모의경진대회 오리엔테이션 및 1차 모의경진대회 과제 특강: 도서자료 검색 효율화를 위한 기계독해

㈜마인즈앤컴퍼니 | 김태훈 매니저

2022.11.14

KOSME 중소벤처기업진흥공단 Korea SMEs and Startups Agency

MINDs@company

Index

실전 캐글 프로젝트 개요

1 AI 이론 및 실습

딥러닝 기초 이론 트랙 별 심화 교육

- 기간 : 10/31 ~ 11/11
- 딥러닝 기초 이론 및 CV/NLP/RS 트랙 별 심화 이론 교육
- 오후 실습 교육 및 일일 과제 수행
- 온/오프라인 멘토링 실시

2 모의경진대회

1차 모의경진대회

- 기간 : 11/14 ~ 11/25
- 플랫폼 : AI CONNECT
- CV/NLP/RS 트랙 별 1문제. 총 3문제

2차 모의경진대회

- 기간 : 11/28 ~ 12/09
- 플랫폼 : AI CONNECT
- CV/NLP/RS 트랙 별 1문제. 총 3문제

3 실전캐글대회

실전 Kaggle 과제 참여

- 기간 : ~ 12월
- 실전 경진대회 과제 소개 및 베이스라인 강의 제공
- 트랙별 집중 멘토링 통한 실전경진대회 수상 지원

3 경진대회 참여 지원

경진대회 과제 특강

- 과제 소개 및 베이스라인 강의 제공
- Kaggle, AI Coonect 플랫폼 활용법 교육

멘토링

- 트랙 별 온/오프라인 멘토링 세션 운영
- Kaggle 과제 집중 멘토링

4 전문가 특강

산업

- 각 분야 산업 전문가 특강
- AI 활용 및 적용 사례 공유

Kaggle

- Kaggle Master 등 Kaggle 전문가 특강
- 경진대회 수상 노하우 및 인사이트 공유

AI 이론 및 실습 / 모의경진대회 일정

2022년 11월

월요일	화요일	수요일	목요일	금요일
31 (오전)AI 이론 및 실습 교육 - SNU 조원국 - Fundamentals of Deep Learning & Essential Mathematicsc (오후)실습 멘토링	1	2 (오전)AI 이론 및 실습 교육 - SNU 조원국 - Convolutional Neural Netwo & Image Classification (오후)실습 멘토링	3	4 (오전)AI 이론 및 실습 교육 - SNU 조원국 - U-Net Architecture & image Segmentation (오후)실습 및 캐글 멘토링
7 (오전)AI 이론 및 실습 교육 - SNU 조원국 - Transposed convolution - Fully convolutional network - UNet (오후)실습 멘토링	8	9 (오전)AI 이론 및 실습 교육 - SNU 조원국 - Transformer - Pretrained Language Model (오후)실습 멘토링	10	11 (오전)AI 이론 및 실습 교육 - SNU 조원국 - Recommender Systems & Factorization Machine (오후)실습 및 캐글 멘토링
14 (오전)1차 모의경진대회 OT - 김태훈, 도성진, 한두희 - AIConnect 플랫폼 활용법 - 1차 모의대회 과제 설명 및 Baseline 실습 (오후)트랙별 멘토링	15	16 (오전)Pytorch 특강 - 박성호 - Autograd 구현 및 Pytorch 기초 (오후)트랙별 멘토링	17 채용 행사	18 (오전)산업 전문가 특강 - SSG.COM 이혜진 DS - 산업 현장 추천 시스템 (오후)캐글 멘토링
21 (오전)실험관리 특강 - 박성호 - Wandb 사용법 - Optuna 사용법 (오후)트랙별 멘토링	22	23 (오전)전문가 특강 - KB 박장원 - 개발자 github 관리 방법 (오후)트랙별 멘토링	24	25 (오전)전문가 특강 - KAIST 이신의 - 메타러닝 튜토리얼 (오후)캐글 멘토링 1차 모의경진대회 종료
28 (오전)2차 모의경진대회 OT - 박성호, 김태훈, 한두희 - 2차 모의대회 과제 설명 및 baseline 실습 (오후)트랙별 멘토링	29	30 (오전)산업 전문가 특강 - 딜리셔스 임정욱 - 기획자와의 커뮤니케이션 (오후)트랙별 멘토링		

1 AI 이론 및 실습

- 기간 : 10/31 ~ 11/11
- AI 기초 및 track별 심화 이론 교육
- 실습 교육 및 멘토링

2 1, 2차 모의경진대회

- 기간 : 11/14 ~ 11/25, 11/28 ~ 12/09
- CV/NLP/RS track별 1문제
- 과제 소개 및 Baseline 특강

3 특강

- 시간 : 월/수/금 오전 9(or 10)시 ~ 12시
- 모의경진대회 및 실전 Kaggle 과제 특강
- 산업 전문가 특강

4 멘토링 세션

- 시간 : 월/수/금 오후 1시 ~ 4시
- CV/NLP/RS 트랙 별 맞춤형 멘토링

모의경진대회 개요

	이미지(CV)	자연어(NLP)	추천(RS)
1차 모의경진대회 (11.14 ~ 11.25)	 사과 이미지를 이용한 사과 품종 분류	 도서자료 데이터를 이용한 기계독해	 고객 및 식당 데이터를 이용한 식당 만족도 예측
2차 모의경진대회 (11.28 ~ 12.09)	 추후 공개	 추후 공개	 추후 공개

