Code States Project 2

AI 9기 NLP 1팀

AI 09 장원서
AI 09 이기하
AI 09 정다윤

Contents

01. 프로젝트 개요

02. 프로젝트 팀 구성 및 역할

03. 프로젝트 수행 절차 및 방법

04. 프로젝트 수행 결과

05. 자체 평가 의견

1. 프로젝트 개요

■ 프로젝트 계획 의도

· 스마트 폰의 증가로 정보가 쏟아지는 현 시대에 **필요한 정보**만 빠르게 탐색가능한 서비스 구현을 목표로 **사용자의 편의성 증진**

■ 프로젝트 목표

· **KLUE-STS** 데이터셋을 활용한 의미적 텍스트 유사도 (Semantic Textual Similarity) 측정

· 두 개의 한국어 문장을 입력 받아 두 문자의 **의미적 텍스트 유사도**를 출력하는 모델 구현

■ 프로젝트 구조

· **KLUE-STS-Datasets** + **KorNLUDatasets**를 통해 데이터 보강

· **Sentence transformers** fine-tuning

· **KLUE-Roberta-Base** 모델 적용하여 pre-train

· **Optuna, WandB**를 통한 hyper-parameter tuning으로 모델 성능 향상

· **Pearson's r score, f1 score**

· **FLASK** 웹 프레임워크를 통한 서비스 시연

■ 기대 효과

· input) 2개의 한국어 문장

· output) 의미적 유사도 점수 출력

2. 프로젝트 팀 구성 및 역할

- 프로젝트 수행 기획안을 바탕으로 효율적인 담당 업무 지정
- 데이터 수집, 데이터 정규화, 모델링 과정 전원 참여

Team member	역할	담당 업무	비고
장원서	팀장	· 데이터 EDA 및 정규화 , AI model pre-train, 학습 보고서 작성	
이기하	팀원	· AI model structure building , hyper parameter tunning	
정다운	팀원	· Sub AI model structure building, API service building	

3. 프로젝트 수행 절차 및 방법

· 프로젝트 사전 기획시 일정표 작성 후 프로젝트 진행현황에 따라 유동적으로 일정 재 배분

구분	기간	활동	비고
Pytorch, NLP lesson	· 05.09(월) ~ 05.20(금)	· Pytorch 및 NLP,NLU 학습	· 2주간의 학습기간
사전 기획	· 05.23(월) ~ 05.24(화)	· 프로젝트 기획 및 주제 선정, 기획안 작성 · 프로젝트 모델 선정 및 해당 논문 리뷰	
데이터 수집	· 05.25(수) ~ 05.26(목)	· 외부 데이터 수집	· 팀 주간보고 실시
데이터 전처리	· 05.27(금) ~ 05.28(일)	· 데이터 정제 및 정규화 진행, 데이터 추가 학습	
모델링	· 05.29(월) ~ 06.01(수)	· 모델 구현, 대안 모델 구현	· 팀 주간보고 실시
서비스 구축	· 06.02(목) ~ 06.03(금)	· API 서비스 구현	· 팀 최종보고 실시 (예정)
총 개발기간	· 05.09 ~ 06.03 (4주)		

4. 프로젝트 수행 결과

4.1 Data EDA (exploratory data analysis)

KLUE-STS-Dataset

STS task를 해결하기 위해 만들어진 한국어 데이터셋

· AIRBNB (구어체 리뷰)

· Policy (격식체 뉴스)

· ParaKQC (구어체 스마트홈 쿼리)

세가지 도메인으로 구성

Over-all Data : 13,224개

Train Data : 10,494개

Valid Data : 1,167개

Test Data : 519개

'Guid, source, sentence1, sentence2, labels, annotations'

