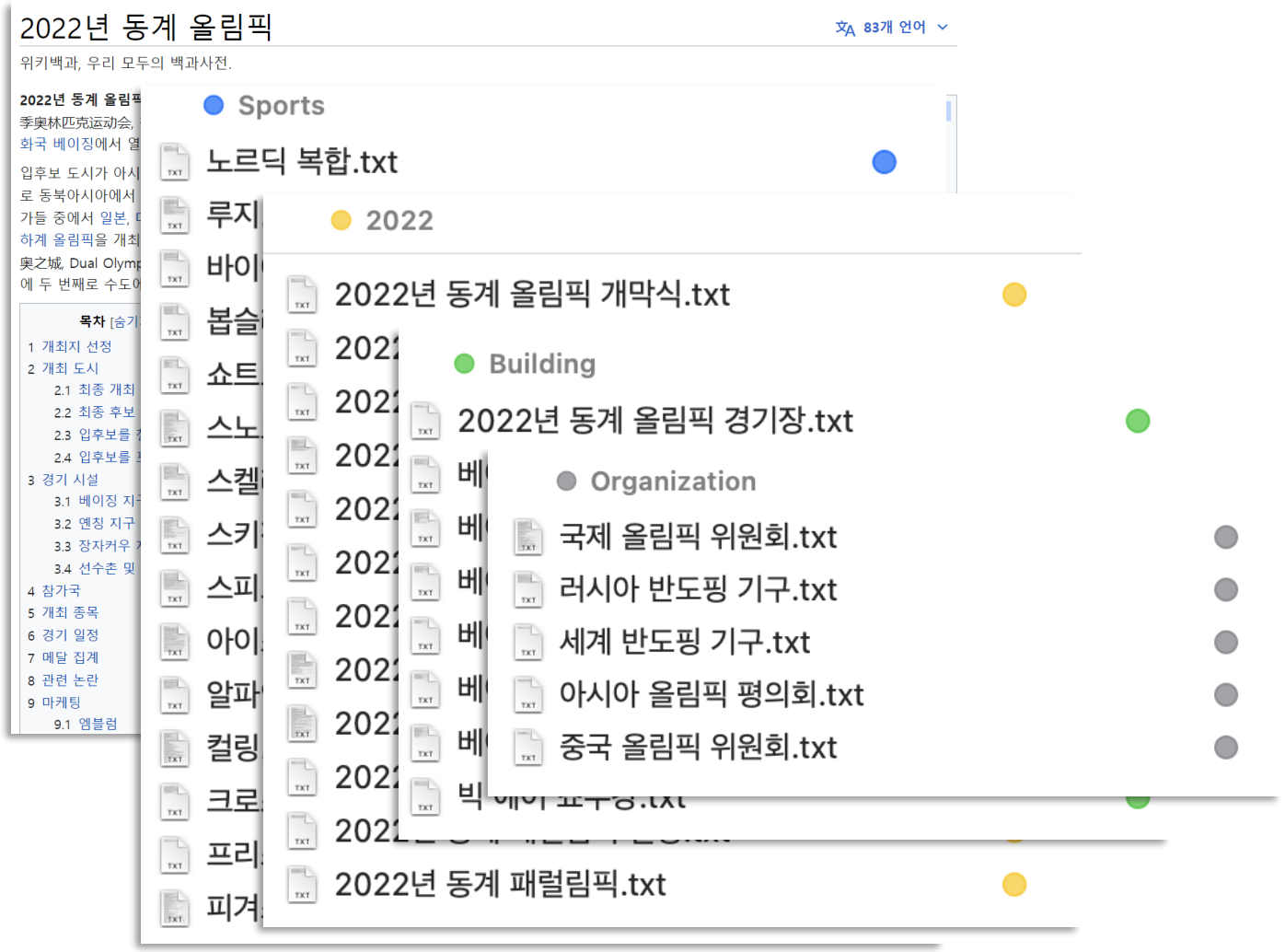
베이징 동계 올림픽 RE 데이터 제작 소개

05조 외않되조

강나경_T3001, 김산_T3031, 김현지_T3069, 정민지_T3196, 최지연_T3224

데이터셋 소개

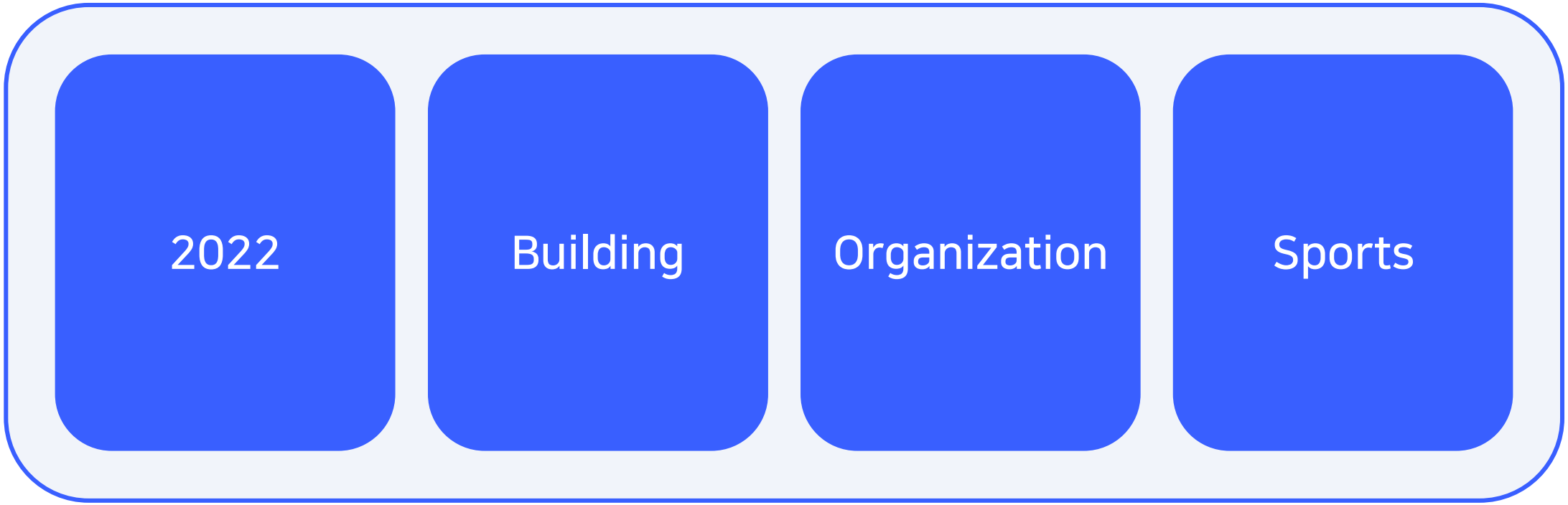


- **베이징 동계 올림픽** 관련 위키 데이터
- **데이터 43개, 문장 1,693개**로 구성
- 크게 4개로 분류:
  2022, Building, Organization, Sports
- **데이터 예시)** 2022년 동계 올림픽은 2022년 2월 4일부터 2월 20일까지 중화인민공화국 베이징에서 열린 동계 올림픽이다.

데이터 제작 과정 소개


데이터 확인
오피스 아워 + 멘토링
Relation 선정 후
entity 고려
Relation 초안 정의
가능한 관계들을 모두 고려한 뒤,
도메인 특화 relation으로 축소
가이드라인 작성 및 RE 체계 정의
RE 체계 초안에서
Relation 및 Entity 확정
파일럿 태깅
전체 데이터셋 태깅
데이터셋 검수
모델링 진행 및
최종 가이드라인 작성

마주한 문제점 및 해결방안 1.여러 주제로 소분류 된 데이터셋



베이징 동계 올림픽

마주한 문제점 및 해결방안 2.여러 명의 작업자간 작업물 일치

