](O)

<오류>

<상호참조 태깅> {[홀렙], ['홀렙(HPLEP)'이라고 부르는 새 수술법]}(O)

2.1.8. 여러 명사들이 반점, 온점 등으로 묶이는 경우

▶ 2.1.6.의 대등 접속으로 연결된 멘션들과 마찬가지로 전체를 하나로 추출하고 헤드부분을 제외한 명사들을 따로 추출한다.

예1)

<예문> 카이사르, 아우구스투스, 바루스 등(etri 지침 예시 33쪽)

<보기> [카이사르, 아우구스투스, 바루스 등](O), [카이사르](O), [아우구스투스](O), [바루스](O)

<오류> [카이사르, 아우구스투스](X), [아우구스투스,바루스](X)

<상호참조 태깅> {∅}(예문 내 멘션들 중 상호참조 관계 없음)

예2)

<예문> 제너럴모터스, 포드, 크라이슬러 등 생사의 기로에 선 미국의 빅3가…월스트리트저널이 5일 보도했다. (표본명: NWRW1800000021-0003(유럽차), 2번 문장)

<보기> [제너럴모터스, 포드, 크라이슬러 등 생사의 기로에 선 미국의 빅3](0), [제너럴모터스, 포드, 크라이슬러 등](O), [제너럴모터스](O), [포드](O), [크라이슬러](O), [미국](O), [월스트리트저널](O), [5일](O)

<오류> [제너럴모터스, 포드](X),[포드, 크라이슬러](X)

<상호참조 태깅> {[제너럴모터스, 포드, 크라이슬러 등], [제너럴모터스, 포드, 크라이슬러 등 생사의 기로에 선 미국의 빅3]}(O)

2.1.9. 지정사가 포함된 경우

▶ 멘션추출이 가능한 단어('이다, 이었다, 로서, 이며')와 불가능한 단어('라면서 등 '이다'이외의 단어)를 구분하여 가능한 단어는 추출, 불가능한 단어는 제외한다.

예1)

<예문> 아프로디테는 그리스 신화에 나오는 미와 사랑의 여신이다.(etri 지침 예시 33쪽)

<보기> [아프로디테](O), [그리스 신화에 나오는 미와 사랑의 여신](O),

[그리스 신화](O), [미와 사랑](O), [미](O),[사랑](O)

<오류>

<상호참조 태깅> {[아프로디테], [그리스 신화에 나오는 미와사랑의 여신]}(O)

2.1.10. 입력된 문장 전체가 명사구인 경우

▶ 안긴 문장이 안은 문장 내에서 명사 역할을 한다면 멘션으로 잡는다.

예1)

<예문> 뉴질랜드는 1893년 여성의 참정권을 세계 최초로 부여한 나라(etri 지침 예시 33쪽)

<보기> [뉴질랜드는 1893년 여성의 참정권을 세계 최초로 부여한 나라](O), [뉴질랜드](O), [1893년](O), [여성의 참정권](O), [여성](O), [세계 최초](O)

<오류>

<상호참조 태깅> {[뉴질랜드], [뉴질랜드는1893년 여성의 참정권을세계 최초로 부여한 나라]}(O)

2.1.11. 지시관형사를 포함한 관형사구와 대명사는 수식어를 제외한다.

예1)

<예문> 별무리처럼 흩어지며 보이던 그것(etri 지침 예시 34쪽)

<보기> [별무리](O), [그것](O)

<오류> [별무리처럼 흩어지며 보이던 그것](X)

<상호참조 태깅> {∅}(예문 내 멘션들 중 상호참조 관계 없음)

예2)

<예문> 나라를 위하여 헌신한 이 사람(etri 지침 예시 34쪽)

<보기> [나라](O), [이 사람](O),

<오류> [나라를 위하여 헌신한 이 사람](X)

<상호참조 태깅> {∅}(예문 내 멘션들 중 상호참조 관계 없음)

2.1.12. 멘션의 중심어가 불용사전에 포함된 경우 추출하지 않는다.

▶ 실제 명사가 아닌 것들(부사류, 형용사류)과 명사지만 부사처럼 사용되는 것들은 제외한다.

