KLUE Relation Extraction Competition

2021/09/28 ~ 2021/10/07

9조 Quarter100

프로젝트 개요

KLUE RE Dataset으로 주어진 문장의 지정된 두 Entity의 관계를 추출, 분류하는 Task이다.

프로젝트 팀 구성 및 역할

조원	Role, Work
김다영	Model Architecture 설계 및 개선
김다인	Data Augmentation 실험 - Subject & Object Entity Masking
박성호	Data Augmentation 실험 - Back Translation으로 Task Adaptive Pre-Training
박재형	Data Augmentation 실험 - Random Delete
서동건	Model Architecture 설계 및 개선
정민지	Data Augmentation 실험 - AEDA
최석민	Model Architecture 설계 및 개선

프로젝트 수행

1. EDA and preprocessing

(1) EDA

주어진 KLUE Relation Extraction 데이터셋은 Train 32470개, Test 7765개로 이루어져 있으며 각 데이터는 Sentence, Subject Entity 정보, Object Entity 정보, Relation Label, Source로 구성되어있다.

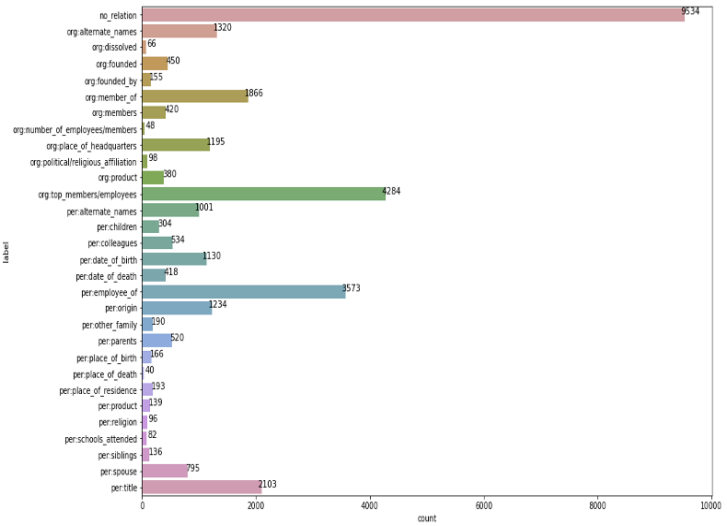
```
Sentence : 〈Something〉는 조지 해리슨이 쓰고 비틀즈가 1969 년 앨범 《Abbey Road》에 담은 노래다.
Subject_entity : {'word': '비틀즈', 'start_idx': 24, 'end_idx': 26, 'type': 'ORG'}
Object_entity : {'word': '조지 해리슨', 'start_idx': 13, 'end_idx': 18, 'type': 'PER'}
Label : no_relation
Source: wikipedia
```

Data example <출처 : KLUE RE benchmark train dataset>

각 Data의 Entity Column에는 해당 Entity Word와 문장 내 Index 위치, Entity Type 정보를 포함하고 있다. Entity Type은 단어의 개체명을 의미하며 이 Dataset에는 PER(사람), ORG(조직), DAT(시간), LOC(장소), POH(기타 표현), NOH(기타 수량 표현) 총 6가지 Type이 존재한다. 이 중 Subject Entity 단어는 PER과 ORG Type 단어만 쓰이며 Object Entity 단어는 모든 6가지 Type의 단어가 사용된다.

Relation Class	Description
no_relation	No relation in between (e_{subj}, e_{obj})
org:dissolved	The date when the specified organization was dissolved
org:founded	The date when the specified organization was founded
org:place_of_headquarters	The place which the headquarters of the specified organization are located in
org:alternate_names	Alternative names called instead of the official name to refer to the specified organization
org:member_of	Organizations to which the specified organization belongs
org:members	Organizations which belong to the specified organization
org:political/religious_affiliation	Political/religious groups which the specified organization is affiliated in
org:product	Products or merchandise produced by the specified organization
org:founded_by	The person or organization that founded the specified organization
org:top_members/employees	The representative(s) or members of the specified organization
org:number_of_employees/members	The total number of members that are affiliated in the specified organization
per:date_of_birth	The date when the specified person was born
per:date_of_death	The date when the specified person died
per:place_of_birth	The place where the specified person was born
per:place_of_death	The place where the specified person died
per:place_of_residence	The place where the specified person lives
per:origin	The origins or the nationality of the specified person
per:employee_of	The organization where the specified person works
per:schools_attended	A school where the specified person attended
per:alternate_names	Alternative names called instead of the official name to refer to the specified person
per:parents	The parents of the specified person
per:children	The children of the specified person
per:siblings	The brothers and sisters of the specified person
per:spouse	The spouse(s) of the specified person
per:other_family	Family members of the specified person other than parents, children, siblings, and spouse(s)
per:colleagues	People who work together with the specified person
per:product	Products or artworks produced by the specified person
per:religion	The religion in which the specified person believes
per:title	Official or unofficial names that represent the occupational position of the specified person