작업자 A

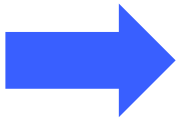
SUB-EVT

2008년 하계 올림픽의 농구종목과 야구 종목 경기를 …

작업자 B

SUB-EVT

2008년 하계 올림픽의 농구종목과 야구 종목 경기를 …



Q2. 문장부호나 연결어로 이어지면서 entity가 온전하지 않고 잘려있다면 어떻게 태깅하나요?

A. entity가 잘려있다면, 원래의 의미가 남아있는 entity만을 태깅하는 것으로 합니다.

<바른 예시>

- **피겨스케이팅**(SUB-SPO): 싱글과 **페어 스케이팅**(OBJ-SPO) 경기에서, 참가자는 '쇼트 프로그램'과 '프리 스케이팅'의 두 가지 정해진 연기를 수행해야 한다.
- **2008년 하계 올림픽의 농구**(SUB-EVT) 종목과 야구 종목 경기를 이 곳 **베이징 우커쑹 스포츠 센터**(OBJ-LOC)에서 개최했다.

<잘못된 예시>

- **피겨스케이팅**(SUB-SPO): **싱글과 페어 스케이팅**(OBJ-SPO) 경기에서, 참가자는 '쇼트 프로그램'과 '프리 스케이팅'의 두 가지 정해진 연기를 수행해야 한다.
- **2008년 하계 올림픽의 농구 종목과 야구 종목**(SUB-EVT) 경기를 이 곳 **베이징 우커쑹 스포츠 센터**(OBJ-LOC)에서 개최했다.

