JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 27 No. 2, pp. 15-23, February 2022
https://doi.org/10.9708/jksci.2022.27.02.015

KorPatELECTRA : A Pre-trained Language Model for Korean Patent Literature to improve performance in the field of natural language processing(Korean Patent ELECTRA)

Ji-Mo Jang*, Jae-Ok Min*, Han-Sung Noh*

*Staff, Korea Institute of Patent Information, Daejeon, Korea
*Part leader, Korea Institute of Patent Information, Daejeon, Korea
*Team leader, Korea Institute of Patent Information, Daejeon, Korea

[Abstract]

In the field of patents, as NLP(Natural Language Processing) is a challenging task due to the linguistic specificity of patent literature, there is an urgent need to research a language model optimized for Korean patent literature. Recently, in the field of NLP, there have been continuous attempts to establish a pre-trained language model for specific domains to improve performance in various tasks of related fields. Among them, ELECTRA is a pre-trained language model by Google using a new method called RTD(Replaced Token Detection), after BERT, for increasing training efficiency. The purpose of this paper is to propose KorPatELECTRA pre-trained on a large amount of Korean patent literature data. In addition, optimal pre-training was conducted by preprocessing the training corpus according to the characteristics of the patent literature and applying patent vocabulary and tokenizer. In order to confirm the performance, KorPatELECTRA was tested for NER(Named Entity Recognition), MRC(Machine Reading Comprehension), and patent classification tasks using actual patent data, and the most excellent performance was verified in all the three tasks compared to comparative general-purpose language models.

▸ **Key words**: Patent, ELECTRA, pre-training, NLP, tokenizer, Language model

• First Author: Ji-Mo Jang, Corresponding Author: Han-Sung Noh
*Ji-Mo Jang (bi02089@kipi.or.kr), Korea Institute of Patent Information
*Jae-Ok Min (okauto@kipi.or.kr), Korea Institute of Patent Information
*Han-Sung Noh (neodream@kipi.or.kr), Korea Institute of Patent Information
• Received: 2021. 12. 14, Revised: 2022. 01. 24, Accepted: 2022. 01. 26.
• This paper is an extension of the paper("Korean Patent ELECTRA : a pre-trained Korean Patent language representation model for the study of Korean Patent natural language processing(KorPatELECTRA)") published at the 64th Summer Conference of the Korea Society of Computer Information in 2021.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

[요 약]

특허 분야에서 자연어처리(Natural Language Processing) 태스크는 특허문헌의 언어적 특이성으로 문제 해결의 난이도가 높은 과제임에 따라 한국 특허문헌에 최적화된 언어모델의 연구가 시급한 실정이다. 최근 자연어처리 분야에서는 특정 도메인에 특화되게 사전 학습(Pre-trained)한 언어모델을 구축하여 관련 분야의 다양한 태스크에서 성능을 향상시키려는 시도가 지속적으로 이루어지고 있다. 그 중, ELECTRA는 Google이 BERT 이후에 RTD(Replaced Token Detection)라는 새로운 방식을 제안하며 학습 효율성을 높인 사전학습 언어모델이다. 본 연구에서는 대량의 한국 특허문헌 데이터를 사전 학습한 KorPatELECTRA를 제안한다. 또한, 특허 문헌의 특성에 맞게 학습 코퍼스를 정제하고 특허 사용자 사전 및 전용 토크나이저를 적용하여 최적화된 사전 학습을 진행하였다. KorPatELECTRA의 성능 확인을 위해 실제 특허데이터를 활용한 NER(Named Entity Recognition), MRC(Machine Reading Comprehension), 특허문서 분류 태스크를 실험하였고 비교 대상인 범용 모델에 비해 3가지 태스크 모두에서 가장 우수한 성능을 확인하였다.

▸ **주제어**: 특허, 일렉트라, 사전학습, 자연어처리, 토크나이저, 언어모델

I. Introduction

특허문헌은 다른 기술 자료보다 신뢰도가 높으며 지금까지의 과학기술 발전단계를 반영하고 있기 때문에 산업별로 기술경쟁력을 높일 수 있는 활용 가치가 높은 자원이다. 그러나 특허문헌은 일반인의 의사소통에서 사용하는 일상적인 문법 구조를 사용한 언어전달이 아닌 과학기술의 기술적이고 창조적 행위에 기반을 둔 언어들이 융합된 텍스트로 다양한 구문 복잡성과 모호한 표현으로 서술되어 있다[1]. 이러한 언어적 특이성에 따라 특허분야에서 자연어처리 문제의 해결을 위한 난이도가 높기 때문에 특허문헌에 최적화된 언어모델 연구가 시급한 실정이다.

최근 자연어처리 분야에서는 Google에서 공개한 BERT (pre-training of Deep Bidirectional Transformers for Language Understanding)[2]와 같은 사전학습 언어모델(pre-training language representation model)의 등장을 시작으로 비약적인 발전을 이루었다. 이후에도 XLNet[3], GPT[4], ALBERT[5], RoBERTa[6] 등 사전학습 언어모델이 꾸준히 등장하면서 기계독해, 문서 분류, 기계번역(machine translation), 텍스트 유사도(text similarity) 등 다양한 응용 태스크에서 우수한 성능을 보이며 산업 전반에 적용하려는 시도가 활발히 이루어지고 있다[7]. 그러나 종래의 공개된 모델들은 신문 기사와 위키피디아(wikipidia)같은 다양한 분야에서 공통적으로 나타날 수 있는 일반적인 형태의 데이터를 사용하여 학습하였기 때문에 일반상식을 다루는 태스크에서는 좋은 성능을 보인다. 하지만 일반적인 형태의 텍스트는 특허문헌 텍스트의 단어 분포와 구문 구조가 상당히 다르기 때문에 특허문헌을 활용한 자연어처리 태스크에서 좋은 성능을 보여주기 어렵다. 그 결과, 특허문헌 텍스트를 사전 학습한 언어모델이 요구되었다.