1차 모의경진대회 일정

일정

경진대회 세부 일정

✓ 오리엔테이션 및 과제 베이스라인 특강 : 11.14(월) 09:00 ~ 12:00

✓ 추론 결과 제출 : 11.14(월) 12:00 ~ 11.25(수) 12:00

✓ 결과 발표 : 11.25(금) 18:00

✓ 과제별 우승팀 코드 리뷰 세션 : 12.16 (금) 09:00 ~ 12:00

온/오프라인 멘토링

	이미지(CV)	자연어 처리(NLP)	추천 시스템(RS)
오프라인 멘토	 **박성호 멘토** - MNC Data Scientist - Upenn Med 뇌과학 연구원 - Upenn 수학과 학사 - SCI, SSCI급 논문 제1저자 - 2022 NIPA 인공지능 온라인 경진대회	 **박성일 멘토** - 서울대 데이터 사이언스 스쿨 - 대한민국 경찰(2019 ~ 2022) - 경찰대 법학 학사 - 카카오톡 챗봇 개발 및 배포	 **유승준 멘토** - 서울대 Learning and Adaptation Lab 인턴 - 서울대 DYROS 로보틱스 부트캠프 수료 - 아주대 기계공학 학사
	 **김태훈 멘토** - MNC Data Scientist - 서울대 경제학부 학사 - 2022 NIPA 인공지능 온라인 경진대회 - 1기 이어드림 스쿨	 **오로훈 멘토** - Megabyte School AI 데이터사이언티스트 취업완성 과정 수료 - 패스트캠퍼스 NLP 오프라인 학습 매니저	 **강하예진 멘토** - Megabyte School AI 데이터사이언티스트 취업완성 과정 수료 - 패스트캠퍼스 RS 오프라인 학습 매니저
온라인 멘토	 **정채연 멘토** - 카이스트 김재철 AI 대학원 석박통합과정 - 고려대 경제학/컴퓨터학 학사 - 삼성전자 AI 교육과정 조교	 **최민석 멘토** - 카이스트 김재철 AI 대학원 석박통합과정 - 일리노이대 컴퓨터학 학사 - 네이버 웹툰 AI Automation 팀 - 글로벌창업사관학교 조교	 **곽대훈 멘토** - 카이스트 김재철 AI 대학원 박사과정 - 카카오엔터프라이즈 AI Lab - 삼성전자 종합기술원 조교 - 삼성전자 DS AI Expert 조교 - 글로벌창업사관학교 조교

Index

AI CONNECT란?

AI Connect 마인즈앤컴퍼니의 경진대회 플랫폼

AI 경진대회 플랫폼

AI 전문인력과 수요자들을 효율적으로 연결하여 AI 문제를 해결할 수 있는 AI 경진대회

특징1 AI 아이디어를 활용하려는 수요자와 다수의 아이디어를 제공하는 제안자들의 연결을 통해 객관적인 가치 평가가 가능함

특징2 크라우드소싱 기반 인공지능 문제해결 협업 플랫폼

AI 생태계 활성화

AI 데이터 및 모델 활용 촉진, 개인/기업의 AI 역량을 증진시켜 전반적으로 AI 생태계 발전

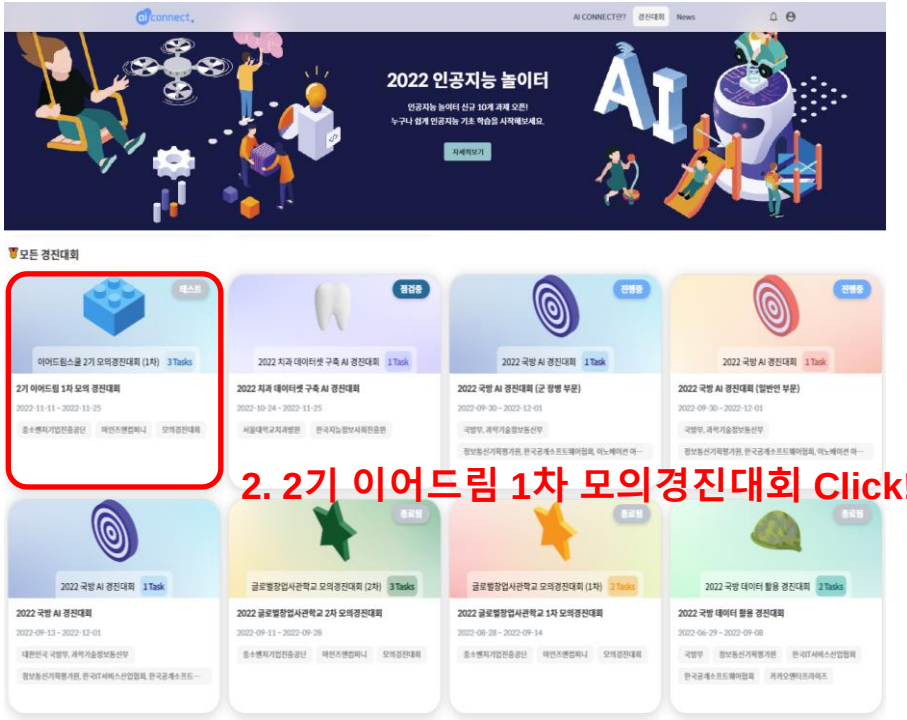
특징3 AI 경진대회 플랫폼은 단순히 순위를 위한 경합의 장이 아닌 모델 제공을 통한 개인/기업 역량 증진과 다양한 모델을 테스트 가능한 기회의 장

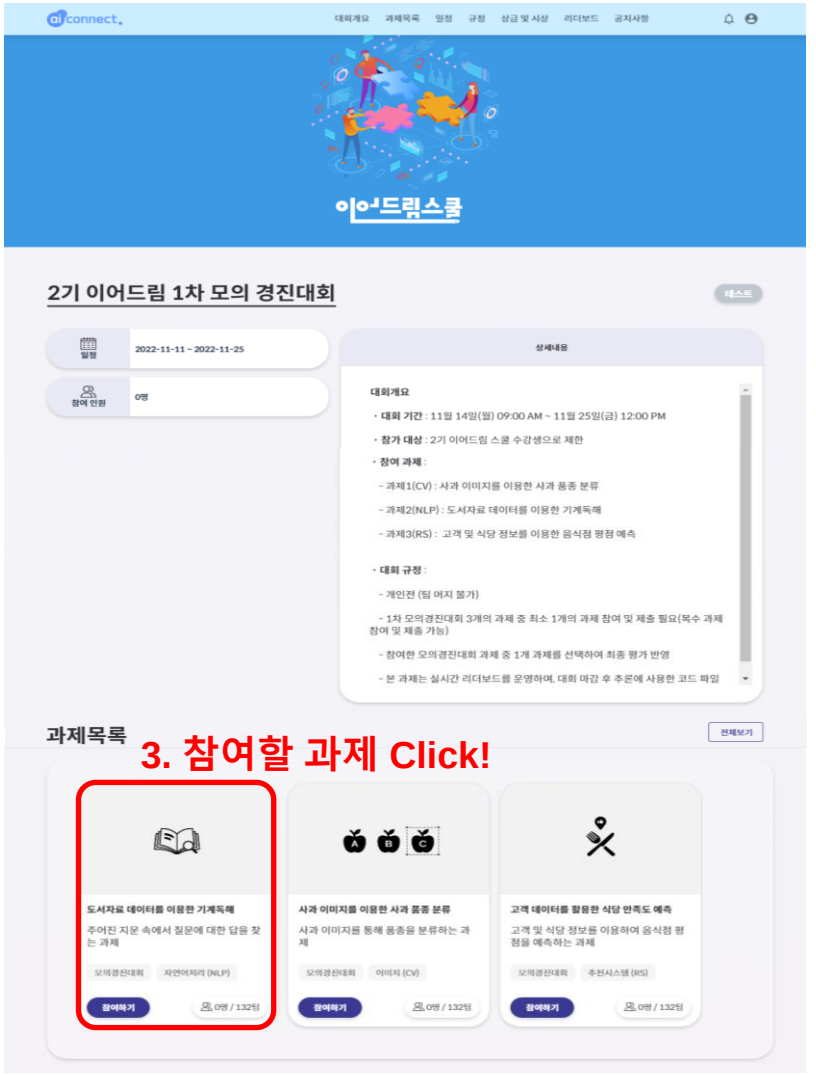
- 문제해결이 필요한 분야에 대한 문제 정의 및 설계
- 데이터셋 제공 (필요에 따라 전처리 및 정제 진행)
- 상금 (수요기업별 협의)
- 인센티브 추가 제공 (상장, 채용기회 등)