▶ [전], [후]가 중심어로 있을 경우 잡지 않는다.

▶ 의성어, 의태어 등이 중심어인 경우 잡지 않는다.

불용사전 목록)

확실히, 아마도, 아마, 도시, 도무지, ~걸, ~수, ~중, ~않기, 제~, 기간 중, 때문에, 중간에, ~사이, ~간, 다음, ~때문, ~인 채, ~이래…, ~기, 또는, 그리고, 신, 그 후, 그 전, 그 가운데, 그 뒤, 그 밖에, ~다음, 그 다음, 그 중간, 아무, 예, 이전, 이후

2.1.13. 중심어인 명사는 다른 멘션과 교차되면 안 된다.

예1)

<예문> 수심 200m 이내의 완만한 경사를(etri지침 예시 34쪽)

<보기> [수심 200m 이내의 완만한 경사](O), [수심 200m 이내](O), [수심](O), [200m](O)

<오류> [이내의 완만한 경사](X)

<상호참조 태깅> {∅}(예문 내 멘션들 중 상호참조관계 없음)

2.1.14. 인용구는 도메인에 따라 처리한다.

▶ ~~QA 도메인 : 한 문장에 발생한 인용구는 인용구 자체를 멘션으로 처리한다.~~

▶ 뉴스 도메인 : 인용구는 무시하고 일반 멘션 태깅하듯 멘션을 추출한다.

→ 뉴스 도메인 규칙을 따른다

2.1.15. 명사절 '-은/는 것', '음/기'도 포함하여 태깅한다.

예1)

<예문> 첼시 피어는 음식물 쓰레기 줄이기, 일회용기 안 쓰기를 실천하고 있다.

<보기> [첼시 피어](O), [음식물 쓰레기 줄이기](O), [음식물 쓰레기](0), [음식물](0), [일회용기 안 쓰기](O), [일회용기](0)

<오류>

<상호참조 태깅> {∅}(예문 내 멘션들 중 상호참조 관계 없음)

2.1.16. 숫자, 날짜 및 수량 표현(지침 추가 사항)

▶ 날짜, 금액, 수치 등 숫자 표현들을 멘션 추출 대상에 포함한다.

예1)　지시하는 날짜, 금액, 수치가 같은 수량 표현의 경우

<예문> 2006년 보다 200만원 하락, …2006년보다 200만원 정도 싼 1500만원 후반대로 결정될 것으로 보인다.표본명: NWRW1800000021-0001(판교 아파트)

<보기> [2006년](O), [하락](O), [200만원](O), [2006년](O), [200만원](O),

[200만원 정도](0),[1500만원](O), [1500만원 후반대](0)

<오류>

<상호참조 태깅>{[2006년], [2006년]}(O), {[200만원], [200만원]}(O)

<예외> 문서 내에 구체적 정보가 없이 대상 또는 시간 명사만 있을 경우 상호참조 태깅을 하지 않는

다.

<예문> 네 오늘은 또 어떤 모범 밥상을…오늘은 또 어떤 요리가 저희를 기다리고…

<보기> [오늘](O), [모범 밥상](O), [오늘](O), [요리](O), [저희](O),

<오류>

<상호참조 태깅> {∅}(예문 내 멘션들 중 상호참조관계 없음

예2) 고정된 숫자값의 경우

<예문> 판교 중대형 분양가 3.3㎡당 1500만원 대, 대우건설 등은 3.3㎡당 분양가를…신청했지만…(후략)표본명: NWRW1800000021-0001(판교 아파트)

<보기> [판교](O), [중대형 분양가](O), [3.3㎡당](O), [1500만원 대](O), [대우건설](0), [대우건설 등](O), [3.3㎡당](O),[3.3㎡당 분양가](O)

<오류>

<상호참조 태깅> {[3.3㎡당], [3.3㎡당]}(O)

예3) 숫자 표현과 문자로 된표현이 같은 것을 지시하는 경우

<예문> 하지만 당초 2008년이던 이전 시기가 차일피일 미뤄지면서 빈집은 절반을 웃돌고 있다… 2004년 7월 기지 이전을 2008년 말까지 끝내기로…(후략)

표본명: NWRW1800000021-0004(동두천)

<보기> [당초2008년이던 이전 시기](0), [당초 2008년](0), [빈집](0), [절반](0),[2004년 7월](0), [기지](0), [기지 이전](0), [2008년 말](0)

<오류>

<상호참조 태깅> {[당초 2008년], [당초 2008년이던 이전 시기]}(0)

2.1.17. 지리 정보(지침 추가 사항)

▶ 지리 정보 전체를 하나의 멘션으로 잡고 그 내부의 개별 지리명을 각각멘션으로 처리한다.

