Github : https://github.com/wogkr810/Dacon_Sentence_Type_Classification

[데이콘] 문장 유형 분류 AI 경진대회

(아최나)

이재학

TABLE OF CONTENTS

0. Usage & Reproduction

데이콘 규칙

코드 검증 내용

1.제출 코드(학습, 추론)으로부터 Private Score 재현 가능 여부 (코드에 Random Seed, Hyperparameter 등 코드 재현을 위한 값들을 꼭 기재해주세요)
-> 15페이지의 arguments & Github에 상세표시

2. 규칙 위반 관련 (Data Leakage, 기타 치팅 요소 등)
-> 위반 X

3. 코드 동작 여부
-> 작동

사전 학습 모델

klue/roberta-large
Link:
- Huggingface : https://huggingface.co/klue/roberta-large
- Github : https://github.com/KLUE-benchmark/KLUE
- Paper : https://arxiv.org/abs/2105.09680

개발 환경

Colab Pro Plus

- CPU : 6C
- GPU : A100-SXM4-40GB(1C)
- 용량 : 100GB(구글 드라이브 One Basic)
- OS : Linux-5.10.133+-x86_64-with-glibc2.27
- Python : 3.8.16.
- W&B CLI Version : 0.13.7
- torch : 1.13.0+cu116
- transformers : 4.25.1
- 실험 기록 : WandB
- 사전 학습 모델 : klue/roberta-large

경로

- 데이터 : ./data
- 학습 : ./train.py
- 추론(단일) : ./inference.py
- 추론(K-Fold) : ./soft_voting.ipynb
- 추론(K-Fold Hard Voting) : ./hard_voting.ipynb

재현 명령어

0-1. 훈련 : python train.py
0-2. 추론(단일 모델) : python inference.py
1. SOTA 재현 : 5개의 5-Fold 단일모델 -> 하드보팅
- 1-1. Colab Pro Plus -> GPU 등급 : 프리미엄 -> Nvidia A100 GPU
- 1-2. pip install –r requirements.txt (python-dotenv, wandb, transformers, tqdm, datasets)
- 1-3. ./arguments.py의 **model_name을** 변경후 학습(python train.py)
  - model_name :
    - 'roberta_document_mean_max'
    - 'roberta_document_weighted'
    - 'roberta_document_concat_hidden'
    - 'roberta_document_sds'
    - 'roberta_document_linear'
- 1-4. 각각의 model_name을 학습 후에 ./soft_voting.ipynb 코드 실행하면 5-fold 결과물(csv파일)들이 './results/soft_ensemble/'경로에 생성
- 1-5. 1-3&1-4의 과정 이후에 csv파일들이 저장됐다면, ./hard_voting.ipynb 코드 실행 후 최종 하드보팅 결과물(csv파일) './results/hard_ensemble/' 경로에 생성됨.
- ++ WandB를 사용하려면, wandb 관련 주석 해제 후 login key값, name, project 설정 후 실행

1. Introduction

대회 개요

대회 기간

2022.12.12 ~ 2022.12.23

대회 설명

문장 유형 분류 AI 모델 개발

평가 방법

Weighted-F1 Score

1. 리더 보드

- 평가 산식 : Weighted F1 Score
- Public score : 전체 테스트 데이터 중 30%
- Private score : 전체 테스트 데이터 중 70%

2. 평가 방식

- 1차 평가 : 리더보드 Private Score
- 2차 평가 : Private Score 상위 10팀 코드 및 PPT 제출 후 코드 평가

데이터셋

데이터 예시

ID	문장	유형	극성	시제	확실성	Label
TRAIN_00000	성균관대는 세계 최고의 대학이다.	사실형	긍정	현재	확실	사실형-긍정-현재-확실