- 해결 가능한 문제에 대하여 과제 참여 신청
- 문제 해결을 위해 최적의 AI 모델 개발 및 구축
- 과제 우승에 따른 상금 및 인센티브 획득
- AI 모델 학습 코드 및 문제 해결 노하우 제공

AI CONNECT 과제 확인 방법

1. AI CONNECT 플랫폼 링크 Click!





AI CONNECT 과제 기능 탭 확인

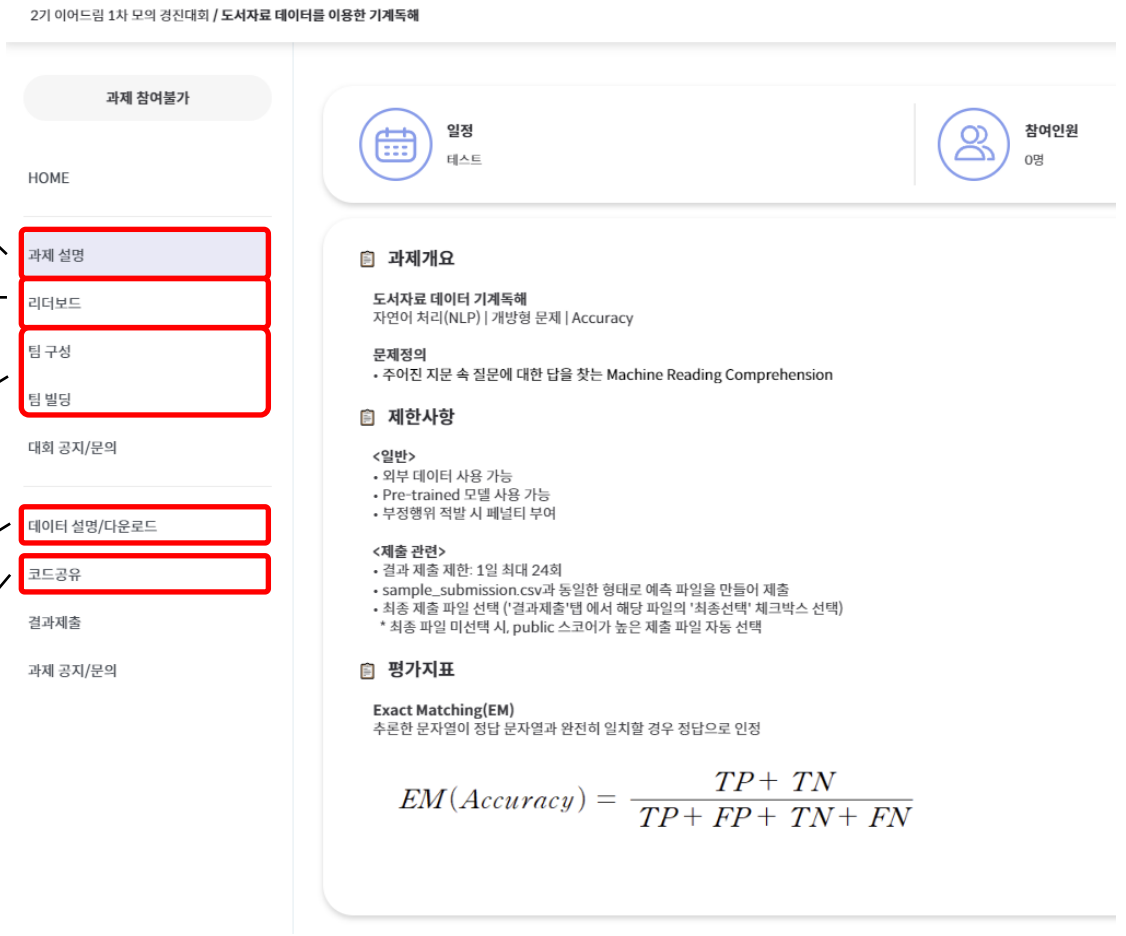
과제 개요, 제한 사항, 평가지표 확인

리더보드 순위 및 score 확인

팀 머지 기능
개인전 과제에서는 사용 X

Train/Test 데이터,
Sample_submission.csv 다운로드

베이스라인 코드 다운로드

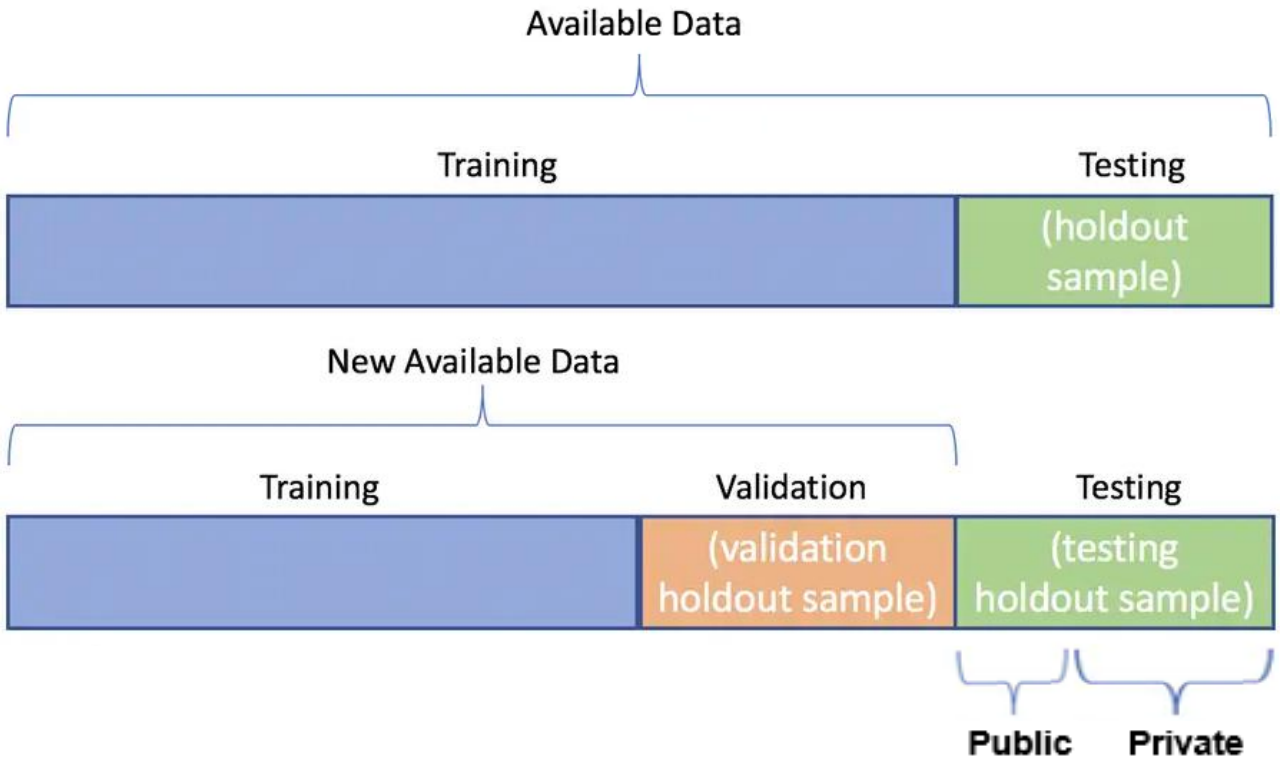


AI CONNECT 결과 제출 방법

‘결과제출’ 탭을 통해
추론 결과 파일 제출

Public score / CV score 등을 고려하여 제출 결과물 중 하나의 결과물 ‘최종선택’
선택하지 않을 시 Public score가 가장 높은 결과물이 자동으로 최종선택됨

Public / Private Score



- **Available Data -> Training / Testing Data** : 학습된 모델의 일반화 성능을 판단하기 위함
- **Training Data -> Training / Validation Data** : 일반화 성능을 높이는 방향으로 모델을 학습하기 위함
