KiSTi
www.kisti.re.kr
한국과학기술정보연구원
Korea Institute of Science and Technology Information

2023 DATA·AI 분석 경진대회

한국어 과학기술 논문의 초록 문장 분류

사이냅 DU (Document Understanding) 프로젝트

사이냅소프트 팀

목차

01

프로젝트 개요

- 문제 및 데이터 설명
- 프로젝트 진행 결과 요약

문제 및 데이터 설명

문제 정의

- 한국어 논문 초록에 존재하는 문장들에 기정의된 태그 중 전/후 문맥을 고려한 가장 알맞은 태그 부여

목적 및 배경

- 순차적 문장 분류를 위한 알고리즘 개발에 도움을 주기 위해 만들어진 생물, 의학 기반 초록 문장 태깅 데이터 PubMed200K 외 과학기술 논문 초록들을 대상으로 한 연구가 매우 적은 실정
- 한국어 과학기술 초록의 자동 분석을 토대로 한 다양한 분석 서비스 제공을 위함

데이터

- 한국어 과학기술 논문 초록 임의로 크롤링한 1천개의 데이터셋 PubKorSci-1k
- 구성: 1000개 초록, 6,728개 문장

태그	정의	개수
SI	현재까지 진행해 온 상황이나 예측되는 내용 기술하는 문장	521
PR	논문에서 풀고자 하는 문제점을 기술한 문장	693
RES	상기 정의된 research를 기술하는 핵심 문장	1,060
RED	상기 정의된 research를 기술하는 내용 중 일부, 단편적 내용을 기술하는 문장	2,265
EX	상기 정의된 EXPERIMENT를 설명하는 문장(개요, 환경, 과정) 등	1,377
DE	대상에 대한 설명, 정의, 특정한 역할 등을 기술하는 문장	806

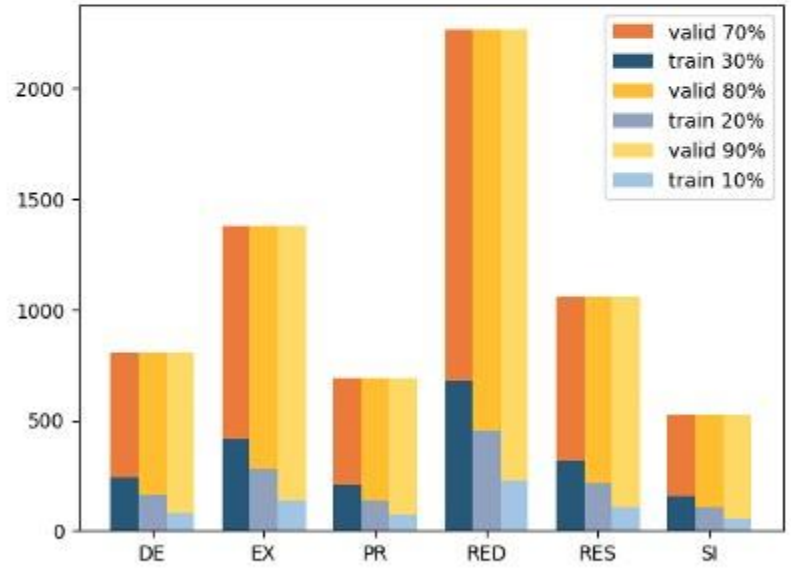
프로젝트 진행 결과 요약

구축 데이터

- 논문 48,405 편 / 107,702 문장
- RISS (학술연구정보서비스) 데이터 수집

평가 결과

PubKorSci-1K 활용 비율		평가 방법	
학습	평가	Weighted F1 Score	Accuracy
30%	70%	98.10%	98.10%
20%	80%	96.79%	96.80%
10%	90%	93.56%	93.58%
0%	100%	89.40%	89.37%



02

활용 데이터

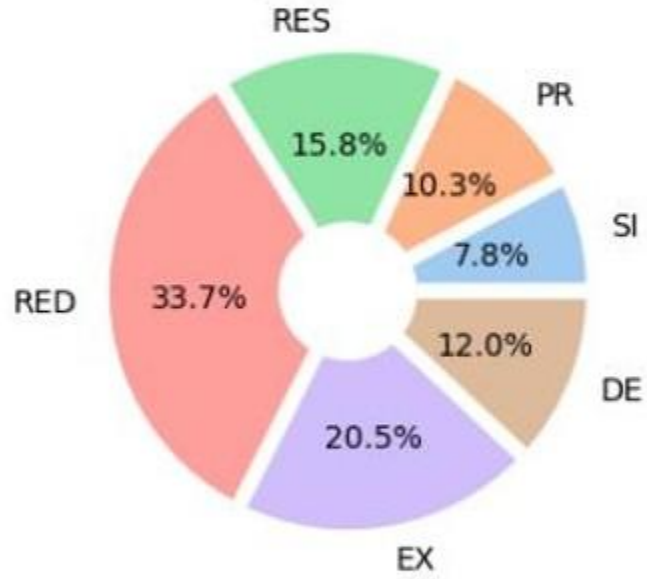
- PubKorSci-1k 분석
- 데이터 수집
- 데이터 생성

PubKorSci-1k 분석

불필요한 태그 데이터 제거 및 클래스 별 데이터 분포 확인

태그	개수
SI	521
PR	693
RES	1,060
RED	2,265
EX	1,377
DE	806
NOT	6

PubKorSci-1k 태그별 개수



NOT 태그 제거 후 데이터 분포

데이터 수집

학술연구정보서비스(RISS) 논문 초록 데이터 수

	분야	세부 분야	논문 수	수집된 문장수
1차	기술	공학	20,000	29,387
2차	기술	공학	12,205	40,193
	과학	자연과학	8,000	22,798
3차	기술	공학	3,900	5,816
	과학	자연과학	4,300	9,508