데이터셋 설명

- 뉴스 기사에서 추출한 총 23631(16541+7090)개의 단문 대화 텍스트 데이터
  - Train : 16541개
  - Test : 7090개
- 데이터 column
  - Train : ID, 문장, 유형, 극성, 시제, 확실성 , label
  - Test : ID, 문장
- Labels(Train_only)
  - 유형 : 사실형, 추론형, 대화형, 예측형
  - 극성 : 긍정, 부정, 미정
  - 시제 : 과거, 현재, 미래
  - 확실성 : 확실, 불확실

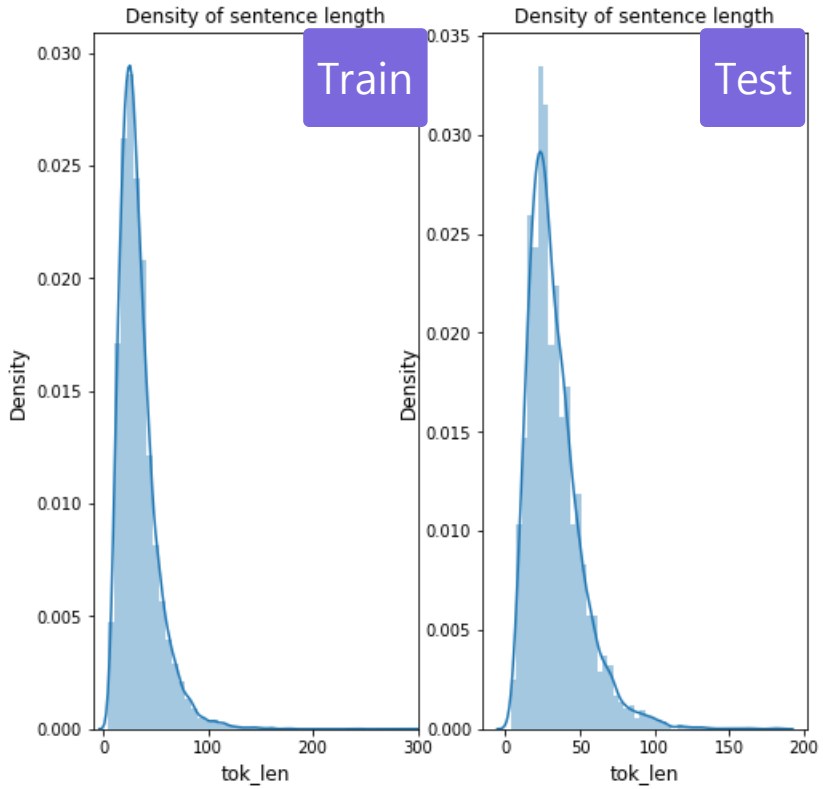
2. EDA

Dacon Talk : https://dacon.io/competitions/official/236037/codeshare/7317?page=1&dtype=recent

EDA

문장의 길이

Train max length : 313
TeST max length : 183



문장의 중복

- train -> 35개 * 2 중복
  - 문장은 같은데 label 이다른경우 ->
    - (14989,07269) -> 14989 삭제
    - (00208,03364) -> 03364 삭제
    - (07099,04670) -> 07099 삭제
    - (2108,15167) -> 02108 삭제
  - 문장,label 같은 경우 -> keep = 'first'
- test - > 11개 * 2 중복
  - label이 없고, 지울수도 없음 그냥 냅둬야함

코드

- pd.set_option('display.max_rows', 100)
- test[test[['문장']].duplicated(keep=False)].sort_values('문장')

