프로젝트 Wrap Up

1. 프로젝트 개요

본 프로젝트에서는 **2022 베이징 동계 올림픽**과 관련된 위키 원시 말뭉치를 활용해 자연어처리 **관계 추출 태스크**에 쓰이는 주석 코퍼스를 제작했습니다. 프로젝트의 의의는 한국어 및 다른 언어에서의 자연어처리 데이터셋의 유형 및 포맷이 어떠한지, 그리고 데이터셋을 구축하는 일반적인 프로세스가 무엇인지 학습하는 것입니다.

2. 프로젝트 팀 구성 및 역할

이름	역할
공통	가이드라인 작성, Entity•Relation 정의, 파일럿 및 메인 어노테이션
강나경	카테고리별 문장 split, fleiss-kappa 계산
김산	가이드라인 FAQ 작성
김현지	데이터셋 전처리, 가이드라인 이미지 제작
정민지	모델 Fine-tuning, 데이터셋 분석
최지연	여러 개의 파일을 카테고리별로 분류하여 통합

3. 프로젝트 수행 절차 및 방법

3-1. 데이터셋 소개

- 베이징 동계 올림픽 관련 위키 데이터
- 총 43개의 문서, 문장 1,693개로 구성
- 예시

> 🏅 2022년 동계 올림픽은 2022년 2월 4일부터 2월 20일까지 중화인민공화국 베이징에서 열린동계 올림픽이다.

3-2. 데이터 제작 과정

3-3. 마주한 문제점 및 해결 방안

1. 여러 주제로 소분류 된 데이터셋
    - 대분류 베이징 동계 올림픽에서 소분류 2022, Building, Organization, Sports로 분할
    - 소분류된 카테고리 내에서 등장하는 관계 선정
2. 여러 명의 작업자간 작업물 일치

