개인회고록 KLUE-RE

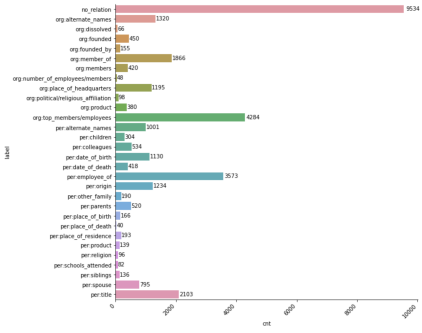
이번 프로젝트의 목표

철저한 근거를 바탕으로한 실험 설계와 기록이 목표였습니다.

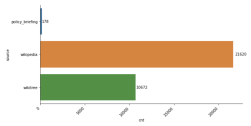
프로젝트 A - Z

1. 기초 EDA

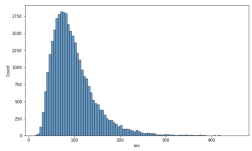
탄탄한 시각화를 통해, 데이터의 관점에서 문제가 무엇인지 파악하려고 노력했습니다.



label 분포도



source별 label

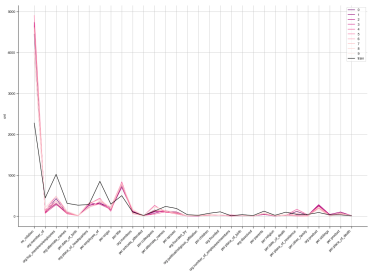


문장길이

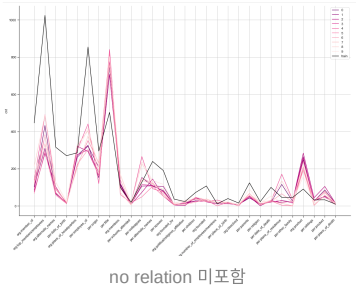
2. Validation score와 test score의 차이

그 다음에는 validation score와 test score의 간극에 대한 이유를 찾으려고 했습니다. 이유를 알면 차이를 줄이는 것 역시 가능할 것이라고 생각했습니다.

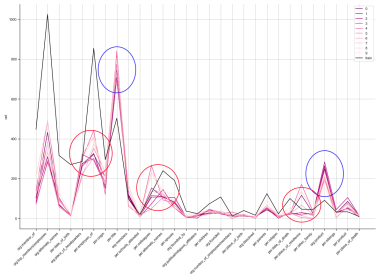
처음으로 생각했던 것은 분포가 다르기 때문일 것이라고 생각을 했습니다. 그래서 분포를 찾아보았습니다.



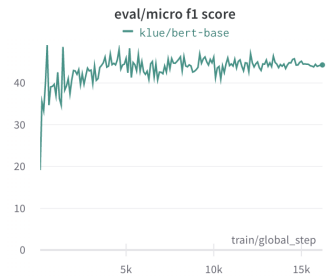
no relation 포함



no relation 미포함



submission마다 차이가 큰 영역(문제가 있을만한 영역

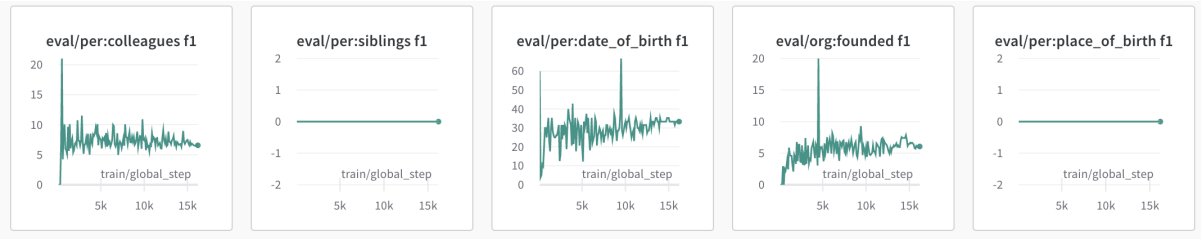


submission 분포의 f1 score

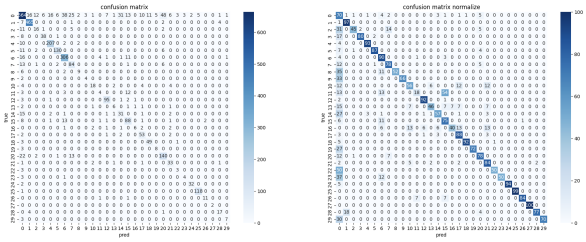
validation 분포를 상위 10개의 submission의 평균 label에서 분포를 가져와서 나타내 보았습니다. test f1 score가 60점 대 였던, bert-base모델이 약 40점 정도의 f1 score를 나타냈습니다. 지금 생각해보면, 여기서 좀 더 실험을 진행하여 좀 더 좋은 validation set을 만들 수 있었을 것 같습니다. 하지만, 생각이 더 나아가지 못하고 재현 실패라는 생각으로 좋은 validation set만들기를 마무리 지었습니다.