Relation class Description <출처 : KLUE 논문>



Train dataset 내 Relation label 분포

Relation에는 총 30가지 Class가 존재하며 그래프를 통해 확인할 수 있듯이 Data의 Class Imbalance가 극심하다. 또한 리더보드의 순위를 결정하는 Metric인 Micro F1 Score가 no_relation class를 제외한 나머지 Class 대해서만 F1 Score를 계산하는 것이기 때문에 일반적인 Classification Task Approach로는 문제를 잘 해결할 수 없다고 판단하였다.

(2) Preprocessing

Entity Word가 무조건 한 단어로만 구성되지 않는 것을 발견하였다. 또한 Entity Word 안에 ' , ' , ' : ' 가 포함되어 있는 경우도 있었다. 따라서 각 Entity Column의 Index 정보를 추출하여 Entity Word(Phrase) 전체를 뽑아내고 Entity Type도 동시에 저장하는 전처리를 진행하였다. 이 정보들은 추후 Input Format을 구성하고 Data Augmentation을 진행하는데 사용된다.

2. Input Format

- Typed Entity Marker with Punctuation : Special Token이나 Embedding Layer를 추가하지 않고 Entity의 위치와 Type을 표시하기 위한 방법으로 @, *, #, ^ 을 이용해 문장의 형태를 바꿔 모델의 Input으로 사용했다.
  Before) 이순신은 조선 시대 중기 무신이다
  After) @ * PER * 이순신 @ 은 조선 시대 중기 # ^ POH ^ 무신 #이다
- Add Query : Subject Entity와 Object Entity의 관계를 QA와 비슷한 형식으로 추가해주었다.
  Example) @*PER*이순신@과 #^POH^무신#의 관계 [SEP] @ * PER * 이순신 @ 은 조선 시대 중기 # ^ POH ^ 무신 #이다 [SEP]

3. Augmentation

- Subject & Object Entity Random Masking : Subject Entity나 Object Entity를 Masking하게 되면 모델이 학습 시 주변 문맥에 더욱 집중할 수 있고, 모르는 단어가 Entity로 들어오는 경우에 더욱 잘 대처할 수 있을 것이라고 가정해 시도하였다. 50% 확률로 Fold내의 Train Data 중 Entity Masking 후보를 정한 후 다시 50%의 확률로 Subject를 가릴 것인지 Object를 가릴 것인지 정하는 방법을 이용하였다.
  Before) @ * PER * 이순신 @ 은 조선 시대 중기 # ^ POH ^ 무신 #이다
  After Subject Entity Masking) @ * PER * [MASK] @ 은 조선 시대 중기 # ^ POH ^ 무신 #이다
  After Object Entity Masking) @ * PER * 이순신 @ 은 조선 시대 중기 # ^ POH ^ [MASK] #이다

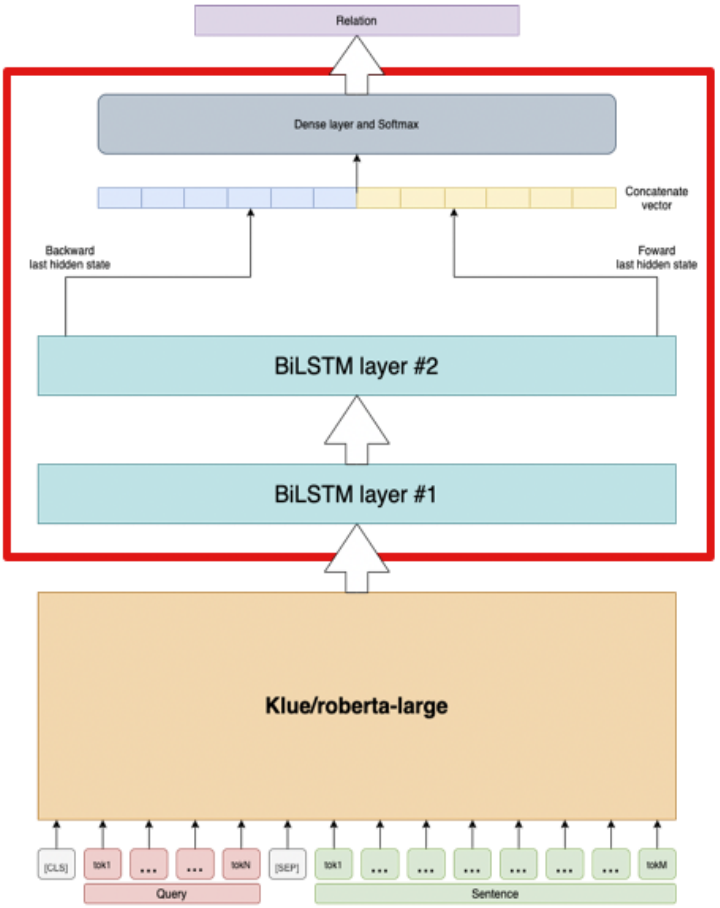
- AEDA: 문장 부호(. / , / ! / ? / ; / :)를 Okt 형태소 분석기 기준으로 토크나이징한 결과에 문장의 길이마다 다른 확률로 삽입하였다. 기존 train 문장에 AEDA로 생성한 한문장을 추가하여 문장을 두배로 늘려서 학습을 시켰다.
  Before) @ * PER * 이순신 @ 은 조선시대 중기 # ^ POH ^ 무신 #이다
  After) ! @ * PER * 이순신 @ 은 조선시대 : 중기 # ^ POH ^ 무신 # ?이다