9	☑	수비수 안셀 살멜라가 첫 홈 개막 경기에서 ~~선제골~~을 기록하여 이는 ~~이 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 경기**를 수도 실내 체육관에서 진행했다.	시범경기(이벤트)의 태깅 범위에 "전미 농구 협회의"가 포함될까요?	ks) ~의 ~~ 가 애매한 것 같은데 저는 왠만하면 ~의까지 포함하고 있긴 했습니다!	지연) '시범경기'라고 하면 그냥 일반명사인데, 앞에 수식을 붙여줘서 특정 경기를 의미하는게 되는거니, 포함하는게 더 적절할 것 같다고 생각합니다!
11	☑	준결승 1조에서 경기를 치른 대한민국 선수 황대헌과 2조에서 경기를 치른 대한민국 선수 이준서가 각각 **조 1위, 조 2위로 결승선을 통과**했다.	이 문장에서 보면 결승선을 통과한 것이 경기 결과인 것 같습니다. 그런데 다음 문장에서는 실격당했다 라는 내용이 나옵니다! 이 문장만 가지고 판단할 때는 조 1위, 조 2위가 경기결과라고 생각할 수 있을까요?	현지) 황대헌-조 1위, 이준서-조 2위 로 태깅하면 될 것 같습니다. 본 문장에선 실격이라는 단어가 나오지 않으니까요! (외부 지식 활용으로 생각할 수 있을 것 같음!)	
12	☑	베이징 수도 체육관: 2022년, 2022년 동계 올림픽 쇼트트랙, 피겨스케이팅 종목 경기를 개최하고 있다.	이 경우 저희가 사전에 정의한 쉼표 등으로 인해 분할되는 케이스에 해당되어 "피겨스케이팅"은 이벤트 태깅 대상이 아닐까요?	민지) 피겨스케이팅만을 이벤트로 보기는 애매한 것 같습니다! 2022년 동계 올림픽 쇼트트랙은 이벤트인 것 같지만요. 그리고 말씀해주신대로 쉼표 분할 케이스인것 같습니다!	
13	☑	한국의 김연아	김연아(per) - 한국(org) 가능하겠죠//.??	현지) 넵! 가능하다고 생각합니다!	
14	☐	캐나다 중부 지역에서 **컬링**하던 사람들은 19세기 초까지 ~~스톤~~보다는 ~~철~~을 종종 사용했다.	1. (컬링-아이템-철) 가능할까요? 19세기 초까지 사용한거라 지금은 사용하지 않는거라서 시점 생각하면 아닌것 같기도 해서 고민입니다! 2. 여기서의 "스톤"이 컬링의 (현재 사용하는)아이템이라는걸 유추할 수 있을까요?	현지) 철,...은 따로 아이템은 아닌 것 같습니다! 구성 요소? 가 더 적합할 것 같아요 ㅋㅋㅋㅋㅋ 여기서 스톤은... 돌을 의미하는 걸까요? 뭔가 문장내 뉘앙스가 돌 vs 철 의 느낌이네요	
15	☐	컬링: 스톤 핸들에 불빛이 녹색이면 적법하게 투구했음을 말해주며 (보류: 아이템 유지?)	"스톤 핸들"은 아이템 vs 전문 용어, 당신의 선택은? (두둥)	민지) 전문용어..? 요 근데 애매한 것 같습니다ㅠ 이거 다시 논의해보면 좋을 것 같습니다!	지연) 저는 아이템이라고... 생각했습니다. 손잡이만 따로 떼기도 하더라고요
16	☑	**베이징 지구**는 주로 **동계 올림픽의 빙상 종목의 경기**를 치를 예정이고, **개막식**과 **폐막식**이 열릴 예정이다.	베이징 지구가 건물은 아니지만, 장소의 용도가 나오기도 하는 것 같습니다! (그냥 발견, 건물:용도 관계의 확장이 필요한가?에 대한 궁금증)		
17	☑	컬링: 이 대회에서 이듬해 ~~**2007년 월드 챔피언**~~이 된 **켈리 스콧**이 1998년 월드 챔피언십 동메달리스트였던 캐시 킹을 상대로 한 경기의 한 엔드에서 8점을 기록했다.	"2007년 월드 챔피언"은 켈리 스콧의 경기_결과에 해당되지 않겠죠? (FAQ 업적 참고) (캐시 킹은 저렇게 태깅했습니다!)	민지) - 우선, 캐시킹에 대해서 <1998년 월드 챔피언십 동메달> 이렇게 경기결과 태깅하기로 했던 것으로 기억합니다! - 월드 챔피언도 경기결과 아닐까요..?	지연) 챔피언이 '우승자'라는 뜻이어서... '우승'만을 떼어서 태깅할 수 있다면 딱 좋겠는데, 이 경우는 좀 애매한듯하네요ㅜㅜ 어렵군요! 그리고 동메달 부분은 민지님 의견과 같습니다!
18	☑	**2022년 동계 올림픽 빅 에어** 종목 개최를 위해 **빅 에어 쇼우강 경기장**을 2018년 착공했다. (건물:용도 없어지는거 예약)	해당 문장은 "이벤트: 개최_장소" 일까요? "건물:용도" 일까요? 아래 중에서 적합한 태깅은 무엇일까요? **2022년 동계 올림픽 빅 에어** 종목 개최를 위해 **빅 에어 쇼우강 경기장**을 2018년 착공했다. > evt:holding_place ~~**2022년 동계 올림픽 빅 에어 종목** 개최를 위해 **빅 에어 쇼우강 경기장**을 2018년 착공했다. > bld:spo (1)~~ ~~2022년 동계 올림픽 **빅 에어** 종목 개최를 위해 **빅 에어 쇼우강 경기장**을 2018년 착공했다. > bld:spo (2)~~	민지) 1번과 2번 중에는 2번이 더 적절하다고 생각합니다! 근데 이벤트:개최_장소, 건물:용도가 명확하게 구분되지 않는 것 같기도 합니다!	지연) 이벤트인경우 '이벤트:개최_장소', 스포츠인경우 '건물:용도'로 하기로 했던걸로 기억합니다. 이 문장의 경우는 이벤트인듯 하면서 스포츠인듯도 해서 어렵군요ㅜㅜㅜㅜ 문맥상으로는 개최를 '위해' 착공한것이니 스포츠라고 보고 '건물:용도'로 태깅하는 것이 좋을 듯 합니다(2번)
19	☑	**빅 에어 쇼우강**: **2022년 동계 올림픽의 스노보드 빅 에어** 종목과 ~~프리스타일 스키 빅 에어 종목~~을 이 경기장에서 진행한다. (건물:용도 없어짐)	12번이랑 동일한 질문 프리스타일 스키 빅 에어 종목은 태깅 대상이 아니겠죠?	지연) 넵, 앞의 '2022년 동계 올림픽의' 부분이 '스노보드 빅 에어 종목'과 '프리스타일 스키 빅 에어 종목'을 모두 수식하는 걸로 보여서 현지님 말씀대로 뒤에 오는 '프리스타일 스키 빅 에어 종목'은 태깅 대상이 아닌 것 같아요 근데 저희 '종목'은 제외하기로 했던 것 같은데.. 그럼 '2022년 동계 올림픽의 스노보드 빅 에어' 부분만 태깅해야 하지 않을까요!	