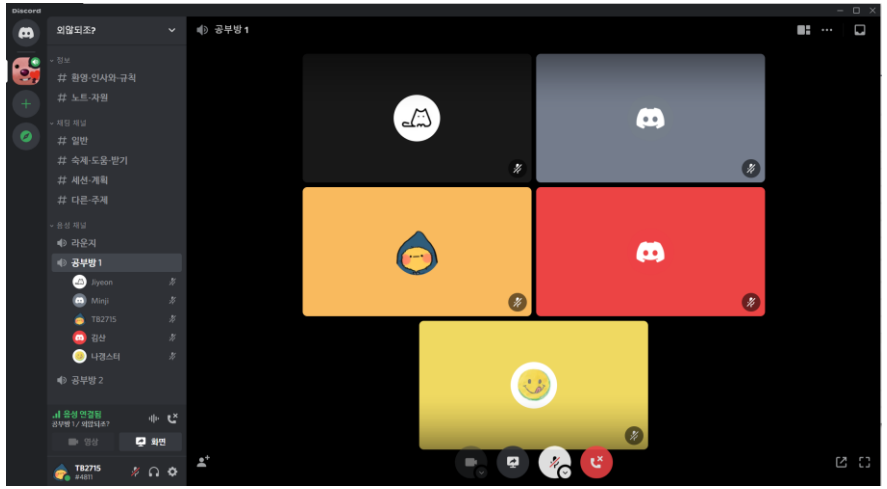
가이드라인 문서 FAQ에 관련 내용 작성

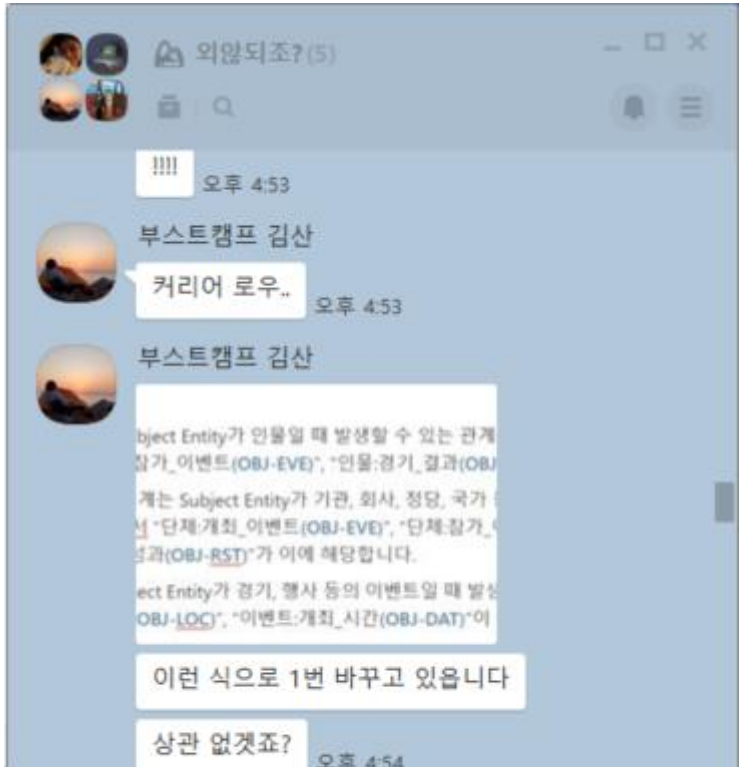
마주한 문제점 및 해결방안 2.여러 명의 작업자간 작업물 일치

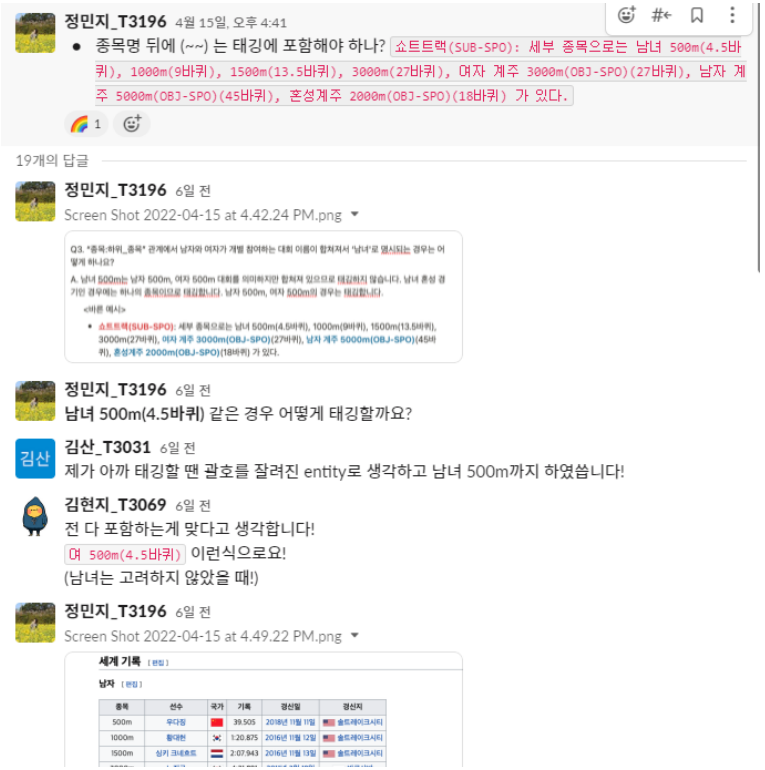
A1 | fx | 해결 유무

	A	B	C	D	E
1	해결 유무	문장	질문	응답1	응답2
2	☑	하우스를 향해 컬링 시트 빙판 위에 락이라고도 부르는 무겁고 광택이 있는 화강암 스톤을 차례대로 미끄러뜨린다.	락'은 용어일까요 아이템일까요?	현지) 저는 "아이템"이라고 생각합니다.	
3	☑	2022년 동계 올림픽 쇼트트랙 남자 5000m 계주: 특히 중국팀이 꼴찌로 결승선을 통과했음에도 진출한 점, 중국팀의 어드밴스드로 실격 처리된 팀도 없었다는 점에서 더욱 논란이 되었다.	중국팀은 태깅을 할 때 "중국"만 태깅하나요? "중국팀"까지 태깅하나요?		
4	☑	2022년 동계 올림픽 쇼트트랙 남자 5000m 계주: 2021-22 ISU 쇼트트랙 월드컵에서는 ~~월드컵 직전 4경기~~에서 캐나다가 1위를 차지했으며, 2위는 대한민국, 3위는 헝가리였다.	"월드컵 직전 4경기"도 태깅 대상일까요?	지연) 태깅할 관계가 있으면 할 수도 있을 것 같은데, 태깅할만한 관계가 없으면 굳이..?	민지) 월드컵 직전 경기를 이벤트로 볼 수도 있을 것 같은데 뭉뜽그려서 4경기라 애매한거 같아요. 캐나다 - 1위 (단체:성과) 이 관계도 중요하지 않을까요?
5	☑	스톤 스위핑은 커브를 줄여주고	"스톤 스위핑"이 용어일까요, "스톤"은 아이템, "스위핑"만 용어일까요?	현지) ~~저거 기술명이 "스톤 스위핑" 아닌가용?~~ 찾아보니까 ㅋㅋㅋ 아니네요 "스톤": 아이템, "스위핑": 기술명 인 것 같습니다 -> #부덕_동공지진	
6	☑		("컬링" - 아이템 - "스톤")만 엄청나게 나오는데 어쩔 수 없는거겠죠?	현지) 컬링,... 그렇더라구요 (지난번 배정이 컬링..ㅎ) 그거 빼면 태깅할 것도 없을껄요? ㅋㅋㅋㅋ	