데이터 생성

PubKorSci-1k를 활용한
임베딩 벡터 분석

Synap-KNN-63K

Synap-KNN-29K

ChatGPT

Fine-tuning

PubKorSci-1k를 활용한
ChatGPT Fine-tuning

Synap-GPT-15K

A

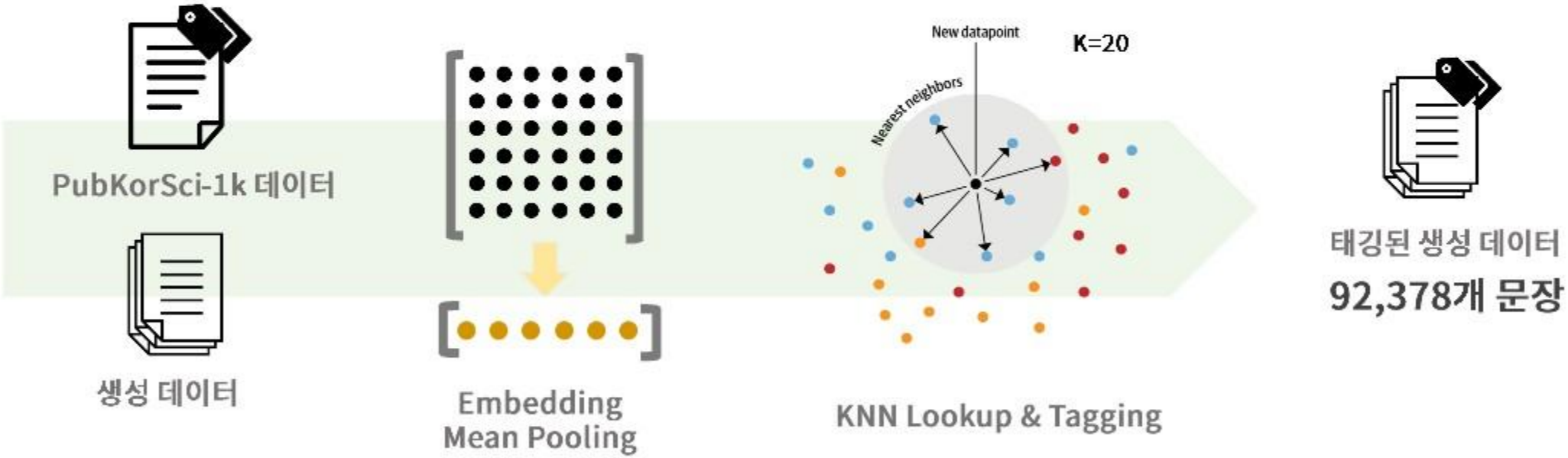
가

PubMed200K를 번역한
학습 데이터 활용

미사용

데이터 생성 – 최근접 임베딩 룩업

PubKorSci-1k 데이터를 학습한 임베딩 모델로 최근접 이웃 기법(K-Nearest Neighbor) 활용한 태깅 진행



데이터 생성 – Open AI Chat GPT Fine Tuning

GPT 3.5 모델에 PubKorSci-1k의 6,000개 초록 문장 태깅 Fine Tuning 진행

Fine Tuning 학습 데이터 예시

```
{"messages": [{"role": "user",
               "content": "다음 문장들을 분류해줘.
                          1. 또한 신뢰도를 평가하기 위해 이미 많은 수의 소프트웨어 신뢰도 성장 모델이 제안되었다.
                          2. 그런데 초기에 수집된 고장 데이터는 미래 고장 예측에 영향을 주지 않을 수도 있고 경우에 따라서는
                             미래 고장 예측 과정에서 왜곡된 결과를 초래할 수도 있다.
                          3. 이를 해결하기 위해서 이 논문에서는 부분 고장 데이터를 이용하여 적합도 평가를 수행하는 방법에
                             기반을 둔 소프트웨어 신뢰도 성장 모델 선택 방법을 제안한다."},
               {"role": "assistant",
                "content": "1. SI
                           2. PR
                           3. RES"}]}
```

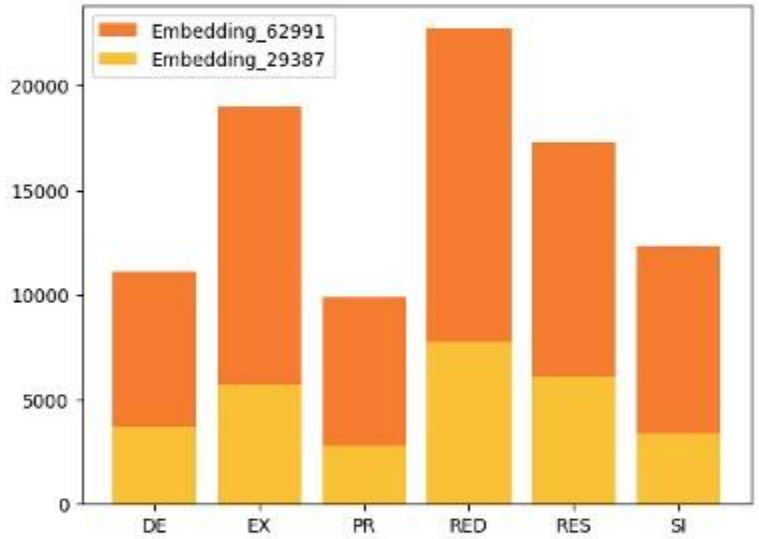
- Fine Tuning 결과
  - Train Loss: 0.117
  - Validation Accuracy: 86.9%