총 6개의 칼럼으로 구성

Labels 칼럼은 real-label, label, binary-label 3가지 값을 가짐

	guid	source	sentence1	sentence2	labels	annotations
0	klue-sts-v1_train_00000	airbnb-rtt	숙소 위치는 찾기 쉽고 일반적인 한국의 반지하 숙소입니다.	숙박시설의 위치는 쉽게 찾을 수 있고 한국의 대표적인 반지하 숙박시설입니다.	{'label': 3.7, 'real-label': 3.714285714285714...	{'agreement': '0:0:0:2:5:0', 'annotators': ['0...
1	klue-sts-v1_train_00001	policy-sampled	위반행위 조사 등을 거부·방해·기피한 자는 500만원 이하 과태료 부과 대상이다.	시민들 스스로 자발적인 예방 노력을 한 것은 아산 뿐만이 아니었다.	{'label': 0.0, 'real-label': 0.0, 'binary-labe...	{'agreement': '5:0:0:0:0:0', 'annotators': ['1...
2	klue-sts-v1_train_00002	paraKQC-sampled	회사가 보낸 메일은 이 지메일이 아니라 다른 지메일 계정으로 전달해줘.	사람들이 주로 네이버 메일을 쓰는 이유를 알려줘	{'label': 0.3000000000000004, 'real-label': 0...	{'agreement': '4:2:0:0:0:0', 'annotators': ['1...
3	klue-sts-v1_train_00003	policy-sampled	긴급 고용안정지원금은 지역고용대응 등 특별지원금, 지자체별 소상공인 지원사업, 취업...	고용보험이 1차 고용안전망이라면, 국민취업지원제도는 2차 고용안전망입니다.	{'label': 0.6000000000000001, 'real-label': 0....	{'agreement': '4:2:1:0:0:0', 'annotators': ['1...
4	klue-sts-v1_train_00004	airbnb-rtt	호스트의 답장이 늦으나, 개선될 것으로 보입니다.	호스트 응답이 늦었지만 개선될 것으로 보입니다.	{'label': 4.7, 'real-label': 4.714285714285714...	{'agreement': '0:0:0:0:2:5', 'annotators': ['1...

EDA

· Train Data의 샘플 수가 부족하다 판단하여 KorNLUDatasets Train 데이터에 추가로 적용하여 데이터 보강

	sentence1	sentence2	real-label	binary-label
0	비행기가 이륙하고 있다.	비행기가 이륙하고 있다.	5.00	1
1	한 남자가 큰 플루트를 연주하고 있다.	남자가 플루트를 연주하고 있다.	3.80	1
2	한 남자가 피자에 치즈를 뿌려놓고 있다.	한 남자가 구운 피자에 치즈 조각을 뿌려놓고 있다.	3.80	1
3	세 남자가 체스를 하고 있다.	두 남자가 체스를 하고 있다.	2.60	0
4	한 남자가 첼로를 연주하고 있다.	자리에 앉은 남자가 첼로를 연주하고 있다.	4.25	1

Train Data : 10,494 → 17,417

· 각종 특수 문자 및 영어, 일본어, 한자 제거

· Py-hanspell 맞춤법 검사

평균 0.45개의 맞춤법 에러를 고침으로 시간 대비 성능이 나오지 않아 데이터셋에 반영하지 않았음

	sentence1	sentence2	real_label	binary_label	spell_checked1	spell_checked2
0	숙소 위치는 찾기 쉽고 일반적인 한국의 반지하 숙소입니다	숙박시설의 위치는 쉽게 찾을 수 있고 한국의 대표적인 반지하 숙박시설입니다	3.714286	1	숙소 위치는 찾기 쉽고 일반적인 한국의 반지하 숙소입니다	숙박시설의 위치는 쉽게 찾을 수 있고 한국의 대표적인 반지하 숙박시설입니다
1	위반행위 조사 등을 거부 방해 기피한 자는 500만원 이하 과태료 부과 대상이다	시민들 스스로 자발적인 예방 노력을 한 것은 아산 뿐만이 아니었다	0.000000	0	위반행위 조사 등을 거부 방해 기피한 자는 500만 원 이하 과태료 부과 대상이다	시민들 스스로 자발적인 예방 노력을 한 것은 아산뿐만이 아니었다
2	회사가 보낸 메일은 이 지메일이 아니라 다른 지메일 계정으로 전달해줘	사람들이 주로 네이버 메일을 쓰는 이유를 알려줘	0.333333	0	회사가 보낸 메일은 이 지메일이 아니라 다른 지메일 계정으로 전달해줘	사람들이 주로 네이버 메일을 쓰는 이유를 알려줘
3	긴급 고용안정지원금은 지역고용대응 등 특별지원금 지자체별 소상공인 지원사업 취업...	고용보험이 1차 고용안전망이라면 국민취업지원제도는 2차 고용안전망입니다	0.571429	0	긴급 고용안정 지원금은 지역 고용 대응 등 특별 지원금 지자체별 소상공인 지원 사...	고용보험이 1차 고용안전망이라면 국민취업지원 제도는 2차 고용안전망입니다
4	호스트의 답장이 늦으나 개선될 것으로 보입니다	호스트 응답이 늦었지만 개선될 것으로 보입니다	4.714286	1	호스트의 답장이 늦으나 개선될 것으로 보입니다	호스트 응답이 늦었지만 개선될 것으로 보입니다