본 연구에서는 Google이 새로운 사전학습 기법을 적용하여 기존 여러 모델들의 성능을 능가한 ELECTRA (Pre-training Text Encoders As Discriminators Rather Than Generators)[8]로 대규모 한국어 특허문헌 데이터만을 사용하여 사전학습 하였다. 뿐만 아니라, 최적의 사전학습을 위해 특허 사용자 사전(vocabulary) 및 전용 토크나이저(tokenizer)를 적용하고 비교 실험하였다. 한국 특허문헌 자연어처리 연구를 위한 사전학습 모델을 KorPatELECTRA(Korean Patent ELECTRA)로 명명하여 제안하며, 사전학습 방법 및 실험을 통한 성능을 소개하고자 한다.

2장에서는 고성능 언어모델 사전학습 관련 연구와 한국어 특허 데이터에 대해 살펴보고, 3장에서는 사전학습에 활용한 특허문헌 데이터 및 사전 구축 내용과 기술들을 소개한다. 이를 토대로 4장에서는 ELECTRA를 사전 학습한 후 전이학습을 통해 성능 검증을 위한 실험 과정을 분석하고, 5장에서는 결론 및 향후 방향을 제시하고자 한다.

II. Related Work

2.1 Pre-trained Language Models

BERT, RoBERTa와 같은 기존 모델의 사전 학습 방법은 입력 데이터의 일부 토큰을 [MASK] 토큰으로 치환하고 원본 토큰이 무엇인지 맞추는 MLM(Masked

Language Modeling) 방식을 사용하면서 높은 성능 향상을 이루었다. 그러나 CLARK, Kevin, et al(2020)은 BERT의 MLM은 [MASK]로 변환된 15%에서만 loss값이 계산되기 때문에 학습 효율성이 떨어진다고 주장하였고, 이를 개선한 RTD 방식을 제안하며 ELECTRA를 공개하였다. RTD는 generator를 이용해 입력의 일부 토큰을 가짜 토큰으로 바꾸고 모든 토큰에 대해 진짜(original)인지 가짜(replaced)인지 discriminator가 예측하는 방식이다. 모든 토큰에 대해 학습하기 때문에 MLM 방식에 비해 효율성이 향상되었다. 또한, ELECTRA는 BERT와 동일한 데이터와 컴퓨팅 자원 등의 조건에서 GLUE(General Language Understanding Evaluation) 벤치마크[9]의 모든 태스크에서 더 높은 성능을 보이며 BERT의 성능을 능가하는 결과를 보였다. ELECTRA-small의 경우 하나의 GPU(nvidia v100)로 단 4일이면 학습이 가능하며 BERT-large 모델과 대비하여 파라미터는 1/20이며 사전학습하는데 1/135의 계산을 하였음에도 불구하고 BERT-small보다 GLUE 벤치마크의 성능이 좋았다. ELECTRA-large는 RoBERTa와 XLNet 계산량의 1/4만을 사용하여 비슷한 성능에 도달하였다. 또한, SQuAD 태스크에서 10배의 계산량이 많은 ALBERT보다 더 높은 성능을 보이면서 SOTA를 달성하였다[8].

한국어 데이터로 학습한 사례로는 KoELECTRA[1]와 Dialog-KoELECTRA[2]가 있으며, KoELECTRA는 크롤링한 뉴스 데이터로 국립국어원[3]에서 제공하는 문어체 형태의 한국어 데이터 등을 사용하여 약 34GB 데이터를 35,000개의 vocab으로 학습하였다. Dialog-KoELECTRA는 대화체와 문어체 데이터의 비율 조합을 통해 대화체의 성능을 향상시킨 언어모델이다. 국립국어원의 대화 말뭉치와 챗봇 데이터 등 약 22GB 데이터를 40,000개의 vocab으로 학습하였다. Dialog-KoELECTRA는 대화체 기반의 데이터를 활용한 fine-tuning 실험에서 KoELECTRA보다 나은 성능을 보였다.

본 연구에서는 일반적인 한국어 텍스트를 학습한 KoELECTRA와 비교실험을 진행하였다.

2.2 Domain specific language models

그간 특정 전문 분야의 텍스트를 활용하여 언어모델을 사전학습 하는 연구는 BERT 공개 이후 활발히 진행되어왔다. BioBERT[10]는 Biomedical 도메인의 코퍼스인 PubMed와 PMC(PubMed Central)를 통해 사전 학습하였고 SciBERT[11]는 과학 도메인의 논문 코퍼스를 사전학습하여 다운스트림 태스크에서 새로운 SOTA를 달성하였다. BioRoBERTa[12]는 biomedical과 clinical 도메인에 특화된 RoBERTa를 학습한 모델이다. biomedical 분야에서 본 연구에서 활용한 ELECTRA를 사전 학습한 연구도 진행되었다. BioELECTRA[13]는 ELECTRA를 biomedical 도메인의 PubMed와 PMC 데이터를 사용하여 사전 학습한 모델이다.

