KorSciDeBERTa: 한국어 과학기술 분야를 위한 DeBERTa 기반 사전학습 언어모델

김성찬°[1,3], 김경민[1], 김은희[1,3], 이민호[1,3], 이승우[1,3], 최명석[2]
[1]한국과학기술정보연구원 인공지능기술연구팀 [2]한국과학기술정보연구원 인공지능데이터연구단, [3]UST-KISTI 응용 AI
{sckim,kkmkorea,ehkim,cokeman,swlee,mschoi}@kisti.re.kr

KorSciDeBERTa: A Pre-trained Language Model Based on DeBERTa for Korean Science and Technology Domains

Seongchan Kim°[1,3], Kyung-min Kim[1], Eunhui Kim[1,3], Minho Lee[1,3], Seungwoo Lee[1,3], Myung-Seok Choi[2]
[1]AI Tech Research Team KISTI, [2]Dept. of AI Data Research KISTI, [3]Applied AI UST-KISTI

요 약

이 논문에서는 과학기술분야 특화 한국어 사전학습 언어모델인 KorSciDeBERTa를 소개한다. DeBERTa Base 모델을 기반으로 약 146GB의 한국어 논문, 특허 및 보고서 등을 학습하였으며 모델의 총 파라미터의 수는 180M이다. 논문의 연구분야 분류 태스크로 성능을 평가하여 사전학습모델의 유용성을 평가하였다. 구축된 사전학습 언어모델은 한국어 과학기술 분야의 여러 자연어처리 태스크의 성능향상에 활용될 것으로 기대된다.

주제어: 과학기술, 한국어, 사전학습 언어모델, DeBERTa, KorSciDeBERTa

1. 서론

GPT-3, PaLM, MT-NLG를 비롯하여 범용 거대언어모델(LLM)이 등장하였지만 환각(Hallucination), 데이터 품질, 데이터 보안과 같은 문제가 이슈가 되고 있다. 이러한 문제를 피하기 위해 소규모의 분야 특화된 언어모델이 필요하며 기업이나 전문기관에서는 특정 도메인 문제를 해결하기 위해 고품질의 자체데이터에 훈련된 도메인 모델을 개발하고 있다. Eckerson Group이 실시한 한 설문조사[1]에서는 약 30%의 회사가 자체 언어모델을 구축하고 있다고 대답하였고, 이러한 맥락에서 과학기술분야에서도 SciBERT, SciFive, Galactica와 같은 도메인 특화 언어모델이 꾸준히 개발되어 발표되고 있다.

본 논문에서는 한국어 과학기술분야의 언어모델인 KorSciDeBERTa[2,3] 모델 개발 과정을 소개한다. 마이크로소프트사의 DeBERTa Base 모델(180M 파라미터)을 활용하여 약 146GB의 한국어 논문, 특허 및 보고서 등이 사전학습된 언어모델이다.

이러한 소규모의 과학기술 특화 언어모델이 개발되면 사전학습(Pre-trained) 모델을 기반으로 미세조정(Fine-tuning) 방식으로 특정한 단어, 문장, 단락, 문서단위의 과학기술문서 분석에 활용될 수 있으며, 과학기술 분야에서 전문 QA(Question & Answer)나 요약(Summarization) 추천(Recommendation)과 같은 태스크에 활용될 수 있다.

[1]https://shorturl.at/eiOPR
[2]https://aida.kisti.re.kr/model/bc928a9b-1574-4952-84c0-0f6529290f3c
[3]https://huggingface.co/kisti/korscideberta

2. 관련연구

과학기술 도메인에 특화된 언어모델로, 영어의 경우에는 SciBERT[1], BioAlbert[2], 그리고 Galactica[3]와 같은 모델들이 있다. 한국어의 경우에는 KorSciBERT[4], KorSciELECTRA[5], KoPatBERT[6], 그리고 KorPatELECTRA[7]와 같은 모델들이 개발되었다. 한국어 과학기술 분야의 언어모델 중에는 과학기술 전 분야를 커버하기 위해 논문, 연구 보고서, 특허 문서 등을 학습한 KorSciBERT[4]와 KorSciELECTRA[5]가 있으며, 주로 특허 문서를 학습시킨 KoPatBERT와 KoPatELECTRA가 공개되어 있다.