4. 프로젝트 수행 결과

4.2 모델 개요

KLUE-RoBERTa-BASE

https://huggingface.co/klue/roberta-base

· BERT 보다 data, sequence, batch size, training period를 더욱 향상시킨 Pre-trained 모델

· 성능이 다소 낮은 NSP objective 대신 MLM task 로 사전학습

· KLUE 데이터(한국어)를 활용하여 사전학습

Model	TC	STS
	F1	Pearsons' r
mBERT-base	81.55	84.66
XLM-R-base	83.52	89.16
XLM-R-large	**86.06**	92.97
KR-BERT-base	84.58	88.61
koELECTRA-base	84.59	*92.46*
KLUE-BERT-base	*85.73*	90.85
KLUE-RoBERTa-small	84.98	91.54
KLUE-RoBERTa-base	85.07	92.50
KLUE-RoBERTa-large	85.69	**93.35**

KLUE-PLMs

-Layers : 12
-Hidden Size : 768
-Embedding Size : 768
-Head : 12

Evaluation Methods

· Sentence Pair Regression Task
2개의 input sentences 의 의미적 유사도를 STS-Dataset 의 'real-label'을 활용하여 regression 을 통해 측정

· Pearson's coreelation coefficient : human-labeled 문장 유사도 점수와 모델 예측 점수 간의 선형상관 관계를 측정

· Binary classification Task
Sentence Pair Regression Task를 통해 측정된 값을 3.0 threshold 기준으로 0과 1로 분류

· F1 score: STS Dataset의 'binary-labe'과 모델의 binary predictions 에 대한 precision 과 recall 에 대한 평균 측정

선정이유

1. KLUE 벤치마크 baseline 모델 중 STS task에서 가장 높은 점수를 기록한 모델은 Pearsons'r 93.35를 기록한 KLUE-RoBERTa-Large 이었으나 개인 hardware 및 Google Colab으로 모델 훈련을 진행하는 환경상 지속적인 메모리 오류가 발생

2. API를 이용하는 사용자의 편의성을 높이기 위해서 API 응답속도를 고려하여 적절한 layer 수를 보유하면서 최대의 성능을 낼 수 있는 KLUE-RoBERTa-BASE 모델을 선정

4. 프로젝트 수행 결과

4.3 모델 하이퍼 파라미터 서치

하이퍼 파라미터 튜닝

· 최적의 하이퍼 파라미터를 찾기위해 wandb에 내장된 sweep을 이용해 하이퍼 파라미터 튜닝을 진행

· 튜닝할 파라미터는 KLUE 논문을 참고하여 선정 빠르고 고성능인 Bayesian을 기반으로 파라미터 튜닝을 진행

튜닝할 파라미터

learning Rate : 1e-5, 2e-5, 3e-5
Train set batch size : 8, 16
Weight decay : 0, 0.01
Warm_Up ratio : 0, 0.1