- 동일한 작업 결과물을 얻기 위해 헷갈릴 수 있는 entity의 영역과 relation 관계 같은 경우는 slack과 구글 스프레드시트를 활용하여 팀원들과 의견을 정리한 후, 방향을 명확하게 잡았습니다.

3-4. Relation 정의 기준 및 방식 소개

- Relation set의 구성 및 정의, 가이드라인 작성, 파일럿 및 메인 어노테이션, 그리고 간단한 모델 Fine-tuning의 과정을 통해 실제 데이터 제작의 workflow를 경험해 볼 수 있습니다.
- 이 과정에서 정밀한 가이드라인 제작의 중요성과 inter-annotator agreement(IAA)의 개념을 체득할 수 있습니다.

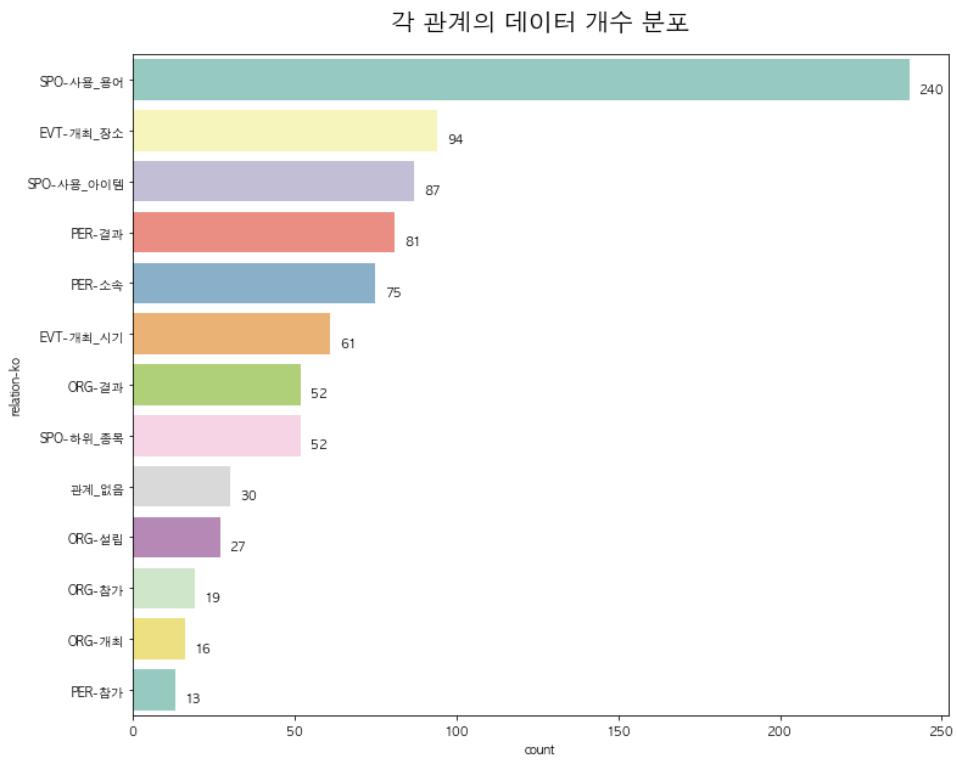
4. 프로젝트 수행 결과

4-1. 가이드라인 소개

가이드라인

4-2. 데이터셋 분석 결과

스포츠에서 사용하는 용어에 대한 관계



4-3. 정의된 Entities

아래와 같은 9개의 entity를 정의하였습니다. 팀원들과의 소통을 통해 각 **entity의 범위를 명확하게** 정의하였고, 헷갈릴 수 있는 부분은 가이드라인의 FAQ에 정리하였습니다.

	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 묶어 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

4-4. 정의된 Relations

초기의 아래와 같은 59개의 relation 관계를 정의하였습니다. 보통 10개 이내의 relation 관계를 정의하는 것이 일반적이므로, 점차 줄여나가는 방향으로 진행하였습니다. 이 과정에서 가장 초점을 둔 부분은, **도메인에 특화된 관계**를 정의하는 것이었습니다. 비록 데이터셋에서 많이 등장하는 relation 관계일지라도 일반화된 관계는 고려하지 않았습니다. 그 결과, 13개의 relation 관계로 좁힐 수 있었습니다. 마찬가지로 팀원들과의 소통을 통해 각 relation 관계의 범위를 명확하게 정의하였고, 헷갈릴 수 있는 부분은 가이드라인의 FAQ에 정리하였습니다.