태깅된 생성 데이터

15,324개 문장

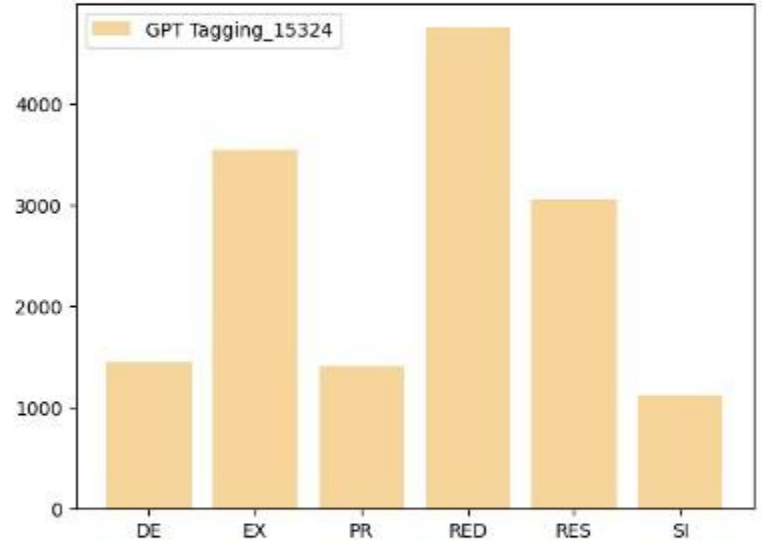
데이터 생성 – 생성 데이터

임베딩을 활용해 태깅한 데이터



DE	EX	PR	RED	RES	SI
11,132	18,980	9,882	22,719	17,336	12,329
12.1%	20.5%	10.7%	24.6%	18.8%	13.3%

ChatGPT를 활용해 태깅한 데이터



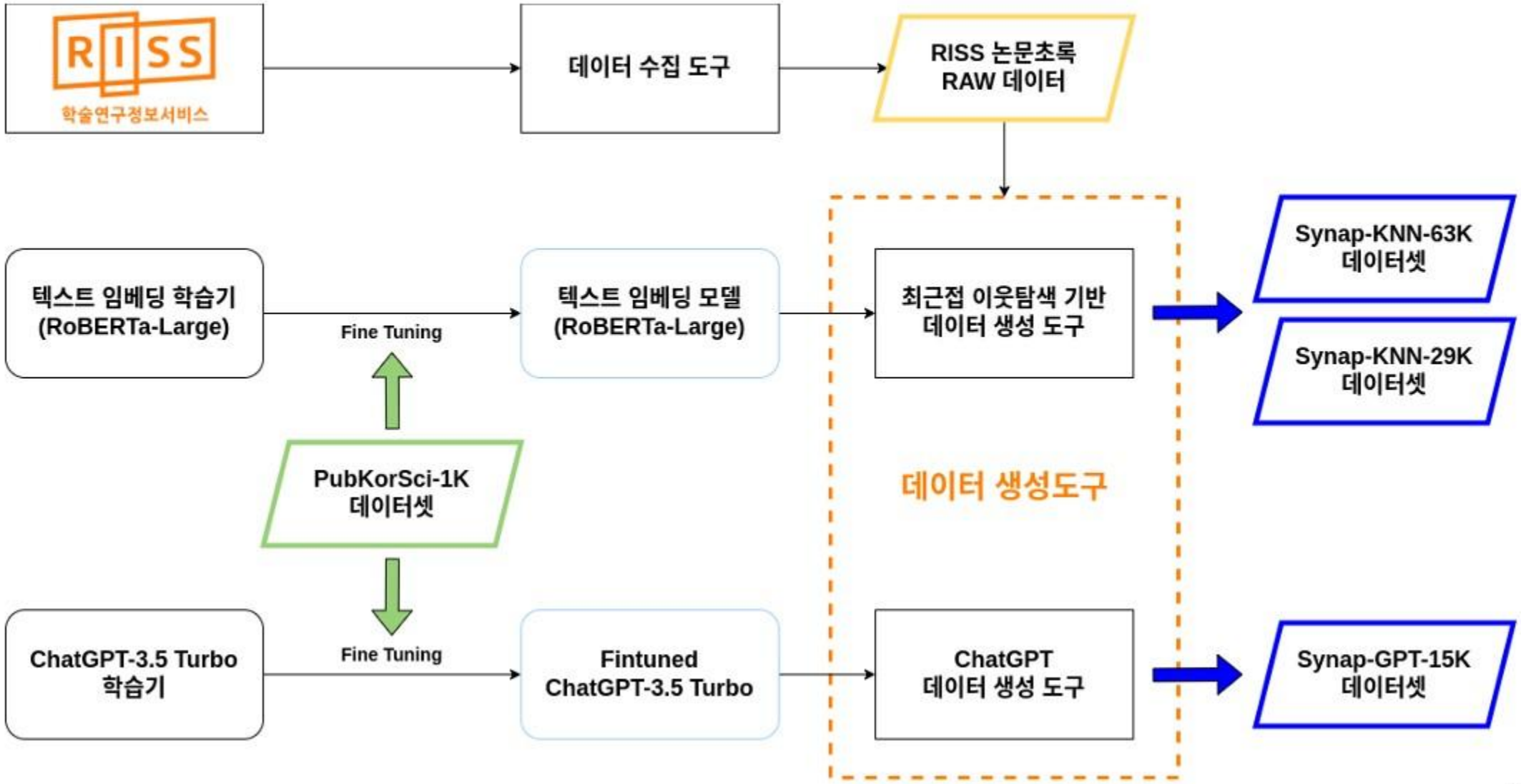
DE	EX	PR	RED	RES	SI
1,448	3,537	1,410	4,749	3,056	1,124
9.4%	23.1%	9.2%	31.0%	19.9%	7.3%

