

Data-Centric: KLUE-Topic Classification benchmark

NLP-09 조(역삼동불나방)

1 프로젝트 개요

1. 프로젝트 주제 및 목적

- 자연어에서 독해 및 분석 과정을 거쳐 주어진 태스크를 수행하기 위해서는 자연어의 주제에 대한 이해가 필수적이다. KLUE-Topic Classification benchmark 는 뉴스의 헤드라인을 통해 그 뉴스가 어떤 topic 을 갖는지를 분류해 내는 task 로, 각 자연어 데이터에서 생활문화, 스포츠, 세계, 정치, 경제, IT 과학, 사회 등 다양한 주제 중 하나로 라벨링한다.
- 본 프로젝트는 Data-Centric 의 목적에 맞게 주어진 데이터셋을 바탕으로 베이스라인 모델의 수정 없이 오로지 데이터의 수정으로만 성능 향상을 이끌어내야 한다.

2. 프로젝트 환경

컴퓨팅 환경	5 인 1 팀, 인당 V100 서버를 VSCode 와 SSH 로 연결하여 사용
협업 환경	Notion, GitHub, Google Drive
의사 소통	Slack, Zoom, 카카오톡

3. 프로젝트 구조

a. 데이터 총 개수

Data-noise 가 섞인 KLUE-YNAT 학습 데이터셋
Train 7,000
Test47,785

b. 데이터셋 구조

Column	설명
ID	데이터 샘플의 고유번호
text	분류의 대상이 되는 연합 뉴스 기사의 헤드라인. 한국어 텍스트에 일부 영어, 한자 등의 단어가 포함
target	정수로 인코딩된 라벨
url	데이터 샘플의 뉴스 url (출처)
date	데이터 샘플의 뉴스가 작성된 날짜와 시간

c. Label Type 설명

id	0	1	2	3	4	5	6
설명	IT 과학	경제	사회	생활문화	세계	스포츠	정치

d. 평가 지표: F1 score, accuracy

e. 대회 중 리더보드 평가 기준: Public Score (Test 의 50%를 바탕으로 평가)

f. 최종 평가 기준: Private Score (Test 의 100%를 바탕으로 평가)

g. 하루 제출 횟수: 10 회

2

프로젝트 팀 구성 및 역할

1. 역할

- 전현욱 : 팀 리더, Label Error Detection, G2P Noise
- 곽수연 : 특수문자 및 한자 처리, Back Translation
- 김가영 : Semantic Similarity Analysis
- 김신우 : Data Augmentation
- 안윤주 : Text Keyword Extraction

3

프로젝트 수행 절차 및 방법

1. 프로젝트 기간

- 2024-01-24 10:00 ~ 2024-02-01 19:00

2. 활용된 기술 및 라이브러리

- 개발 언어: Python
- 데이터 전처리 및 증강: Pandas, Numpy, kakaobrain/Pororo, hanja
- 모델링: PyTorch, HuggingFace

3. 프로젝트 세부 수행 절차

- 1) 2024-01-24 (수)까지 대회 기초 강의 전부 수강 완료
- 2) 서버 수령 후 GitHub 및 작업 환경 설정
- 3) 대회 개요 및 데이터 성질, 베이스라인 코드 분석
- 4) EDA 진행 후, 분석 내용을 바탕으로 가설 설립
- 5) 데이터 전처리 및 증강
- 6) 가설 실험 및 검증 (5 번과 동시 진행)
- 7) 검증 결과를 바탕으로 최종 방법 선택 후 모델 학습 및 평가
- 8) 최종 결과물 제출

4

프로젝트 수행 결과

1. 작업 환경 설정

- GitHub : branch 별로 버전 관리
- V100 서버 : 코드 작성 및 모델 학습 (GPU CUDA 버전 : 11.4)

2. EDA

- Labeling Error:

Label 이 잘못 Tagging 된 데이터 샘플이 존재

```
input text: 주말 N 여행 제주권 만설 한라산...그 순수한 아름다움 속으로
label: 세계
input text: 리듬체조 유망주 서고은, '집사부일체' 깜짝 등장
label: 스포츠
input text: '복수해라' 윤현민 "좋은 사람들과 동행...행복했던 시간" 종영 소감
label: IT과학
input text: "죄질이 좋지 않다"...자가격리 두번 이탈 20대 첫 징역형
label: 스포츠
input text: '애로부부' 박철민·유경진 부부, 싸한 분위기 "오는 길에도 싸웠다"
label: IT과학
input text: 무릎 부상 기성용 선두 첼시전 출전 명단서 제외
label: 경제
```

- Text Noise (G2P):