7	☑	관계 없음 -> 개인이 알아서~	"관계없음"을 태깅할 만한 것도 없을 경우에만 문장 버리는건가요?	현지) 저는 "관계없음"은 따로 태깅을 안하고 나중에 저희가 엑셀로 확인할 때 애매한 것들을 "관계없음"으로 태깅한다고 생각했습니다! (질문하신대로 태깅하게 되면 굉장히 많은 관계없음이 나와서 데이터 불균형이 나올거라고 생각했어요!)	-> entity wnd에 "SUB-SPO-관계_없음"이 있어서 태깅하라고 있는건줄 알았어요! 태깅 안할게요ㅎㅎㅎㅎㅎ ~~(안하면 훨씬 편하죠)~~
8	☑	2022년 동계 올림픽 메달 집계: 주최국인 중화인민공화국은 연보라색으로 표시되어 있다	여기서 "2022 동계 올림픽"을 개최 이벤트 라고 볼 수 있을까요? (메달 집계까지가 제목이니까)	지연) 오,,, 저는 볼 수 있을 것 같다고 생각합니다	
9	☑	수비수 안셀 살멜라가 첫 홈 개막 경기에서 선제골을 기록하여 이는 ~~아 경기장의 첫 골~~로 기록되었다.	안셀 살멜라 - 이 경기장의 첫 골 (per - rst) 가능할까요?	현지) ~~선제골... 만 경기_결과,... 라고 해야하지 않을까 싶지만~~ "선제골"이 경기_결과 (승패, 메달 등) 인가...? 싶네요 (부덕:혼란)	민지) 음 선제골을 보통 경기의 결과로 보지는 않는 것 같아요..! 경기 결과 선제골을 했습니다 라는 걸 못들어본 거 같아요
10	☑	베이징 수도 체육관: 2004년 10월 17일, 전미 농구 협회(NBA)의 2004-05시즌 시범 경기를 수도 실내 체육관에서 진행했다.	시범경기(이벤트)의 태깅 범위에 "전미 농구 협회의"가 포함될까요?	ks) ~의 ~~ 가 애매한 것 같은데 저는 웬만하면 ~의까지 포함하고 있긴 했씁니다!	지연) '시범경기'라고 하면 그냥 일반명사인데, 앞에 수식을 붙여줘서 특정 경기를 의미하는게 되는거니, 포함하는게 더 적절할 것 같다고 생각합니다!

