Lv2. NLP Data Centric: Topic Classification

1. 프로젝트 개요

주제

주제 분류 프로젝트: 모델 구조의 변경 없이 Data-Centric 관점으로 텍스트의 주제를 분류하는 태스크

일정

2023.05.24 (월) 10:00 ~ 2023.06.01 (목) 19:00

데이터

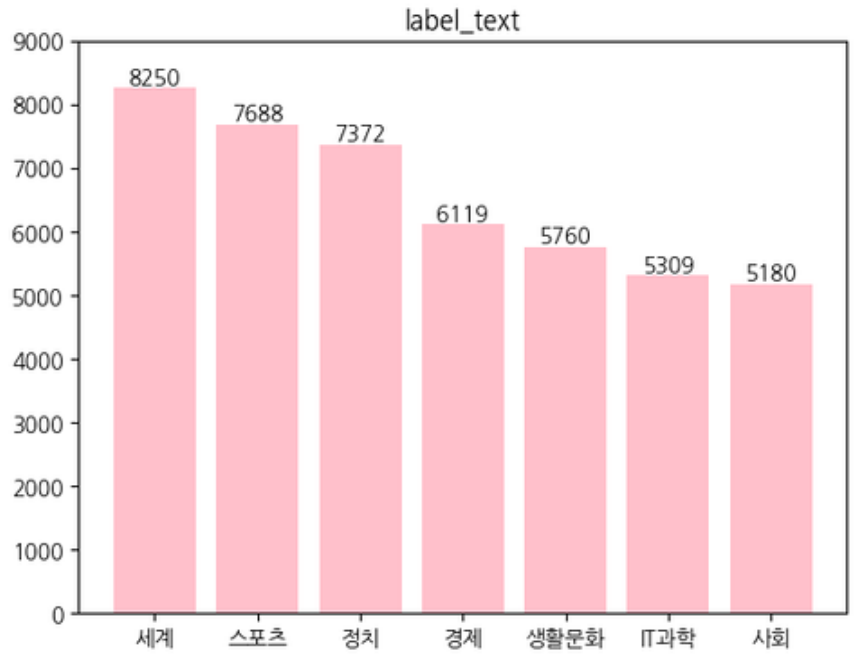
KLUE(Korean NLU Benchmark) TC(Topic Classification) dataset + noise data

- train : 45,678 개
- test : 9,107 개
- 데이터 구성

ID	text	target	url	date
ynat-v1_train_08697	배구 남자대표팀 감독 공모에 임도헌 코치 단독 지원	스포츠	https://sports.news.naver.com/news.nhn?oid=001...	2019.05.24 17:17

- target

target	meaning	count
0	IT과학	5309
1	경제	6119
2	사회	5180
3	생활문화	5760
4	세계	8250
5	스포츠	7688
6	정치	7372



평가 방법

1. Macro F1-Score: KLUE Topic classification의 공식 리더보드와 동일한 평가 방법 사용
   - F1 score: precision과 recall의 조화 평균

$$F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

   - Macro f1-score: 각 class에 동일한 가중치를 부여하여 계산한 class별 f1 score의 평균
   - Public data와 Private data는 dev dataset에서 무작위로 50:50으로 선정되며, 각각을 통해 Puplic F1과 Private F1을 평가.
     - Private F1: KLUE-TC dev 데이터에서 무작위로 50% 선정
     - Public F1: KLUE-TC dev 데이터에서 무작위로 50% 선정
2. Baseline model 변경 여부 검토

협업

- 데이터 버저닝 - HuggingFace Datasets
- 코드 버저닝 - GitHub
- 실험 관리 및 기록 - WandB & Notion

2. 프로젝트 진행 과정

Data Cleaning

- Labeling error를 더 잘 탐지하기 위해서 G2P Noise Cleaning을 먼저 진행하고, 그 다음 Labeling error를 제거하는 작업을 진행.