- **Testing Data -> Public / Private Data** : 리더보드를 통해 정답을 유추해내는 행위를 막기 위함

Index

1. 실전 캐글 프로젝트 커리큘럼 및 일정
2. AI CONNECT 플랫폼 소개 및 사용법 안내
3. 1차 모의경진대회 – (자연어) 도서자료 검색 효율화를 위한 기계독해
    - 과제 소개 (데이터셋 / 베이스라인 모델 / 평가지표 / 제한사항)
    - 베이스라인 코드

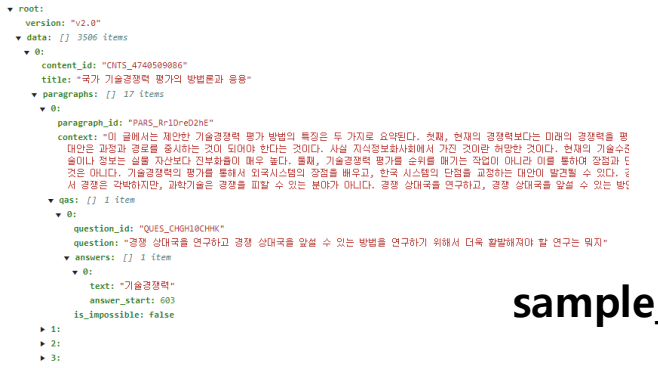
데이터셋

과제 개요

주어진 지문 속에서 질문의 답을 찾는 **기계독해(Machine Reading Comprehension)** 과제

Data Set

- **Input** : 본문(content)의 문단(paragraph) 별 텍스트 및 질문
- **Output** : 각 질문에 대한 답(paragraph 내에 있는 단어 일부, 각 질문 당 답은 하나)
- **수량**
  - train.json 내에는 3506개의 content가 있고, 각 content는 여러 개의 paragraph로 구성되며, 각 paragraph 당 하나의 질문이 있음
    cf) 질문에 대한 답이 해당 paragraph에 있는 경우(is_impossible:false)도 있고, 없는 경우(is_impossible:true)도 있음
  - test.json 내에는 1038개의 content가 있고, 각 content는 여러 개의 paragraph로 구성되며, 각 paragraph 당 하나의 질문 있음