예1)

<예문> 중국 저장성 동부 타이저우시 해안에 있는 다천다오에서

<보기> [중국 저장성 동부 타이저우시 해안에 있는 다천다오](O), [중국 저장성 동부 타이저우시해안](0),[중국 저장성 동부 타이저우시](O), [중국 저장성 동부](O), [중국 저장성](O), [중국](O)

<오류>

<상호참조 태깅> {∅}(예문 내 멘션들 중 상호참조 관계 없음)

예2)

<예문> 서울 종로구 신문로 1가 197금호아시아나 신사옥

<보기> [서울 종로구 신문로 1가 197금호아시아나 신사옥](O), [서울 종로구 신문로 1가 197 금호아시아나](O), [서울 종로구 신문로 1가 197](O), [서울 종로구 신문로 1가](O), [서울 종로구 신문로](O),[서울 종로구](O), [서울](O)

<오류>

<상호참조 태깅> {∅}(예문 내 멘션들 중 상호참조 관계 없음)

2.2. 상호참조해결 태깅 규칙

2.2.1. AND 규칙

▶ 각각 추출된 멘션들을 서로 다른 상호참조 관계로 취급한다.

예1)

<예문> 삼국지의 유비와 관우와 장비는 삼형제이다. 유비가 첫째, 관우가 둘째, 장비가 셋째이다.(etri지침 예시 36쪽 발췌 후 변형)

<보기> [삼국지의 유비와관우와 장비](O), [삼국지](O), [유비](O), [관우](O), [장비](O),[삼형제](O), [유비](O), [첫째](O), [관우](O), [둘째](O),[장비](O), [셋째](O)

<오류> {삼국지의 유비와 관우와 장비는, 유비, 관우, 장비}(X)

<상호참조 태깅> {[삼국지의 유비와 관우와 장비], [삼형제]}(O),{[유비], [유비], [첫째]}(O),{[관우], [관우], [둘째]}(O),{[장비], [장비], [셋째]}(O)

▶ 반점, 온점으로 묶인 명사들도 AND 규칙에 따라 상호참조해결한다.(2.1.8의 예 참조)

2.2.2. OR 규칙

▶ 각각 추출된 멘션들을 서로 같은 상호참조 관계로 취급한다.

예1)

<예문> 운전을 도와주는 장치나 프로그램(etri 지침 예시 36쪽)

<보기> [운전을 도와주는 장치나 프로그램](O), [장치](O), [프로그램](O)

<오류> [운전을 도와주는 장치](X)

<상호참조 태깅> {[운전을 도와주는 장치나 프로그램], [장치], [프로그램]}

2.2.3. 지정사가 포함된 경우

▶ 해당 멘션의 선행사를 찾아 상호참조 한다. 지정사의 명사가 상호참조 되는 경우에 그 명사가 뜻하는 것이 명확할 경우(개체명일 경우)이다.

예1)

<예문> 바이마르는 독일 튀링겐 주에 있는 문화도시이다.(etri지침 예시 37쪽)

<보기> [바이마르](O), [독일](O),[독일 튀링겐 주](O), [독일 튀링겐 주에 있는 문화도시](O)

<오류> [문화도시](X)

<상호참조 태깅> {[바이마르], [독일 튀링겐 주에 있는 문화도시]}

▶ ‘NP1은(는) NP2 이다’와 같은 지정사 구문에서 NP1, NP2가 모두 일반 명사구일 때는 상호참조 태깅을 하지 않는다.

예2)

<예문> 어~ 우리의 집밥은요 세계적으로도 알아주는 모범 밥상이라고 합니다.