KorSciBERT는 구글의 BERT모델을 아키텍쳐를 기반으로 논문과 특허 등 총 97GB 약 3억 8천문장을 학습시킨 모델로 사용된 코퍼스를 기반으로 명사 및 복합명사 약 600만개의 사용자사전이 추가된 Mecab-ko Tokenizer와 기존 BERT WordPiece Tokenizer가 병합된 토크나이저를 사용하였으며 15000개의 어휘가 사용되었다. KorSciELECTRA는 Google ELECTRA base 모델의 아키텍쳐를 기반으로 논문, NTIS 연구과제, 특허, 뉴스, 한국어 위키 코퍼스 총 141GB (문서 407만건)를 사전학습한 언어모델이다. 16200개의 어휘를 사용하였으며 KorSciBERT와 마찬가지로 전문용어 사용자사전이 추가된 Mecab-ko Tokenizer와 기존 BERT WordPiece Tokenizer가 병합된 토크나이저를 사용하였다. 연구보고서의 과학기술표준분류 145개의 연구분야 분류 태스크에서 top 3 Micro F1기준으로 88.58의 성능을 보였으며 특허의 경우 Macro-F1기준으로 73.0의 성능을 보였다.

[4]https://doi.org/10.23057/46
[5]https://doi.org/10.23057/51

KoPatBERT는 특허분야에 특화된 언어모델로 특허문헌 400만건 120GB를 활용하여 학습하였으며 특허문서를 주제분류체계에 따라 구분하는 CPC(Cooperative Patent Classification) 분류 태스크에서 정확도 76.32를 기록하였다. KorPatELECTRA는 데이터를 추가하여 특허공보 466만건 130GB로, 어휘수를 35000으로 늘리고 ELECTRA base 모델을 학습하였다. 그 결과 화학특허 NER과 MRC(PatQuAD)와 같은 태스크에서 KoPatBERT보다 좋은 성능을 보여주었다.

3. KorSciDeBERTa: 모델 및 데이터

3.1 모델구조

KorSciDeBERTa는 MS의 DeBERTa[8]에 기반을 두고 있다. DeBERTa는 각 단어의 컨텐츠와 위치를 각각 인코딩하여 두개의 벡터를 사용하고 단어간 어텐션 가중치를 상대위치에 기반하여 연산하는 Disentangled attention 방식과, 마스크 언어모델링(MLM)을 수행할때 컨텐츠 및 위치정보를 활용하는 Enhanced mask decoder 방식을 사용하는 것을 특징으로 한다. KorSciDeBERTa의 주요 특징은 다음과 같다:

- 모델구조: Microsoft DeBERTa-V2
- 모델규모: Base (180M)
- 토크나이저: 형태소 분리(Mecab-ko) + Word Piece (6백만개의 명사 & 합성명사 사용자 사전 추가)
- 어휘 수: 128K
- train_batch_size: 4,096 * 4 accumulative update = 16,384 13.2억 샘플
- num_train_steps: 1,600,000 (4.97 epochs)
- max_seq_length: 512

3.2 학습데이터 및 전처리

학습에 사용한 데이터는 표 1에 기술되어 있다. HTML, latex, img 등 각종 태그 제거, 괄호문자 및 특수문자 제거 등과 논문, 보고서, 특허 데이터에 따라 각각 정제를 실시하였다. 상대적으로 작은 규모의 데이터 및 작은 수의 어휘를 사용하여 사전학습하기 위해 저 빈도의 덜 중요한 문자들을 많이 정제하였다. 유니코드 정규화 후 유니코드 기준 한글자 이상 포함한 한글문장만 필터링하였다. Mecab-ko[6]으로 형태소분석을 실시하고 조사와 어미를 어간으로 분리하였으며 전문용어가 잘못 분리되는 경우를 줄이기 위해 과학기술 전문용어사전을 추가하였다. 즉 과학기술 말뭉치에서 추출한 명사&합성명사 (6백만개)를 Mecab-ko 사용자 사전에 추가하여 형태소 분석을 실시하였다. 예를들어 '구동부'를 Mecab-ko 사전에 추가하여 '구동', '부'로 분리하지 않고 한 단어로 인식되게끔 하였다. 서브워드 토크나이징으로는 SentencePiece 알고리즘[7]을 사용하였다. 토크나이저 학습시에는 146GB 말뭉치로 RAM 800G 이상 사용하여 32코어급 서버에서 하루정도 소요되었다.