[G2P Noise Cleaning] - 3가지 방법으로 진행

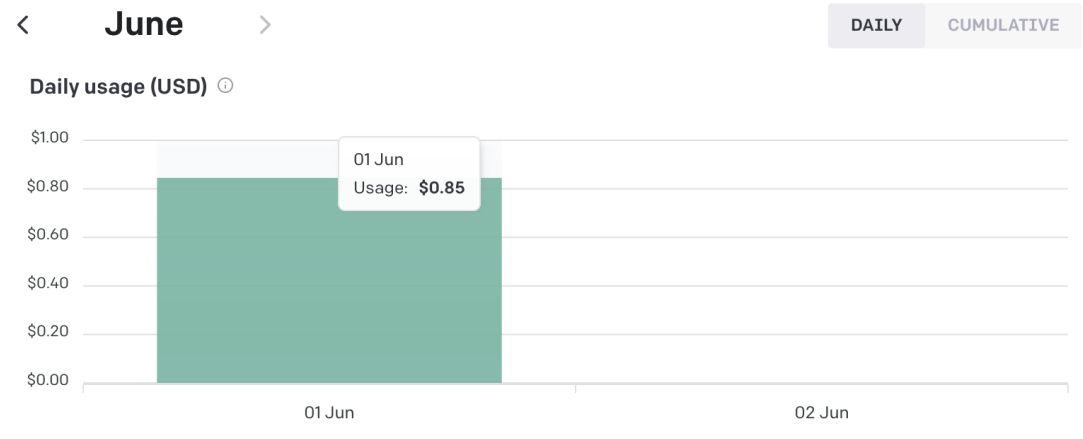
1. **GPT-3.5-turbo**
    - 비용(시간 및 무료 API call 자본) 부족으로 '사회'로 레이블링된 3600여 개 데이터에 대해서만 GPT-3.5-turbo의 도움으로 데이터 클리닝을 진행.
    - 프롬프트 예시

```
full_prompt = """아래 각 text에 대해 노이즈가 포함되어 있다고 판단되면 원 문장으로 수정해줘.

이때, 아래에 있는 rules를 만족해야만 해. Examples를 참고해서 문장을 잘 가다듬어 줘.

[Texts]
11,천년의 비상…부안군 고을 이름 600주년 기념사업 추진
16,여수 윤화류 보관 창고 화재현장서 치손는 거므 년기
32,코레일 산불 피해지역 상생발전 위한 여행사 간담회 개최
36,총선 D7 진해 후보들 소외 공감·부산편입 이견
76,에쓰오일 서울 사회복지협의회에 1억3천700만원 기부

[Rules]
1. 어법 및 어문 규정을 준수할 것
2. 노이즈가 포함된 문장은 수정될 원 문장과 어절 및 음절 수가 최대한 같을 것
3. 알맞은 문맥을 추론할 것
4. 구개음화, 된소리 되기와 같은 현상이 일어난 노이즈 문장이 많음을 인지할 것
5. '종합'과 같은 단어가 문장 끝에 붙은 경우, 의미가 없을 확률이 높으니 생략할 것
6. 불필요한 특수문자는 최대한 생략하며, 문장 끝에 마침표를 생략할 것

[Example 1]
(수정 전) 방통시믜위 불법 때부업 쩡보 등 오백꾸시보건 삭쩨 등 요구
(수정 후) 방통심의위 불법 대부업 정보 등 595(오백구십오)건 삭제 등 요구

[Example 2]
(수정 전) 여수 윤화류 보관 창고 화재현장서 치손는 거므 년기
(수정 후) 여수 윤활유 보관 창고 화재 현장서 치솟는 검은 연기

Provide them in JSON format with the following key:
index and text.
"""

chat_model = ChatOpenAI()
response_chat_open_ai = chat_model.predict(full_prompt)
```

- GPT-3.5-turbo 답변 예시

```
[
    {"index": 11, "text": "천년의 비상… 부안군 마을 이름 600주년 기념사업 추진"},
    {"index": 16, "text": "여수 윤활유 보관 창고 화재 현장서 치솟는 검은 연기"},
    {"index": 32, "text": "코레일 산불 피해 지역 상생발전 위한 여행사 간담회 개최"},
    {"index": 36, "text": "총선 D-7 진해 후보들 소외 공감·부산편입 이견"},
    {"index": 76, "text": "에쓰오일 서울 사회복지협의회에 1억 3천700만원 기부"}
]
```

- 결과
    - 3628개 행에 대해 데이터 노이즈 교정을 진행하였으며, 4시간 정도 소모. API 호출 비용은 0.85 달러.