마주한 문제점 및 해결방안 2.여러 명의 작업자간 작업물 일치







Relation 정의 기준 및 방식 소개

Object \ Subject	PERSON (딱 사람 이름! 직업 ㄴㄴ)	ORGANIZATION	EVENT (경기)	SPORT (종목)	~~ARTIFACTS~~ (건축물) BUILDING
Person	경쟁	대표 / 멤버 / (IOC 위원장 홍길동은 ~)	-	-	(개인) 건축자, 디자이너 / 이용자
Organization	**소속** (대한민국 선수단 코치 A씨)	**경쟁 / 파생, 유래 / 포함, 소속 (subj>obj)** (중국 올림픽 위원회는 중국을 대표하는 국가 위원회이다.)	-	-	(단체) 건축자, 디자이너
Event (경기)	**참가** (스피드 스케이팅에 참가한 선수 곽윤기)	**개최, 주도, 주관 / 참가 /** 불참	-		
Sport (종목)	주 종목, 참가 대회 (스피드 스케이팅 선수 곽윤기)	공인 (A단체에서 공인한 a 종목)	-	하위 종목 (수영 종목에 배영, 평영, 접영, 자유형, 100m, 50m)	용도 (애매) ..
Location			**장소** 개최지, 행사 건물, 행사장 **(Artifacts 포함)**	발생지, 유래지	위치
Date		설립	**행사 시간 (시간, 날짜, 기간 포함)** (4월 17일"까지" 포함되게 태깅)	정식 종목 채택 년도 / 종목을 하는 시기, 때 / 시작 년도	건축일 / 개장일
~~Prize (Score)~~ Result	**개인 (최고) 기록, 수상 실적(등수)** (금메달), **득점 수**	**성과** (대한민국 대표단이 금메달 8개로 전체 3위를 달성했다.)	참가 자격 (상위 3위까지만 올림픽에 진출)		
Quantity (수량)		-	-		크기 (면적, 층수 등)
Term	별명, 타이틀(금메달리스트)	별칭, 약자 (러시아 올림픽 위원회의 명칭을 따서 'ROC'라는 약자로 출전)	-	**전문용어, 기술 용어** (스노보드에 관심이 있다면 본인이 레귤러인지 구피인지부터 알아야 한다.)	별칭, 이전 이름? (제목 필요 + 경기장의 애칭은 부채를 닮았다고 하여, 더 팬(,)이다.)
Sport_Item		-	-	**사용 되는 아이템**	
Role (직업)	**직위, 직책** (IOC 위원장 홍길동은 ~~ 작가 A씨의 작품 ~~ 대한민국 선수단 코치 A씨 ~)		-	-	건축물의 역할 (제목 필요 + 또한, 차오양구 지역 주민들의 다목적 스포츠 센터의 역할을 맡고 있다.)
Artifacts (건축물)	-		-	-	
count	7	10	4		13