Object \ Subject	PERSON	ORGANIZATION	EVENT (경기)	SPORT (종목)	SPORT_ITEM (스포츠 용품, 장비)	ARTIFACTS (건축물)
Person	경쟁	대표 / 멤버 /	-	-	-	(개인) 건축자, 디자이너 / 이용자
Organization	소속 (대한민국 선수단 코치 A씨)	경쟁 / 파생, 유래 / 포함, 소속 (subj->obj) (중국 올림픽 위원회는 국기를 대표하는 징기 위원회이다.)	-	-	-	(단체) 건축자, 디자이너
Event (경기)	참가 (스피드 스케이팅에 참가한 선수 짝궁가)	개최, 주도 / 참가 / 불참	-	+ 나중에	-	공모전
Sport (종목)	주 종목, 참가 대회 (스피드 스케이팅 선수 주종가)	관련 (A단체에서 관리한 a 종목)	-	하위 종목 (수영 종목에 배영, 평영, 접영, 자유형, 100m, 50m)	-	용도
Location	출신지 (~에 태어났다)	위치	개최지	발생지, 유래지	-	위치
Date	생일 / 데뷔 년도 / 사망년도	설립일 / 해체일	행사 시간 (시간, 날짜, 기간 포함) (예 1개월 여자 포함하여 표현하여 태깅)	정식 종목 채택 년도 / 종목을 하는 시기, 때 / 시작 년도	사용 시작 시기	건축일 / 개장일
Prize (Score)	개인 (최고) 기록, 수상 실적(등수) (금메달, 특점 수	설과 (대한민국 대표팀에 금메달 공적로 인해 크게를 달성했다.)	참가 자격 (상위 3위까지만 올림픽에 진출)	+ 데이터를 보면서 생각	-	건축물이 받은 상
Quantity (수량)	나이	-	-	-	규칙, 규격 / 수영	크기 (면적, 층수 등)
Term	별명, 타이틀(금메달리스트)	별칭, 약자 (리사이 올림픽 위원회의 영향을 미쳐 '캐나다'라는 약자로 출전)	-	별칭, 약자 / 전문용어 (예제처 스피드 스케이팅을 표현하 스피드 스케이팅이라 부르기도 하는)	다른 이름 (동의어)	별칭 (제목 필요 + 경기장의 예정은 부제를 달았다고 하여, 더 원,)이다.)
Sport_Item	애장품	-	-	사용 되는 아이템	함께 사용 / 부착물	-
Role (직업)	직위, 직책 (IOC 위원장 토마스 --- 역자 A씨의 직업 --- 대한민국 선수단 코치 A씨 -)	+ 향후 추가	-	-	-	건축물의 역할 (제목 필요 + 토론, 차오양구 지역 주민들의 다목적 스포츠 센터의 역할을 알고 있다.)
Artifacts (건축물)	-	홈 구장	행사 건물	-	-	하위 건축물 (소속)

