

KBIGTA 신입기수 프로젝트 (2021년 4월)



[베이스라인] kobert_classification_model (public 0.695)



④ Pororo API로 NL API Pretrained only

pororo로 초보자도 5분만에 결과 제출하는 방법 (public : 0.817)



③ Hugging Face 활용한 Modeling(public: 0.841) Finetuning
Huggingfaceclassifier finetuning why?



허깅페이스 학습 파이프라인(klue/roberta-base, LB 0.805)



(Albert등...) Roberta/ Kobert 아키텍처 요약

KoElectra



② kobert 베이스라인 수정 public score 0.736

BERT 복제기
classifier만
finetuning



현재 우리 GitHub에 올린 코드
pretrained 모델을 이용하여 LB 0.83 받기

roberta pretrained

원본
GitHub 코드
참고



① 데이터 살펴보기



데이터 가볍게 살펴보기 (각 label, feature 별 Word Cloud)

① 데이터셋 설명

• KLUE - NLI 데이터셋

• 한국어 테이터셋

SPC: wikitree, policy, wikinews, wikipedia, NSMC, Airbnb

formal news
articles

영화리뷰

여행·캐스트등의
전야가 많은 애호취성

encyclopedia
movie&trip
reviews

• 10,000 premises → elicit hypothesis (제작되는 것)

• valid hypothesis의 세 가지 조건

⇒ premise가 proposition어야 한다. (창·거짓을 판별 가능한 선언적 문장)

단, 수학적 공식이나 사례·영어 등의 예외 존재 ↴ But

⇒ 최소 1개 이상의 predicate 포함해야 한다.

(서술)

⇒ length of premise : 20 character ~ 90 character

(광역 표현)

데이터셋의 분포 Train/Dev/Test 분포는 무엇인가?

Source	Train	Dev	Test	Total
Wikitree	3838	450	450	4738
Policy	3833	450	450	4733
wikinews	3824	450	450	4724
Wikipedia	3780	450	450	4680
NSMC	4899	600	600	6099
Airbnb	4824	600	600	6024
Overall	24998	3000	3000	30998

여기서 전부는 아니고

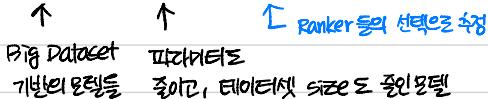
90% 정도 가져온

것으로 보임.

slightly
imbalanced ↴ 약간의
비

→ KOBERT 등의 모델들은 Huggingface의 transformer API를 활용해 훈련할 수 있는 (Trainable)

② ROBERTA/ALBERT/ELECTRA



Pretrained 모델 feature는 fix

• BERT

- ELMo의 경우 unidirectional + feature-based approach
- BERT는 bidirectional + fine-tuning approach (\rightarrow , \leftarrow)
- Masked Language Model (Input Token은 Random Masking)
- Pretrain의 장점은 어떤 LM representation이 모두 학습되도록
bidirectional + multi-task학습 가능하다는 것이다.
- Task-specific input, fine-tune 할 때는 task마다 다른

• ROBERTA: A Robustly Optimized Pretraining Approach

- BERT 모델은 "underfitted"되어 있다고 주장, 예측 Tuning

① 현재 BERT는 랜덤 masking 한 단계를 강화

⇒ ROBERTA에서 100% 모든 mask 적용 + 4 epoch (Dynamic Masking)
그래서 같은 dataset을 "多重화" ... few-shot learning

② Large Batch Size로 성능향상 256 → 2K → 8K

• ALBERT : A Lite BERT

- Pretrain Large Model \Rightarrow distill it to smaller models
- Distilled 70% input embedding matrix \Rightarrow matrix factorization
- cross-layer parameter sharing
- dropout 70%까지 성능향상 있음

• ELECTRA : Pre-training text encoders as discriminators rather than generators

- source 문장 \Rightarrow mask step \Rightarrow Generator (GAN-like pipeline) \Rightarrow mask된 문장 \Rightarrow Discriminator \Rightarrow original replaced ELECTRA
- ELECTRA 역시 Task별 fine-tuning
- 배운面具형 리스너에게 차지! ↩

④ Pororo Framework ← Train first & then do inference & finetuning

Platform of neural models for natural language processing

PORORO Library

- Code Snippet

```
from pororo import Pororo ### Lookup for available tasks
Pororo.available_tasks() >> Available tasks are ... ['mrc', 'rc', 'qa',
'question_answering', 'machine_reading_comprehension',
'reading_comprehension', 'sentiment', 'sentiment_analysis', 'nli',
'natural_language_inference', 'inference', 'fill', 'fill_in_blank', 'fib',
'para', 'pi', 'cse', 'contextual_subword_embedding', 'similarity', 'sts',
'semantic_textual_similarity', 'sentence_similarity', 'sentvec',
'sentence_embedding', 'sentence_vector', 'se', 'inflection',
'morphological_inflection', 'g2p', 'grapheme_to_phoneme',
'grapheme_to_phoneeme_conversion', 'w2v', 'wordvec', 'word2vec',
'word_vector', 'word_embedding', 'tokenize', 'tokenise', 'tokenization',
'tokenisation', 'tok', 'segmentation', 'seg', 'at', 'machine_translation',
'translation', 'pos', 'tag', 'pos_tagging', 'tagging', 'const',
'constituency', 'constituency_parsing', 'cp', 'pg', 'collocation',
'collocate', 'col', 'word_translation', 'wt', 'summarization',
'summarisation', 'text_summarisation', 'text_summarisation', 'summary',
'gec', 'review', 'review_scoring', 'lemmatization', 'lemmatisation', 'lemma',
'ner', 'named_entity_recognition', 'entity_recognition', 'zero-topic', 'dp',
'dep_parse', 'caption', 'captioning', 'asr', 'speech_recognition', 'st',
'speech_translation', 'ocr', 'srl', 'semantic_role_labeling', 'p2g', 'aes',
'essay', 'qg', 'question_generation', 'age_suitability'] ### Lookup for
available models for specific task Pororo.available_models("collocation") >>
Available models for collocation are ... ([lang]: ko, [model]: kcollocate),
([lang]: en, [model]: collocate.en), ([lang]: ja, [model]: collocate.ja),
([lang]: zh, [model]: collocate.zh) ### Get Pretrained Specific Task Model
ner = Pororo(task="ner", lang="en")
```

- Retrain or Fine-tuning is impossible!