<보기> [우리의 집밥](O), [세계적으로도 알아주는 모범 밥상](O

<오류>

<상호참조 태깅> {∅}(예문 내 멘션들 중 상호참조 관계 없음)

2.2.4. 상/하위어 관계

▶ 수식어로 인해 상위어의 범위가 좁혀져 의미가 명확하게 일치하는 상/하위어만 상호참조해결 태깅한다.

예1)

<예문> 사과와 배는 대표적인 추석 과일이다.…(중략)…가을 단풍처럼 붉게 물든 과일은…

<보기> [사과](0), [배](0), [사과와 배](0), [대표적인 추석 과일](0),[추석](0), [가을 단풍처럼 붉게 물든 과일](0), [가을 단풍](0), [가을](0)

<오류>

<상호참조 태깅> {[사과와 배], [대표적인 추석 과일]}(0), {[사과],[가을 단풍처럼 붉게 물든 과일]}(0)

2.2.5. 관형격 조사가 포함된 멘션

▶ 관형격 조사가 포함된 멘션에서 중심어와 관형격 조사가 결합된 명사는 서로 참조하지 않는다.

예1)

<예문> 이 제품의 무게는 230g이다.

<보기> [이 제품](O), [이 제품의 무게](O), [230g](O)

<오류>

<상호참조 태깅> {[이 제품의 무게], [230g]}(O), {[이 제품], [230g]}(X)

<예외> '~의 이름, ~의 뜻, ~의 단어, ~별명, ~라는 이름' 등은 제외

<예문> 이 책의 이름인 반지의 제왕.

<보기> [이 책](O), [이 책의 이름](O), [이 책의 이름인 반지의 제왕](O)

<오류>

<상호참조 태깅>{[이 책], [이 책의 이름], [이 책의 이름인 반지의 제왕}(O)

2.2.6. 교차발생 상호참조해결

▶ 두 개의 상호참조해결 사이에 교차되는 멘션이 있는 경우 상호참조에 대한정의가 필요하다.

예1) *위첨자 숫자는 구분을위함

<예문> 유비1는 촉나라의 황제이다. 유비2의 자는 현덕이다.…(후략)(etri 지침 예시 38쪽 발췌 후 변형)

<보기> [유비1](O), [촉나라의 황제](O), [촉나라](O),[유비2](O), [유비의 자](O), [현덕](O)

<오류>

<상호참조 태깅> {[유비1], [촉나라의 황제], [유비2], [유비의 자], [현덕]}(O)

2.2.7. 동일 어구 반복 상호참조해결(지침 추가 사항)

▶ 동일한 어구로 동일한 대상을 지칭하는 경우 가장 명시적인 상호참조 현상이다.

예1)

<예문> 폴크스바겐은 2018년까지 미국 시장에서의 판매량을 3배로 늘리기로 하고20년 만에 처음으로 미국 공장에 10억 달러를 투자하는 한편 미국 시장을 겨냥한 새로운 모델들을 개발하고 있다…세계 최대 자동차 시장인 미국 시장은 2000년대 초반만 해도 매년 1700만 대가 팔렸으나 지난해에는 1300만 대로 시장이 크게 축소됐다.'

<보기> [폴크스바겐](O), [2018년](O), [미국](O), [미국 시장](O),

[미국시장에서의 판매량](O),[3배](O),[20년](0), [20년 만](O), [미국](O),

[미국 공장](O), [10억 달러](O), [미국](O), [미국 시장](O),

[미국 시장을 겨냥한 새로운 모델들](O), [세계](O), [최대](O), [세계 최대 자동차](O),

[세계 최대 자동차 시장](O), [세계 최대 자동차 시장인 미국 시장](O), [미국](O), [2000년대](O),[2000년대 초반](O), [매년](O), [1700만 대](O), [지난해](O), [1300만 대](O), [시장](O)

<오류>

<상호참조 태깅>{[미국], [미국], [미국], [미국]}(O),

{[미국시장], [미국 시장], [세계 최대 자동차 시장], [세계 최대 자동차 시장인 미국 시장], [시장]}(O)

2.2.8. 줄임말 및 약어(지침 추가 사항)

▶ 동일한 대상을 지칭하는 명사구와 줄임말 혹은 약어 사이에도 상호참조 현상이나타난다.

