Wrap-up Report

대회 개요

- Relation extraction 데이터셋 제작
    - 관계 추출 태스크에 쓰이는 주석 코퍼스를 만들어 본다.
    - 문장내에서 두 단어(subject, object)와 둘의 관계 쌍의 데이터셋을 제작하는 과정을 통해 데이터에 대해 이해한다.

팀 구성 및 역할

- 문석암_T2075
- 박마루찬_T2078
- 박아멘_T2090
- 우원진_T2137
- 윤영훈_T2142
- 장동건_T2185
- 홍현승_T2250

데이터 수행 절차와 경과

전처리 (KSS,Filter)

- Sentence Segmentation 을 진행하기 위해 KSS 를 사용함.
    - KSS Github
- mecab을 이용한 morph filter를 진행.

Relation, Entity type 선정

- Pororo NER 을 이용하여 Entity 데이터 결과를 바탕으로 아래 Relation 을 선정 후 새롭게 Entity를 다시 태깅함

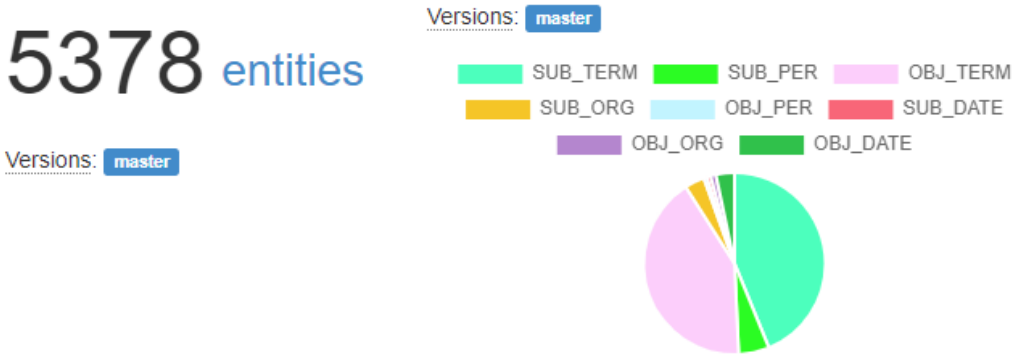
가이드라인 작성

- 관계_없음
    - 하위 모든 경우에 포함되지 않는 모든 경우
- 용어:별칭
    - (TERM,TERM) Object는 Subject의 또 다른 이름
- 용어:등장시기
    - (TERM,DATE) Object는 Subject의 등장 시기
- 제품:부품
    - (TERM,TERM) Object는 Subject의 부품
- 용어:행위
    - (TERM, TERM) Object는 Subject의 행위
- 용어:일종
    - (TERM, TERM) Object는 Subject의 일종
- 사람:고안물/제작물
    - (PER, TERM) Object는 Subject의 제품/창작물/고안물/작품

- 직업:도구
  - (PER,TERM) Object는 Subject의 도구
- 단체:제품
  - (ORG,TERM) Object는 Subject의 제품
- 용어:도구
  - (TERM, TERM) Object는 Subject의 동종업계

엔티티 태깅 (tagtog.net)

- 문장의 가능한 relation을 고려하여 entity 태깅
- 전체 문장을 나눠서 각자 진행 후 병합



라벨링

- 라벨링은 각 문장-엔티티 쌍 당 5명이 수행했다.
- SpreadSheet에 작성해서 데이터를 취합하였다.

최종 프로젝트 결과

IAA (inter-annotator agreement)

- 0.7128749615856718

Model fine-tuning

- 라벨링 완료된 스프레드 시트 저장하여 데이터셋으로 사용
- Stratification 적용하여 train,val set 분리

결과


```
"epoch": 35.09,
"eval_accuracy": 0.7533039647577092,
"eval_auprc": 79.1367011846813,
"eval_loss": 1.7397611141204834,
"eval_micro f1 score": 70.0228832951945,
```


- dataset에서 test_set의 no_relation 비중을 줄이면 점수가 소폭 하락.

자체 평가 의견

잘했던 것, 좋았던 것, 계속할 것

- 토의해야 할 점이라는 생각이 든다면 카톡으로 바로 공유한 점.
- 의견을 내고 회의 시간 내에 방법을 정한 것.

잘못했던 것, 아쉬운 것, 부족한 것 -> 개선방향

- 적당히 느낌적인 느낌으로 가이드라인을 마무리 한 점. 끝까지 명확하게 하려 노력했어야 했다.
- '관계없음'으로 태깅된 데이터들이 너무 많은 것 같다. entity들을 더 많이 포함할 수 있는 relation을 고려했어야할 것 같다.
- 최대한 라벨링 규칙을 세웠음에도 개인적인 편향이 많이 사용된 것을 보면 좀 더 구체적으로 해야했을 것 같다.
- 일정이 어려운 것이라 예상되었음에도 충분한 시간을 두지 않았던 점이 아쉽다. 더 여유있게 구성했어야 했다.
- RE task의 application을 알았다면, 데이터 제작에서 어떤 부분을 개선해야 할 지 알 수 있을 것 같다.

도전할 것, 시도할 것

- 더 정확한 가이드라인.
- 파일럿 라벨링 시간을 더 길게 가져서 라벨링 과정에서 생기는 이슈를 더 찾아 가이드라인에 반영하기
- 데이터 필터링을 더 세세하게 진행하여 데이터로서의 가치가 없는 것들을 최소화해야할 것 같다.
- 더 많은 데이터를 다루기 위한 일괄 처리 방식.