3. data의 관점에서 개선점 찾아보기

분포를 찾아보고서 label별 학습이 되는 정도가 다를 것이라는 생각에 binary f1으로 모든 label을 살펴보았습니다.



f1이 낮게 유지되는 label들은 data가 적은 label들이었습니다. micro f1을 사용했기 때문에 무시해도 된다고 생각했습니다. 어떤 label에서 얼만큼 맞고 틀리는지를 알기 위해 confusion matrix를 그려보았습니다.



(좌) count 수 (우) 백분율

알 수 있는 점

1. 비율적으로 많이 틀리는 label은 보통 수가 적습니다.→ score적인 측면에서 무시해도 될 것으로 생각됩니다.
2. imbalancing이라고 생각했던 부분들은 no relation vs 나머지 label 간의 문제였습니다.
    a. 아마도 imbalancing과 noise의 문제로 생각되었습니다.

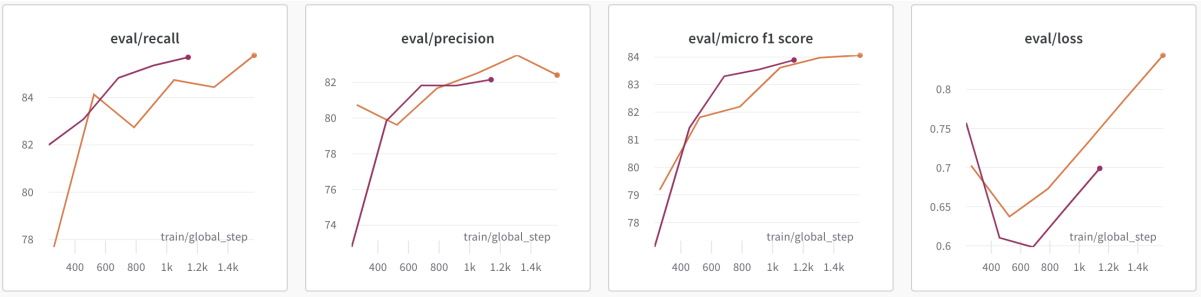
여기서 떠올린 실험 아이디어

f1 score는 precision과 recall의 조화 평균, 그리고 조화평균은 낮은 score에 가중치를 주는 평균 연산, 그리고 현재 baseline인 roberta-small의 precision, recall, f1은 각각 82.1, 87.5, 83.9로 recall을 조금 줄이면서 precision을 늘릴 수 있다면 f1에 상당히 긍정적인 영향을 끼치지 않을까요?

confusion matrix에서 1행(no relation이 다른 label로 예측), 1열(다른 label을 norelaiton으로 예측)이 문제라고 생각하였습니다. 그리고 1행은 precision의 영역, 1열은 recall의 영역이 될 것 이라고 생각했습니다. (이 부분이 헷갈리는데, 저희 평가방식에서는 no relation을 제외한 f1을 구하기 때문에 1행이 다른 label들에 대한 precision으로 작용하는 것으로 이해했습니다.)

no relation을 늘리면 다른 label을 no relation으로 예측을 많이 하는 (recall 하락) 대신, no relation을 다른 label로 예측을 덜할 것이기 때문에 precision의 개선이 있을 것으로 생각했습니다.