특허분야에서의 태스크 성능 향상을 위해 특허데이터를 사용하여 언어모델을 사전 학습한 연구도 진행되어왔다. Google에서는 미국과 다른 나라들의 특허 문헌 1억 개 이상에 대해 BERT를 사전 학습한 모델[4]을 공개하였다. 특허는 아이디어를 새로운 방식으로 기술하기 때문에 특히 특허 검색에 있어서는 용어를 식별하는 것이 어렵다는 것을 예로 들며 transformer[14] 기반의 BERT를 사전 학습하는 것에 대한 필요성을 강조하였다. Min, Jae-Ok, et al(2020)는 BERT-base를 기반으로 한국어 특허 데이터와 한국어 위키백과 데이터를 추가 학습하여 특허상담 질의에 대한 정답을 결정하는 기계독해 연구를 진행하였다[15].

언어모델을 사전학습 하진 않았지만 공개된 모델을 활용하여 특허 데이터를 활용한 태스크에서 fine-tuning 실험을 통한 성능을 연구한 사례도 있다. Park, Joo-Yeon, et al(2020)는 BERT를 사용하여 fine-tuning 실험을 통해 미국 특허 데이터에서 청구항의 독립항과 종속항을 구분하는 연구를 진행하였다[16]. LEE, Jieh-Sheng; HSIANG, Jieh(2020)는 BERT-base 모델과 USTPO 특허 데이터에서 청구항만을 사용하여 CPC(Cooperative Patent Classification)코드 분류 태스크를 수행하여 특허가 인공지능 분야에서의 혁신 과제를 해결하기 위한 유용한 자원이라고 주장하였다[17].

과학, 의학 등 특정 전문 분야에 특화된 BERT에 대한 연구 사례는 많은 반면 ELECTRA에 대한 연구 사례는 많지 않았고 특히 특허 데이터를 활용한 사전학습 모델은 찾아보기 어렵다. 따라서 본 연구에서는 BERT보다 우수한 ELECTRA 모델을 활용하고 단일 필드가 아닌 제목, 배경기술, 요약 등 다양한 필드로 구성된 대용량의 한국어 특허 데이터를 사전 학습하고자 한다.

1) https://github.com/monologg/KoELECTRA

2) https://github.com/SKplanet/Dialog-KoELECTRA

3) https://corpus.korean.go.kr/

4) https://github.com/google/patents-public-data

III. Methods

3.1 Models

ELECTRA는 RTD방식을 사용하여 BERT보다 효과적으로 학습할 수 있고 다른 모델들과 비교하였을 때 동일한 컴퓨팅 자원과 데이터 환경에서 적은 비용과 시간이 소요되는데에도 불구하고 성능이 우수하다. 이러한 이유로 효율적인 학습 및 서비스 측면에서 판단하여 사전학습 언어모델로 ELECTRA를 선정하였다. 대용량 한국어 특허문헌 텍스트를 학습하여 특허분야에 최적화된 KorPatELECTRA를 구축하고 NER, MRC 그리고 분류 태스크에서 KLUE(Korean Language Understanding Evaluation)[5] 벤치마크, KorQuAD와 같은 일반적인 데이터 셋 뿐만 아니라 특허 데이터 셋을 활용하여 성능 검증을 위한 실험을 진행하였다.

KorPatELECTRA를 제안함에 있어 모델, 데이터, vocab 크기 및 토큰화 방식에 대한 3가지 관점으로 성능 검증 실험을 하고 분석하였다.

3.2 Corpus

특허는 IPC(International Patent Classification)코드와 CPC코드 분류 체계[6]를 사용하는데 CPC코드는 IPC코드보다 세분화된 특허분류체계로 문헌에 다수의 CPC코드가 부여됨으로써 산업 기술 분야를 구분할 수 있다. 효율적인 선행기술조사를 위해 미국 특허청과 유럽 특허청의 주도로 2012년 개발되어, 2021년 현재 IP5를 포함한 주요국들이 특허문헌을 CPC 코드로 분류하고 있는 만큼 중요한 지표이다. 국내 출원된 특허문헌 중 2,879,585건을 수집하였고 CPC코드 섹션별 수집 결과는 Table 1과 같다.

Table 1. Result on sentence classification

Section	Field	Documents
A	Human Necessities	460,000
B	Performing Operations, Transporting	460,000
C	Chemistry, Metallurgy	416,332
D	Textiles, Paper	69,896
E	Fixed Constructions	185,017
F	Mechanical Engineering, Lighting, Heating, Weapons, Blasting	367,676
G	Physics	460,000
H	Electricity	460,000
Y	General Tagging of new technological developments	664
All Documents		**2,879,585**

수집한 특허문헌은 구조적으로 분리가 가능한 XML 형태로 되어 있기 때문에 문장 텍스트로 사용할 수 있는 16개의 필드 영역(XML element)을 선별하고 텍스트를 추출하였다. 선별한 필드 영역은 Table 2와 같다.