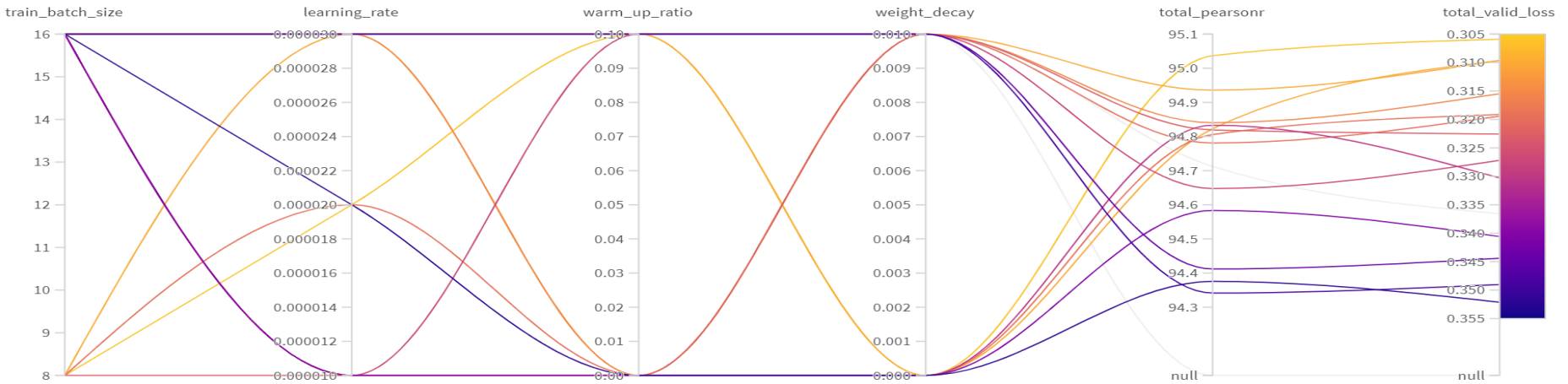
위 조합으로 총 15회 학습을 진행

최종 파라미터

learning Rate : 2e-5
Train set batch size : 8
Weight decay : 0
Warm_Up ratio : 0.1

최종 파라미터를 적용한 모델의 성능

Pearson : 91.88
F1 Score : 86.18
Total Loss : 0.37



4. 프로젝트 수행 결과

4.4 Data Augmentation

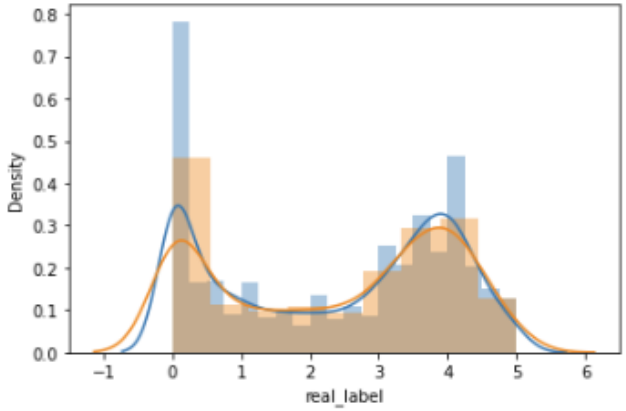
Why need?

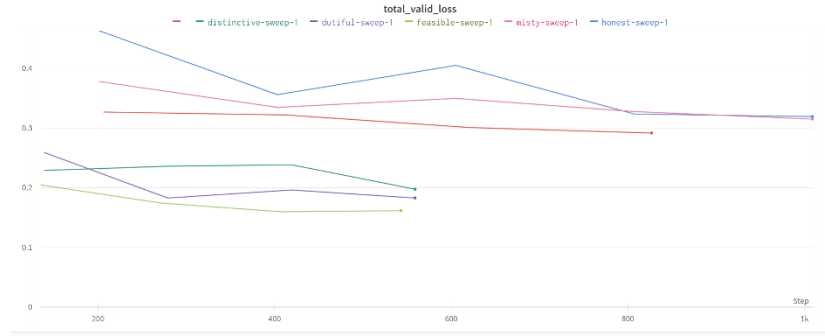
Train 데이터의 라벨 분포가 불균등

· 1 ~ 3 사이 label 개수가 매우 적음

· 따라서 Valid 데이터 또한 불균등 함으로 fine-tuning을 통한 성능 개선이 이루어 지지 않음

· 일반화 성능을 확인할 수 있는 valid 데이터 보강 필요





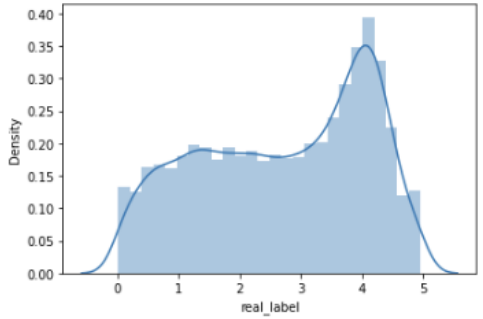
How?