- 성능 비교 측정

  대조군: baseline 데이터

  실험군 1: baseline 데이터의 일부 행 교체('사회' 분류 데이터에 대해서만 데이터 노이즈 교정한 데이터)

  실험군 2: baseline 데이터 + 증강 데이터('사회' 분류 데이터에 대해서만 데이터 노이즈 교정한 데이터)

	Validation F1	Public LB F1	Private LB F1
실험군	**0.838**	**0.8729**	**0.8597**
대조군 1	0.8278	0.8675	0.8537
대조군 2	0.8317	0.8603	0.8418

- 결과 해석 및 논의
  - GPT-3.5-turbo를 사용하여 일부 데이터를 클리닝한 데이터로 학습한 것의 성능이 오히려 떨어지는 문제가 발생.
  - 실제 원본 '사회' 레이블 데이터 vs. GPT-3.5-turbo로 클리닝된 '사회' 레이블 데이터 비교.

```
원본 || 클리닝

여수 윤화류 보관 창고 화재현장서 치손는 거므 년기 || 여수 윤활유 보관 창고 화재 현장서 치솟는 검은 연기
총선 D7 진해 후보들 소외 공감·부산편입 이견 || 총선 D7 진해 후보들 소외 공감 부산 편입 이견
추락한 日 F35A 전투기 이전에도 2차례 긴급착륙 전력 || 추락한 일본 F35A 전투기 이전에도 2차례 긴급착륙 전력
방정오 측 故장자연과 통화 보도 사실무근…법적대응 || 방정오 측 故장자연과 통화 보도 사실무근…법적 대응
폭언 논란 권용원 금융투자협회장 사퇴 안한다종합 || 폭언 논란 권용원 금융투자협회장 사퇴 안한다
게시판 EBS대전방송 인재 양성 위한 업무협약 || 게시판 EBS 대전방송 인재 양성 위한 업무협약
게시판 과천과학꽈 니천시뷩년 최우수 채기무녕기관 선정 || 과천과학관 니천시 빅뱅 최우수 채기무녕기관 선정
與 檢과 전면전 태세…특검카드 꺼내고 공정수사특위 꾸리고종합 || 여당 검찰과 전면전 태세…특검 카드 꺼내고 공정수사특위 꾸리고
카드뉴스 방금 사용한 제 스마트기기 정보 동의 구하셨나요 || 방금 사용한 제 스마트 기기 정보 동의 구하셨나요
정부 부동산시장 선별적·단계적 때응…내주 대책 빨표두보 || 정부 부동산 시장 선별적·단계적 대응…내주 대책 빨리 보도
신문방송편지빈혀푀 전국 찌역씬문 편집꾹짱 포럼 || 신문방송편지빈혀푀 전국 지역종합 뉴스 포럼
폴리뉴스 19주년 창간 기념식… 상생과통일포럼 박원순 초청 특강 || 폴리뉴스 19주년 창간 기념식…상생과 통일 포럼 박원순 초청 특강
황창규 KT 회장 국쩡농단 스캔드 렬루에 유감 표명 || 황창규 KT 회장 국정 농단 스캔들에 유감 표명
타인 며녁쎄포로 암치료 임상일상 결과 국쩨학쑬지에 || 타인 며느리 쎄포로 암치료 임상일상 결과 국제학술지에
KBS 이사회 여당 우위로 재편…경영진 해임 수순 밟나종합 || KBS 이사회 여당 우위로 재편 경영진 해임 수순 밟나
집창촌 유착 폭로에 비리 의혹 경찰 명단까지…파문 확산종합 || 집착촌 유착 폭로에 비리 의혹 경찰 명단까지 파문 확산
롯데 檢수사로 자금조달 계획 차질 || 롯데 검수 사로 자금 조달 계획 차질
```