Table 2. Text fields in the patent

No	Field
1	Title of invention
2	Technical field
3	Background art
4	Technical problem
5	Solution to problem
6	Advantageous effects
7	Brief description of drawings
8	Detailed description of the invention
9	Embodiment
10	Reference sign
11	Citation
12	Patent citation
13	NPL(Non-Patent Literature) citation
14	Claim
15	Keyword
16	Abstract

사전 학습을 진행하기 위해 자연어로 기술된 텍스트 파일을 문장 간의 순서를 유지하여 1줄에는 1문장씩 위치하고, 단락이 변경되는 문단 간은 1줄의 빈 줄이 삽입된 형태로 코퍼스(Corpus)를 구축하였다.

추출한 데이터 중 문단 간의 구분은 각 필드로 구분하였고, 한 문장을 구분 지을 수 없거나 너무 길어지는 경우가 발생하여 문장을 분리하였다. 한나눔[7], Open Korean Text[8], 코모란[9] 등 다양한 형태소 분석기에서 문장 분리기(sentence splitter)를 지원하지만 형태소 분석기마다 특징과 성능이 다르다. 본 연구에서는 KSS(Korean Sentence Splitter)[10] 문장 분리기를 사용하여 문장을 분리하였다. 특허문헌의 문장은 사전 학습에 불필요한 특수문자나 수식이 많은 특징에 따라 문장 분리 이후 예외 규칙을 더 추가하여 코퍼스의 학습 품질을 높였다. 공백을 포함한 글자 수가 10 이하인 문장은 제거하고 수식, 한자 그리고 특수문자들은 학습의 성능을 떨어뜨리는 노이즈라고 판단하여 내용을 변질시키지 않는 수준에서 제거하였다. 최종적으로 약 82GB의 한국어 특허 코퍼스를 구축하였다.

5) https://github.com/KLUE-benchmark/KLUE

6) https://www.cooperativepatentclassification.org/index

7) https://sourceforge.net/projects/hannanum/

8) https://github.com/open-korean-text/open-korean-text

3.3 Vocabulary and Tokenizer

모델이 입력 텍스트에 대해 다차원 공간의 실수 벡터로 표현하는 임베딩 과정을 효과적으로 진행하기 위해서는 입력 텍스트가 의미를 가지는 용어 단위의 토큰(token)으로 인식되도록 해야 한다. 그러기 위해선 vocab이 필요하고 토큰 단위로 분리할 수 있는 토크나이저의 역할이 매우 중요하다고 알려져 있다. RUST, Phillip, et al(2020)는 monolingual BERT와 multilingual BERT를 대상으로 단일 언어로 최적화된 토크나이저로 학습하였다[18]. 또한, 해당 언어에 대한 다운스트림 태스크(downstream task)를 수행하여 단일 언어 토크나이저가 multilingual 모델의 monolingual 성능 향상에 도움이 된다는 점을 시사함으로써 모델 성능 향상에 언어모델의 종류와 데이터의 크기뿐만 아니라 토크나이저의 역할이 중요하다는 것을 실험적으로 증명하였다.

영어의 경우에는 띄어쓰기를 기준으로 단어를 기본 토큰 단위로 사용하지만 한국어는 교착어이기 때문에 단순히 BPE(Byte Pair Encoding)[19]만을 사용하여 토큰화할 경우 동일한 의미를 가진 단어에 조사가 붙는 형태로 토큰화가 되어 불필요한 변화가 많이 발생하기 때문에 학습이 잘 되지 않는다. PARK, Sungjoon, et al(2021)의 KLUE에서는 형태소 기반의 서브워드(subword) 토큰화 기법을 제안하여 단순히 BPE만을 사용하지 않고 raw text를 mecab-ko[11] 형태소분석기를 사용하여 미리 토큰화한 후 vocab을 생성할 때 토큰을 더 잘 분리하는 것을 증명하였다[20]. 따라서 본 연구에서는 vocab을 생성하기 전 코퍼스를 mecab-ko 형태소 분석기로 문장을 형태소 단위로 토큰화를 하였다.

다음으로 vocab을 만들기 위해서 2가지 BPE방식의 토크나이저를 사용하였다. 첫 번째는 KoELECTRA를 구축할 때 사용한 방식으로 mecab-ko 형태소 분석기로 분절한 코퍼스에서 word piece 토크나이저로 학습한 MWP (Mecab-ko Wordpiece Patent tokenizer)이며, 두 번째는 한국어 특허 코퍼스를 대상으로 좋은 성능을 보이는 토크나이저 방식을 제안한 MSP(Mecab-ko Sentencepiece Patent tokenizer)[21]방식으로 vocab을 구축하였다. MSP 방식은 2단계의 토큰화 과정을 거치는데 먼저, 코퍼스 내 명사와 복합명사를 추출하고 mecab-ko 형태소 분석기의 사용자 사전으로 등록하여 1차 토크나이저로 사용하였다. 토크나이저의 성능을 평가하는 주요 태스크인 기계번역 BLEU(Bilingual Evaluation Understudy) 평가로 토큰 타입별 성능 비교를 하여 토큰 개수를 도출하고 전체 코퍼스에 대한 vocab 크기를 결정하였다. 이때 사용한 토큰화 방식이 사용자 사전이 포함된 2차 토크나이저이다.

IV. Experiments

사전학습 최적화를 위해 vocab과 토크나이저를 결정하는 과정과 사전학습 방법을 설명하고 NER, MRC, 문서분류 태스크를 수행하여 KorPatELECTRA의 성능을 분석해보고자 한다.