Relation 정의 기준 및 방식 소개

	Entities	Tag	설명	주의사항	예제
2	Person	PER	실존 인물을 나타내며, 축약어로 PER(Person)으로 표시한다	해당 사항 없음	곽윤기, 조나단 테일러, 카타리나 알트하우스
3	Organization	ORG	기관/단체 명칭을 나타내며, 축약어로 ORG(Organization)으로 표시한다. 스포츠 기관, 정치/정부 기관 등이 포함된다.	- 기관/단체명과 그 기관/단체의 별칭, 약어가 동시에 쓰였을 경우 이를 묶에 전체를 기관/단체명으로 보며, 전체를 태깅한다. - "~ 대표팀"의 일반 명사도 ORG로 태깅한다 (예. 대한민국 선수단) - 우승팀, 준우승팀, 주최국과 같은 일반 명사는 태깅에서 제외	세계반도핑기구(WADA), 국제 올림픽 위원회(IOC), 독일, 노르웨이 현지 언론
4	Event	EVT	특정 행사, 경기의 명칭을 나타내며, 축약어로 EVT(Event)로 표시한다.	- 2022년 동계 올림픽과 같이 년도와 행사, 경기명이 동시에 쓰였을 경우 이를 묶어 전체를 이벤트로 보며, 전체를 태깅한다. - 스포츠 경기명의 경우 종목명과 동일하게 사용되므로 문맥 상 사용된 뜻으로 구분하도록 한다. (예. **쇼트트랙 남자 5000m 계주 준결승전(EVT)**에서 리원룽이 넘어지며 최종 4위를 기록했다. / 과연 **쇼트트랙(SPO)**에서 나올 수 있는 판정인가라는 생각까지" 했음을 밝혔다.)	2022년 동계 올림픽 쇼트트랙, 2022년 동계 패럴림픽, 쇼트트랙 남자 1000m, 쇼트트랙 남자 5000m 계주 경기, 2021-22 ISU 쇼트트랙 월드컵 1000m 종목
5	Sport	SPO	스포츠/레포츠 종목명을 나타내며, 축약어로 SPO(Sport)로 표시한다.	- 종목명의 경우 스포츠 경기명과 동일하게 사용되므로 문맥 상 사용된 뜻으로 구분하도록 한다. (예. **쇼트트랙 남자 5000m 계주 준결승전(EVT)**에서 리원룽이 넘어지며 최종 4위를 기록했다. / 과연 **쇼트트랙(SPO)**에서 나올 수 있는 판정인가라는 생각까지" 했음을 밝혔다.)	스키 점프, 쇼트트랙, 남자 500m, 바이애슬론, 크로스컨트리 달리기, 사격
6	Result	RST	우승 결과, 금/은/동메달, 세계 기록 결과, 경기 결과를 나타내며 축약어로 RST(Result)로 표시한다.	- 특정 인물이 가지고 있는 기록인 경우 인물을 지칭하는 타이틀(인격적인 의미가 담긴 부분)은 제외하고 태깅한다. (예. **세계기록** 보유자 황대헌이 **2등**)	세계 기록, 금메달, 2등, 올림픽 기록, 실격
7	Date	DAT	날짜 표현, 기간을 나타내며 축약어로 DAT(Date)로 표시한다.	- 기간을 태깅할 경우 ~까지를 포함하여 태깅한다. (예. **2022년 2월 5일부터 2월 7일까지** 중화인민공화국)	2022년 2월 5일부터 2월 7일까지, 2021년 2월 19일
8	Location	LOC	국가명, 지역/장소를 나타내며 축약어로 LOC(Location)으로 표시한다.	- 지역명이 기관/단체명과 동일하게 사용되므로 주의하여 태깅한다. - 지역명과 "지역"이 함께 등장한 경우, 지리/지역명만을 태깅한다. (예. **차오양구** 지역)	중화인민공화국 베이징, 중화인민공화국 베이징시 차오양구, 차오양구, 노르웨이 릴레함메르
9	Term	TRM	스포츠에서 사용하는 전문 용어, 기술 용어 또는 사용되는 아이템이나 도구 등을 나타내며 축약어로 TRM(Term)으로 표시한다.	- 명사구 형태의 기술 용어일 경우 이에 해당되는 기술 모두를 태깅한다. (예. 트리플 점프) - 해당 스포츠에서만 사용되는 아이템이 아니더라도 사용되는 아이템은 모두 여기에 해당하는 것으로 본다.	트리플, 트리플 점프, 업라이트 스핀, 스파이럴, 팔꿈치 보호대
11	Item	ITM	스포츠에서 사용하는 도구 명칭	해당 사항 없음	스케이트화, 스톤

