

한국어 중첩 개체명 말뭉치 구축의 실제

KONNE: A Korean Nested Named Entity Corpus

정유남(중앙대) 송영숙(경희대) 유현조(서울대)

2022 세계 한국어 한마당

2022.10. 07. 프레스센터 20층

차례

- ① 연구 목적과 내용
- ② 연구 방법: 중첩 개체명 주석
- ③ 세분류 개체명 말뭉치 구축
- ④ 중첩 개체명 말뭉치 구축
- ⑤ 중첩 개체명 말뭉치 활용

차례

① 연구 목적과 내용

- 연구 목적
- 연구 내용

연구 목적

개체명 인식(NER, Named Entity Recognition)

- 비정형의 자연어 텍스트에서
- 고유명, 시간, 날짜, 수량 표현을 중심으로 특정한 개체를 지시하는 표현을
- 인식하고 해당 표현에 분류 표지 부착(유현조 외 2020:804)

예: 개체명 주석

이 영화는 봉준호 감독의 영화가 아닌 한류붐과 같이 철저하게
CV_ART PS_NAME CV_POSITION CV_ART
PS_NAME

한국을 알리기 위해 정부에서 만든 영화 같이 보이네요.
LCP_COUNTRY OGG_POLITICS CV_ART

개체명 말뭉치의 활용

- 질의응답, 정보 추출, 상호 참조 해결, 개체 연결 등에 이용

연구 내용



- KLUE-NER은 수정과 재배포가 가능한 라이선스
- KLUE-NER의 원시 문장을 주석 대상으로 선택

Source	Train	Dev	Total
WIKITREE	11,435	3,534	13,969
NSMC	9,573	2,466	12,039
Overall	21,008	5,000	26,008

문장수 (참조: <https://klue-benchmark.com/tasks/69/data/description>)

- KONEC (Korean Named Entity Corpus)
 - KLUE-NER의 원시 문장을 대상
 - 150개 세분류 개체명 주석
 - 국립국어원 2021년 개체명 주석 지침 버전 2.1을 따름
-
- KONNE (Korean Nested Named Entity Corpus)
 - KONEC 세분류 개체명 주석을 바탕으로
 - 개체명 중첩 구조를 주석
 - 국립국어원 개체명 지침을 바탕으로 중첩 주석으로 확장

② 연구 방법: 중첩 개체명 주석

- 중첩 개체명 주석
- 기존 개체명 말뭉치
- 개체명 분류 체계

중첩 개체명 주석 (1)

중첩 개체명(NNE, Nested Named Entity) 주석

- 개체명 표현의 구조적 분석
- 개체명 표현 내부의 성분 개체명 사이의 통사 의미적 연관성 확인 가능
- 자연어처리 활용에 유용함

‘nested’와 ‘overlapping’

- 본 연구의 ‘중첩’은 한 개체명 안에 다른 개체명이 ‘내포(nested)’되는 구조를 의미
- 서로 다른 두 개체명이 부분적으로 ‘겹치는(overlapping)’ 분석을 하는 시도도 존재

예: nested vs. overlapping

① nested

한국대사관 측은 밝혔다.

LCP_COUNTRY

OGG_POLITICS

② overlapping

송재훈 삼성서울병원장은

OGG_MEDICINE

CV_OCCUPATION

중첩 개체명 주석 (2)

예: 기존 개체명 주석

<u>노벨</u>	<u>경제학상</u>	<u>수상자</u>	<u>조지</u>	<u>에커로프</u>	<u>교수다.</u>
	CV_PRIZE	CV_POSITION		PS_NAME	CV_OCCUPATION

예: 중첩 개체명 주석

<u>노벨</u>	<u>경제학상</u>	<u>수상자</u>	<u>조지</u>	<u>에커로프</u>	<u>교수다.</u>
PS_NAME	FD_SOCIAL_SCIENCE	CV_POSITION	PS_NAME		CV_OCCUPATION
	CV_PRIZE		PS_NAME		
	CV_POSITION				
			PS_NAME		

예: 중첩 개체명 주석의 유용성

- 1 <윤석열:PS_NAME> <대통령:CV_POSITION>
- 2 <2002년:DT_YEAR> <한일월드컵:EV_SPORTS>
- 3 <서울:LC_CAPITALCITY> <남산타워:AF_BUILDING>

기존 개체명 말뭉치 (1)

한국어 개체명 말뭉치 구축 동향

- 국립국어원, AI Hub, KLUE 등에서 주석 지침 개발과 말뭉치 구축이 이루어짐
- 개체명 분류 표지는 대분류 6~15개에서 세분류 150개로 상세화됨
- 원시 문장으로는 신문 기사 문어, 학술 및 법률 문어, 구어, 웹 언어 등이 주석 대상
- 중첩 개체명 주석 말뭉치는 공개된 사례 없음

말뭉치	자료	태그 개수	크기
2021 국어원 개체명 말뭉치 분석 말뭉치구축	2019년 대분류 개체명 분석 말뭉치 세분화(구어 200, 문어100) , 2021년 신규 추출(문어 100만, 구어 200만 어절) IT과학, 문화, 건강 분야	150	600백만 어절
KLUE-NER	위키�트리(16654), NSMC(14354)	6	31008건
AI Hub 전문분야 말뭉치	학술, 법령, 판결문, 특허	15	150만 건
2020 국어원 개체명 분석 말뭉치 구축	웹언어	150	약 500만 어절
2019 국어원 개체명 분석 말뭉치구축	문어(100만) , 구어(200만)	15	300백만 어절
2018 국어원 개체명 말뭉치 분석 말뭉치구축	문어, 구어	10	

기존 개체명 말뭉치 (2)

해외 중첩 개체명 말뭉치 구축 사례

- Genia(J-D Kim et al., 2003): 영어 의학 논문 초록. 전문 용어 의미 주석.
- ACE-2004(Doddington et al., 2004): 영어, 아랍어, 중국어. 개체, 관계 주석.
- NNE(Ringland et al., 2019): 영어 신문기사. 중첩 개체명 주석.