4.1 Pre-training KorPatELECTRA

4.1.1 Make Vocabulary

3.3에서 소개한 바와 같이 BERT, ELECTRA에서 사용한 word piece 토크나이저를 기반으로 한국어 특성에 맞춰 개선한 MWP 토크나이저와 특허문헌 데이터를 대상으로 사용자 사전이 적용된 MSP 토크나이저를 사용하였다. Table 3는 각 모델에서 사용한 토크나이저에 따른 vocab의 크기를 나타낸 표이다.

Table 3. The size of the vocabulary according to the tokenizer used in the pretrained models.

Model	Tokenizer	Vocab size
mBERT	WPM	119,547
KoELECTRA	MWP	35,000
KorPatELECTRA	MWP	35,000
	MSP	19,400

vocab의 크기는 그 개수에 따라 문장이 분리되는 토큰의 개수가 달라지기 때문에 토크나이저의 성능에 영향을 미친다. KoELECTRA와 동일한 성능비교를 위해 MWP 토크나이저의 vocab 크기는 35,000개로 하였다.

PARK, Jinwoo, et al(2020)는 특허문서 초록 데이터를 기반으로 기계번역 태스크를 통해 vocab 크기에 따른 MSP 토크나이저 성능을 비교하는 실험을 진행하였고 본 연구에서도 해당 방법을 적용하여 실험하였다[21]. 수집한 한국어 특허 데이터에 대해 한자, 특수문자 제거 등 전처리를 거친 학습 코퍼스에 맞는 vocab의 크기를 계산

9) https://github.com/shineware/KOMORAN

10) https://github.com/hyunwoongko/kss

11) https://bitbucket.org/eunjeon/mecab-ko-dic/src/master/

하였고 19,400개의 토큰으로 이루어진 vocab을 사용하여 사전학습을 진행하였다.

4.1.2 Train data Transformation

사전학습은 Tensorflow API를 사용하여 진행되기 때문에 앞서 생성한 vocab을 활용하여 데이터 포맷을 TFRecord 포맷으로 변경한다. TFRecord는 일반적인 텍스트, 이미지 등의 데이터를 바이너리 코드의 시리즈를 저장하기 위한 단순한 형식이다. 이를 사용하는 것으로 대규모 데이터를 효율적으로 학습할 수 있게 된다.

사전학습을 진행하기에 앞서 mecab-ko 형태소분석기로 분절하지 않은 원본 특허문헌 코퍼스(82GB)를 ELECTRA 입력형식에 맞춰진 TFRecord로 변환하였다. ELECTRA-base의 조건과 동일하게 max seq len을 512로 설정하여 학습 데이터를 변환하였고 그 결과 MWP 토크나이저는 약 82GB의 TFRecord가 생성되었고, MSP 토크나이저는 약 78GB의 TFRecord가 생성되었다.

4.1.3 Experimental Setup and pre-training

사전학습에서 사용한 모든 하이퍼 파라미터(hyper parameter)는 ELECTRA-base 하이퍼 파라미터와 동일한 조건으로 설정하였다. 512개의 입력 토큰 중 15%를 마스킹하였고 768개의 embedding size, 12개의 attention head로 1M steps 까지 학습하였다. KoELECTRA-base가 train batch size를 256으로 설정한 것과 다르게 KorPatELECTRA는 8개의 V100 GPU[12]를 분산학습을 사용하여 총 128 train batch size로 사전 학습을 진행하였다.

4.2 Fine-tuning KorPatELECTRA

자연어처리의 대표적인 3가지의 다운스트림 태스크를 KorPatELECTRA로 파인튜닝을 통해 모델 성능을 평가하였다.

평가 결과를 바탕으로 3가지 관점에서 모델 성능을 비교하였다. 첫 번째는 동일한 토크나이저에서 BERT 및 KoELECTRA와 성능을 비교하였다, 두 번째는 다양한 분야의 일반적인 형태를 가진 비특허 데이터와 실제 특허분야의 데이터를 활용하여 KoELECTRA와 우리 모델간의 성능 변화를 비교하였다. 세 번째는 전용 토크나이저와 vocab이 언어모델의 성능과 각 태스크에서 어떠한 영향을 미치는지에 대해 분석하였다.

4.2.1 Tasks & Dataset

Table 4는 각 NER, MRC, 문서분류 3가지 태스크에서 사용한 특허와 비특허 데이터 종류이다. 특허 데이터는 실제 특허 문헌을 기반으로 구축된 데이터 셋이며 해당 데이터를 활용함으로써 KorPatELECTRA가 특허 분야에서 실질적으로 우수한 성능을 보이는지 판단할 수 있다.

Table 4. Data list of patent and non-patent fields used for downstream tasks.

	NER	MRC	CLS
Non Patent	KLUE, NAVER-CW Challenge	KorQuAD	K-NSCC
Patent	Patent (Chemistry)	PatQuAD	CPC Code

NER 태스크는 미리 정의해 둔 개체명을 텍스트에서 인식하여 추출하고 분류하는 태스크이다. 본 연구에서는 3종류의 데이터 셋를 사용하여 fine-tuning 실험을 진행하였고 KLUE, 어절기반 NAVER-창원대 NLP Challenge에서 제공된 데이터(이하, NAVER-CW Challenge), 형태소 기반 특허 데이터를 사용하였다.