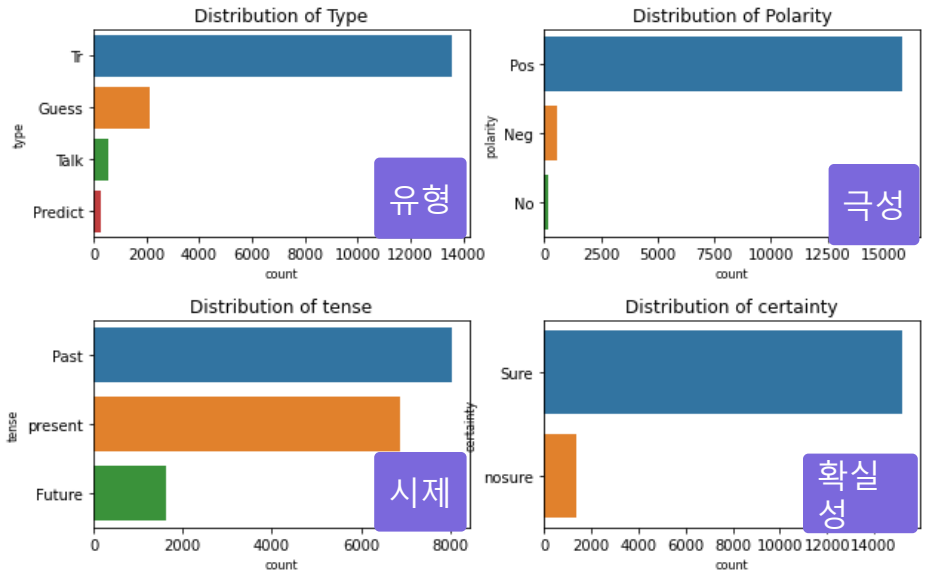
바꾸기

- df = df.loc[df.ID != 'TRAIN_14989']
- df = df.loc[df.ID != 'TRAIN_03364']
- df = df.loc[df.ID != 'TRAIN_07099']
- df = df.loc[df.ID != 'TRAIN_02108']
- df = df.drop_duplicates('문장', keep = 'first')

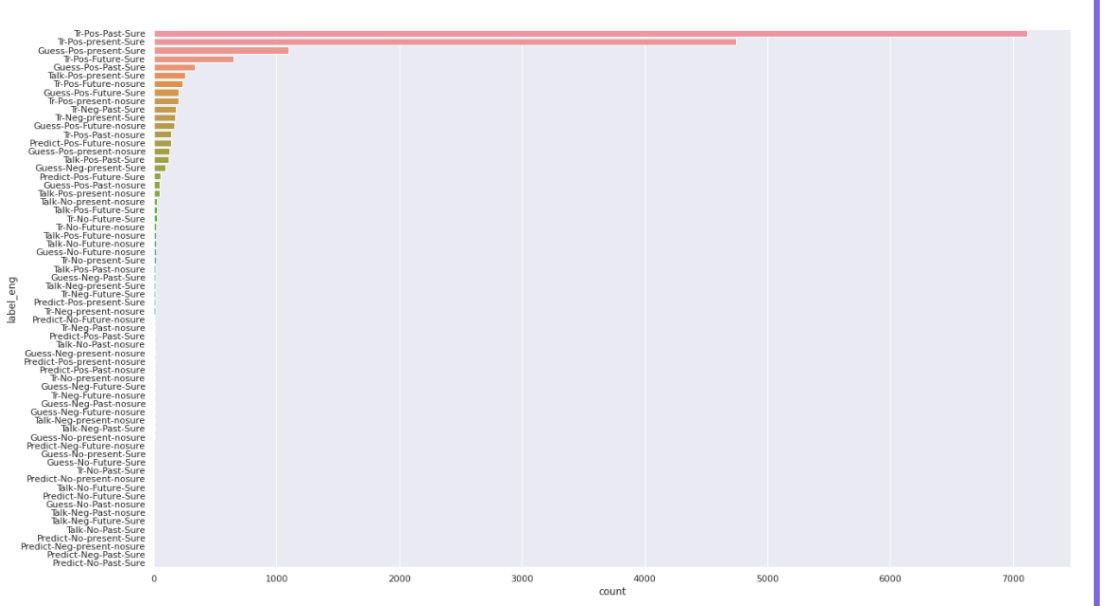
Dacon Talk : https://dacon.io/competitions/official/236037/codeshare/7317?page=1&dtype=recent

EDA

각 label



전체 label



Summary

- 통계치 : 대부분 짧다.
- 결측치 : 없다. 하지만 중복은 존재.
- 데이터 : 뉴스 기사의 일부임을 알 수 있음. 정제된 데이터다 보니, 맞춤법 및 띄어쓰기에 관한 전처리는 하지 않아도 된다고 판단.