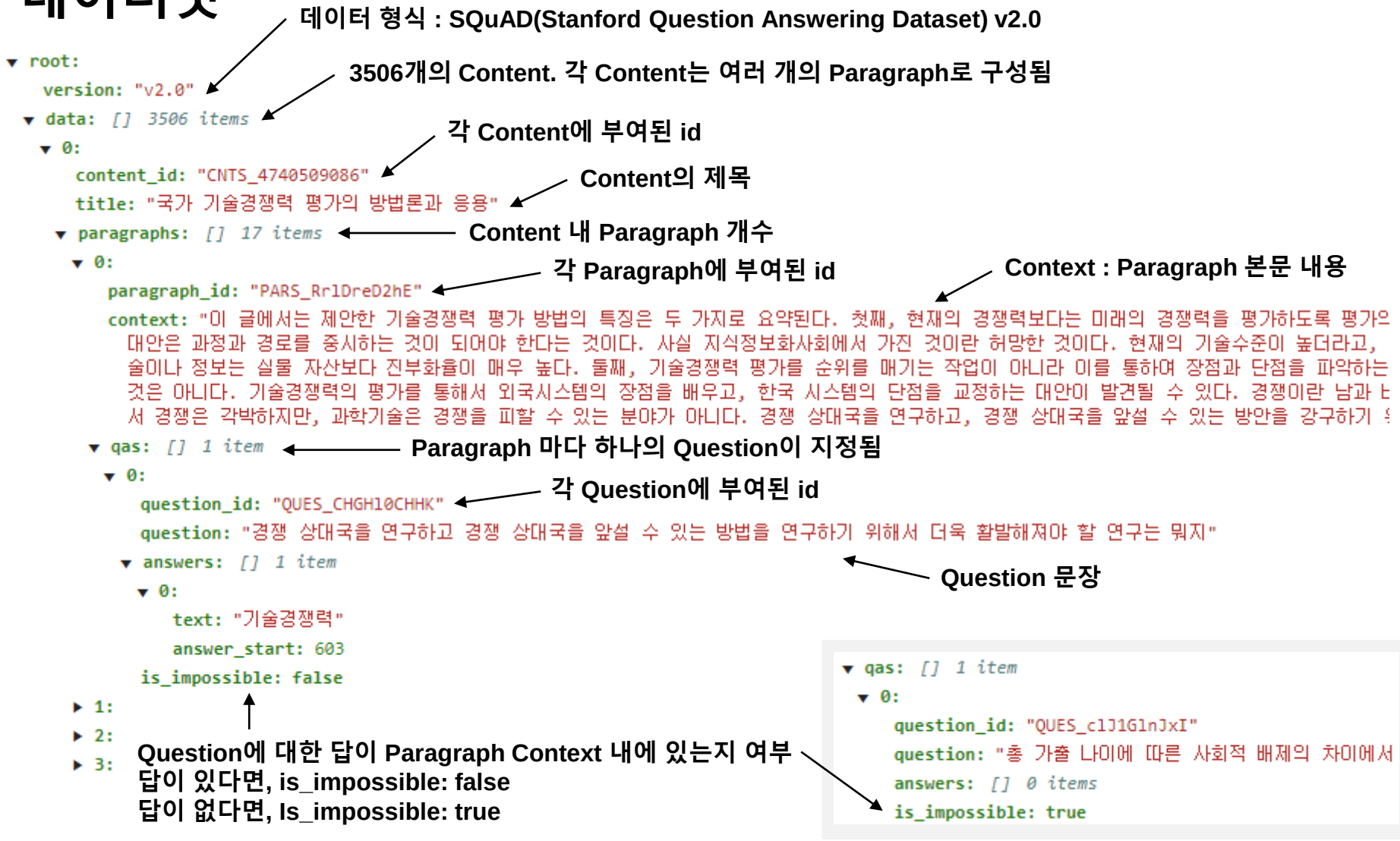


← train.json

sample_submission.csv →

	question_id	answer_text
1	QUES_cyOI2451l1	한국원자력안전기술원
2	QUES_pz2vbWpWWo	가출청소년 문제
3	QUES_1g3jl4y7eo	탐색기
4	QUES_qzwOZwaeeY	Prime Air
5	QUES_hfdtXCtdzf	장애인케어서비스

데이터셋

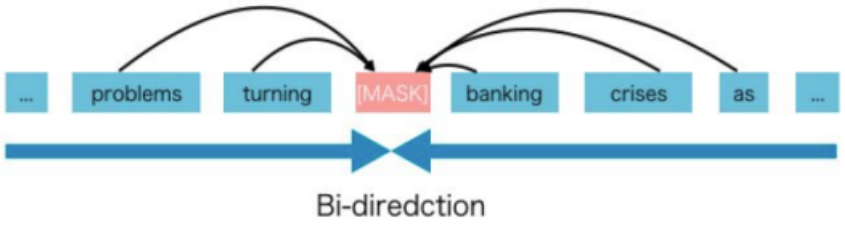


베이스라인 모델 : BERT

- **2018년 10월**에 등장 (~~**벌써 4년....**~~)
- 당시 ELMO, GPT 등보다 높은 점수를 보임
- **Pretraining + Finetuning** 방식을 유행시킴
- **4년이 지난 지금도 여전히 많이 쓰이고 있는 모델**

베이스라인 모델 : BERT

Masked Language Model (MLM)



- 일단 문장 전체를 모델에게 알려줌
- 주변 문맥을 통해 **빈 칸을 예측**함
- **(기존의 GPT 등과 비교하면) 더 많은 정보를 이용한다는 장점**
  - “양방향”으로 고려!

Masked Language Model = 빈칸 맞추기

베이스라인 모델 : BERT

“Pretraining”

베이스라인 모델 : BERT

Pretraining한 걸로는 빈 칸만 잘 맞춘다...

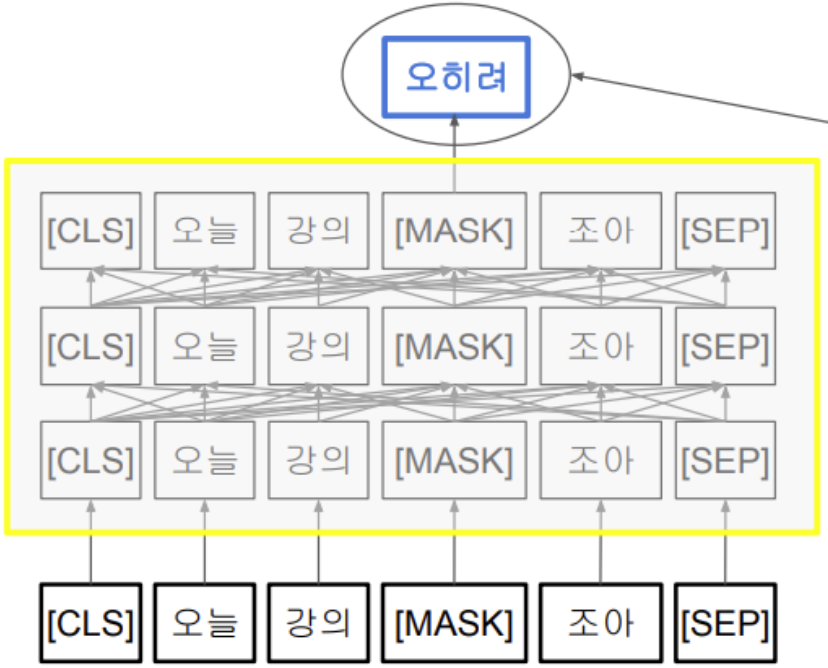
우리는 빈칸 맞추기가 아닌 다른 Task를 하고 싶다!