예1)

<예문> 도널드 트럼프 미국 대통령은…(후략)…트럼프 대통령은…(후략)(자체 생성 예시문)

<보기> [도널드 트럼프](O), [미국](O), [도널드 트럼프 미국 대통령](O), [트럼프](O), [트럼프 대통령](O)

<오류> [미국 대통령은](X)à2.1.4. 중심어중복 제거 규칙 적용

<상호참조 태깅> {[도널드 트럼프], [도널드 트럼프 미국 대통령], [트럼프], [트럼프 대통령]}(O)

2.2.9. 생략된 명사구(지침 추가 사항)

▶ 상호참조 관계의 멘션 중 하나가 생략된 명사구로 나타날 경우에도 상호참조해결처리한다.

예1)

<예문> '노근리 59주년'인권평화캠프 열린다…6.25전쟁 초기에 발생한 '노근리 사건'의 현장인…노근리 사건 발생 59주기를 맞아…(후략)(NWRW1800000021-0189)

<보기> [노근리](O), [노근리 59주년](O), ['노근리 59주년'인권평화캠프](O), [6.25전쟁 초기에 발생한 '노근리 사건'의 현장] [6.25 전쟁](O),[6.25 전쟁 초기](O), [노근리 사건 발생 59주기](O), [노근리 사건 발생](0), [노근리 사건](0)

<오류>

<상호참조 태깅> {[노근리], [노근리 사건],[노근리 사건]}(O)

2.2.10. 기관/단체와 소속자/소속물의 관계(지침 추가 사항)

▶ 특정기관/단체와 소속자 혹은 소속물 사이는 상호참조 대상이 아니다. 단, 대명사 표현으로 인해 동일성이 존재할 경우 상호참조해결한다.

예1)

<예문> 삼성전자는…삼성직원들은…삼성관계자는 “저희는…했습니다”라고 대답했다.

(기존 지침에서 발췌 후변형)

<보기> [삼성전자](O), [삼성직원들](O), [삼성관계자](O), [저희](O)

<오류> {[삼성전자], [삼성직원들]}(X)

<상호참조 태깅> {[삼성전자], [저희]}(O)

2.2.11. 명사(구)와 관형사구(지침 추가 사항)

▶ 명사(구)와 ‘지시관형사(이/그/저)+ 명사’또는 ‘이런/저런+명사’패턴 사이에 상호참조가 일어나기도 한다.

예1) 이/저/그+명사의 예

<예문> 제10호 태풍 크로사가…, 이 태풍은…, 크로사는…(기존 지침에서 발췌 후변형)

<보기> [제10호 태풍 크로사](O),[이 태풍](O), [크로사](O)

<상호참조 태깅> {[제10호 태풍 크로사],[이 태풍], [크로사]}(O)

예2) 이런/저런+명사의 예

<예문> 00일 새벽 0시 규모 9의 지진이…, 이런 규모의 지진은…

(기존 지침에서 발췌 후 변형)

<보기> [00일 새벽 0시 규모 9의 지진](0), [00일 새벽 0시](O), [규모 9](O), [이런 규모의 지진](0), [이런 규모](O)

<상호참조 태깅> {[규모 9], [이런 규모]}(0),{[00일 새벽 0시 규모 9의 지진], [이런 규모의 지진]}(O)

2.2.12. 유의어 및 일반화/구체적 표현

▶ 다음과 같은 유형들은 상호참조 대상으로 본다.

예1) 유의어관계

<예문> 칠레에서 시위가 이어지는 가운데… 칠레의 소요 사태는…(후략)

<보기> [칠레](0), [시위](0),[칠레](0), [칠레의 소요 사태](0)

<상호참조 태깅> {[칠레], [칠레]}, {[시위], [칠레의 소요 사태]}

<예문> 지방자치단체장이 독단적으로 지방의료원을 폐업하지 못하도록 하는 이른바 ‘진주의료원법’…(증략)…‘지방의료원의 설립 및 운영에 관한 법’개정안을 6일 법안심사 소위원회로 넘겼으나…(증략)…일부 새누리당 의원들은 ‘지방자치권 침해’를 이유로 개정안 처리에 반대하고 있는 것으로 전해졌다.