03

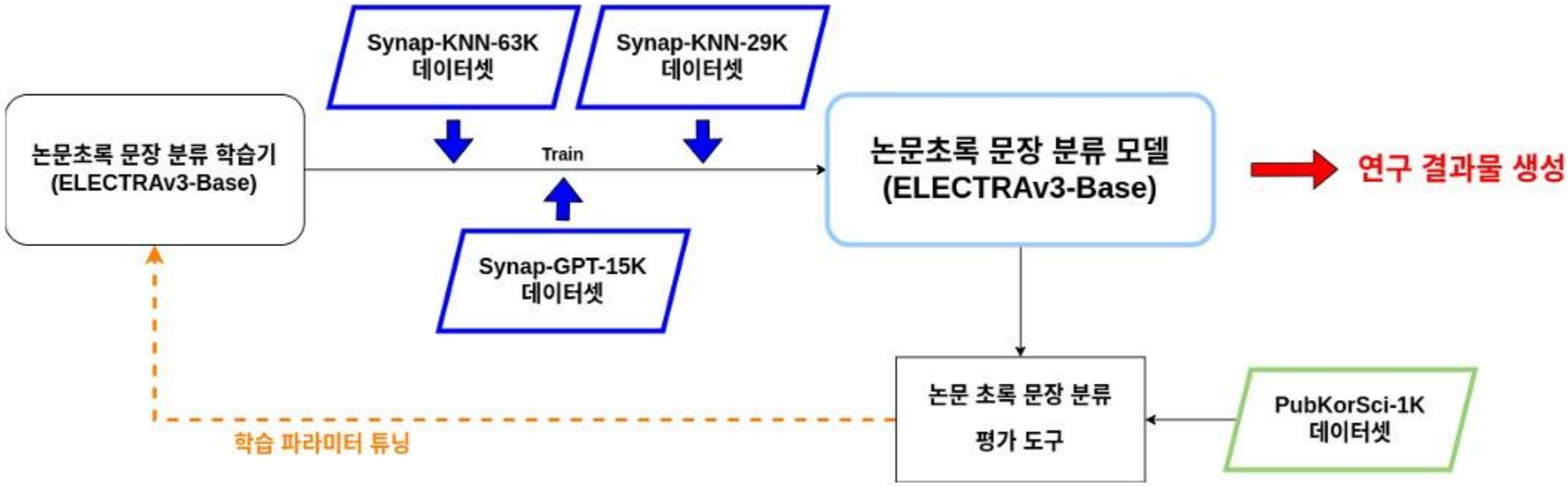
모델 개발 방법

- 학습 시스템 구성
- 백본 모델 선정
- 학습 모델 구성
- 학습 기법

학습 시스템 구성 – 데이터 구축



학습 시스템 구성 – 모델 학습



백본 모델 선정

ELECTRA Efficiently Learning an Encoder that Classifies Token Replacements Accurately

MLM의 비효율적인 학습을 개선하기 위해 RTD를 사용하며, GAN과 유사한 Generator 및 Discriminator를 구성하여 학습하고, 높은 성능과 최적화 속도를 갖춘 모델.

DeBERTa Decoding-enhanced BERT with Disentangled Attention

디코딩 과정을 개선하여 임베딩 생성 유연성을 제공하며, Disentangled Self-Attention 메커니즘을 사용하여 정확한 상호 작용 모델링과 언어 관계 이해에 용이하며, 언어 특성에 민감한 작업에서 높은 성능을 보임.