KLUE 데이터는 한국어 NLP 분야에서 사전 학습된 모델의 성능을 정량적으로 평가할 수 있는 벤치마크 데이터 셋이며 사람, 위치, 기관 등 6개의 개체명으로 구성되어 있다. NAVER-CW Challenge 데이터는 NAVER가 창원대학교와 함께 개최한 한국어 자연어처리 기술 대회[13]에서 창원대학교가 마련한 대량의 한국어 데이터로 인물명, 학문 분야, 지역명칭 등 14개의 개체명으로 구성되어 있다. Patent 데이터는 화학 특허 분야의 데이터 셋으로 고분자 복합수지 분야의 화학 전문가를 통해 수작업으로 화학용어 사전과 BIO(begin-inside-outside) 태깅으로 구축하였다. 학습 데이터는 조성, 물성, 조성 단위, 물성 단위, 조성 값, 물성 값으로 총 6개의 개체명으로 구성되어 있다.

NER 태스크의 성능평가 지표는 F1 점수를 사용하였다.

MRC 태스크는 정답이 존재하는 문서에서 사용자 질의에 대해서 기계가 문서를 이해하여 정답의 위치를 스스로 찾아내는 태스크이다. 본 연구에서는 MRC 실험을 위해 KorQuAD와 PatQuAD 2종류의 데이터 셋을 사용하였다.

KorQuAD는 위키백과 문서를 대상으로 질의와 정답을 생성한 데이터 셋으로 학습 데이터 60,407건, 평가 데이터 7,774건으로 구성되어 있다. PatQuAD는 Min, Jae-Ok, et al(2020)가 구축한 특허상담 질의응답 데이터

12) 국가슈퍼컴퓨팅센터(KISTI)로부터 슈퍼컴퓨팅 자원을 지원받아 수행하였다.

13) https://github.com/naver/nlp-challenge

셋으로 작업자가 특허법, 상담 사례 등 수집된 문서에서 질의-정답 셋을 수작업으로 구축한 데이터이다[15]. 총 6,011건으로 실험을 위해 학습 데이터 4,808건, 평가 데이터 1,203건으로 실험을 진행하였다.

MRC 태스크의 성능평가 지표는 EM(Exact Match)와 F1 점수를 사용하였다.

Document Classification 태스크는 입력된 문서가 어떤 범주에 속하는지 분류하는 태스크이다. 본 연구에서는 2종류의 데이터 셋을 이용하는데, 국가과학기술지식정보서비스(NTIS) 시스템에 제출되어 국가과학기술표준분류체계[14](K-NSCC)를 따르고 있는 국가 R&D보고서 데이터와 특허문헌 분류체계인 CPC 코드가 부여된 특허 데이터이다.

국가 R&D보고서는 주제 범위가 33개의 대분류, 371개의 중분류, 2,898개의 소분류로 이루어져 있고 해당 데이터에 대한 자동 분류 태스크는 연구자가 복잡한 분류체계를 모두 이해하지 않고 연구 문헌의 핵심 주제를 쉽게 파악할 수 있다는 점에서 의미가 크다고 할 수 있다. 2013년부터 2018년까지의 데이터 중 학습 데이터 130,515건, 평가 데이터 14,502건을 샘플링 하였고, 중분류 86개의 분류 체계로 평가를 진행하였다. 특허문헌 데이터는 390,540건을 샘플링 하였고, CPC코드의 subclass 기준으로 144개의 분류 체계로 평가를 진행하였다.

문서분류 태스크의 성능평가 지표는 정확도를 사용하였다.

4.2.2 Non-Patent Domain

Table 5는 다양한 분야의 일반적인 형태에 대한 데이터 셋을 활용하여 fine-tuning 실험을 진행한 결과이다. 모든 태스크에서 BERT 모델보다는 ELECTRA를 사전 학습한 모델들이 더 나은 성능을 보였다.

Table 5. Performance of tested model on the Non-Patent downstream tasks. Overall best scores are underlined and highlighted in bold. Document Classification Task is indicated by CLS.

Model	CLS (ACC)	NER (F1)		MRC (EM/F1)
	K-NSCC	KLUE	NAVER-CW Challenge	KorQuAD
mBERT	62.68	83.4	83.23	70.12/90.19
KoELECTRA	66.82	**83.5**	**86.43**	**84.74/93.48**
KorPatELECTRA	**68.68**	80.67	84.63	81.88/90.75

NER과 MRC 태스크에서는 KoELECTRA가 모든 데이터 셋에서 KorPatELECTRA의 성능을 능가하였다. 반면에, 국가 R&D보고서 분류 태스크에서 KorPatELECTRA가 KoELECTRA보다 정확도가 1.68% 상승하였다. 이는 국가 R&D보고서와 특허문헌은 다양한 과학기술 지식을 기반으로 한 문서로서 두 데이터가 기술적 내용을 포함하고 있기 때문에 해당 태스크에서 좋은 성능을 보였다고 판단된다. NER 태스크에서 KLUE 데이터 셋을 사용한 실험에서는 KoELECTRA가 KorPatELECTRA보다 F1이 2.83% 상승하였고 NAVER-CW Challenge 데이터 또한 1.8% 상승하였다. KorQuAD 데이터 셋을 사용한 MRC 태스크에서도 KoELECTRA가 EM과 F1이 각각 2.86%, 2.73% 높은 점수를 보이고 있다. 결과적으로, KLUE, NAVER-CW Challenge, KorQuAD 데이터 셋을 활용한 태스크에서는 법적 요건에 맞춘 형식과 복잡한 기술적인 내용으로 구성되어 있는 특허와는 다른 일반적인 형태의 데이터를 사전 학습한 KoELECTRA가 좋은 성능을 보였다.