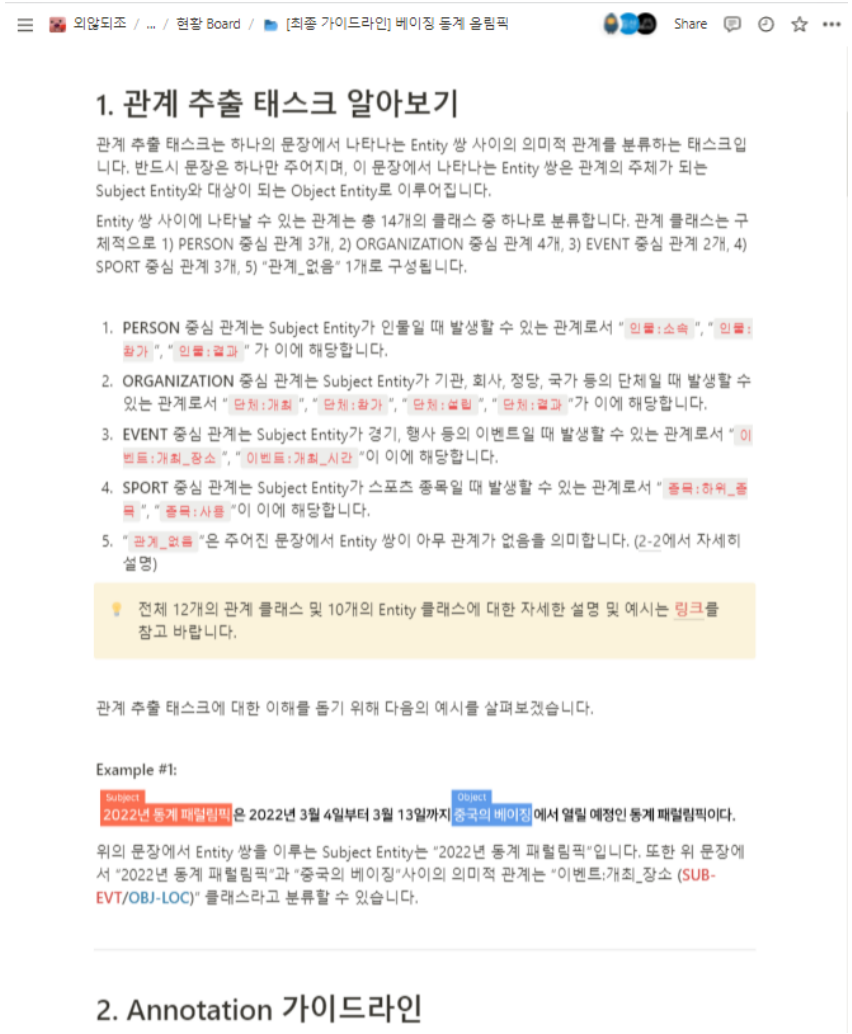
Entity Class

Relation 정의 기준 및 방식 소개

Object / Subject	PERSON	ORGANIZATION	EVENT (경기)	SPORT (종목)
Person				
Organization	인물:소속			
Event (경기)	인물:참가	단체:개최 / 단체:참가		
Sport (종목)				종목:하위_종목
Location			이벤트:개최_장소	
Date		단체:설립	이벤트:개최_시기	
Result	인물:결과	단체:결과		
Term				종목:사용_용어
Sport_Item				종목: 사용_아이템

Relation Class

가이드라인 소개



외않되조 / ... / 현황 Board / [최종 가이드라인] 베이징 동계 올림픽 Share

1. 관계 추출 태스크 알아보기

관계 추출 태스크는 하나의 문장에서 나타나는 Entity 쌍 사이의 의미적 관계를 분류하는 태스크입니다. 반드시 문장은 하나만 주어지며, 이 문장에서 나타나는 Entity 쌍은 관계의 주체가 되는 Subject Entity와 대상이 되는 Object Entity로 이루어집니다.

Entity 쌍 사이에 나타날 수 있는 관계는 총 14개의 클래스 중 하나로 분류합니다. 관계 클래스는 구체적으로 1) PERSON 중심 관계 3개, 2) ORGANIZATION 중심 관계 4개, 3) EVENT 중심 관계 2개, 4) SPORT 중심 관계 3개, 5) "관계_없음" 1개로 구성됩니다.

1. PERSON 중심 관계는 Subject Entity가 인물일 때 발생할 수 있는 관계로서 "인물:소속", "인물:창가", "인물:결과" 가 이에 해당합니다.
2. ORGANIZATION 중심 관계는 Subject Entity가 기관, 회사, 정당, 국가 등의 단체일 때 발생할 수 있는 관계로서 "단체:개최", "단체:창가", "단체:설립", "단체:결과"가 이에 해당합니다.
3. EVENT 중심 관계는 Subject Entity가 경기, 행사 등의 이벤트일 때 발생할 수 있는 관계로서 "이벤트:개최_장소", "이벤트:개최_시간"이 이에 해당합니다.
4. SPORT 중심 관계는 Subject Entity가 스포츠 종목일 때 발생할 수 있는 관계로서 "종목:하위_종목", "종목:사용"이 이에 해당합니다.
5. "관계_없음"은 주어진 문장에서 Entity 쌍이 아무 관계가 없음을 의미합니다. (2-2에서 자세히 설명)

전체 12개의 관계 클래스 및 10개의 Entity 클래스에 대한 자세한 설명 및 예시는 링크를 참고 바랍니다.

관계 추출 태스크에 대한 이해를 돕기 위해 다음의 예시를 살펴보겠습니다.

Example #1:

Subject 2022년 동계 패럴림픽은 2022년 3월 4일부터 3월 13일까지 Object 중국의 베이징에서 열릴 예정인 동계 패럴림픽이다.

위의 문장에서 Entity 쌍을 이루는 Subject Entity는 "2022년 동계 패럴림픽"입니다. 또한 위 문장에서 "2022년 동계 패럴림픽"과 "중국의 베이징"사이의 의미적 관계는 "이벤트:개최_장소 (SUB-EVT/OBJ-LOC)" 클래스라고 분류할 수 있습니다.