Augmented - SBERT

· 기존 train 과 valid 데이터 문장의 가능한 모든 조합 생성

· 각 문장별 가장 유사도가 높은 문장으로 data pair 구성

· Sementic Search 과정에 SRoBERTa 활용하여 최종적으로 얻은 data pair (Silver dataset) 사이 유사도를 통한 labeling



· Validation에 활용할 dataset 15,000개 생성

· 비교적 균등한 Valid dataset 확보

· Valid loss가 완만하게 줄었음 (Test set에서의 성능 향상은 미비함)

4. 프로젝트 수행 결과

4.5 모델 평가

최종 모델 선택

Model	Model Size	Pre-trained Data	Evaluation
KLUE-BERT-base	- Layers: 12 - Hidden Size: 768 - Embedding Size: 768 - Head: 12	KLUE	Pearson'r: 91.89 F1 : 85.57
KoELECTRA-base-v3	- Layers: 12 - Hidden Size: 768 - Embedding Size: 768 - Head: 12	Korean News, Wiki, 나무위키, 모두의 말뭉치	Pearson'r: 91.49 F1 : 85.77
KLUE-RoBERTa-base	- Layers: 12 - Hidden Size: 768 - Embedding Size: 768 - Head: 12	KLUE	Pearson'r: 92.36 F1 : 86.01

· 결과적으로 KLUE-RoBERTa-base 가 두 모델과 비교하여 가장 높은 점수를 얻음

· KLUE-RoBERTa-base 와 KoELECTRA-base-v3 비슷한 성능을 보이지만 Sentence pair regression task 에서 regression을 사용하기에 Pearson 점수가 더 높은 KLUE-RoBERTa-base 모델을 선정

· 데이터 증강을 적용한 모델의 경우 Pearson과 Loss에서 가장 높은 성능을 나타냄

· Valiation 보강을 위한 데이터 증강의 경우 검증 단계에서 약 2배정도 적은 Loss를 기록함

· 하이퍼 파라미터 튜닝을 진행하고, 데이터 증강을 적용한 데이터셋으로 학습한 Klue-RoBERTa-base 모델을 최종 모델로 선택

	Pearson	F1	Loss
Fine-Tuning(base)	91.27	85.77	0.42
Hyper Parameter Tuning	91.88	86.18	0.39
Data Augmentation	92.36	86.01	0.37

4. 프로젝트 수행 결과

4.6 서비스 구축 및 시연

Flask

Directory Structure

```
— flask_app
  ├──── __init__ .py
  ├──── main.py
  ├──── module.py
  ├──── requirements.txt
  └──── templates
        ├──── index.html
        └──── result.html
```

Requirements

```
pandas==1.1.5
torch==1.9.0
transformers==4.15.0
Flask==2.0.1
```

5. 자체 평가 의견

한계 및 개선방향

- 촉박한 프로젝트 일정과 자원의 한계로 모델의 성능을 baseline 모델에서 충분히 만족할 만큼 끌어올리지 못하였다.
- 파라미터 튜닝과 데이터 증강으로 유의미한 성능향상이 이루어지지 않았다.
- Pre-trained 모델을 사용하는 대신 tokenizer 부터 설계하여 해당 프로젝트에 다시 적용해볼 예정이다.
- Wandb에서 제공하는 다양한 시각화 기능을 심도있게 사용하여 해당 파라미터가 얼마나 모델 성능에 영향을 주는지 확인할 예정
- 데이터 증강 중 SBERT 외에 round-trip translation, back translation 방식 등을 추후에 적용

프로젝트 진행 후 Acknowledgement

- Pytorch 및 SOTA 모델에 대한 논문 비교 및 정리법 학습
- 논문 구글링부터 시작해서 Scratch 부터 서버코드까지 팀단위로 수행하여 프로젝트 모든 과정에 참여

감사합니다