예: NNE Annotation Guidelines

- ① Courtaulds Chairman and Chief Executive Sir Christopher Hogg
- NAME ROLE ROLE HON FIRST NAME
- ORGCORP ROLE PER
- ROLE
- PER
- ② the 2000 Sydney Olympics
- YEAR CITY
- SPORTS-EVENT
- ③ Rosencrantz and Guildenstern Are Dead
- NAME NAME
- NAME
- PLAY

개체명 분류 체계 (1) 15개 대분류

개체명 분류 체계: 15개 대분류와 150개 세분류

- 국립국어원 개체명 주석 지침 버전 2.1의 15개 대분류, 150개 세분류를 따름
- 고유명인 인물, 장소, 조직과 날짜 및 시간표현, 수량표현을 핵심으로
- 사건, 제품, 동식물 등의 명칭과 다양한 전문 영역의 용어들을 포함

태그	설명	태그	설명	태그	설명
PS	인물(Person)	QT	수량(Quantity)	AM	동물(Animal)
LC	장소(Location)	CV	문명(Civilization)	PT	식물(Plant)
OG	조직(Organization)	EV	사건(Event)	MT	물질(Material)
DT	날짜(Date)	AF	인공물(Artifact)	TR	이론(Theory)
TI	시간(Time)	TM	용어(Term)	FD	학문분야(Field)

개체명 분류 체계 (2) 150개 세분류

1. PERSON(PS)	PS_NAME	5. ORGANIZATION(OG)	OGG_ECONOMY OGG_EDUCATION OGG_MILITARY OGG_MEDIA OGG_SPORTS OGG_ART OGG_MEDICINE OGG_RELIGION OGG_SCIENCE OGG_LIBRARY OGG_LAW OGG_POLITICS OGG_FOOD OGG_HOTEL OGG_OTHERS	7. CIVILIZATION(CV)	CV_CULTURE CV_TRIBE CV_LANGUAGE CV_POLICY CV_LAW CV_CURRENCY CV_TAX CV_FUNDS CV_ART CV_SPORTS CV_SPORTS_POSITION CV_SPORTS_INST CV_PRIZE CV_RELATION CV_OCCUPATION CV_POSITION CV_FOOD CV_DRINK CV_FOOD_STYLE CV_CLOTHING CV_BUILDING_TYPE	9. TIME(TI)	TI_DURATION TI_HOUR TI_MINUTE TI_SECOND TI_OTHERS	12. ANIMAL(AM)	AM_INSECT AM_BIRD AM_FISH AM_MAMMALIA AM_AMPHIBIA AM_REPTILIA AM_TYPE AM_PART AM_OTHERS
	PS_CHARACTER PS_PET								
2. STUDY_FIELD(FD)	FD_SCIENCE FD_SOCIAL_SCIENCE FD_MEDICINE FD_ART FD_HUMANITIES FD_OTHERS	6. LOCATION(LC)	LCP_COUNTRY LCP_PROVINCE LCP_COUNTY LCP_CITY LCP_CAPITALCITY LCG_RIVER LCG_OCEAN LCG_BAY LCG_MOUNTAIN LCG_ISLAND LCG_CONTINENT LC_SPACE LC_OTHERS	8. DATE(DT)	DT_DURATION DT_DAY DT_WEEK DT_MONTH DT_YEAR DT_SEASON DT_GEOAGE DT_DYNASTY DT_OTHERS	10. QUANTITY(QT)	QT_AGE QT_SIZE QT_LENGTH QT_COUNT QT_MAN_COUNT QT_WEIGHT QT_PERCENTAGE QT_SPEED QT_TEMPERATURE QT_VOLUME QT_ORDER QT_PRICE QT_PHONE QT_SPORTS QT_CHANNEL QT_ALBUM QT_ADDRESS QT_OTHERS	13. PLANT(PT)	PT_FRUIT PT_FLOWER PT_TREE PT_GRASS PT_TYPE PT_PART PT_OTHERS
	TR_SCIENCE TR_SOCIAL_SCIENCE TR_MEDICINE TR_ART TR_HUMANITIES TR_OTHERS								
4. ARTIFACTS(AF)	AF_BUILDING AF_CULTURAL_ASSET AF_ROAD AF_TRANSPORT AF_MUSICAL_INSTRUMENT AF_WEAPON AFA_DOCUMENT AFA_PERFORMANCE AFA_VIDEO AFA_ART_CRAFT AFA_MUSIC AFW_SERVICE_PRODUCTS AFW_OTHER_PRODUCTS					11. EVENT(EV)	EV_ACTIVITY EV_WAR_REVOLUTION EV_SPORTS EV_FESTIVAL EV_OTHERS	14. MATERIAL(MT)	MT_ELEMENT MT_METAL MT_ROCK MT_CHEMICAL
								15. TERM(TM)	TM_COLOR TM_DIRECTION TM_CLIMATE TM_SHAPE TM_CELL_TISSUE_ORGAN TMM_DISEASE TMM_DRUG TMI_HW TMI_SW TMI_SITE TMI_EMAIL TMI_MODEL TMI_SERVICE TMI_PROJECT TMIG_GENRE TM_SPORTS