<보기> [지방자체단체장](0), [지방의료원](0), [진주의료원법](0), [지방의료원의 설립 및 운영에 관한 법](0), [지방의료원](0), [지방의료원의 설립 및 운영](0), [‘지방의료원의 설립 및 운영에 관한 법’개정안](0),[6일](0), [법안심사 소위원회](0), [새누리당](0), [일부 새누리당 의원들](0), [지방자치권](0), [지

방자치권 침해] (0), [개정안] (0), [개정안 처리](0)

<상호참조 태깅> {[진주의료원법], ['지방의료원의 설립 및 운영에 관한 법'개정안], [개정안]}(0), {[지방의료원], [지방의료원]}(0)

예2) 일반화및 구체화

<예문> 교통 사고 희생자 수는…(중략)… 2016년부터 지난 해 동안 교통 사고로 사망한 사람들의 수는…(중략)…금년들어서도 이미 154명이 교통 사고로 사망하면서 가장 희생자가 많았던 지난 해 같은 기간의146명을 넘어섰다.

<보기> [교통 사고](0),[교통 사고 희생자](0), [교통 사고 희생자 수](0),[2016년](0), [지난 해](0), [지난 해 동안](0), [교통 사고](0), [교통 사고로 사망한 사람들](0), [교통 사고로 사망한 사람들의 수](0), [금년](0), [154 명](0),

[교통 사고](0), [희생자](0),[가장 희생자가 많았던 지난 해](0), [가장 희생자가 많았던 지난 해 같은 기간](0), [146 명](0)

<상호참조 태깅> {[교통 사고 희생자], [교통 사고로 사망한 사람들], [희생자]}(0)

2.2.13. 일반대상을 지칭하는지 구체적인 특정 대상을 지칭하는지 판단이 어려운 경우

▶ 일반 대상을 가리키는지 특정한 대상을 가리키는지 판단이 어려운 경우에는다음과 같은 기준들에 따라 처리한다.

- 이어진문장들에서 동일한 어구의 형태로 계속 반복적으로 나타나는가?
- 이어진문장들에서 줄임말, 유의어 등을 사용한 동일 대상 지칭 변용 표현들이 나타나는가?
- 텍스트의내용과 밀접히 관련된 핵심적인 명사구인가?

예1)

<예문> 세법 바뀌었는데 내 연금저축 어떡하나 세금 환급액…세법 개정으로 연금저축 연말정산이… 연금저축 가입으로 인해 돌려받는 세금이 절반 이하로 쪼그라들기 때문이다. 반면 …연금저축 가입에 따른 세금 환급액이 더 늘어난다.

<보기> [세법], [내 연금저축], [세금 환급액], [세금], [세법], [세법 개정], [연금저축], [연금저축 연말정산], [연금저축], [연금저축 가입],[연금저축으로 인해 돌려받는 세금], [절반 이하], [연금저축], [연금저축 가입], [연금저축 가입에 따른 세금 환급액]

<상호참조 태깅> {[세법], [세법]}(0),{[연금저축], [연금저축], [연금저축]}(0), {[세금 환급액], [연금저축 가입에 따른 세금 환급액], [연금저축으로 인해 돌려받는 세금]}(O)

사업 책임자　곽용진((주)이르테크 대표이사)
사업 참여자　이석재(연세대학교 언어정보연구원 원장)
최지선((주)이르테크 대리)
윤영민(연세대학교 언어정보연구원 HK교수)
최지명(연세대학교 언어정보연구원 연구원)
담당 연구원　이승재(국립국어원 언어정보과장)
이수미(국립국어원 언어정보과 학예연구사)

발행인: 국립국어원장
발행처: 국립국어원
서울시 강서구 금낭화로 154
전화 02-2669-9718, 전송 02-2669-9727
인쇄일: 2020년 1월 15일
발행일: 2020년 1월 15일
인　쇄: 세종기획

※ 이 책은 국립국어원의 용역비로 수행한 '상호 참조 해결 말뭉치 구축' 사업의 결과물을 발간한 것입니다.