  - 첫 문장은 two-shot example로 준 문장 중 하나이며, 원하는 대로 데이터 클리닝이 된 것을 확인.
  - 특징
    - 문장 끝에 '종합'이 붙는 것을 제거해달라는 Rule 5 요구를 충실히 수용하고 반영하는 것을 확인.
    - 띄어쓰기는 요구 사항에 직접적으로 없었으나, Rule 1의 어법 및 어문 규정을 신경 써달라는 요구를 반영한 것으로 추측.
    - 다만 GPT 모델이 어법 및 어문 규정에 맞는 문장 생성에 있어서의 성능이 다른 벤치마크에 비해 점수가 떨어지는 편임을 상기할 필요는 있음.
    - Rule 2처럼 음운 요소를 고려해야 한다는 요구가 없는 경우, 원 문장을 명사형이 아닌 동사형으로 끝맺음 해준다든가 문장 길이가 짧아지거나 길어지는 답변을 내는 경우가 종종 있었는데, rule 2가 추가되는 경우 최대한 음절 수를 비슷하게 맞춰주려고 하는 일관성을 보임.

        - Rule 3의 경우, 된소리 되기(ㄱ → ㄲ, ㄷ → ㄸ, ㅂ → ㅃ, ㅅ → ㅆ, ㅈ → ㅉ)와 구개음화(굳이[구지] 등)에 대한 예시를 더 길게 주면 조금 더 노이즈 수정 성능이 좋아지는 것으로 보였으나, 과금 요인 등을 고려해 중간 수정 등이 포함된 것.
        - Rule 5의 불필요한 특수 문자를 제거해달라는 요구를 이해하여 직접적으로 요구하지 않은 특수문자도 생략하는 것을 확인.
    - 한계
        - 한자의 경우, 與를 여당으로, 檢을 검찰로 제대로 치환해주는 경우도 있었으나, 단순히 檢을 검으로 치환하는 경우도 존재하였음.
        - 데이터 클리닝을 진행한 문장 단위가 bsz = 40이었는데, 모델의 답변이 40개씩마다 어느 정도의 경향성이 보이는 것을 확인.
- 결론
    - 프롬프트 요구를 더 구체적으로 가다듬을 수 있는 능력의 필요성 확인.
    - API call을 고려하여 프롬프트 토큰 수를 최소화할 기술의 필요성.
    - 이번 실험 통해 GPT-3.5-turbo를 활용한 데이터 클리닝의 가능성을 확인.
    - 특히 노이즈가 있는 데이터를 깔끔하게 정리하는 데에 있어서 어느 정도의 효과를 확인.
    - 그러나 클리닝된 데이터로 학습한 모델의 성능이 오히려 하락하는 문제가 발생한 것을 확인.
    - GPT-3.5-turbo가 데이터 클리닝을 완벽하게 수행하지 못하거나, 또는 너무 과도한 클리닝으로 인해 중요한 정보를 잃었을 수도 있음을 시사.
    - GPT-3.5-turbo가 일관성을 유지하면서도 다양한 규칙을 충족시키려는 모습을 확인. 이는 높은 유연성과 적응력을 가진 모델임을 보여줌. 그러나 어법 및 어문 규정에 맞는 문장 생성에 있어서의 성능이 아직은 완벽하지 않음이 나타남.

2. **정상 text와 G2P text로 이루어진 pair를 학습하여 P2G 모델 개발**
    - T5 모델을 사용해 정상 텍스트와 발음열을 학습 후 변환하고자 함.
    - 적절하지 않은 사전훈련모델 선택으로 인해 학습 실제로 프로젝트에 적용하기에는 학습 결과가 미흡했음.
    - Huggingface `transformers.Seq2SeqTrainer` 를 사용해 학습.