③ 세분류 개체명 말뭉치 구축

- 세분류 개체명 주석 방법
- 세분류 개체명 말뭉치 구축
- 개체명 중의성
- 개체명 주석 통계

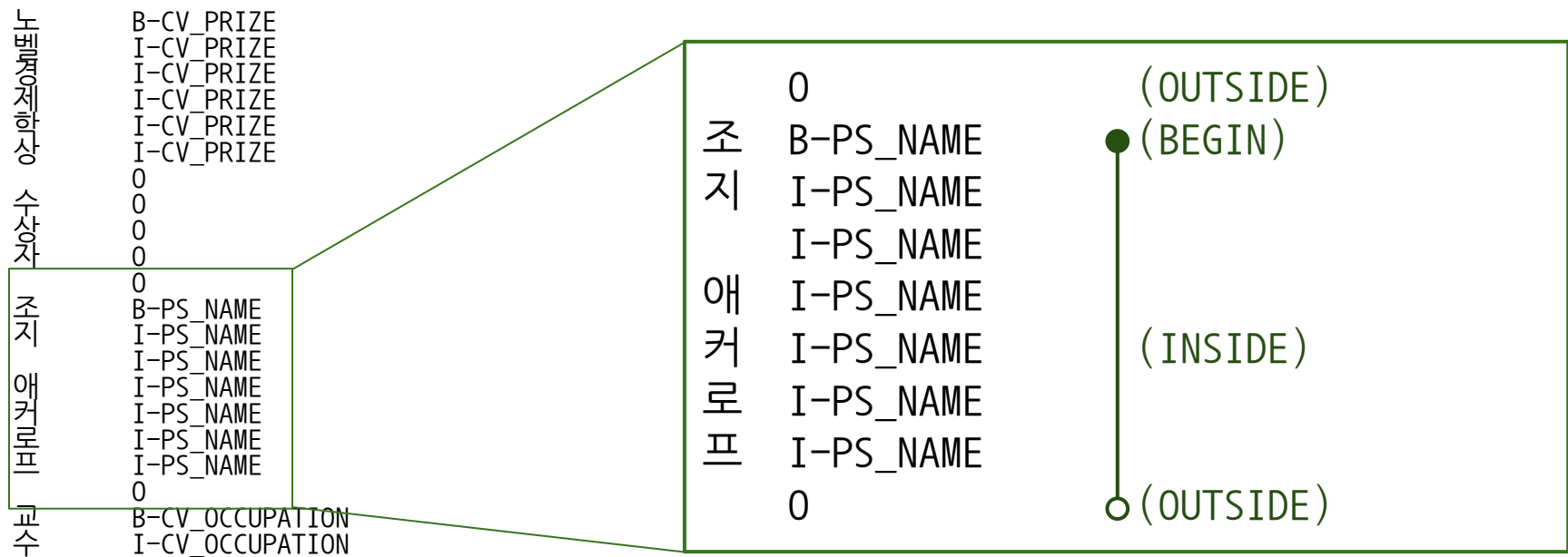
세분류 개체명 주석 방법

세분류 개체명 주석 방법과 도구

- 국립국어원 2021 개체명 주석 지침 버전 2.1을 기준으로 주석.
- KLUE-NER의 BIO 파일 형식을 수용. KLUE-NER과 호환 가능한 형식으로 배포.

예: 개체명 주석 BIO 형식

〈노벨경제학상:CV_PRIZE〉 수상자 〈조지 애커로프:PS_NAME〉 〈교수:CV_OCCUPATION〉



세분류 개체명 말뭉치 구축

세분류 개체명 말뭉치 구축 과정

- 원시 말뭉치 정비
- 국립국어원 개체명 말뭉치로 학습한 BERT 기반 개체명 인식기로 자동 태깅
- 형식적 오류 수정: 유니코드 문자로 인한 오류, BIO 형식 오류 수정
- KLUE-NER의 6개 대분류 주석과 대조하여 자동 태깅에서 누락된 개체명 추가
- 개체명 문맥색인을 이용한 일관성 검토 및 수정

예: 문맥색인

8359		박범계	PS_NAME		, 정청래, 김현 등 민주당 의원 3명은 '국회에서의 증언
8360		태조 이성계	PS_NAME		역을 맡았던 김무생의 압도적인 카리스마가 돋보였던
8361	연 대하드라마, 당시 이방원역을 맡았던 유동근이 15년이 지나	이성계	PS_NAME		역을 한다는데 기대되긴한다.
8362	날 '택시'는 디자이너 황재근과 댄스 마스터 박지우가 출연한 '	직업의 세계	AFA_VIDEO		2탄으로 꾸며졌다.
8363	월드디제이페스티벌'은 일렉트로닉 음악을 중심으로 록과 힙합,	제3세계	QT_ORDER		음악까지 장르를 불문한 국내외 뮤지션들의 파격 퍼포
8364		신세계	AFA_VIDEO		재밌게 보고 기대하고 봤는데 너무..... 박감독 2년새
8365	옛날에 1만 봤을때 재미있었는데...	신세계	OGG_ECONOMY	AFA_VIDEO	보고 3개 다보니까 별로다
8366	박성웅	신세계	AFA_VIDEO		때 황정민이 부러웠나보네
8367		신세계	OGG_ECONOMY	AFA_VIDEO	에서의 황정민은 여기서 시작됐다고 생각한다
8368	긴 러닝타임에 지루하지 않은 편집.	신세계	OGG_ECONOMY	AFA_VIDEO	의 미국판이 아닌 도니브래스코의 한국판이 신세계
8369	않은 편집. 신세계의 미국판이 아닌 도니브래스코의 한국판이	신세계	OGG_ECONOMY	AFA_VIDEO	
8370		신세계	AFA_VIDEO		는 황정민의 영화지만 무간도는 300분이 넘는 시간 온
8371	0년도 더 지난 영화인데 정말 기가 막히네 최고라는 말밖에는.	신세계	OGG_ECONOMY	AFA_VIDEO	는 그저 아류일뿐이었다

개체명 중의성 (1)

개체명 주석과 중의성

- 개체명 인식에서 중의성 문제는 핵심적인 문제 중의 하나
- 중의적 개체명 오류는 인간 주석과 자동 태깅에서 모두 빈번하게 나타남
- 개체명 판별 오류, 분류 오류, 경계 오류의 원인이 됨
- 국립국어원 지침 근거, 우선 적용 범주, 서술어 호응에 의해 판단 가능
- 개체명 문맥색인을 이용하여 일관성 검토 및 주석자간 교차 검수 필요
- 문맥에 따라 장소/조직으로 분류되는 개체명들의 비중이 높은 편 (①②)
- 세분류 체계로 상세화됨에 따라 중의성 해소의 난이도가 높아짐 (③)

예: 문맥에 따른 개체명 중의성 해소

- ① <한국:LCP_COUNTRY>에서 공연을 볼 수 있는 마지막 기회
- ② <한국:OGG_POLITICS>은 개발목표 달성을 위해 적극 기여할 것입니다
- ③ <한국:OGG_SPORTS>과 브라질의 축구 경기

개체명 중의성 (2)

개체명 판별 오류와 분류 오류

- 문맥에 따라 동일 언어 형태의 개체명 여부가 달라짐. 개체명 판별 오류의 원인.
- 문맥에 따라 동일 개체명 표현의 분류가 달라짐. 개체명 분류 오류의 원인.

예: 개체명 판별 오류

- ① 모든 장면 모든 대사 그 <하나하나:QT_COUNT>가 전부 소중하게 느껴지는 영화
- ② 블랙스완 <한번:QT_COUNT>도 눈을 떴지 않고 봤다.
- ③ 감동과 재미도 있고 <한마디로:QT_COUNT>로 짝입니다.