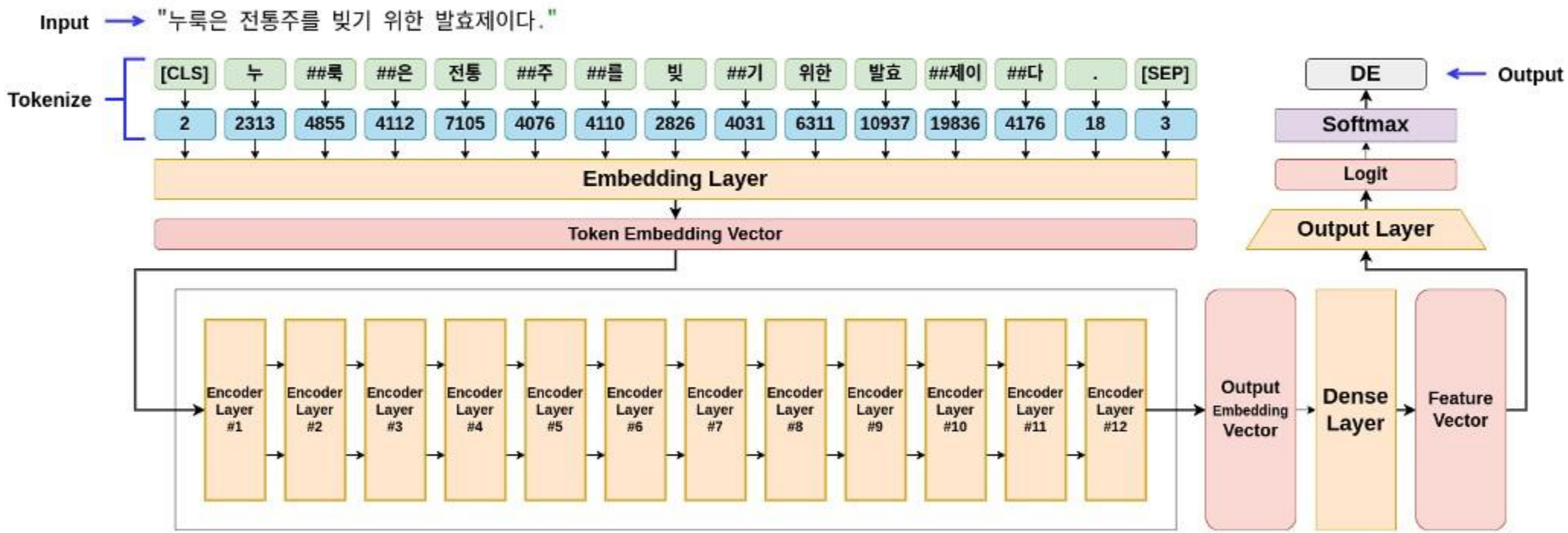
RoBERTa A Robustly Optimized BERT Pretraining Approach

BERT의 변형 모델로, NSP 생략 및 Dynamic masking 전략을 사용하며, 더 많은 훈련 데이터와 큰 배치 크기로 학습하여 뛰어난 성능을 달성함.

BERT Bidirectional Encoder Representations from Transformers

양방향 Transformer 모델로, MLM 및 NSP를 통해 문맥을 고려하여 단어 및 문장 임베딩을 생성하며 다양한 자연어 처리 작업에서 높은 성능을 보임.

학습 모델 구성


Input
"누룩은 전통주를 빚기 위한 발효제이다."
Tokenize
[CLS] 누 ##룩 ##은 전통 ##주 ##를 빚 ##기 위한 발효 ##제이 ##다 . [SEP]
2 2313 4855 4112 7105 4076 4110 2826 4031 6311 10937 19836 4176 18 3
Embedding Layer
Token Embedding Vector
Encoder Layer #1
Encoder Layer #2
Encoder Layer #3
Encoder Layer #4
Encoder Layer #5
Encoder Layer #6
Encoder Layer #7
Encoder Layer #8
Encoder Layer #9
Encoder Layer #10
Encoder Layer #11
Encoder Layer #12
Output Embedding Vector
Dense Layer
Feature Vector
Output Layer
Logit
Softmax
DE
Output

학습 기법

Multi-GPU 학습
복수의 GPU를 사용한 학습 지원
(RTX3090 * 2, RTX 2080Ti * 4)