4.2.3 Korean Patent Domain

Table 6는 특허 데이터 셋을 활용하여 4.2.2에서의 태스크와 동일한 자원과 파라미터로 fine-tuning 실험을 진행한 결과이다.

Table 6. Performance of tested model on the Korean patent downstream tasks. Overall best scores are underlined and highlighted in bold. Document Classification Task is indicated by CLS.

Model	CLS (ACC)	NER (F1)	MRC (EM/F1)
	CPC Code	Patent (Chemistry)	PatQuAD
mBERT	72.33	87.98	51.48/81.79
KoELECTRA	72.98	87.48	72.46/88.09
KorPatELECTRA	**73.45**	**91.2**	**75.79/88.64**

KorPatELECTRA는 모든 태스크에서 BERT를 능가하였고 4.2.2에서의 실험 결과와 반대로 모든 태스크에서 KoELECTRA보다 좋은 성능을 보였다. CPC코드 분류 태스크에서 KoELECTRA보다 정확도가 0.47%, NER 태스크에서는 3.72% 향상된 성능을 보였다. 또한, MRC 태스크에서 EM은 3.13%, F1은 0.55% 상승하였다. 이렇게 특허 데이터만을 학습한 KorPatELECTRA가 특허데이터 셋을 활용한 모든 fine-tuning 태스크에서 KoELECTRA와 BERT의 성능을 능가하며 특정 전문분야의 데이터를

14) https://www.msit.go.kr/SYNAP/skin/doc.html?fn=c41890287bdde8054310eb89c78ed285&rs=/SYNAP/sn3hcv/result/

사전 학습한 언어모델이 그 분야의 태스크에 미치는 긍정적인 결과를 확인할 수 있었다.

4.2.4 Tokenizer Comparison

Table 7는 특허 데이터를 사용한 다운스트림 태스크에서 MWP와 MSP 토크나이저에 따른 KorPatELECTRA 성능 결과를 나타낸 표이다.

Table 7. Performance of KorPatELECTRA model according to tokenizer on the various Korean patent downstream tasks. Overall best scores are underlined and highlighted in bold. Document Classification Task is indicated by CLS.

Tokenizer	Vocab size	CLS (ACC)	NER (F1)	MRC (EM/F1)
		CPC Code	Patent (Chemistry)	PatQuAD
MWP	35,000	73.45	**91.2**	**75.79/88.64**
MSP	19,400	**73.64**	90.91	62.71/85.12

MSP vocab의 크기는 4.1.1에서 언급한 바와 같이 특허 코퍼스에 맞게 계산한 19,400개로 실험하였다. MSP 토크나이저를 사용한 모델이 CPC코드 분류 태스크에서만 0.19% 더 높은 점수를 보였다. NER과 MRC 태스크에서는 MWP 토크나이저를 사용한 모델이 더 나은 성능을 보였는데 NER 태스크에서는 MWP 토크나이저를 사용한 모델에서 0.29% 상승하였다. 또한, MRC 태스크에서는 F1은 3.42% 차이로 NER 태스크에 비해 더 큰 성능 차이를 보였으며, EM은 13.08% 차이로 다른 태스크와 비교하여 토크나이저에 따른 성능 차이가 더 큰 결과를 보였다.

V. Conclusions and Future Work

본 연구에서는 특허분야에 특화된 사전학습 모델인 KorPatELECTRA를 학습하는 방법을 제안하고 성능을 검증하였다. 또한, 학습을 위해 대용량 한국어 특허 데이터 특성에 맞는 전처리 과정을 거치고 토크나이저에 따른 성능변화도 확인하였다.

4장에서의 실험결과들을 분석해 보았을 때, 대용량 특허데이터를 학습한 KorPatELECTRA가 결과적으로 특허 데이터를 활용한 fine-tuning 실험에서 타 분야의 데이터 셋을 학습한 사전학습 언어모델보다 성능 향상이 있음을 확인하였다. 그 결과 각 분야별 데이터 특성에 대해 분석하고 그에 특화된 전처리와 토크나이저를 활용하여 vocab을 생성하는 과정이 언어모델의 성능에 영향을 미치고 특화된 언어모델의 필요성을 확인하였다.

향후 연구 방향은 다음과 같다. 첫째, 토크나이저와 vocab의 크기에 따른 언어모델의 성능이 NER, MRC 등 서로 다른 태스크에 어떤 영향을 끼치는가에 대해 상세히 분석하는 연구가 필요하다. 둘째, 대용량의 한국어 특허 데이터를 가공하고 KorPatELECTRA를 구축한 경험을 바탕으로 향후 더 좋은 성능의 사전학습 언어모델 생성 연구를 지속하고자 한다. 셋째, 기존의 특허 데이터뿐만 아니라 논문, 연구보고서 등 과학기술 분야에 대한 데이터를 포함한 과학기술 분야에 특화된 고성능의 언어모델이 개발되어 과학기술 분야 전반에 활용되기를 기대해 본다.