예: 개체명 분류 오류

- ① <송새벽:TI_DURATION>은 거품인듯... 주인공은 아닌거같아요
- ② 나는야.. <멕 라이언:PS_CHARACTER> 누나 광팬... 허나..
- ③ <헤이든 처치:LC_OTHERS>, <도널드 서덜랜드:LC_OTHERS>도 마찬가지.
- ④ 영화 보는동안 97년작 <볼케이노:CV_FOOD>가 훨씬 낫다고 내내 생각했다
- ⑤ <50센트:QT_PRICE>는 경찰이 도착하기 전 현장을 떠났으며

개체명 중의성 (3)

개체명 경계 오류

- 개체명은 한 어절의 일부일 수도 있고 여러 어절에 걸쳐 있을 수도 있음.
- 개체명 표현이 아닌 것을 개체명의 내부에 포함하는 경계 오류 (❶)
- 개체명의 범위(span)을 잘못 탐지하는 것은 분류 오류를 동반하기도 함 (❷)
- 개체명의 일부만 인식하고 일부를 개체명 표현 아닌 것으로 인식하는 오류 (❸)

예: 개체명 경계 오류와 분류 오류

- ❶ 이들은 <이어 **모라:PS**>가 쓰러지자 구덩이를 파 매장하고 풀더미 등으로 덮어 ...
- ❷ 2015년에 개봉한 영화보다 10년전 <이**영화:PS**>가 훨씬 나음
- ❸ <**홍콩:LC**> 금융 <**관리국:OG**>(HKMA)는 시위대가 점거한 지역에 있는 17개 은행의 29지점이 일시적 휴업에 들어갔다고 밝혔다.

개체명 중의성 (4)

예: 국가명(LCP_COUNTRY) - 조직명(OGG_POLITICS)

- ① 오바마 <미국:LCP_COUNTRY> 대통령이 2기 행정부
- ② 핵 관련 언급 역시 타깃은 <한국:LCP_COUNTRY>이 아닌 <미국:LCP_COUNTRY>과 <중국:LCP_COUNTRY>일 가능성이 크다.
- ③ 이날 회담은 <한국:OGG_POLITICS>과 <미국:OGG_POLITICS>의 새로운 정부 간 정책협력의 기본 틀

예: 건물명(AF_BUILDING), 조직명(OGG_POLITICS)

- ① <청와대:AF_BUILDING>와 동십자각이 있는 4번 출입구에선 대한민국 대통령사(史)를 소개하고, 경복궁과 국립고궁박물관이 있는 5번 출입구에는 왕궁 건축과 혼례 등 역사를 담는다.
- ② <청와대:OGG_POLITICS> 페이스북이 공개한 박 대통령과 빌 게이츠의 대화 내용입니다.

개체명 중의성 (5)

예: 스포츠 수량 표현(QT_SPORTS) - 스포츠 사건명(EV_SPORTS)

- ① 브라질 월드컵에서 네덜란드가 칠레를 꺾고 B조 1위로 <16강:QT_SPORTS>에 올랐다.
- ② 특히 북한과 치른 <8강전:EV_SPORTS>에서 혼자서 <4골:QT_SPORTS>을 쏟아내며 팀의 <5-3:QT_SPORTS> 승리를 이끈 것은 에우제비우 최고의 경기로 팬들에게 기억되고 있다.

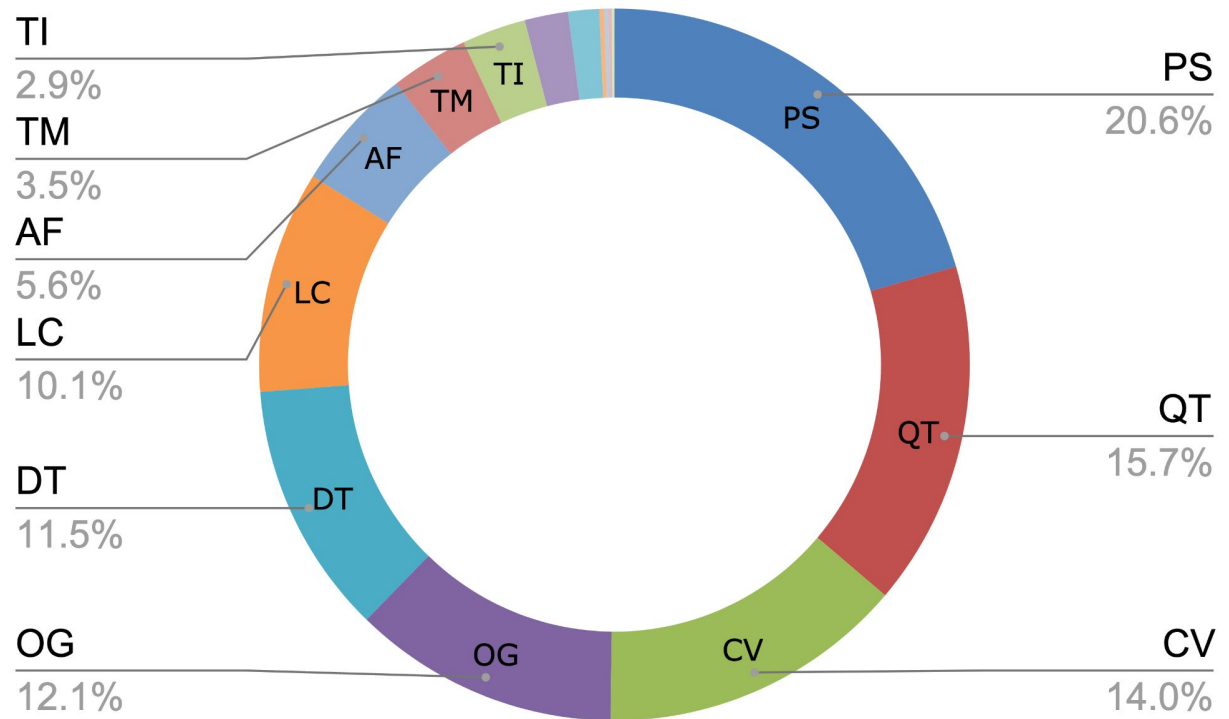
예: 순서/순차 수량 표현(QT_ORDER) - 기간(DT_DURATION)

- ① 스티브 잡스 <3주기:QT_ORDER>를 맞아 미국 경제매체 ‘비즈니스 인사이더’가 그를 재조명하는 보도를 내놨다.
- ② 이 블랙홀은 ‘빅뱅’ 이후 <9억 년:DT_DURATION>이 지난 시점에 만들어진 것으로 추정된다고 WP는 전했다.