텍스트의 발음 표기법 상으로 변환해 주는 G2PK 패키지를 통해 생성된 노이즈 텍스트가 존재

```
0      개포이단지 부낭 압두고 개포지구 재건축 뿔부터
1      삼성전자 KBIS 이영일팔서 셰프컬렉션 선보여
2      L지 지육 싸면 보 이어포니 단도 노쳐뵤
3      신간 불록체인형명 이영사명·남자의 고독싸
4      이스라엘 정보당국 팔레스타이니 노심영 테러 혐의로 체포
5      배구연맹 이영일구 순천 KOVO커 부녕 대항업체 입찰 공고
6      콜마비애네이치 장녀 녀어빅 쌈배고시비어권...오사.오퍼센트↑
7      소이 킬조이처너권 경유·납싸·항공유 공급계약
8      카카오·삼성화재 디지털 손해보험사 설립 추진
9      긴급배상대채귀 참서카는 김병준과 나경원
10     포겨 미기는 도심 소 고펀라 공연
Name: text, dtype: object
```

3. 데이터 전처리 및 증강

1) 특수 기호 및 한자 한글로 변환

- 특수 기호 처리

전	U+	%	↑	...	~	↓	cm	→	..	(주)	mm	%
후	유폴 러스	-	-	-	퍼센 트	상승	-	-	하락	센티 미터	에서	-	-	밀리 미터	퍼센 트
전	μm	km	m²	+	'	kg	°C	○	×	GHz	"	"	'	↔	
후	마이 크로 미터	킬로 미터	제곱 미터	플러 스	-	킬로 그램	도	-	-	기가 헤르 츠	"	"	-	-	

- 한자 음 변경

변경 전	朴대통령 한미일 대북압박 연대강화 X 北도발시 더 강력제재
변경 후	박대통령 한미일 대북압박 연대강화 X 북도발시 더 강력제재

2) 불용어 제거

단어 빈도수 분석을 통하여 모든 target 에서 불필요하게 다수 등장한 단어 제거

3) Label 의 Keyword 단어 추가

- 각 카테고리별 빈도 10 이상 단어 리스트를 추출 후 해당 카테고리에만 존재하는 고유 키워드 vocab 생성
- 해당 키워드 vocab 을 각 카테고리의 텍스트에 이어 붙여 모델이 강한 확신을 갖고 학습

4. 데이터 증강

1) 역번역(Back Translation)

- Kakao Brain 의 "Pororo"를 활용해 한국어를 영어로 번역 후, 번역된 문장을 다시 한국어로 번역
- 토큰을 원래 단어로 바꾸고, 번역 문장 중 '다'로 끝나지 않는 문장(명사로 끝나는 문장 등) 제거
- 문장의 의미 유사도를 비교하여 유사도가 낮은 하위 25% 문장을 제외
- [출처] [kakaobrain/pororo](https://kakaobrain.com/pororo)

예시	Sentence	target
원본 문장	개포 2 단지 분양 앞두고 개포지구 재건축 불붙어	경제
변경 문장	개포 2 단지 분양을 앞두고 개포지구 재건축이 불붙고 있다.	경제

2) 외부 데이터 (AI-Hub) 사용

- 뉴스 기사 기계 독해 데이터
- 국내 종합일간지 및 지역신문 기사의 제목, 카테고리, 본문으로 이루어진 데이터셋
- 자체적으로 카테고리를 재분류 후 학습 데이터로 활용
- [출처] [AIHub / 뉴스 기사 독해 데이터](https://aihub.or.kr/)

3) G2P 노이즈 생성


- Train set 에 Text Noise 가 포함되어 Test Set 에도 Text Noise 가 존재한다고 가정
- 노이즈에 대한 데이터 케이스를 다양하게 만들어서 노이즈에 대한 추론 능력 강화
- 모든 train 텍스트 데이터를 g2pk 패키지를 통해 노이즈로 변환 후 증강
- [출처] [g2pK: g2p module for Korean](#)

5. Label Error Detection

- Label 이 잘못 Tagging 된 데이터 샘플이 존재하기 때문에 모델의 학습에 방해가 될 것으로 판단.
- cleanlab 패키지로 Label Error Detection 을 진행
- 검출한 결과 중에서 model 의 Label Prediction Probability 가 99% 이상인 데이터들을 Labeling Error 로 판단하고 해당 Label 을 모델의 Prediction 값으로 변경 (총 137 개)

6. 최종 결과

[Public Score] F1-score: 0.8454 / accuracy: 0.8484

5 (-)	NLP_09조		0.8454	0.8484	27	14m
----------	---------	---	--------	--------	----	-----

[Private Score] F1-score: 0.8414 / accuracy: 0.8443

7	NLP_09조		0.8414	0.8443	27	15m
---	---------	---	--------	--------	----	-----

[최종 채택한 방법론 및 데이터]

- 데이터 전처리 및 증강
 - 데이터 전처리 기법 :
 - 특수 기호 및 한자 한글로 변환
 - 불용어 제거
 - 데이터 증강 기법 :
 - G2P Noise
 - 사용한 최종 데이터의 개수 : 14,000 개
 - train valid 비율 : 0.75:0.25
 - Batch Size : 8
 - Tokenizer Max Sequence Length : 80