초기 relation (59개)_

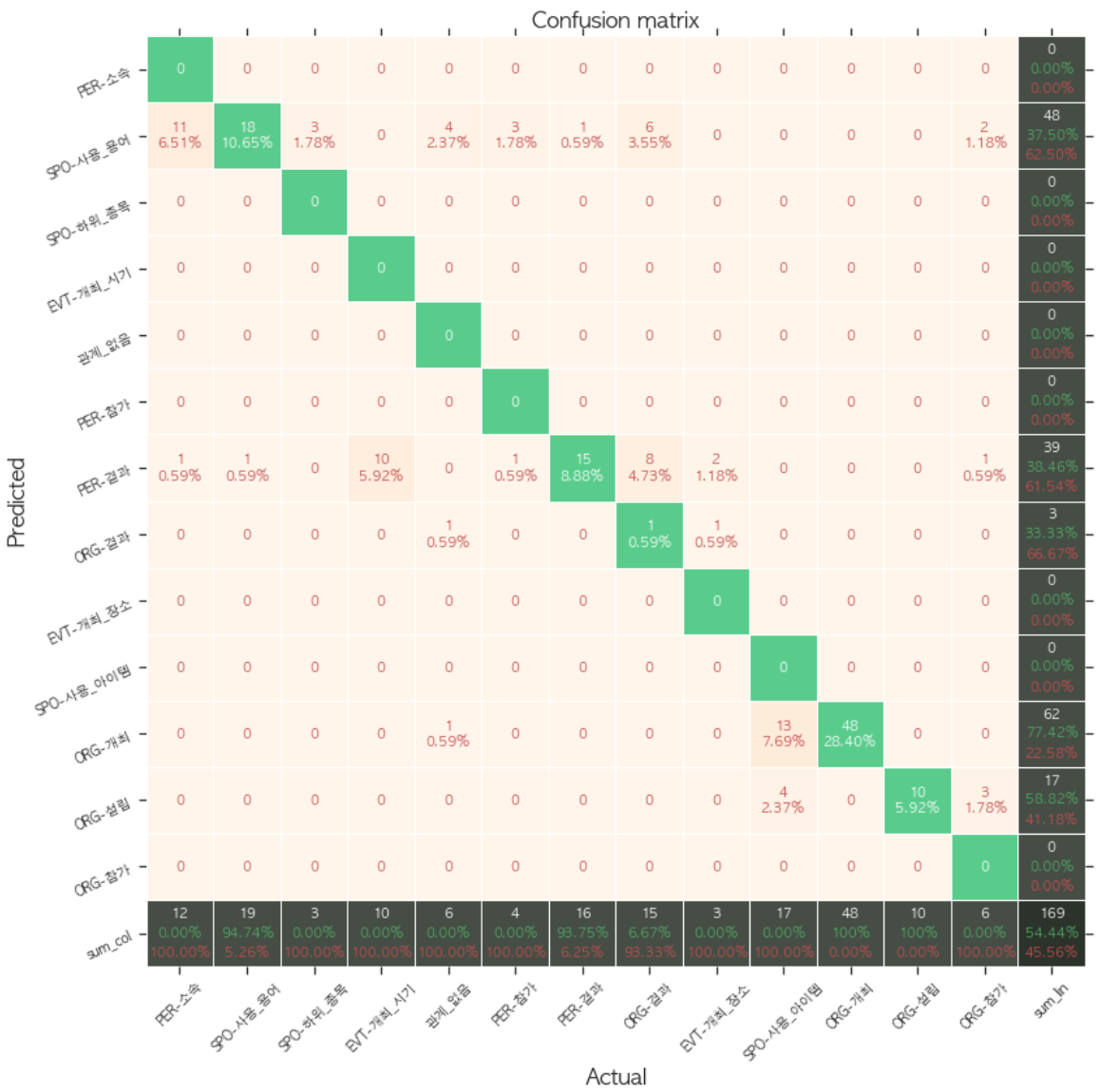
Object / Subject	PERSON	ORGANIZATION	EVENT (경기)	SPORT (종목)
Person				
Organization	인물:소속			
Event (경기)	인물:참가	단체:개최 / 단체:참가		
Sport (종목)				종목:하위_종목
Location			이벤트:개최_장소	
Date		단체:설립	이벤트:개최_시기	
Result	인물:결과	단체:결과		
Term				종목:사용_용어
Sport_Item				종목: 사용_아이템

Relation Class

최종 relation (13개)

4-5. 모델 Fine-tuning

제작한 데이터셋을 scikit learn의 stratified shuffle split을 활용해 train(0.6), validation(0.2), test(0.2)로 분리했습니다. 그리고 huggingface 프레임워크로 klue/roberta-large 사전 학습 모델을 fine-tuning 했습니다(hyperparameter는 다음과 같습니다. epochs: 20, learning rate: 2e-5, batch size: 32). validation micro f1 score는 59이며, test set evaluation micro f1 score는 55입니다. 그림에서는 test set의 ground truth와 predict label을 confusion matrix로 표현해서 모델이 어떤 라벨을 잘 예측하는지 확인해봤습니다. 모델은 13개 라벨 중 “종목:사용_용어”, “인물:결과”, “단체:결과”, “단체:개최”, “단체:설립" 라벨만 예측했으며, 단체 entity의 관계(“단체:개최”, “단체:설립")를 잘 예측하는 경향을 보였습니다.



5. 자체 평가 의견

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 데이터셋을 구축 방향으로 이끌기 위해 많은 소통이 필요하다는 것도 알 수 있는 값진 기회였던 것 같습니다!