개체명 주석 통계 (1): 개체명 대분류 표지 빈도

대분류	빈도
PS	18449
QT	14056
CV	12525
OG	10874
DT	10311
LC	9055
AF	5033
TM	3180
TI	2615
EV	1773
AM	1268
TR	200
MT	161
FD	140
PT	113
합계	89753

문장	26008
개체명 토큰	89753
개체명 타입	30250



개체명 주석 통계 (2): 고빈도 개체명 표현

형태	세분류	빈도
이날	DT_DAY	808
경찰	OGG_POLITICS	734
배우	CV_OCCUPATION	720
A	PS_NAME	594
감독	CV_POSITION	563
한국	LCP_COUNTRY	556
김	PS_NAME	515
미국	LCP_COUNTRY	510
서울	LCP_CAPITALCITY	472
대통령	CV_POSITION	456
10점	QT_OTHERS	398
일본	LCP_COUNTRY	380
중국	LCP_COUNTRY	354
박	PS_NAME	345
의원	CV_POSITION	326
이	PS_NAME	316
정부	OGG_POLITICS	305
1점	QT_OTHERS	301
하나	QT_COUNT	300
검찰	OGG_POLITICS	289

형태	세분류	빈도
북한	LCP_COUNTRY	230
1편	QT_ORDER	216
오후	TI_DURATION	191
대표	CV_POSITION	188
가수	CV_OCCUPATION	182
오전	TI_DURATION	179
가족	CV_RELATION	177
회장	CV_POSITION	175
B	PS_NAME	173
학생	CV_OCCUPATION	171
오늘	DT_DAY	160
새누리당	OGG_POLITICS	153
국회	OGG_POLITICS	153
아들	CV_RELATION	152
선수	CV_OCCUPATION	146
앞	TM_DIRECTION	138
눈	AM_PART	134
MBC	OGG_MEDIA	132
트위터	TMI_SERVICE	130
지난해	DT_YEAR	125

형태	세분류	빈도
딸	CV_RELATION	124
올해	DT_YEAR	123
영국	LCP_COUNTRY	122
징역	CV_LAW	122
2명	QT_MAN_COUNT	117
박근혜	PS_NAME	116
러시아	LCP_COUNTRY	115
한번	QT_COUNT	113
북한	OGG_POLITICS	113
3명	QT_MAN_COUNT	112
윤	PS_NAME	111
민주당	OGG_POLITICS	110
전날	DT_DAY	109
2편	QT_ORDER	109
1위	QT_ORDER	107
얼굴	AM_PART	106
17일	DT_DAY	106
최	PS_NAME	105
한국영화	CV_ART	102
정	PS_NAME	102

개체명 주석 통계 (3): 개체명 세분류 표지 빈도

세분류	빈도	세분류	빈도	세분류	빈도	세분류	빈도
PS_NAME	17466	AF_TRANSPORT	1023	DT_MONTH	438	CV_DRINK	169
OGG_POLITICS	4492	AM_PART	1011	OGG_EDUCATION	428	AM_MAMMALIA	165
DT_DAY	4356	QT_PRICE	999	OGG_MILITARY	422	LCG_CONTINENT	154
CV_POSITION	4290	OGG_SPORTS	970	OGG_LAW	400	AF_ROAD	154
LCP_COUNTRY	3705	PS_CHARACTER	969	QT_SPORTS	372	OGG_MEDICINE	143
CV_OCCUPATION	3329	LCP_PROVINCE	936	CV_FOOD	360	EV_FESTIVAL	139
QT_COUNT	3166	LCP_COUNTY	919	TMM_DISEASE	341	CV_SPORTS_POSITION	130
QT_ORDER	2633	LC_OTHERS	905	TM_SPORTS	339	LCG_ISLAND	124
AFA_VIDEO	2287	EV_OTHERS	826	TI_HOUR	331	MT_CHEMICAL	114
QT_MAN_COUNT	2224	CV_LAW	763	AFW_OTHER_PRODUCTS	292	TM_CELL_TISSUE_ORGAN	113
QT_OTHERS	1946	LCP_CAPITALCITY	738	QT_LENGTH	286	TM_COLOR	112
OGG_ECONOMY	1897	TMI_HW	733	AF_WEAPON	282	AF_CULTURAL_ASSET	97
DT_DURATION	1879	QT_PERCENTAGE	730	CV_SPORTS	233	CV_POLICY	94
DT_YEAR	1758	EV_SPORTS	720	CV_TRIBE	214	OGG_RELIGION	92
DT_OTHERS	1712	TMI_SERVICE	699	CV_CLOTHING	193	LCG_OCEAN	85
TI_DURATION	1703	CV_ART	674	OGG_FOOD	181	DT_SEASON	83
CV_RELATION	1679	TM_DIRECTION	619	AFA_DOCUMENT	181	TR_MEDICINE	82
QT_AGE	1334	TI_OTHERS	577	AFA_MUSIC	180	TR_SCIENCE	79
LCP_CITY	1262	OGG_OTHERS	531	QT_TEMPERATURE	178	LCG_MOUNTAIN	76
OGG_MEDIA	1194	AF_BUILDING	447	CV_PRIZE	175	TMI_SW	69