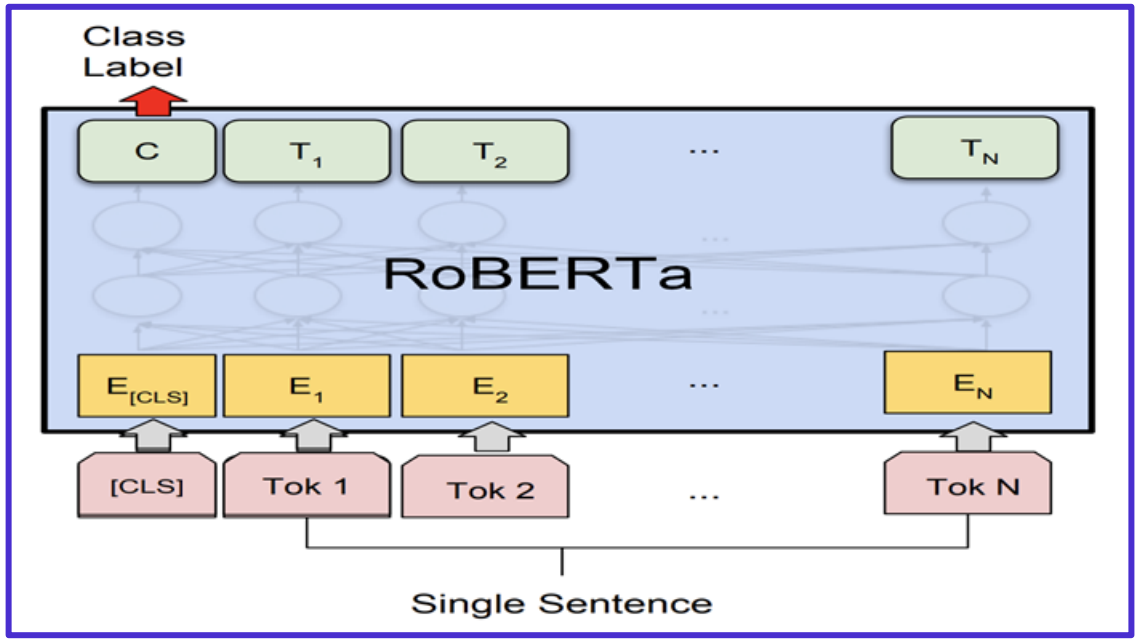
3. Proposed Method

모델 설명

공통

Roberta -> custom layer -> classification layer(fc layer)

- 1. Klue/roberta-large pretrained model을 이용
- 2. Roberta 모델의 sequence_output, pooled_ouput, hidden_states 추출
- 3. 2번의 출력을 이용하여 custom layer & heads & pooling layer 통과
- 4. 3번의 최종 출력을 각 성분의 label 개수에 맞춰 4개의 classification layer 통과 후 logit 추출



모델 성능

Arguments

Hyper Parameter	Defualt
seed	41
Batch_size	64
epochs	6
scheduler	ReduceLRonPlateau
optimizer	Adam
Learning_rate	3e-5
PLM	Klue/Roberta-large
loss	Cross-entropy
Max_input_length	128

성능

Model(5-Fold)	Public	Private
Weighted layer Pooling	0.7509	0.7524
Concat Last 4 Hidden states	0.7476	0.7537
Mean-Max Pooling	0.7555	0.7550
Custom layers(SDS) head	0.7475	0.7537
Roberta Classification head	0.7452	0.7499

5개 결과물 Hard Voting

최종 제출

Public : 0.7582
Private : 0.75746

4. Experiments

WandB : https://wandb.ai/wogkr810/sen_class

Experiments

Model 실험

Various Custom Models with Roberta-Large

감사합니다!

Q & A