표 1. 학습 데이터

구분	말뭉치명	출처	용량(GB)
논문	국내 과학기술 논문 전문	KISTI	6.3
	전문분야 말뭉치(논문)	AIHub	3.5
	논문자료 요약(전체/섹션)	AIHub	2.9
특허	특허/실용 공개/등록공보	특허청	7.3
	전문분야 말뭉치(특허)	AIHub	13.2
	논문자료 요약(특허명세서)	AIHub	2.7
보고서	국가R&D보고서(메타,본문)	KISTI	24.6
일반	한국어 위키	Web	0.7
	뉴스(일반, IT)	Web	60.2
	신문 말뭉치	국립국어원	16.2
	대규모 웹 한국어 말뭉치	AIHub	9.6
합계			146.3

3.3 사전학습(Pre-training)

문장중 랜덤하게 비어있는 단어를 예측하는 MLM 태스크로 사전학습을 실시하였다. 학습을 관찰하기 위해 Train/eval loss 를 추적하였다(그림 1). 약 100만 스텝 이후에 loss의 변화폭이 크지 않음을 확인하였으며 160만 스텝에서 학습을 중지하였다. 학습은 한국과학기술정보원의 슈퍼컴퓨터 뉴런[8]의 전용 3 노드를 사용하였다. 각 노드에는 1TB의 RAM과 8개의 A100(80GB) GPU가 탑재되어 있으며 160만 iteration (4.97 epoch) 학습에 약 2달의 시간이 소요되었다. 참고로 32.5만, 72.5만 112.5만 스텝에서 슈퍼컴퓨터 정기점검 및 알수 없는 오류로 학습이 중단되었으며 이때마다 완료된 checkpoint부터 재학습을 실시하였다.

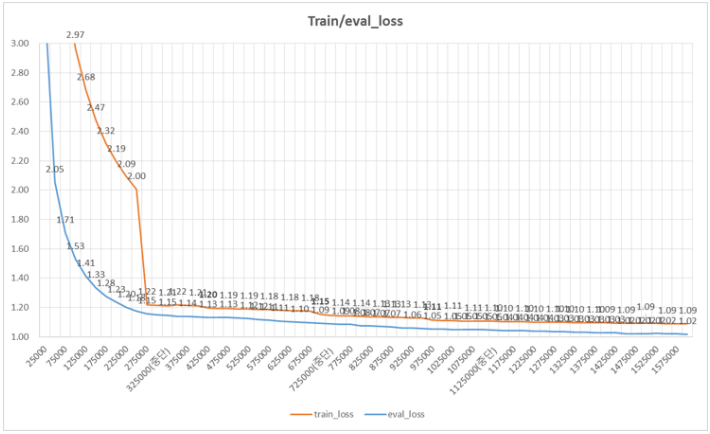

그림 1. 학습과정

[6]https://bitbucket.org/eunjeon/mecab-ko

[7]https://github.com/google/sentencepiece

[8]https://www.ksc.re.kr/byjw/neuron

4. 평가

본 언어모델의 성능평가는 국내논문 과학기술표준분류 태스크에 미세조정(Fine-tuning)하여 평가하는 방식을 사용하였다. 160만 스텝에서 학습이 완료된 모델(checkpoint)을 가지고 논문 연구분야 분류데이터셋으로 미세조정하여 주제분류 성능을 평가하였다.