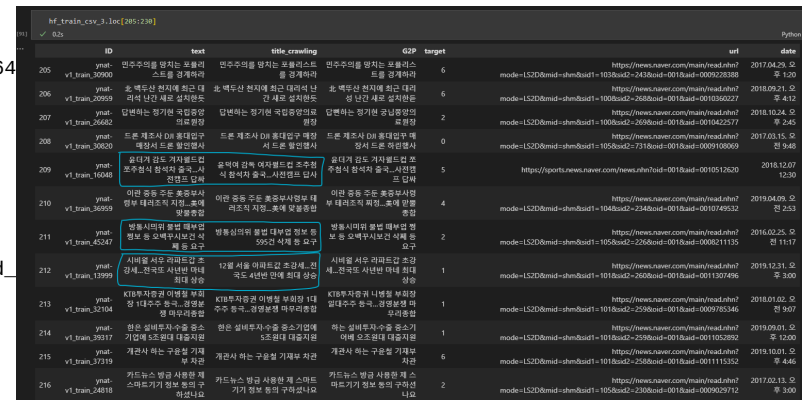
    Dataset 구성

    - 대회에서 주어진 `train.csv` 에서 추출한 G2P가 가해지지 않은 일반 텍스트 40353개와, AI Hub에서 제공하는 뉴스 기사 기계독해 데이터셋에서 추출한 헤드라인 텍스트 109323개를 사용(총 149676개).
    - 한국어 G2P 변환을 위한 모듈인 g2pk를 사용해 일반 텍스트를 발음열로 변환.
    - 발음열을 일반 텍스트로 변환하기 위해 발음열을 `input`, 일반 텍스트를 `target` 으로 구성.
    - 데이터 예시

```
input,target
유튜브 내다 리일까지 크리에이터 지원 공가 누녕,유튜브 내달 2일까지 크리에이터 지원 공간 운영
내년부터 국까RD 평가 때 논문건수는 바녕 안는다,내년부터 국가RD 평가 때 논문건수는 반영 않는다
김명자 시님 과총 회장 월로와 절믄 과학짜 지혜 모을 껃,김명자 신임 과총 회장 원로와 젊은 과학자 지혜 모을 것
⋮
```

3. **URL 이용해서 기사 제목 크롤링 후 text 전처리(주어진 데이터와 최대한 유사하게 특수문자를 제거)**

```
def title_crawling(url) :
    news = requests.get(url, headers={"User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64
    news_html = BeautifulSoup(news.text,"html.parser")

    if 'sports' in url :
        title = news_html.select_one("h4.title")

    else :
        title = news_html.select_one("#ct > div.media_end_head.go_trans > div.media_end_
        if title == None:
            title = news_html.select_one("#content > div.end_ct > div > h2")

    # html태그제거 및 텍스트 다듬기
    pattern1 = '<[^>]*>'
    title = re.sub(pattern=pattern1, repl='', string=str(title))
    pattern2 = r'[,\[\]\(\)\"\'\-~\?+/:]'
    title = re.sub(pattern = pattern2, repl='', string=title)
    title = title.replace('&lt;','')
    title = title.replace('&gt;','')
    title = title.replace('&amp;','')

    return title
```



[Labeling Error Cleaning]

- Cleanlab 라이브러리를 이용하여 라벨링 에러를 제거.
    - label_quality 0.011 보다 낮은 example 모두 대체.

	데이터 설명	public 리더보드 f1	private 리더보드 f1	eval f1
원본 데이터	라벨링 에러 있음, G2P noise 있음	-	-	0.8371
(baseline) G2P **노이즈 제거 후 데이터**	라벨링 에러 있음	-	-	0.8455
라벨링 에러 제거 후 데이터 version 1	약 1200개의 라벨링 오류가 있는 데이터를 대체함	**0.8798**	0.8542	0.8632
라벨링 에러 제거 후 데이터 version 2	1,310개의 라벨링 오류가 있는 데이터를 대체함	0.8750	**0.8658**	**0.8750**

Data Preprocessing

실험명	예시	validation	leaderboard
baseline	묘비명 알리…故무하마드 알리 10만명 추모받으며 영면종합	0.8371	0.8738
한자 제거	묘비명 알리…무하마드 알리 10만명 추모받으며 영면종합	0.8265	
한자 to 한글	묘비명 알리…고무하마드 알리 10만명 추모받으며 영면종합	0.839	0.8685
특수문자 제거	묘비명 알리故무하마드 알리 10만명 추모받으며 영면종합	0.8365	0.8590
특수문자 to 공백	묘비명 알리 故무하마드 알리 10만명 추모받으며 영면종합	0.8379	
‘종합~’ 제거	묘비명 알리…故무하마드 알리 10만명 추모받으며 영면	0.8367	

