

Data-Centric 랩업 리포트

NLP 07조 양서현 | 이상경 | 이승백 | 이주용 | 정종관 | 정지영

1. 프로젝트 개요

- 모델 구조의 변경 없이 **Data-Centric** 관점으로 텍스트의 주제를 분류하는 태스크

2. 프로젝트 수행 절차 및 방법

- 데이터 재 라벨링

1) 전체 라벨 직접 검수

- 1000개씩 7개의 label이 붙은 train data
- 각 label 별로 정렬하여 수작업으로 라벨에 문제가 있는 데이터 탐지
- 라벨에 문제가 있는 데이터의 경우 두 가지 방법으로 재 라벨링
 - Open Ko-LLM 중 SOTA모델(LDCC SOLAR 10.7B) 사용하여 예측 :0.8360
 - 사람이 직접 재 라벨링 :0.8441
 - 사람이 직접 라벨링한 결과가 더 성능이 좋았음

2) Cleanlab 라이브러리 이용

- train 데이터셋으로 학습 후 train 데이터셋으로 evaluate한 결과와 비교하여 cleanlab 내 package를 이용하여 라벨링 이슈를 탐지
- mislabel 의심 데이터 삭제, evaluate 데이터로 재학습, Pseudo-labeling 적용
- 실험 결과, 0.8384 → 0.8323, 8380 로 f1 score 하락

- 데이터 클리닝

1) g2p 제거

- g2p (Grapheme-to-Phoneme, 글자를 발음나는 대로 적는 방법) 형태로 train data에 추가된 노이즈를 반대의 p2g (Phoneme-to-Grapheme) 모델을 통해 원상복구
- 원상복구된 학습 데이터로 실험 결과, 0.8384 → 0.8016로 f1 score 하락
- Test data에 여전히 존재하는 g2p 노이즈에 모델이 효과적으로 대응하지 못한 것으로 추측

2) UNK 토큰 단어 제거

- 토큰나이징했을 때 unknown 토큰되는 단어 제거
- 실험 결과, 0.8384 -> 0.8376 로 f1 score 하락

3) 한자 변환

- 한자를 한글로 치환, “北 -> 북”, 0.8384 -> 0.8344
- 한자와 한글을 혼합, “北 -> 北(북)”, 0.8384 -> 0.8363
- Test data에 있는 한자 데이터는 모델이 학습한 형식과 다르므로 모델이 효과적으로 처리할 수 없다고 추측

- 데이터 전처리

1) 형태소 기반 Subword 토큰나이징

- 강의 내용 중 Morpheme-based Subword Tokenization 기법 기반
- 형태소 분석기(Mecab-ko)를 통해 토큰라이저에 들어가기 전에 한 번 더 형태소 단위의 사전 토큰라이징을 진행 후, 본 토큰라이저에서 BPE(Byte Pair Encoding) 적용
- 실험 결과, 0.8384 → 0.8280로 f1 score 하락
- Test data는 동일 방식으로 전처리를 할 수 없어서 점수가 소폭 하락한 것으로 추측

● 데이터 증강

1) g2p로 노이즈 데이터 생성

- p2g 모델을 통해 g2p된 데이터를 원상 복구 한 후에 모든 데이터를 다시 g2p 모델을 통해 노이즈 데이터 생성
- 기존 일부데이터로만 생성되었던 노이즈 데이터를 전체 데이터로 생성 하면서 노이즈 데이터의 양 증가

2) AI hub 뉴스 기사 기계독해 데이터

- AI hub에 있는 외부 데이터셋을 통해 데이터 증강
- train 데이터의 라벨과 규칙이 다를 것이라고 예상하여 train 데이터셋으로 만든 모델로 다시 라벨링
- 실험 결과, 0.8437 → 0.8414로 f1 score 하락
- 기존 모델을 통해 다시 라벨링 한것이므로 애매한것을 더욱 못 맞추게 되어 오히려 성능 하락한 것으로 보임


3) Back Translation

- 기존 train 데이터 셋을 back translation(영어, 영어-스페인어, 일본어)하여 데이터 증강
- 실험 결과, 0.8454 → 0.8425로 f1 score 하락
- 번역이 정확하지 않아 전체적인 데이터의 질이 떨어져 오히려 성능이 저하

4) AEDA

- 기존 train 데이터셋에 특수문자(. , ? : ! ,)를 문장 중간에 삽입하여 데이터 증강
- train 데이터를 p2g 데이터로 클리닝할 때, ‘...(3개 문자)’와 ‘...(1개 문자)’가 의미는 같지만 다른 단어로 간주하여 제거하지 않았을 때의 f1 score가 높았던 것에 착안
- 실험 결과, public score 기준 0.8455 → 0.8461로 f1 score 상승

4. 프로젝트 수행 결과

순위	팀 이름	팀 멤버	f1 ↕	accuracy ↕	제출 횟수	최종 제출
3 (1 ▼)	NLP_07조		0.8441	0.8474	66	3d

- Private macro f1 score 0.8441로 최종 순위 3위로 마감