학습 속도 향상
컴퓨팅, 통신, 메모리, IO 효율
최적화로 학습 소요시간 47% 감소

대형 배치 사이즈
메모리 효율 향상으로
학습 배치 사이즈 2.2배 증가

04

실험 및 평가

- 평가 지표
- 실험 결과

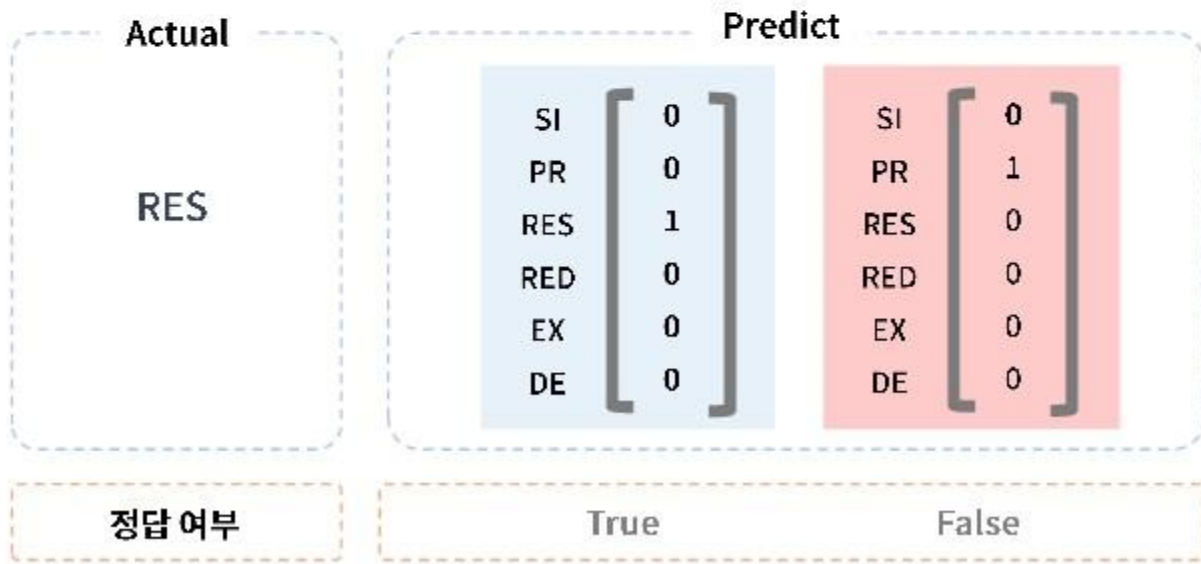
평가지표

Accuracy

모델이 분류한 클래스가 실제 정답과 얼마나 일치하는지 확인하는 지표

Accuracy 계산 방법

$$\text{Accuracy} = \frac{\text{정답으로 분류된 샘플 수}}{\text{전체 샘플 수}}$$



평가지표

Weighted F1 Score

클래스의 수가 여러 개인 분류 모델의 클래스 불균형을 고려한 평가 지표
각 클래스 별로 F1 Score를 계산한 후에 클래스에 따라 가중치를 적용하여 계산

Weighted F1 Score 계산 방법

$$\text{Weighted F1 Score} = \sum_{k=1}^{n} F1\_k * weight\_k$$

Class
SI
PR
RES
RED
EX
DE

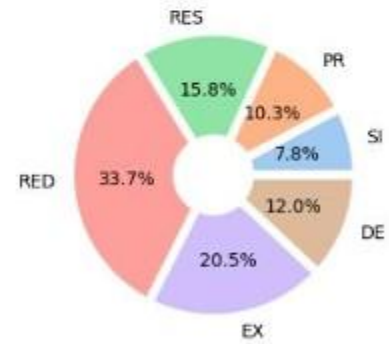
Predict

Actual		
	TP	TN
	FP	FN

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

F1-Score
SI
PR
RES
RED
EX
DE

클래스별 F1-Score 계산

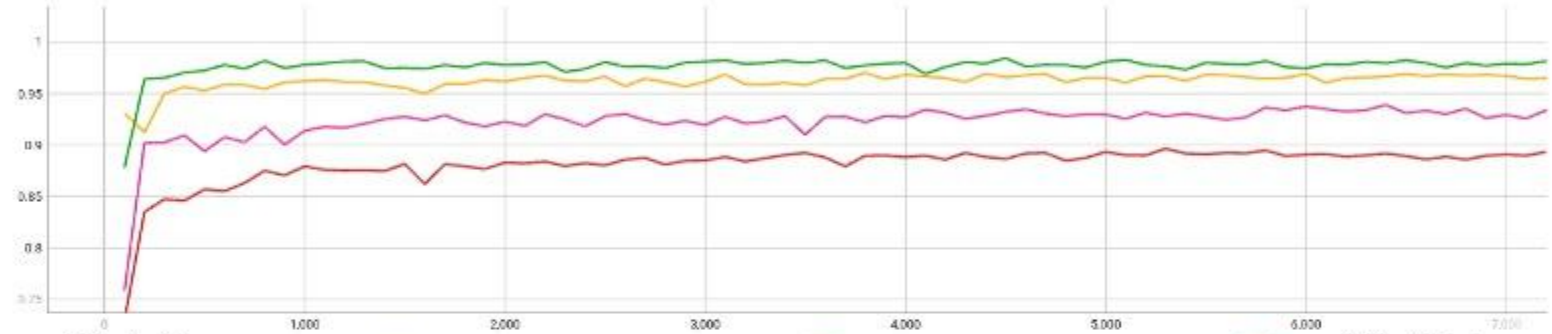


클래스 분포에 따른 가중치(weight)

실험 결과 - PubKorSci-1k 사용 비중 별 학습 결과

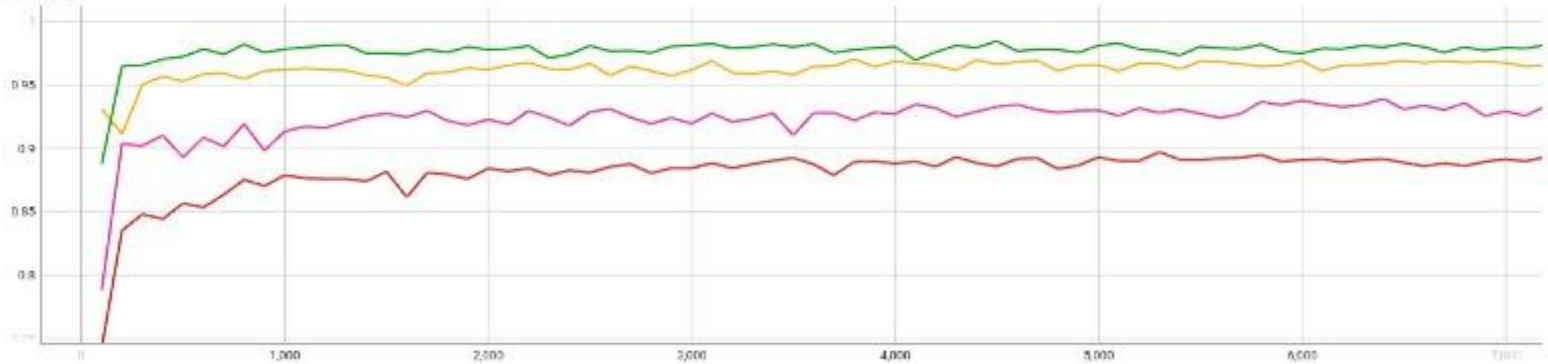
백본 모델 : ELECTRA v3
학습 데이터 : Synap-GPT-15K + Synap-KNN-(29K + 63K) + PubKorSci-1K n%