- 한자 제거: baseline에 대한 사후 분석 결과 실제값이 `사회` 인데 `정치` 로 잘못 예측한 데이터가 ‘한자’를 포함하는 경우가 많았음. 이에 따라 한자를 제거하는 전처리를 진행한 후 학습을 진행해보았으나, `정치` label을 구분하는 지표 역할을 했던 한자가 사라짐으로써 성능이 저하됨.
- 한자 to 한글: 한자를 단순히 제거했을 때 성능이 저하된 것에서 착안하여, 한자를 한글로 바꾸는 전처리를 진행한 후 학습을 진행. 목표했던 `사회` label의 지표들이 개선되며 전체 validation set에 대한 f1 score가 향상됨.
- ‘종합~’ 제거: text의 마지막 어절에 기사 제목에는 포함되지 않는 것으로 보이는 ‘종합’, ‘종합2보’ 등의 글자가 포함되어 있어 이를 제거하는 전처리를 진행한 후 학습을 진행해보았으나, 성능의 유의미한 변화는 없었음.
- 무엇보다 데이터 전처리와 관련한 시도들 중 일부는 validation f1 score의 향상에 기여했으나, test 데이터셋에는 전처리를 진행할 수 없었기 때문에 모든 경우에 대하여 Leaderboard에서는 오히려 성능이 저하되는 모습을 보임.

Data Augmentation

[Easy Data Augmentation]

- koeda 라이브러리를 이용하여 Random Deletion 방법과 Random Swap 방법으로 데이터를 증강. Baseline은 증강 전 데이터(data cleaning 후)를 의미.

데이터 설명	leaderboard	validation
Baseline	0.8798	0.8632
RD training dataset만 10% 증강	0.8687	0.8611
RS training dataset만 10% 증강	0.8675	0.8616

- 실험결과 및 추론
  - 데이터 증강 후 baseline의 성능보다 하락.
  - 데이터 증강 후 성능이 개선되지 않은 이유는 모델이 잘 맞추지 못하는 label을 고려하지 않고 무작위로 데이터를 증강시켰기 때문일 것으로 추정.

[AI Hub 데이터]

- AI Hub내 공개데이터 **뉴스 기사 기계독해 데이터** 중 기사 제목(doc_title)과 분류(doc_class.code)를 필터링해 사용.
- 데이터셋 예시 (109,323 rows X 2 columns)

```
{
"doc_id": "01100201.20210502175653001",
"doc_title": "민주당 새 당대표에 86맏형 송영길…최고위원은 ‘친문’ 강화",
"doc_source": "국민일보",
"doc_published": 20210502,
"doc_class": {
  "class": "한국언론진흥재단 빅카인즈 뉴스기사",
  "code": "정치"
},
```

번호	데이터 설명	leaderboard	validation	비고
1	기본(baseline)	0.8769	0.8371	
2	기본 + ai-hub(89,000개)	0.8441	0.8056	가장 많은 사회라벨을 2만개 랜덤하게 없앴다.
3	기본 + ai-hub(9,000개)	0.8545	0.8258	[1500, 1500, 1500, 1500, 1000, 1000, 1000] 비율로 추가
4	기본 + ai-hub(9000개, cleanlab 기준 label_quality 높은 순서대로)	0.8705	0.85766	[1500, 1500, 1500, 1500, 1000, 1000, 1000] 비율로 추가
5	기본 + ai-hub 사회 1500개만 추가	0.8768	0.8624	cleanlab 기준 label_quality 높은 순서대로