Text Classification
Named Entity Recognition
Question Answering
...

베이스라인 모델 : BERT

“Finetuning”

베이스라인 모델 : BERT

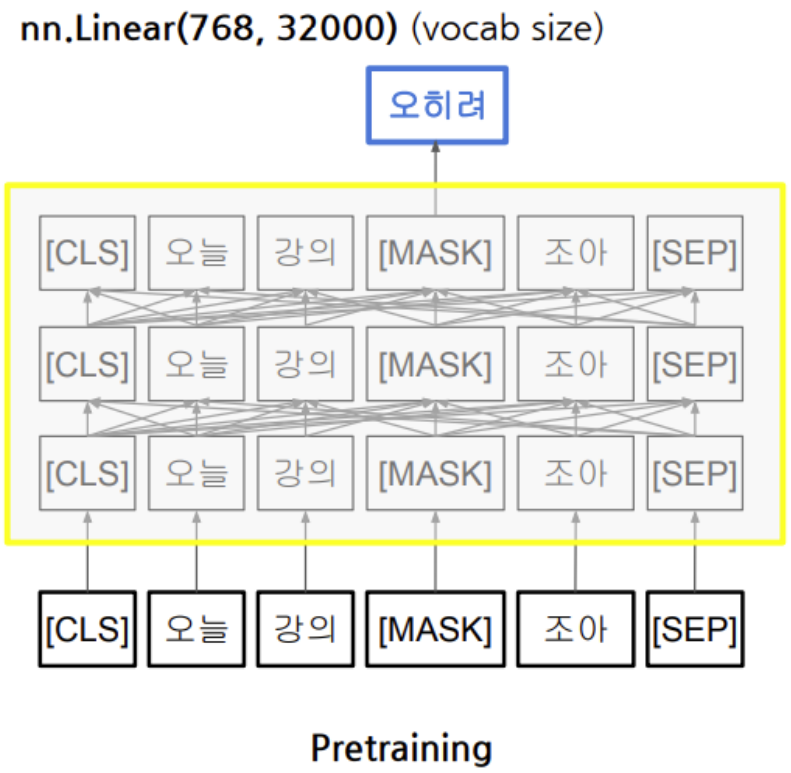


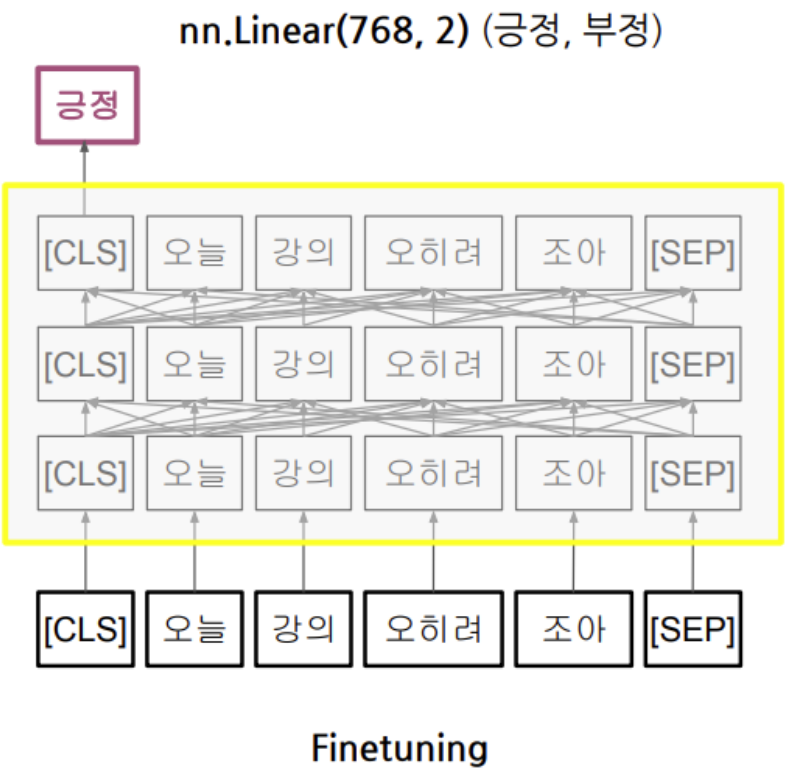
Pretraining

Vocab (단어)을 예측하던 부분을 떼어내고

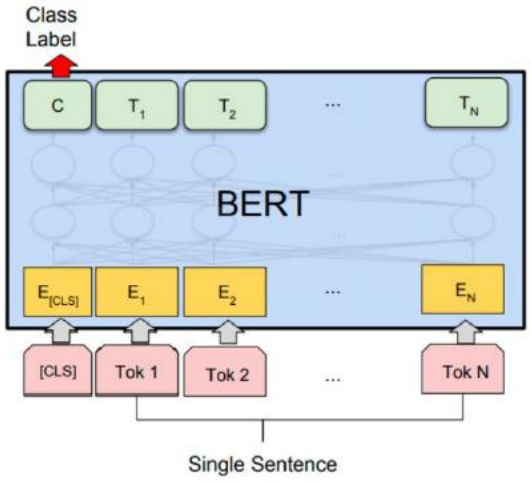
다른 Task를 적용하자!!