F1 score



30% 학습에 사용 : 98.10%　20% 학습에 사용 : 96.79%　10% 학습에 사용 : 93.56　0% 학습에 사용 : 89.40%

Accuracy

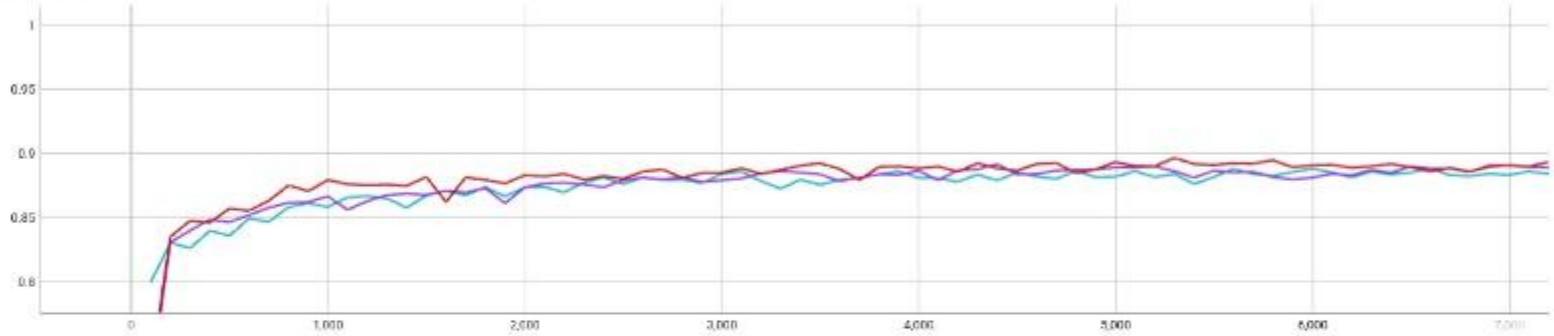


30% 학습에 사용 : 98.10%　20% 학습에 사용 : 96.80%　10% 학습에 사용 : 93.58%　0% 학습에 사용 : 89.37%

실험 결과 – 학습 데이터 구성에 따른 학습 결과

백본 모델 : ELECTRA v3
PubKorSci-1k 학습 사용 비중 : 0%

F1 score



Synap-GPT-15K+Synap-KNN-(29K+63K) 89.40 : %

Synap-KNN-29K+63K : 89.06%

Synap-KNN-63K : 88.31%

Accuracy

Synap-GPT-15K+Synap-KNN-(29K+63K) : 89.37 %

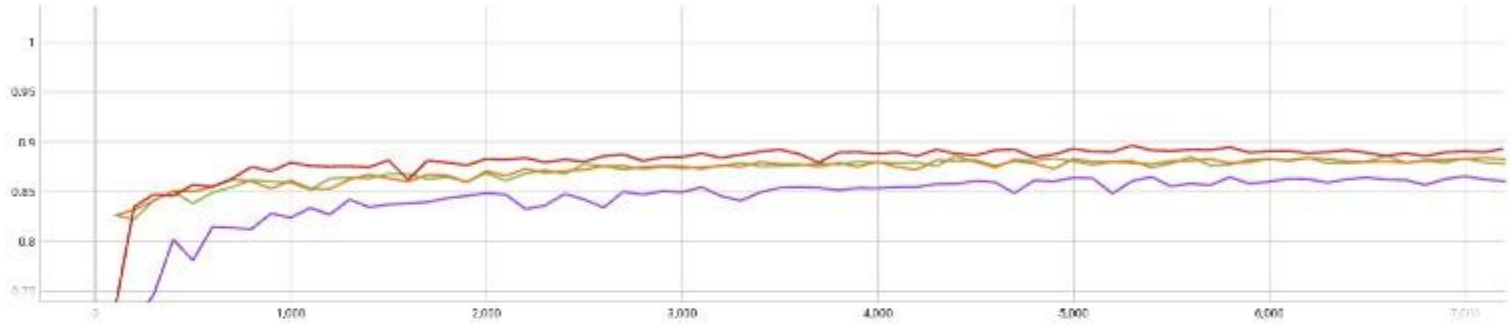
Synap-KNN-29K+63K : 89.04%

Synap-KNN-63K : 88.25%

실험 결과 – 백본 모델에 따른 학습 결과

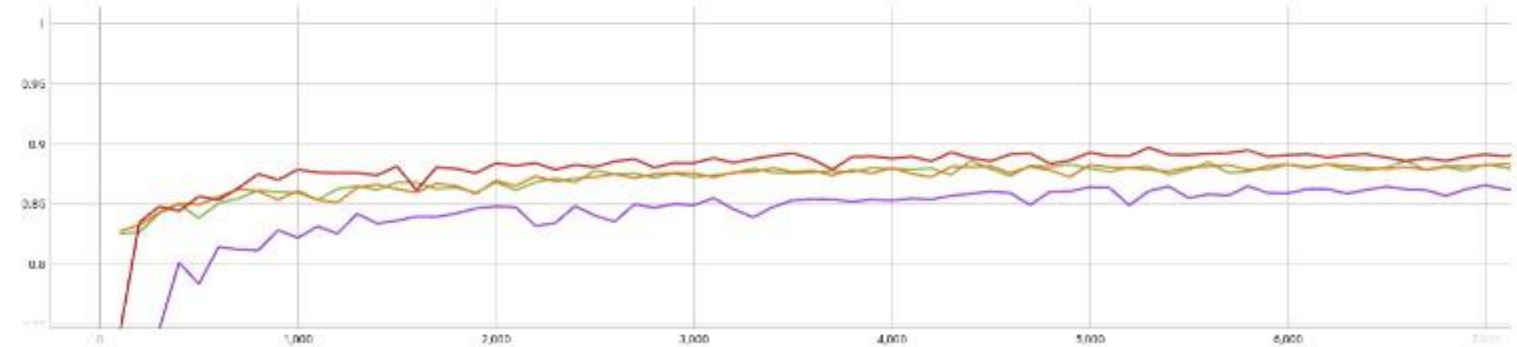
학습 데이터 : Synap-GPT-15K + Synap-KNN-(29K + 63K) + PubKorSci-1K n%
