🪐 NLP-5조 데이터제작 Report

1. Task 설명 및 주제 선정 이유

- Relation Extraction을 위한 데이터 셋을 제작한다. Relation Task를 위한 Data Annotion에서는 Subject Entity, Object Entity, Relation에 대한 Annotation을 진행한다.
- Annotation이 된 Data를 바탕으로 IAA score를 계산, KLUE 모델로 훈련을 하여 기본적인 성능을 측정한다.
- `태양계의 형성과 진화` 주제 선정 이유

  천문학은 사실을 바탕으로 작성된 문서다. 데이터에 편견, 편향 등이 나타나지 않기에 Data annotation 작업이 수월할 것이라 판단했다.

  기존 KLUE RE 데이터는 PER과 ORG를 위주로 관계가 구성되고 있다. 천문학은 기존 KLUE 데이터와 달리, 사람보다 천체가 더 많이 등장한다. 새로운 도메인에서는 어떤 Entity와 Relation이 가능한지 알아보고자 KLUE RE 데이터와 가장 관련이 없는 `태양계의 형성과 진화` 주제를 선정했다.
- 대상 리스트(아래의 단어로 검색하면 나오는 위키피디아 전문을 크롤링한 txt 파일)

  `과학/금성/달/명왕성/목성/물리학/블랙홀/성운/소행성`
  `수성/수소/온도/왜성/위성/은하/중력/지구/천문학/천왕성/초신성`
  `태양/태양계/토성/항성/해왕성/핵융합/행성/헬륨/화성`

2. Entity

- **PER (PERSON) :** PERSON은 사람 또는 신을 의미하는 개체.
- **CLO (CELESTIAL OBJECT) :** 천체를 의미하며 한국어 위키피디아 문서에서 정의하는 의미로 한정
- **CON (CONCEPT) :** 법칙, 이론, 개념을 의미한다. (상대성 이론 등)
- **DAT (DATE) :** 시기를 의미한다. “1920년”, “19세기” 등 특정 시기를 의미하는 단어로 제한
- **ELM (ELEMENT) :** 원소, 암석 등의 물질을 의미한다. 대기와 같이 영역은 제외한다.
- **MET (METRIC) :** 온도, 무게, 밀도, 자전주기 등 단위로 나타낼 수 있는 한 물체의 수치적인 특성을 의미한다.

3. Relation

Relation	Description
no_relation	Entity간의 관계를 정의할 수 없을 때
clo:revolves	<SUBJ-CLO>가 <OBJ-CLO>를 공전할 때 또는 위성일 경우
clo:exists_in	<SUBJ-CLO>가 <OBJ-CLO>에 위치하는 경우/(물리적 개념)
clo:contains	<SUBJ-CLO>가 <OBJ-CLO>에 개념적으로 포함되는 경우.
clo:turn_into	<SUBJ-CLO>가 <OBJ-CLO>로 변화했을 경우 태깅한다. (SUBJ → OBJ)
clo:alias_of	<SUBJ-CLO>가 <OBJ-CLO>의 별칭일 경우 태깅한다. (OBJ ↔ SUBJ 양방향 가능)
clo:composed_of	<SUBJ-CLO>가 <OBJ-ELM>로 이루어졌을 때(구성되었을 때) 태깅한다.
met:feature_of	<SUBJ-MET>가 <OBJ-CLO치>의 수치적 특성일 때(크기, 온도, 속도, 밀도 등)
per:propose	어떤 사람(PER)이 개념, 이론, 법칙, 현상 등을 제안/제시했을 경우
dat:date_of_discovery	<SUBJ-DAT>가 <OBJ-CLO>의 발견 날짜일 때(년도, 세기, 연월일, ~년 전 등)
per:origin_of	<SUBJ-PER>가 <OBJ-CLO>이름의 기원 또는 유래일 때(PER은 사람 또는 신 이름)

4. Dataset annotation process

- 데이터 전처리
  - 정의된 Entity가 추출되지 않는 문장은 삭제
  - Entity를 추출하기 편하게 문장을 분리하거나 뒤에 이어지는 문장과 붙임
- Annotation
  - 300개 문장으로 pilot tagging ⇒ 결과 분석 후 변경 필요한 점 반영
  - Tagtog을 이용하여 Entity tagging 실행
  - 구글 스프레드시트에서 모든 조원이 Annotation 실행

5. Dataset and Train result

- Dataset

Relation Class	Train(Ratio)	Dev(Ratio)	Test(Ratio)
no:relation	316(0.38)	39(0.39)	39(0.39)
clo:composed_of	127(0.15)	15(0.15)	15(0.15)
clo:contains	80(0.09)	10(0.10)	10(0.10)
clo:exists_in	66(0.08)	8(0.08)	8(0.08)
clo:revolves	53(0.06)	7(0.07)	7(0.07)
clo:turn_into	29(0.03)	3(0.03)	4(0.04)
clo:alias_of	17(0.02)	2(0.02)	2(0.02)
met:feature_of	70(0.08)	8(0.08)	8(0.08)
per:propose	37(0.04)	5(0.05)	4(0.04)
per:origin_of	10(0.01)	1(0.01)	1(0.01)
dat:date_of_discovery	15(0.01)	2(0.02)	2(0.02)
Total	**820**	**100**	**100**

- IAA Score
  - **Fleiss' Kappa = 0.85**

- KLUE/roberta 모델을 이용한 데이터 훈련 결과

model	f1-micro	auprc
klue/roberta-small 9 epoch	**74.07**	63.05
klue/roberta-large 5 epoch	69.92	**68.88**