- Random Delete : Text Data Augmentation 중 Easy Data Augmentation인 EDA에서 제시한 방법에서 Random Delete를 사용했다. 이 기법을 이용해서 노이즈를 가진 데이터들을 생성해 모델의 일반화에 도움을 줄 것이라고 가정해 시도하였다. Random Delete는 30%의 확률로 문장 안에서 Subject Entity와 Object Entity를 제외한 단어들 중 1개를 랜덤으로 삭제한다.
  Before) 이순신은 조선 시대 중기의 무신이다.
  After) 이순신은 조선 중기의 무신이다.

- Entity swap: "Subject, Object를 서로 바꿔도 문제없는 라벨", "Subject, Object를 바꾸면 라벨을 바꿔야 하는 경우", "Subject, Object 한쪽 방향만 바꿔야 하는 경우"의 경우에 대한 Swap을 기존 Data에 추가했다.
  ex) 〈 no_relation label 〉
  Before) 〈Something〉는 #^PER^조지 해리슨#이 쓰고 @*ORG*비틀즈@가 1969년 앨범 《Abbey Road 》에 담은 노래다.

After) 〈Something〉는 @*PER*조지 해리슨@이 쓰고 #^ORG^비틀즈#가 1969년 앨범 《Abbey Road》에 담은 노래다.

4. Model Architecture



Query + Sentence 형식의 Input Sequence를 Klue/roberta-large 모델에 Feed하여 출력된 모든 Output을 BiLSTM Layer에 Feed한다. Top Layer에서 출력된 Forward, Backward Last Hidden State Vector를 Concatenate하여 Dense layer와 Softmax 연산을 거쳐 최종 Entity Relation을 Classify하게 된다.

- Loss Function : Label Smoothing Loss
- Optimizer : AdamW Optimizer
- Scheduler : Warmup and Linear Decay

5. Ensemble

a. Stratified K-Fold & OOF(Out-of-Fold) Prediction:
레이블 분포가 불균형함을 EDA로 확인하여 원본 데이터와 유사한 분포를 Train, Val Set에서도 유지하기 위해 Stratified K-Fold를 채택했다. 훈련 후 OOF Prediction으로 모델을 평가했다.

b. Soft voting
4가지 Augmentation을 각각 적용한 모델 4개와 적용하지 않은 기본 모델 1개를 각각 K-fold Ensemble한 후 0.4, 0.15, 0.15, 0.15, 0.15 의 Weight로 Soft Voting Ensemble하였다.

6. 자체 평가 의견(프로젝트 소감 및 느낀점 등)

팀원 모두가 각자 하고자 하는 Method와 Approach들을 주도적으로 실험하고 결과를 공유하며 협업과정에서 적극적인 업무분담의 자세를 보여 좋은 결과를 얻었다. 다만 아이디어를 논의하고 결정하는 단계에서 시간을 많이 소요했고 제시된 방법론들의 Develop를 다양하게, 깊게 진행하지 못한 점에서 아쉬움을 느꼈다.

협업의 과정과 결과를 공유하는 것에 미숙해 기록을 제대로 남기지 못한 점도 아쉬웠다. 이러한 부분을 다음 MRC 프로젝트에서는 보완하고자 한다.