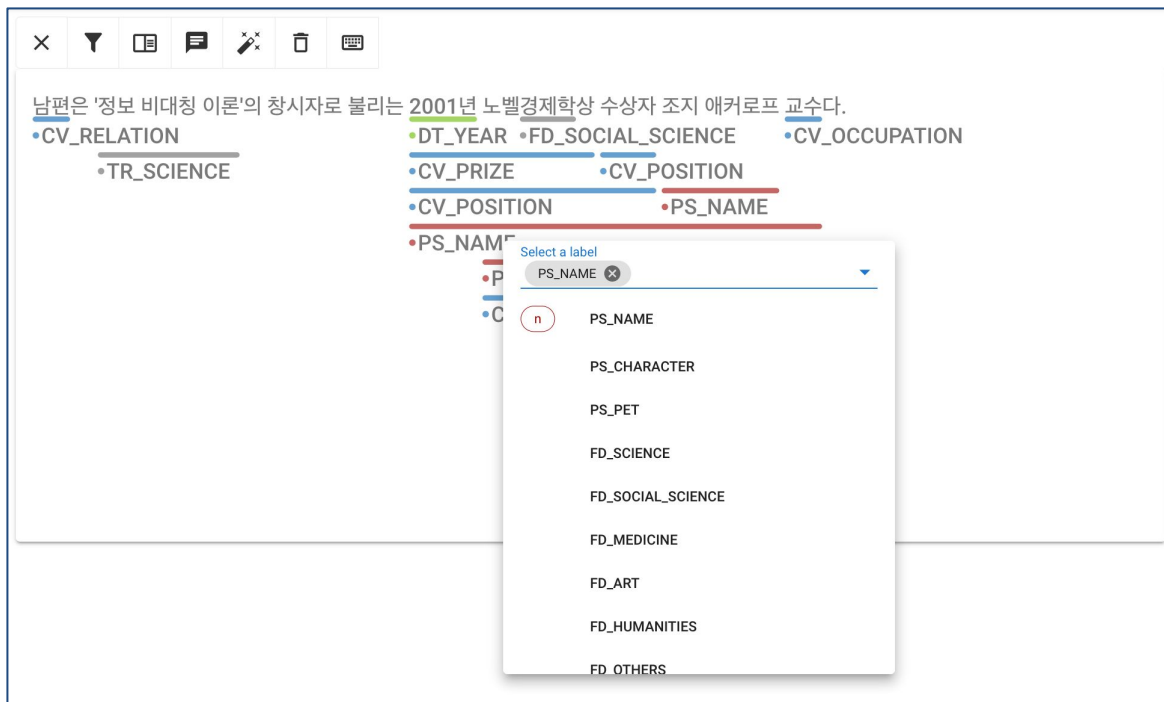
④ 중첩 개체명 말뭉치 구축

- 중첩 개체명 말뭉치 구축 방법과 도구
- 중첩 개체명 주석의 원칙
- 주요 중첩 개체명 주석의 실제

중첩 개체명 말뭉치 구축 방법과 도구 (1)

doccano: 오픈소스 텍스트 주석 도구

- 서버 설치형 웹기반 주석 도구로 온라인 실시간 협업 가능
- 개체명 주석을 위해 필요한 주석 방법인 'sequence labeling' 지원
- 중첩 개체명 주석을 위해 필요한 'overlapping entity' 허락



남편은 '정보 비대칭 이론'의 창시자로 불리는 2001년 노벨경제학상 수상자 조지 애커로프 교수다.

•CV_RELATION •DT_YEAR •FD_SOCIAL_SCIENCE •CV_OCCUPATION
•TR_SCIENCE •CV_PRIZE •CV_POSITION
•CV_POSITION •PS_NAME
•PS_NAME

Select a label

PS_NAME x

n PS_NAME
PS_CHARACTER
PS_PET
FD_SCIENCE
FD_SOCIAL_SCIENCE
FD_MEDICINE
FD_ART
FD_HUMANITIES
FD_OTHERS



<https://github.com/doccano/doccano>

중첩 개체명 말뭉치 구축 방법과 도구 (2)

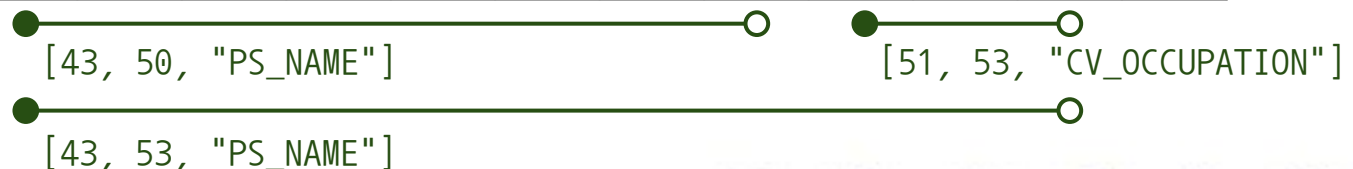
말뭉치 파일 형식

- 말뭉치 파일은 JSONL과 JSON 두 가지로 형식으로 공개
- JSONL 말뭉치: 행 단위로 구분하여 간결하고 압축적인 형식으로 정보를 제공
- JSON 말뭉치: 명시적이고 상세하나 다소 잉여적인 방식으로 정보를 제공

예: JSONL 파일 형식 (doccano 스타일)

```
{"id": 20488, "klue_id": "klue-ner-v1_train_20487_wikitree", "text": "남편은 '정보 비대칭 이론'의 창시자로 불리는 2001년 노벨경제학상 수상자 조지 애커로프 교수다.", "label": [[0, 2, "CV_RELATION"], [5, 14, "TR_SCIENCE"], [26, 31, "DT_YEAR"], [26, 38, "CV_PRIZE"], [26, 42, "CV_POSITION"], [26, 53, "PS_NAME"], [32, 34, "PS_NAME"], [32, 38, "CV_PRIZE"], [34, 37, "FD_SOCIAL_SCIENCE"], [39, 42, "CV_POSITION"], [43, 50, "PS_NAME"], [43, 53, "PS_NAME"], [51, 53, "CV_OCCUPATION"]]}
```

index	...	42	43	44	45	46	47	48	49	50	51	52	53	54
token	...		조	지		애	커	로	프		교	수	다	.



중첩 개체명 말뭉치 구축 방법과 도구 (3)