논문 연구분야 분류데이터셋[9]은 논문내용과 주제에 따라 연구분야를 분류한 데이터셋으로 전문가가 국내 논문 3만건을 대상으로 연구분야를 분류(Annotation)한 데이터셋이다. 국가과학기술표준분류(2018년 개정)의 소분류 기준으로 레이블(정답)이 최대 3개까지 달려있다. 예를들어 2013년도에 한국지구과학지에 발간된 '풍력-기상자원지도에 기반한 제주 행원 풍력발전단지 효율성 평가'라는 제목의 논문에는 EF0606[대분류(EF): 에너지/자원, 중분류(EF06): 신재생에너지, 소분류(EF0606): 풍력], ND0501[대분류(ND): 지구과학, 중분류(ND05): 기상과학, 소분류(ND0501): 기상관측/분석기술], ND0503[대분류(ND): 지구과학, 중분류(ND05): 기상과학, 소분류(ND0503): 기상예보기술]의 레이블이 달려있다. 참고로 과학기술표준분류는 대분류 33개, 중분류 372개, 소분류 2898개로 이루어져있다.

이중 대분류 레벨(33개 카테고리)에서 분류를 예측하였으며 이를 위해 3만건 중 영문 제목과 초록만 있는 데이터(약 20%)는 제외하고 23,937건을 이용하였다. 학습, 검증, 테스트 데이터의 비율은 각각 0.9:0.5:0.5이다. 평가지표로 정답 Top 3 중에서 최소 1개를 예측하는 기준인 F1-micro/macro 및 정답 Top 3 중 모든 정답을 예측하는 기준인 F1-strict을 사용하였다. 이로써 F1-micro는 0.85, F1-macro는 0.52, F1-strict는 0.71의 성능을 달성하였다.

현재는 제한적인 평가만 이루어진 상태이고 KorSciBERT, KorSciELECTRA, KLUE RoBERTa와 같은 타 언어모델과 성능비교를 진행하고 있다.

5. 결론 및 향후계획

본 논문에서는 점차 수요가 많아지는 인공지능 관련한 과학기술분야의 자연어 관련 태스크의 성능을 높이기 위해 과학기술 전문 텍스트를 수집하여 180M의 파라미터수를 가진 한국어 버전의 과학기술전문 언어모델 KorSciDeBERTa를 소개하였다. 이 사전학습 언어모델은 과학기술분야의 주제분류, 단락순위화, MRC등의 기반 모델로 활용될 수 있다. 추후 KorSciBERT, KorSciELECTRA와 같은 타 언어모델과 성능비교 및 회귀언어모델인 GPT를 활용한 생성형 모델(KorSciGPT)을 개발하는 계획을 가지고 있다.

[9]https://doi.org/10.23057/50

Acknowledgement

본 연구는 2023년도 한국과학기술정보연구원 주요사업의 지원을 받아 수행되었습니다.(Data/AI 기반 문제해결 체계 구축, K-23-L04-C05-S01)

참고문헌

[1] I. Beltagy, K. Lo, and A. Cohan, "Scibert: A pretrained language model for scientific text," 2019.

[2] U. Naseem, M. Khushi, V. Reddy, S. Rajendran, I. Razzak, and J. Kim, "Bioalbert: A simple and effective pre-trained language model for biomedical named entity recognition," 2020.

[3] R. Taylor, M. Kardas, G. Cucurull, T. Scialom, A. Hartshorn, E. Saravia, A. Poulton, V. Kerkez, and R. Stojnic, "Galactica: A large language model for science," 2022.

[4] 한국과학기술정보연구원, "과학기술분야 bert 사전학습 언어모델 (korscibert)," 2021. [Online]. Available: https://doi.org/10.23057/46

[5] 한국과학기술정보연구원, "과학기술분야 electra 사전학습 언어모델 (korscielectra)," 2022. [Online]. Available: https://doi.org/10.23057/51

[6] 한국특허정보원, "특허분야 특화된 한국어 ai언어모델 korpatbert," 2022. [Online]. Available: https://github.com/kipi-ai/korpatbert

[7] 장지모;민재옥;노한성, "Korpatelectra : 자연어처리 분야에서의 성능 향상을 위한 한국어 특허 문헌 사전학습 언어모델(korpatelectra)," *한국컴퓨터정보학회논문지*, Vol. 27, No. 2, pp. 15–23, 2022.

[8] P. He, X. Liu, J. Gao, and W. Chen, "Deberta: Decoding-enhanced bert with disentangled attention," 2021.