베이스라인 모델 : BERT

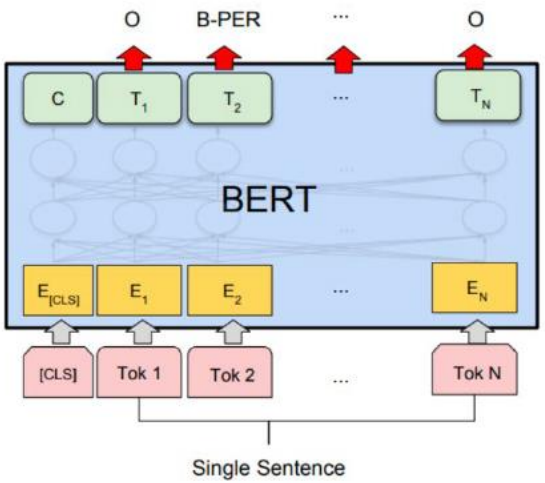




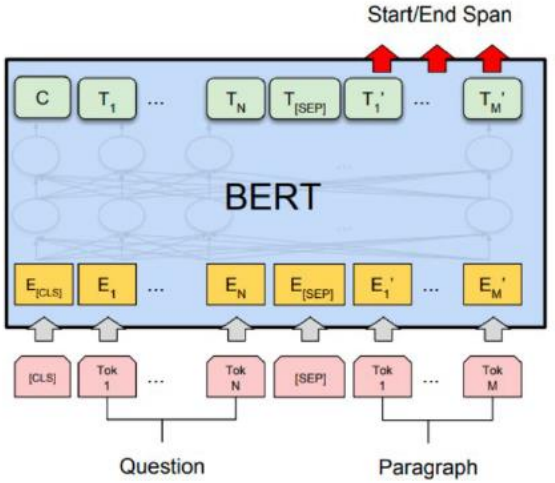
베이스라인 모델 : BERT



Text Classification

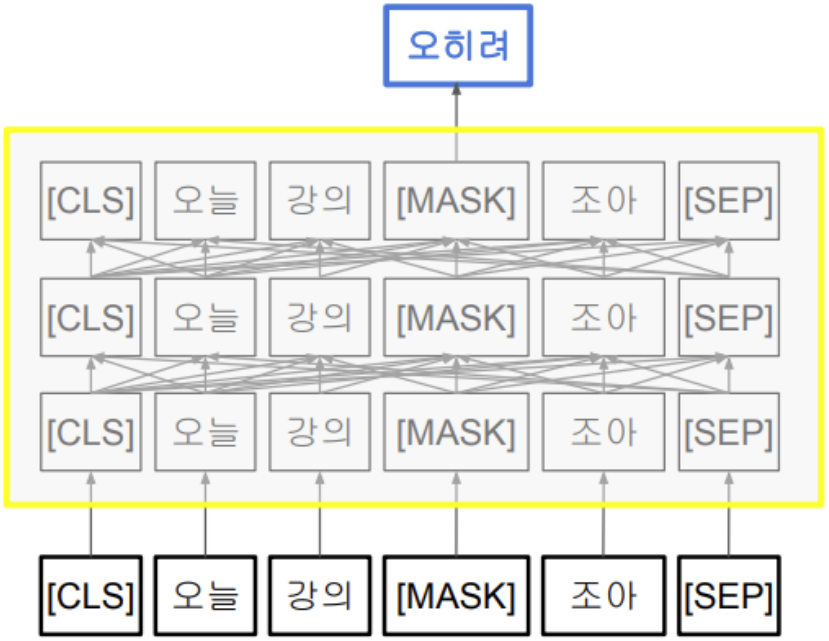


Named Entity Recognition



Question Answering

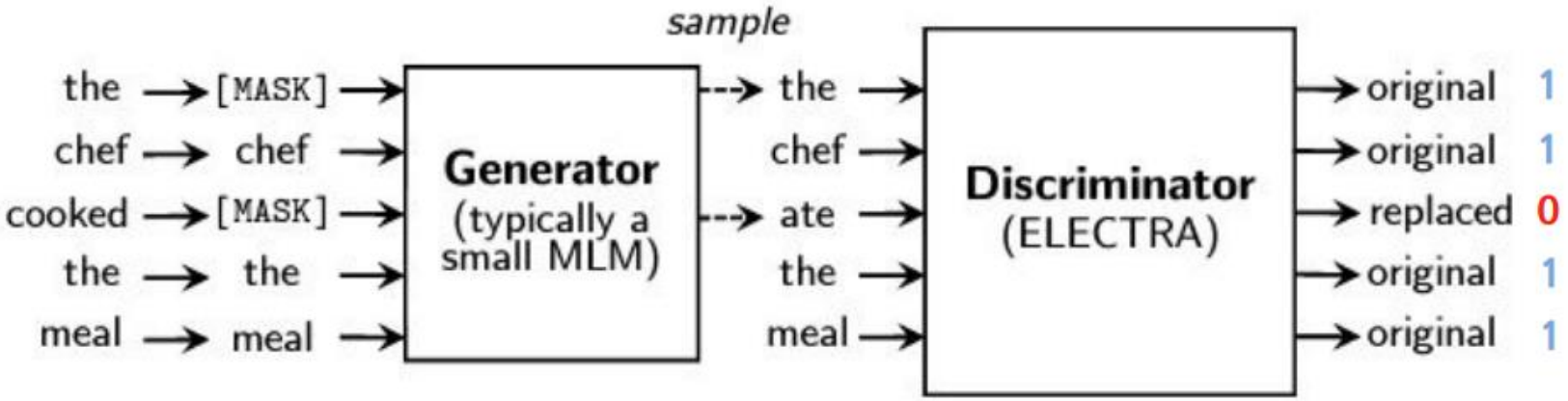
베이스라인 모델 : KoELECTRA v3



기존 BERT의 빈 칸 맞추기

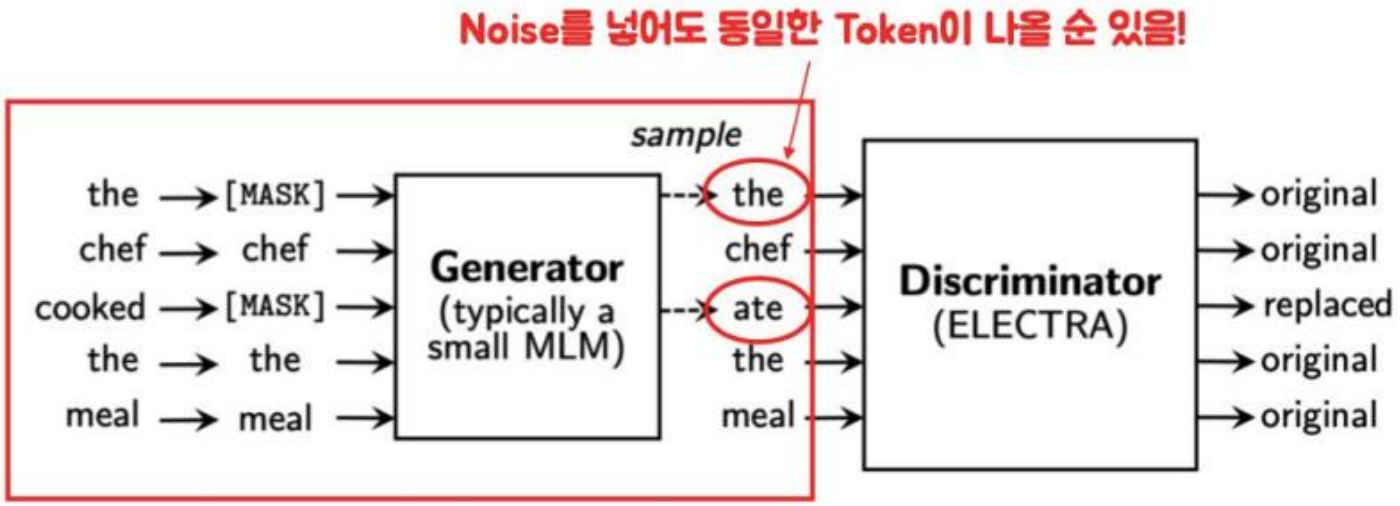
- ELECTRA는 기존 BERT에 대해 새로운 문제점을 제기
- **"오히려"** 라는 단어만 예측
  - "오늘", "강의", "조아"는 예측에 쓰이지 않음
  - [MASK] 토큰은 전체 토큰의 불과 15%...
- **이로 인해 더 많은 training step이 필요하다!**
- 어떻게 하면 **모든 Token**에 대해 예측하는 Task를 만들 수 있을까?

베이스라인 모델 : KoELECTRA v3



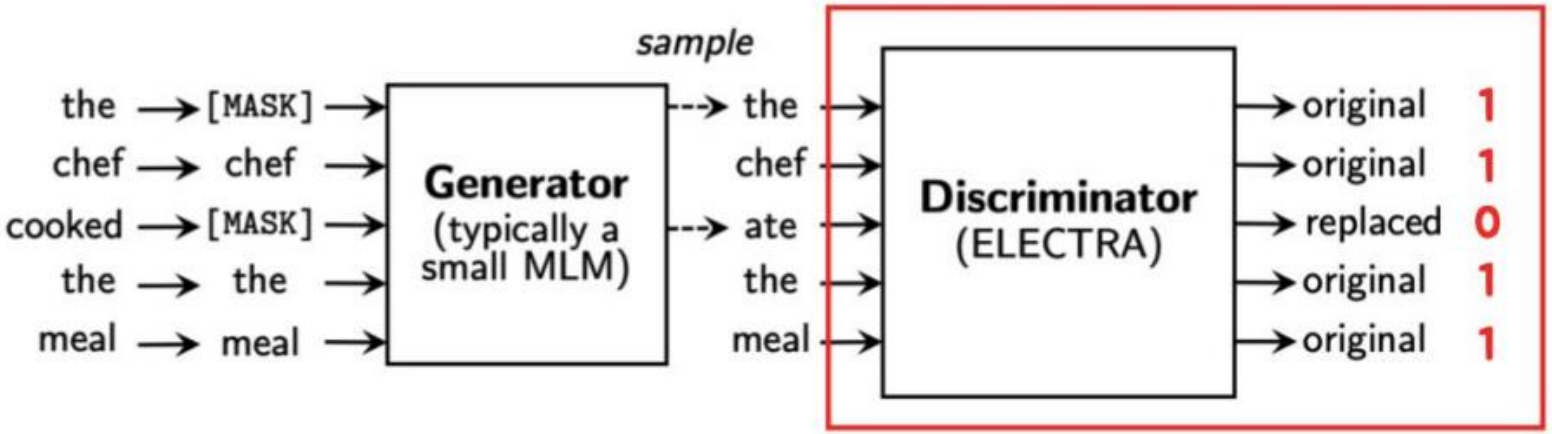
-> 모든 Token에 대하여 Fake인지 확인!

베이스라인 모델 : KoELECTRA v3



- 기존 BERT와 동일 (최적의 단어를 예측)
- 그 후 input에 noise를 넣어 fake token을 생성

베이스라인 모델 : KoELECTRA v3



- Sigmoid로 처리하여 Binary Classification
- 모든 Token에 대하여 계산!

베이스라인 모델 : KoELECTRA v3

ELECTRA는 학습 방법 등이 살짝 다르지만,

아키텍쳐는 BERT와 동일합니다!!

한국어 데이터셋에 pre-train한 ELECTRA 모델이 바로