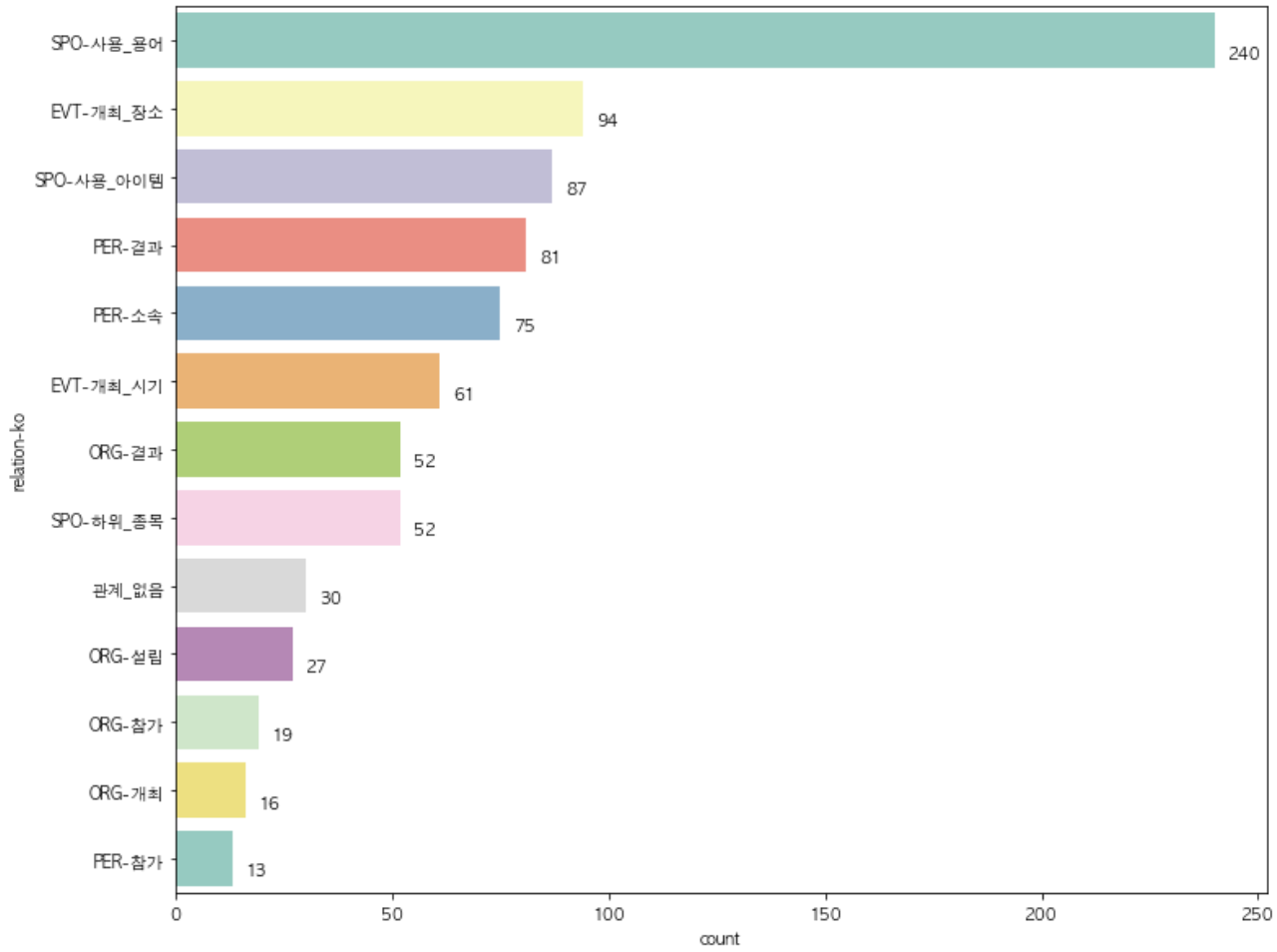
2. Annotation 가이드라인

2-1) 관계의 방향성

Subject에 대한 Object의 관계와 Object에 대한 Subject의 관계는 다를 수 있습니다. 즉, 경우에 따라서 관계의 방향성이 존재합니다. 다음의 예시를 통해서 자세히 알아보겠습니다.

Example #2:

데이터셋 분석



관계별 데이터 개수 분포

데이터셋 분석

Dataset	Fleiss' Kappa
베이징 동계 올림픽	0.937

전체 데이터셋에서 카테고리별(2022, Building, Organization, Sports)
약 50개 문장을 랜덤하게 선택하여
총 847개의 문장 중 123개의 문장 검수 진행

느낀점

데이터 제작을 하면서 좋은 품질의 데이터를 제작하기 위해 많은 노력과 손길이 필요한 것을 느낄 수 있는 경험이었습니다.

파일럿 태깅과 데이터 제작 가이드라인의 FAQ를 작성했음에도
실제 데이터 태깅 과정에서 수많은 예외처리가 등장해 일관된 태깅을 하는 것이 어려웠습니다.

그리고 이번 대회를 통해 제작자가 원하고자 하는 데이터셋 구축 방향으로 이끌기 위해
많은 소통이 필요하다는 것도 알 수 있는 값진 기회였던 것 같습니다!

감사합니다