4. no relation undersampling



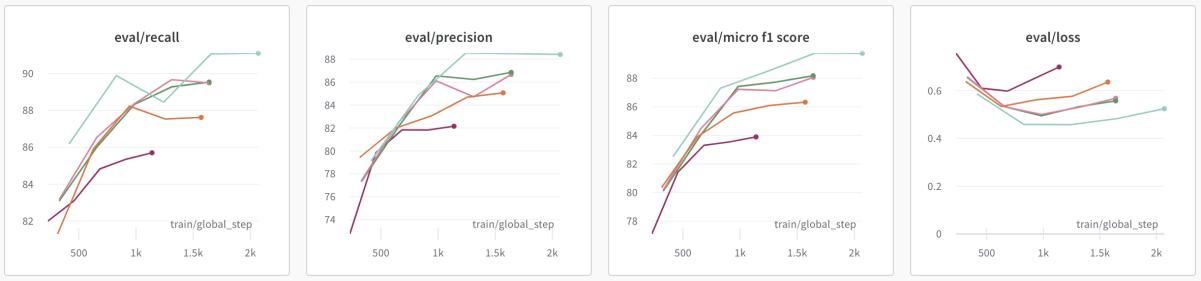
(보라색) baseline (주황색) oversampling

예측했던대로, 다른 label을 no relation으로 예측하는 비율은 늘었지만, no relation을 다른 label로 예측하는 비율이 줄어들었습니다. 결과적으로 precision은 다소 향상, recall은 다소 감소하여 f1 score가 약 1점정도 오르는 결과가 나왔습니다.

하지만 loss가 너무 빨리 올라가는 양상이 보였고, 이는 test에서 사용하기 적절하지 못한 방법으로 생각되었습니다. 같은 데이터를 중복적으로 사용하게 되어 과적합이 심하게 일어난 것으로 생각되었습니다. 다음으로 든 생각은 양질의 데이터가 필요하다는 생각이 들었습니다.

5. back translation

eda, aeda등의 augmentation보다 시간은 오래걸리지만, 가장 유사한 의미를 가지면서 다른 문장구조를 만들어줄 augmentation은 back translation이라고 생각하였습니다.



(보라색) baseline (주황색) en backtranslation (핑크색) ja backtranslation (청록색) en+ja / ja:japanese en:english

validation에서는 augmentation을 넣을 때마다 성능이 크게 올라갔습니다.

또한 test score는 base line 61.20 → ja back translation 62.50으로 약 1.3정도의 f1 스코어 향상이 있었습니다. 하지만 en + ja를 추가해주었을 때는 61.04로 오히려 0.15정도 떨어진 모습을 보였습니다.

큰 모델로 실험했을 때, 이상하게 스코어가 오르기도 하고 떨어지기도 하는 경험을 하였는데, 작은 모델로 실험을 하는 것이 의미가 없어서인가? 하는 의문이 들었습니다. (이유경 멘토님께서 더 큰 모델은 더 많은 데이터를 넣어주어야한다는 답을 주셨습니다.)

이 부분은 어떤 기준으로 떨어지고 오르는지에 대해서 좀 더 실험이 필요한 것으로 보입니다.

이전 stage에 비해서 나아진 점

- 더 체계적으로 접근했습니다.
- 더 많이, 더 세세히 기록하였습니다. → 팀원들 간의 공유로 이어져 공동의 지식을 키워낼 수 있었습니다.
- 가벼운 baseline을 설정하여 더 많은 실험을 더 빠르게 할 수 있었습니다.
- 대회의 흐름을 좀 더 이해했기 때문에, 시간을 비교적 효율적으로 사용할 수 있었습니다.

아쉬웠던 점

- data로부터 몇 가지의 insight를 찾았으나, 이 것이 모델의 성능개선으로 이어지지 못했습니다. 시간이 부족했고, 다양한 해결책에 대한 생각을 하지 못했습니다.
- 좋은 validation set이 필요하다고 저번 대회 때 반성했으나, 이번 대회도 좋은 validation을 만들지 못했습니다. straigtify하게 뽑은 validation이 쓸만하다고 생각하여 더 개선하지 않았습니다.
- model을 제대로 다뤄보지 못했습니다. 모델에 대한 막연한 두려움이 있었고, 다른 팀원들이 모델링을 했기 때문에 제 역할을 해야겠다고 생각했습니다. 다음 대회 때는 긴 기간들을 이용하여 모델에 관련된 개선도 해봐야겠다는 생각을 합니다.

다음 P-Stage에서는?

- 찾아낸 insight를 성능개선으로 발전시키고 싶습니다.
- 가벼운 모델을 초기에 설정하고, 다양한 실험에 대한 세팅을 할 것입니다.
- data에 관한 많은 것들에 대한 실험과 검증을 대회 초반에 해야겠습니다.
- 모델링을 꼭 해봐야겠습니다. 기간이 긴 만큼 더 배우고, 더 다양하게 해보고 싶습니다.