- 실험 결과
  - 결론적으로 baseline보다 높은 성능을 보이지 못함.
  - 뉴스기사 데이터가 원래 데이터인 KLUE TC 데이터와 거의 유사하여 성능향상에 크게 기여할 것이라 생각했으나, 실제로는 다른 결과가 나타남.
  - 많은 데이터를 추가할수록 f1 score가 낮았고 원본 데이터에 가까울수록 f1 score가 높았음.
- 예상되는 이유
  - 대회의 특성 상 데이터 외에 모든 조건(모델 크기, 하이퍼파라미터, 학습 FLOPs 등)이 고정되어 있었음.
  - 그런데 학습 데이터가 증가함에 따라 적절한 performance를 낼 수 있는 조건이 달라지기 때문에 이런 차이가 반영된 것.
  - 또한 AI Hub 데이터 자체의 라벨링 오류도 꽤 존재함.
    - 실제 실험에서도 3번 실험에서 **랜덤추출**하지 않고 cleanlab 라이브러리 학습을 통해 labeling error가 가장 적은 순서대로 층화추출한 데이터가 valid, leaderboard score가 훨씬 높았음.
  - **데이터 크기에 알맞은 조건 설정**과 **공공데이터를 그대로 믿지 말고 정제를 거쳐야한다**는 점을 깨달았음.

[역번역(Back translation) 데이터 증강]

- 방식
  - 동적 웹 스크레이핑으로 Papago에서 역번역을 진행
  - 원본 데이터 31,974개 모두 한국어 → 영어 → 한국어를 거쳐 역번역한 데이터를 얻은 뒤, 실험군 2가지를 구성.
    - 실험군 1: 역번역된 데이터로 원본 데이터를 대체.
    - 실험군 2: 원본 데이터 + 역번역 데이터. 즉 31,974 x 2 = 63,948개 데이터.
- 결과

	Validation F1 at last epoch	Public LB F1	Private LB F1
기본 (대조군)	**0.838**	**0.8729**	0.8597
역번역 대체 (실험군 1)	0.8266	0.8602	0.8495
기본 + 역번역 증강 (실험군 2)	0.8352	0.8709	**0.8599**

- 해석 및 고찰
  - Validation 및 public LB에서의 F1 점수가 baseline보다 낮았기 때문에 private LB F1 점수도 기대하지 않았으나, 결과적으로 baseline보다 0.0002 만큼 높았던 점을 예측하기 힘들었음
  - 중간 번역 언어를 일본어인 데이터를 추가해보고 싶었으나 시간 제약으로 실행하지 못 함.
  - 32000여 개의 데이터에 대해 역번역을 하는 것이 동적 웹 스크레이핑 과정 중 IP 밴 등을 당하지 않게 하기 위해 delay를 더 줄이기 힘들다보니 1.5일 정도씩 소요되었기 때문에 발생한 문제.
  - 다만 역번역을 보다 일찍 시작했다면 더 많은 데이터를 증강시켜보는 시도를 할 수 있었을 것.

3. 프로젝트 고찰

잘한 점 및 아쉬운 점

잘한 점

- Git과 Huggingface Datasets 관련하여 여러 convention을 수립하고 지켜가며 프로젝트를 진행함.
- 데이터 증강 및 클리닝을 위해 다양한 방법들을 시도함.
- 좋은 데이터는 무엇인가에 대한 근본적인 고민을 하고, 이것을 다음 프로젝트에 적용하기 위한 준비를 함.

아쉬운 점

- 원본 데이터를 정제하는데 더 많은 연구가 필요함.
- 데이터 증강으로 생성한 데이터의 품질을 일정하게 유지하기 어려웠음.
- 마지막에 리더보드에 제출할 두 가지를 제대로 선택하지 않아서 private 리더보드 상으로 순위가 크게 낮아졌는데, 긴장감을 유지했다면 마지막 제출 항목을 잘 고를 수 있지 않았을까 회고함.

프로젝트 시작 전 설정한 목표의 달성 여부

목표	달성 여부
ChatGPT와 대화하는 스페셜 미션을 즐겁게 수행하기	O
Debugger를 적극적으로 활용하기	X

목표	달성 여부
실험 전 문제에 대한 사전 조사를 통해 보다 다양한 논문을 소화하고 프로젝트에 반영하기	😑
코드 리뷰 철저히 하기	O
깃 저장소에서 더 완성된 컨벤션 활용하기	O
가능하다면 여러 대의 GPU를 동시에 효율적으로 활용하는 방법 적용해보기	X