**BERT, KoELECTRA를 이용한 질의응답 시스템**

1. **프로젝트 개요**

1950년 앨런 튜링은 그의 논문 < Computing Machinery and Intelligence >에서 튜링 테스트를 제안하였다. 튜링 테스트는 기계에 지능이 있는지를 판별하고자 하는 것이 주 목적이며, 질의자 한명과 둘의 응답자, 이때 응답자 중 하나는 사람, 다른 하나로는 컴퓨터를 준비하여 질의자가 두 답변 중 더 사람 같은 답변을 선택하는 것이다. 이러한 튜링 테스트처럼 질문에 대해 응답하는 것을 Question Answering (QA) task 라고 부른다. 해당 프로젝트에서는 딥러닝 모델을 이용한 질의응답 시스템을 구현하고자 한다.

1. **프로젝트 목적**

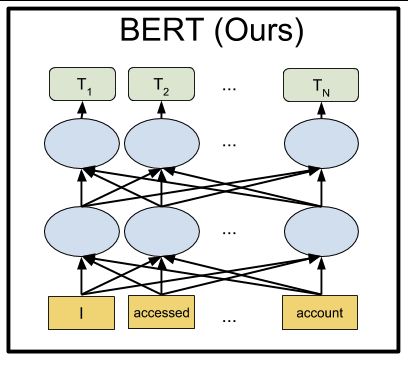
해당 프로젝트에는 2가지 목적이 있다. 우선 첫번째는 QA task에 대한 이해이다. QA는 1950년 앨런 튜링에 의해 제안된 튜링 테스트로부터 시작된 유구한 역사를 지닌 분야이다. 또한, 현 시대에 와서는 인공지능 기반의 기계에 사용자가 질문을 했을 때 그에 대한 답변을 하는 것, 비디오 검색 등 여러 분야에 사용되고 있다. 간단한 QA 프로젝트 시스템을 진행함으로써 해당 분야들까지 확장, 응용할 수 있는 계기를 제공하고자 한다. 그 다음 두번째는 transformers 라이브러리 사용법을 터득하는 것이다. 2018년 BERT가 제안된 이후로, 자연어 처리에서는 대용량 pretrained language model (PLM) 의 사용이 늘어갔다. 하지만 이러한 모델들은 각자의 UI를 갖고 있어서 다루기가 매우 까다롭고 힘들었다. 하지만, 최근 huggingface에서 배포한 transformers 라이브러리는 이러한 대용량의 모델들을 다루기 쉽게 재구성하였다. 바꿔 말하면 transformers를 잘 다룰 수 있다면 자연어처리의 최신 알고리즘들의 태반을 차지하는 PLM을 자유자재로 다룰 수 있다. 해당 프로젝트에서는 transformers 라이브러리를 기반으로 PLM들을 다룸으로써 transformers 라이브러이의 사용법을 숙달하게 하고, 이렇게 함으로써 PLM 모델들에 대한 이해도와 응용을 극대화하고자 한다.

1. **프로젝트 내용**

해당 프로젝트는 다량의 코퍼스에서 이미 학습된 BERT 혹은 KoELECTRA를 불러와서 KorQuAD 데이터에 finetuning하는 방식으로 진행된다. 이후에, 임의의 Wikipedia data를 이용하여 직접 질의 응답을 할 수 있도록 구성하였다. 해당 프로젝트에 사용된 BERT와 KoELECTRA 모델에 대한 간략한 설명을 추가하면 아래와 같다.

1) BERT

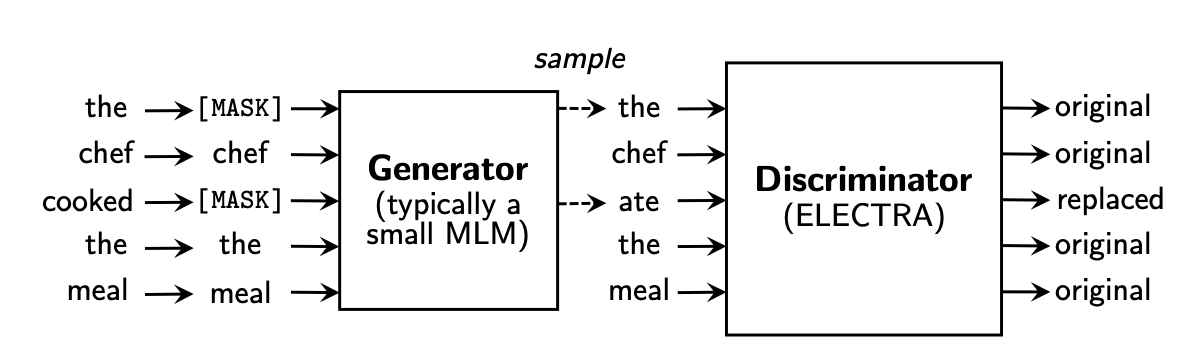
BERT는 2018년 구글에 의해 제안된 모델로서, 발표 당시 glue 리더보드의 대부분의 task를 휩쓸었던 모델이다. BERT는 Transformer의 인코더와 masked language model과 next sentence predictiond을 이용하여 unsupervised learning으로 학습한 모델이다. Masked language model은 input token 중의 일부를 masking하여 모델로 하여금 masking된 token을 맞추게 함으로써 모델이 다양한 문맥의 정보를 이용할 수 있게 한다. 또한 next sentence prediction을 이용하여 문장 내의 문맥적 정보뿐만 아니라, 문장과 문장 사이의 정보들까지도 학습할 수 있도록 하였다.