PubKorSci-1k 학습 사용 비중 : 0%

F1 score



ELECTRA v3 : 89.40%　RoBERTa : 88.35%　BERT : 88.29%　DeBERTa : 86.57%

Accuracy



ELECTRA v3 : 89.37 %　RoBERTa : 88.32%　BERT : 88.25%　DeBERTa : 86.58%

05

활용 계획 및 기대 효과

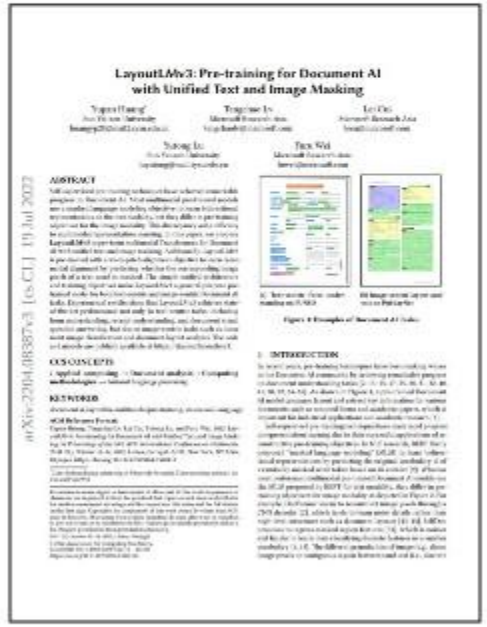
- 분류 모델 활용 및 후속 연구 계획
- 기대효과

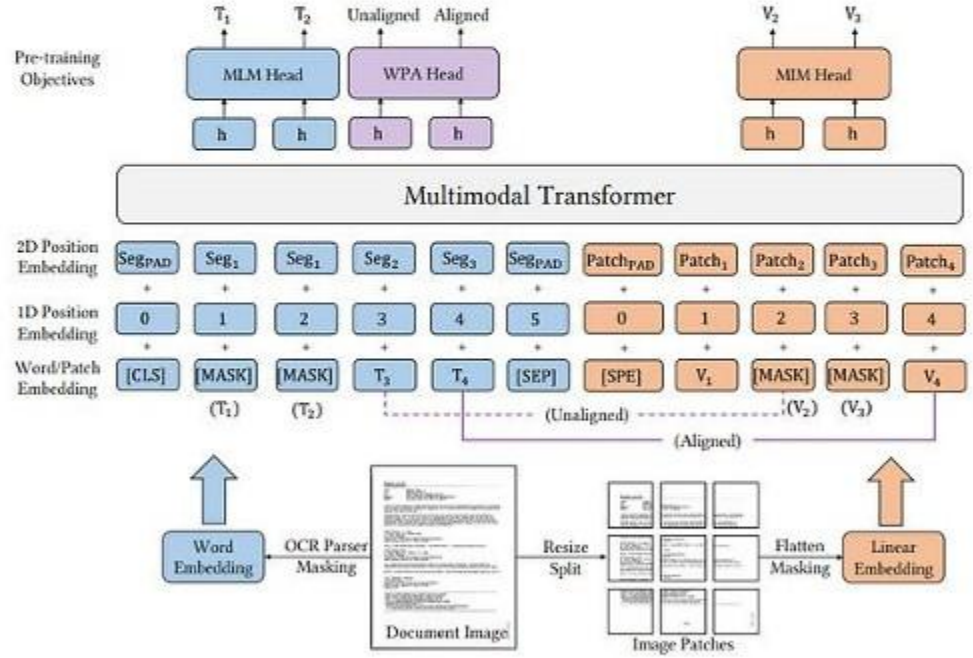
분류 모델 활용 및 후속 연구 계획

텍스트와 시각적 정보를 함께 활용하는 문서 이해 모델로 확장

문서를 정확하게 이해하려면 글자 서식, 배치 등의 시각적 요소도 함께 고려해야 정확한 인식 가능

따라서, 이미지 특징, 텍스트, 레이아웃 정보 모두를 고려한 모델링이 필요함


LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking



LayoutLM v3

기대효과

학술 정보 제공 플랫폼 개발

연구 동향 분석 서비스 제공

우수 연구 지원 및 자원 할당

문헌 조사 지원 서비스 제공

06

시연

감사합니다

사이냅소프트 팀