ACKNOWLEDGEMENT

This research was results of a 「the National Supercomputing Center with supercomputing resources including technical support (KSC-2021- CRE-0200)」 and 「study on the "HPC Support" Project, supported by the 'Ministry of Science and ICT' and NIPA」, and 「With the funding of the Ministry of Trade, Industry and Energy in 2021, the Korea Institute of Industrial Technology Evaluation and Management's big data construction of thermal and electrical plastic complex resins and development of a platform for composition/physical properties over 90% using artificial intelligence technology (No. 200391).」

REFERENCES

[1] Jeong, Su-Jeong, "Zur Analyse von mehr oder weniger festen Wortverbindungen in Patentschriften im Deutschen und Koreanischen," German Literature, Vol. 26, No. 3, pp. 360-361. 2016.

[2] DEVLIN, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[3] Yang, Zhilin, et al. "Xlnet: Generalized autoregressive pretraining for language understanding," Advances in neural information processing systems, 32 , 2019.

[4] RADFORD, Alec, et al, "Improving language understanding by generative pre-training", 2018.

[5] LAN, Zhenzhong, et al. "A Lite BERT for Self-supervised

Learning of Language Representations," arXiv preprint arXiv:1909.11942, 2019.

[6] LIU, Yinhan, et al. "Roberta: A robustly optimized bert pretraining approach," arXiv preprint arXiv:1907.11692, 2019.

[7] LIM, J. H.; KIM, H. K.; KIM, Y. K. "Recent R&D Trends for Pretrained Language Model," Electronics and Telecommunications Trends, Vol. 35, No. 3, pp. 9-19, 2020.

[8] CLARK, Kevin, et al. "Electra: Pre-training text encoders as discriminators rather than generators," arXiv preprint arXiv:2003.10555, 2020.

[9] WANG, Alex, et al. "GLUE: A multi-task benchmark and analysis platform for natural language understanding," arXiv preprint arXiv:1804.07461, 2018.

[10] LEE, Jinhyuk, et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," Bioinformatics, Vol. 36, No. 4, pp.1234-1240, 2020.

[11] BELTAGY, Iz; LO, Kyle; COHAN, Arman. "Scibert: A pretrained language model for scientific text," arXiv preprint arXiv:1903.10676, 2019.

[12] LEWIS, Patrick, et al. "Pretrained Language Models for Biomedical and Clinical Tasks: Understanding and Extending the State-of-the-Art," Proceedings of the 3rd Clinical Natural Language Processing Workshop, pp. 146-157, 2020.

[13] RAJ KANAKARAJAN, Kamal; KUNDUMANI, Bhuvana; SANKARASUBBU, Malaikannan. "BioELECTRA: Pretrained Biomedical text Encoder using Discriminators," Proceedings of the 20th Workshop on Biomedical Language Processing, pp. 143-154, 2021.

[14] VASWANI, Ashish, et al. "Attention is all you need. In: Advances in neural information processing systems," pp. 5998-6008, 2017.

[15] Min, Jae-Ok, et al, "Korean Machine Reading Comprehension for Patent Consultation Using BERT,/ Software and Data Engineering, Vol. 9, No. 4, pp. 145-152, 2020.

[16] Park, Joo-Yeon, et al. "Improving Recognition of Patent's Claims with Deep Neural Networks," Collection of papers from Korea Information Processing Society, Vol. 27, No. 1, pp. 500-503, 2020.

[17] LEE, Jieh-Sheng; HSIANG, Jieh. "Patent classification by fine-tuning BERT language model," World Patent Information, 61: 101965, 2020.

[18] RUST, Phillip, et al. "How good is your tokenizer? on the monolingual performance of multilingual language models," arXiv preprint arXiv:2012.15613, 2020.

[19] SENNRICH, Rico; HADDOW, Barry; BIRCH, Alexandra. "Neural machine translation of rare words with subword units," arXiv preprint arXiv:1508.07909, 2015.

[20] PARK, Sungjoon, et al. "KLUE: Korean Language Understanding Evaluation," arXiv preprint arXiv:2105.09680, 2021.

[21] PARK, Jinwoo, et al. "Patent Tokenizer: a research on the optimization of tokenize for the Patent sentence using the Morphemes and SentencePiece," Annual Conference on Human and Language Technology. Human and Language Technology, pp. 441-445, 2020.

Authors

Ji-Mo Jang received the B.S. degree in Information and Communication Engineering from Yeungnam University, Korea in 2019. She is currently a staff of Intelligent Information Strategy Team in Korea Institute of Patent Information. She is interested in patent analysis, natural language processing, deep learning and machine learning.

Jae-Ok Min received the B.S. degree in Computer Engineering from Myongji University, Korea, in 2010 and received the MBA degree in AI BigData from Sogang University, Korea, in 2021. He is currently a R&D part leader of Intelligent Information Strategy Team in Korea Institute of Patent Information. He is interested in patent analysis, natural language processing, text mining and machine learning.

Han-Sung Noh received the B.S. degree in Information and Communications Engineering from Chungnam National University and M.A. degree in Future Strategy from KAIST, Korea, in 2002 and 2019 respectively. Han-Sung Noh is currently a team leader of Intelligent Information Strategy Team in Korea Institute of Patent Information. He is interested in natural language processing, network analysis and technology forecasting.