KoELECTRA입니다.

KoELECTRA 배포자 박장원님 참고자료

- 깃헙

https://github.com/monologg/KoELECTRA/tree/master/finetune

- 블로그

https://monologg.kr/2020/05/02/koelectra-part1/

평가지표 : Accuracy

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1-Score} = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

		실제 정답	
		Positive	Negative
실험 결과	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

제한 사항

- **외부 데이터 사용 가능**
- **Pre-trained 모델 사용 가능**
  - 모델 라이브러리를 통해 이미 학습이 진행된 모델을 로드하여 사용할 수 있음
  - .pt, .pth, .h5 등의 pre-trained 가중치 파일을 업로드하여 사용할 수 있음
  - Transfer learning, Fine tuning 가능함
- **결과 제출 제한 : 1일 최대 24회**
- **Sample_submission.csv와 동일한 형태로 예측 파일을 만들어 제출**
- **최종 제출 파일 선택('결과제출' 탭에서 해당 파일의 '최종선택' 체크박스 선택)**
  * 최종 파일 미선택 시, public 스코어가 가장 높은 제출 파일 자동 선택

Index

1. 실전 캐글 프로젝트 커리큘럼 및 일정
2. AI CONNECT 플랫폼 소개 및 사용법 안내
3. 1차 모의경진대회 – (자연어) 도서자료 검색 효율화를 위한 기계독해
    - 과제 소개 (데이터셋 / 아키텍쳐 / 평가지표 / 제한사항)
    - 베이스라인 코드

End of document