예: JSON 파일 형식 (국립국어원 개체명 말뭉치 스키마 준수)

```
{
  "id": 20488,
  "klue_id": "klue-ner-v1_train_20487_wikitree",
  "form": "남편은 '정보 비대칭 이론'의 창시자로 불리는 2001년 노벨경제학상 수상자 조지 애커로프 교수다.",
  "word": [
    {"id": 1, "form": "남편은", "begin": 0, "end": 3},
    {"id": 2, "form": "'정보", "begin": 4, "end": 7},
    {"id": 3, "form": "비대칭", "begin": 8, "end": 11},
    {"id": 4, "form": "이론'의", "begin": 12, "end": 16},
    {"id": 5, "form": "창시자로", "begin": 17, "end": 21},
    {"id": 6, "form": "불리는", "begin": 22, "end": 25},
    {"id": 7, "form": "2001년", "begin": 26, "end": 31},
    {"id": 8, "form": "노벨경제학상", "begin": 32, "end": 38},
    {"id": 9, "form": "수상자", "begin": 39, "end": 42},
    {"id": 10, "form": "조지", "begin": 43, "end": 45},
    {"id": 11, "form": "애커로프", "begin": 46, "end": 50},
    {"id": 12, "form": "교수다.", "begin": 51, "end": 55}
  ],
  "NE": [
    {"id": 1, "form": "남편", "label": "CV_RELATION", "begin": 0, "end": 2},
    {"id": 2, "form": "정보 비대칭 이론", "label": "TR_SCIENCE", "begin": 5, "end": 14},
    {"id": 3, "form": "2001년", "label": "DT_YEAR", "begin": 26, "end": 31},
    {"id": 4, "form": "2001년 노벨경제학상", "label": "CV_PRIZE", "begin": 26, "end": 38},
    {"id": 5, "form": "2001년 노벨경제학상 수상자", "label": "CV_POSITION", "begin": 26, "end": 42},
    {"id": 6, "form": "2001년 노벨경제학상 수상자 조지 애커로프 교수", "label": "PS_NAME", "begin": 26, "end": 53},
    {"id": 7, "form": "노벨", "label": "PS_NAME", "begin": 32, "end": 34},
    {"id": 8, "form": "노벨경제학상", "label": "CV_PRIZE", "begin": 32, "end": 38},
    {"id": 9, "form": "경제학", "label": "FD_SOCIAL_SCIENCE", "begin": 34, "end": 37},
    {"id": 10, "form": "수상자", "label": "CV_POSITION", "begin": 39, "end": 42},
    {"id": 11, "form": "조지 애커로프", "label": "PS_NAME", "begin": 43, "end": 50},
    {"id": 12, "form": "조지 애커로프 교수", "label": "PS_NAME", "begin": 43, "end": 53},
    {"id": 13, "form": "교수", "label": "CV_OCCUPATION", "begin": 51, "end": 53}
  ]
}
```


중첩 개체명 주석의 원칙 (2)

중첩 주석의 제한

- 기능적 요소(❶)와 통사적 구조(❷)를 개체명 내부에 포함하지 않음
- 둘 이상의 외연을 지칭하는 표현을 묶어 하나의 개체명으로 주석하지 않음 (❸)
- 다중 주석(❹)과 겹침 주석(❺)은 허용하지 않음

예: 중첩 주석의 제한

- ❶ <노벨상 수상자:CV_POSITON>인 <오무라 사토시 교수:PS_NAME>
- ❷ 여성으로는 처음 건축계 <노벨상:CV_PRIZE>으로 불리는 <프리츠커상:CV_PRIZE>을 받은 <자하 하디드:PS_NAME>
- ❸ 연제구는 <4월 1일부터 3일까지:DT_DURATION><온천천 시민공원:LC_OTHERS><세병교:AF_BUILDING>와 <연안교:AF_BUILDING>사이에서 연제 한마당축제를 연다.
- ❹ <<스파이더맨:PS_CHARACTER>:AFA_VIDEO>
- ❺ <경기<도:LCP_PROVINCE>지사:CV_POSITON>:CV_POSITION>

중첩 개체명 주석의 원칙 (3)

개체명의 내부 요소

- 개체명의 내부가 성분 개체명의 연쇄로 이루어지는 경우가 전형적임 (①)
- 개체명 표현 내부에 개체명이 아닌 요소 일부 포함 (②③)
- 특정 인물 정보와 관련되어 직업, 나이, 소속 등의 괄호 표기도 포함 (④)

예: 개체명의 내부 요소

- ① <<프랑스:LC_COUNTRY><사회학자:CV_OCCUPATION> <다비드 르브르통:PS_NAME>:PS_NAME>
- ② 한 다가구주택에서는 <57살 안모:PS_NAME> 씨와 <아내 55살 이모:PS_NAME> 씨가 안방에서 숨진 채 발견되었다.
- ③ <<오바마:PS_NAME> 케어:CV_POLICY>
- ④ <강동희(47) 감독:PS_NAME>, <배우 노현희(42):PS_NAME> 씨, <가수 싸이(본명 박재상·35):PS_NAME>

주요 중첩 개체명 주석의 실제

개체명	세분류
노벨경제학상 수상자 조지 애커로프 교수 이광종 감독 조 교육감	PS_NAME PS_NAME PS_NAME
주리비아 한국대사관 한국과학기술원(KAIST) 인공위성연구센터 2014 브라질 월드컵 한국 국가대표팀	OGG_POLITICS OGG_SCIENCE OGG_SPORTS
인천 문학경기장 미국 캘리포니아주 로스앤젤레스 다저스타디움	AF_BUILDING AF_BUILDING
로미오와 줄리엣 CBS 라디오 '김현정의 뉴스쇼'	AFA_VIDEO AFA_VIDEO
2014 인천 아시안게임 2014 국제축구연맹(FIFA) 브라질월드컵 조별리그 H조 2차전	EV_SPORTS EV_SPORTS

인물명(PS)

인물(PS)

- 인물명(PS_NAME)은 내부 성분, 인물 관련된 소속, 나이, 직업
- 상위 개체명 분류가 중첩 개체명의 최종 의미 정보 결정함

예: PS_NAME

- ① 노벨 경제학상 수상자 조지 에커로프 교수다.
PS_NAME FD_SOCIAL_SCIENCE CV_POSITION PS_NAME CV_OCCUPATION
CV_PRIZE PS_NAME
CV_POSITION PS_NAME
- ② 지난 2001년에 오바마 케어의 개인 의무가입 조항이
DT_YEAR PS_NAME
CV_POLICY

조직명(OG)

조직(OG)

- 조직명(OGG_POLITICS)은 기관명, 단체명, 하위 세부 조직명

예: OGG_POLITICS

① 16일 주리비아 한국대사관과 AP통신에 따르면
LCP_COUNTRY LCP_COUNTRY
OGG_POLITICS
OGG_POLITICS

중첩 구조 분석으로 풍부한 정보 제공 가능

- ② <고려대학교:OGG_EDUCATION> 체육교육학과
- ③ <<고려대학교:OGG_EDUCATION> 체육교육학과:OGG_EDUCATION>

인공물(AF)

인공물(AF)

- 인공물(AF_BUILDING)은 특정 지역명, 건물명 함께 실현됨

예: AF_BUILDING

- ① 이광종 감독이 이끄는 한국은 28일 인천 문학경기장에서 열린
- LCP_CITY LCP_COUNTY
- AF_BUILDING
- AF_BUILDING
- ② 유현진은 12일 미국 캘리포니아주 로스앤젤레스 다저스타디움에서 벌어진
- LCP_COUNTRY LCP_PROVINCE LCP_CITY AF_BUILDING
- LCP_PROVINCE
- LCP_CITY
- AF_BUILDING

인공물(AF)

인공물(AF)