****

참조 : <https://www.aclweb.org/anthology/N19-1423.pdf>

2) ELECTRA

ELECTRA는 BERT와 비슷한 형태이지만, BERT와 비교했을 때 token 결과물이 masking되었던 token인지 아닌지를 맞추는 task를 추가하여 성능을 향상시켰다.



참조 : <https://openreview.net/pdf?id=r1xMH1BtvB>

본 프로젝트에서는 BERT, ELECTRA 모델 중에 하나를 골라서 Finetuning을 진행한 후, 실제 Wikipedia data를 이용하여 QA를 진행해볼 수 있도록 하였다.

1. **주요 기술**

- 언어 : Python

- 모델 : BERT, ELECTRA

- 필요 패키지 : 실습 가이드 참조

1. **주요 소스 코드 및 시연방법**

**1. 설치**

# pip install로 설치합니다.

torch==1.6.0  
transformers==3.3.1

seqeval

fastprogress

attrdict

$ git clone <https://github.com/monologg/KoELECTRA.git>

**2. Finetune**

$ cd KoELECTRA/finetune

#모델 2개 다 하지 마시고 하나만 골라서 하세요.

#KoELECTRA

$ python3 run\_squad.py --task korquad --config\_file koelectra-small-v3.json

#KoBERT

$ python3 run\_squad.py --task korquad --config\_file kobert.json

**3. Inference**

import torch

from transformers import AutoTokenizer, AutoModelForQuestionAnswering

#tokenizer = AutoTokenizer.from\_pretrained("monologg/koelectra-base-v2-finetuned-korquad")

#model = AutoModelForQuestionAnswering.from\_pretrained("monologg/koelectra-base-v2-finetuned-korquad")

tokenizer = AutoTokenizer.from\_pretrained(Finetune 모델 경로)

model = AutoModelForQuestionAnswering.from\_pretrained(Finetune 모델 경로)

question="2011년 3월 11일 14시 46분에 일어난 지진이 뭐야?"

context='도호쿠 지방 태평양 해역 지진은 2011년 3월 11일 14시 46분(JST) 미야기 현 앞바다에서 일어난 지진이다. 진원 깊이는 약 24km이며, 지진이 발산한 에너지 크기인 릭터 규모는 Mw9.0-9.1로 일본 국내에서 관측한 지진 중 가장 규모가 크다. 전세계적으로 봐도 1900년 이후 일어난 지진 중 1960년 발디비아 지진, 1964년 알래스카 지진, 2004년 인도양 지진 해일에 이어 1952년 세베로쿠릴스크 지진과 같은 규모의 초거대지진이다.[7][주해 1] 도호쿠 지방 태평양 해역 지진은 동북일본이 얹혀진 대륙지각[주해 2]이 일본 해구에서 가라앉고 있는 태평양 판에 끌려가 생긴 변형이 한꺼번에 방출되면서 생긴 태평양 판 얕은 곳에서 일어난 판 경계형 지진이다.[8] 진원 지역이 되는 단층 영역은 가로 150-200km, 세로 400-500km[주해 3]이나 다른 규모 M9급 지진에 비해서는 미끄러진 단층 크기가 작은 편이다.[2] 한편 미야기 현 앞바다의 일본 해구의 해구축 부근에서는 국지적으로 50m 이상 지각이 이동한 것으로 추측하고 있는데,[주해 4][3][2][4] 이정도의 양은 지금까지 관측된 초거대지진의 미끄럼양보다 훨씬 크다.[9]'

inputs = tokenizer.encode\_plus(question, context, return\_tensors="pt")

with torch.no\_grad():

answer\_start\_scores, answer\_end\_scores = model(\*\*inputs)

answer\_start = torch.argmax(answer\_start\_scores)

answer\_end = torch.argmax(answer\_end\_scores) + 1

tokenizer.convert\_tokens\_to\_string(tokenizer.convert\_ids\_to\_tokens(inputs["input\_ids"][0][answer\_start:answer\_end]))