- 작품명(AFA_VIDEO)은 영화명, TV, 라디오 프로그램명

예: AFA_VIDEO

① 로미오와 줄리엣
PS_CHARACTER PS_CHARACTER
AFA_VIDEO

② 조 교육감은 4일 CBS 라디오 '김현정의 뉴스쇼' 인터뷰에서
OGG_MEDIA THI_HW PS_NAME
AFA_VIDEO
AFA_VIDEO

사건명(EV)

사건(EV)

- 사건명(EV_SPORTS)은 특정 시간, 장소에서 일어난 사건명, 사고, 행사

예: EV_SPORTS

① 2014 인천 아시안게임 첫날

DT_YEAR LCP_CITY EV_SPORTS

EV_SPORTS

기존 지침의 일관성 문제를 중첩 구조 분석으로 해결 가능

- ② <1988년:DT_YEAR> <서울 올림픽:EV_SPORTS>
- ③ <88 서울올림픽:EV_SPORTS>

⑤ 중첩 개체명 말뭉치 활용

- 데이터 구축과 활용 방안
- 연구의 의의
- 남은 문제
- 정리

데이터 구축과 활용 방안

데이터 구축과 활용 방안

- ① 데이터 규모뿐만 아니라 데이터 품질, 데이터 활용이 AI 모델 구현에 기여
- ② 데이터 구축 설계, 데이터 정제, 가공, 품질 관리, 데이터 아카이빙 지속화
 - KLUE-NER 원천 데이터
 - KONEC 150개 세분류 개체명 말뭉치 구축
 - KONNE 한국어 중첩 개체명 주석 말뭉치 구축
- ③ 한국어 데이터 공공재화로 활용, 표준화 방안을 모색함
 - 깃허브(Github)에 CC-BY-SA 라이선스로 공개
 - <https://github.com/korean-named-entity/konec>
 - <https://github.com/korean-named-entity/konne>

연구의 의의

연구의 의의

- ① 기존 오픈 데이터 활용도 높임
- ② 데이터 구축 설계 및 과정, 논의사항을 밝혀 설명 가능성, 투명성 확보
- ③ 한국어 중첩 개체명 말뭉치 구축 시도, 방법론의 타당성 모색함
- ④ 정보 손실 없이 품질 향상에 본 말뭉치가 활용되도록 함
- ⑤ 한국어 개체명 말뭉치 연구, AI 모델 개발, 실용 목적으로 다양하게 활용되기를 기대함

남은 문제

남은 문제

- ① 중첩 개체명 내부 구조 허용 범위 논의 : 통사 구조, 일부 명사 허용
- ② 중첩 개체명 적용 가능한 기준 : 특정성
 - 봉준호 영화
- ③ 중첩 개체명 범위와 통사 구조 범위가 다른 경우
 - <시진핑 중국 국가 주석> 겸 <공산당 총서기>
 - [시진핑][중국 국가주석 겸 공산당 총서기]

정리

정리

- ① 한국어 중첩 개체명 말뭉치 구축시 발생 문제 언어학적으로 분석
- ② 오픈 데이터 구축시 설계, 정비, 가공, 품질 관리 설명 가능성, 투명성
- ③ 한국어 세분류 개체명 주석 및 중첩 개체명 말뭉치 구축
- ④ 중첩 개체명 분석 방법론 일관성, 체계성, 정보 풍부성 제공 확인
- ⑤ 한국어 중첩 개체명 말뭉치 공개, 다양한 목적으로 활용도 높임

참고 문헌

- 송영숙, 정유남, 유현조(2022), 한국어 중첩 개체명 분석을 위한 연구, 《한국어 의미학》 76, 65-101.
- 유현조, 송영숙, 김민수, 윤기현, 정유남(2020), 딥러닝 기반 한국어 개체명 인식의 평가와 오류 분석 연구, 《언어》 46-3, 803-828.
- 차정원 외(2020), 2020년 개체명 말뭉치 연구 분석, 국립국어원.
- 차정원 외(2021), 2021년 개체명 분석 및 개체 연결 말뭉치 연구 분석, 국립국어원.
- Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S. M., and Weischedel, R. M.(2004), “The Automatic Content Extraction (ACE) program: tasks, data, and evaluation,” in LREC, volume 2, page 1. Lisbon.
- Kim, J.-D., Ohta, T., Tateisi, Y., and Tsujii, J.(2003), “Genia corpus - a semantically annotated corpus for bio-textmining”, Bioinformatics, 19(suppl 1):i180-i182.
- Park, S., El Arass, M., and Souissi, N.(2018), “Data Lifecycle: From Big Data to SmartData”, 2018 IEEE 5th International Congress on Information Science and Technology (CiSt), 80-87, doi:10.1109/CIST.2018.8596547.
- Park, S., Moon, J., Kim, S., Cho, W., Han, J., Park, J., Song, C., Kim, J., Song, Y., Oh, T., Lee, J., Oh, J., Lyu, S., Jeong, Y., Lee, I., Seo, S., Lee, D., Kim, H., Lee, M., Jang, S., Do, S., Kim, S., Lim, K., Lee, J., Park, K., Shin, J., Kim, S., Park, L., Oh, A., Ha, J., Cho, K.(2021), KLUE: Korean Language Understanding Evaluation, arXiv:2105.09680.
- Ringland, N., Dai, X., Hachey, B., Karimi, S., Paris, C., and Curran, J. R.(2019), NNE: A dataset for Nested Named Entity Recognition in English newswire. arXiv:1906.01359.
- Walker, Ch., Strassel, S., Medero, J., and Maeda, K.(2006), ACE 2005 Multilingual Training Corpus. Linguistic Data Consortium, <https://doi.org/10.35111/mwxc-vh88>.

<말뭉치와 자료>

- TTAK_[2].KO-10.0852(2015), 개체명 태그 세트 및 태깅 말뭉치,
<https://www.tta.or.kr/tta/ttaSearchView.do?key=77&rep=1&searchStandardNo=TTAK.KO-10.0852&searchCate=TTAS>
- 국립국어원, 국립국어원 개체명 분석 말뭉치 개체 연결 버전(2022), 2021(1.0). URL: <https://corpus.korean.go.kr>
- NSMC (Naver Sentiment Movie Corpus) URL: <https://github.com/e9t/nsmc>
- KLUE-NER (KLUE Named Entity Recognition) URL: <https://github.com/KLUE-benchmark/KLUE>

경청해 주셔서 감사합니다.

질의 응답