https://doi.org/10.15813/kmr.2021.22.3.006

# 추가 사전학습 기반 지식 전이를 통한 국가 R&D 전문 언어모델 구축

# Building Specialized Language Model for National R&D through Knowledge Transfer Based on Further Pre-training

유은지 (Eunji Yu) 국민대학교 비즈니스IT 전문대학원[1)]
서수민 (Sumin Seo) 국민대학교 비즈니스IT 전문대학원[2)]
김남규 (Namgyu Kim) 국민대학교 비즈니스IT 전문대학원[3)]

## 〈 국문초록 〉

최근 딥러닝 기술이 빠르게 발전함에 따라 국가 R&D 분야의 방대한 텍스트 문서를 다양한 관점에서 분석하기 위한 수요가 급증하고 있다. 특히 대용량의 말뭉치에 대해 사전학습을 수행한 BERT(Bidirectional Encoder Representations from Transformers) 언어모델의 활용에 대한 관심이 높아지고 있다. 하지만 국가 R&D와 같이 고도로 전문화된 분야에서 높은 빈도로 사용되는 전문어는 기본 BERT에서 충분히 학습이 이루어지지 않은 경우가 많으며, 이는 BERT를 통한 전문 분야 문서 이해의 한계로 지적되고 있다. 따라서 본 연구에서는 최근 활발하게 연구되고 있는 추가 사전학습을 활용하여, 기본 BERT에 국가 R&D 분야 지식을 전이한 R&D KoBERT 언어모델을 구축하는 방안을 제시한다. 또한 제안 모델의 성능 평가를 위해 보건의료, 정보통신 분야의 과제 약 116,000건을 대상으로 분류 분석을 수행한 결과, 제안 모델이 순수한 KoBERT 모델에 비해 정확도 측면에서 더 높은 성능을 나타내는 것을 확인하였다.

**주제어:** 국가 R&D, 지식 전이, 사전학습 모델, BERT, 추가 사전학습

1) 제1저자, eunjiu@kisti.re.kr
2) 제2저자, syg3793@kookmin.ac.kr
3) 교신저자, ngkim@kookmin.ac.kr

논문접수일 2021년 08월 22일 | 1차 수정 2021년 09월 04일 | 게재확정일 2021년 09월 05일

## 1. 서론

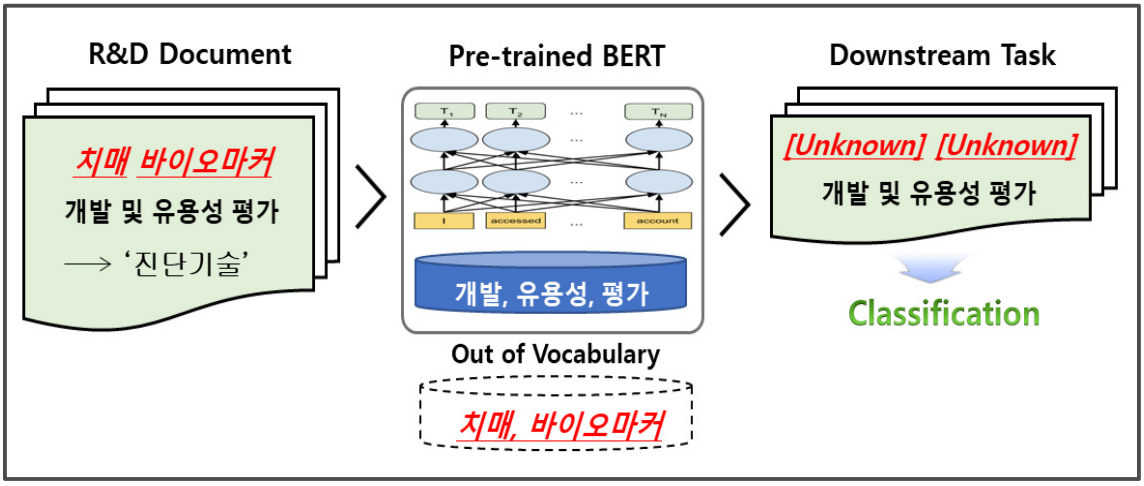
에너지, 감염병과 같은 이슈가 심각한 사회문제로 떠오르면서 과학기술 연구개발(R&D)을 통해 이를 극복하고자 하는 노력이 이어지고 있다. 이러한 노력은 2021년 국가 R&D 총예산이 27.4조로 책정되어 전년도 대비 약 13%가 증가함과 동시에, 생명, ICT 그리고 SW 등 성장 잠재력이 큰 중점 과학기술 분야에 대한 투자가 집중된 현상에서도 확인할 수 있다. 이렇듯 R&D 투자 규모가 증가함에 따라, 국가 R&D 데이터를 분석하여 현황을 파악하고, 투자 방향성을 설정하기 위한 데이터 분석, 구체적으로는 새로운 연구 분야를 발굴하고, 연구 동향을 예측하기 위한 다양한 관점의 분석 수요가 증가하고 있다. 궁극적으로 이를 통해 새로운 지식(Knowledge)을 창출하고 나아가 국가적 과학기술 경쟁력을 확보하고자 하는 기대가 커지고 있다.

국가 R&D 분야의 주 분석 대상인 과제(Project), 논문(Paper) 그리고 특허(Patent) 데이터는 대부분 텍스트(Text)로 구성되어 있어, 이를 분석하기 위해서는 이들 문서를 컴퓨터가 처리할 수 있는 형태로 구조화하는 자연어 처리(Natural Language Processing) 작업이 선행되어야 한다. 최근 자연어 처리 기술은 다양한 분야에서 활용되고 있는 딥러닝(Deep Learning) 기술을 기반으로 상당한 개선을 이루었다. 과거 단순히 단어의 의미를 파악하던 수준을 뛰어넘어 문장과 문맥의 의미를 파악하기 위한 다양한 방법론이 등장하였으며, 특히 방대한 텍스트 데이터를 미리 학습한 사전학습 모델(Pre-trained Language Model)의 활용과 개선에 대한 연구가 활발하게 이루어지고 있다.

대표적인 사전학습 언어모델인 BERT(Bidirectional Encoder Representations from Transformers)(Devlin et al., 2018)는 MLM(Masked Language Model) 기법을 통해 마스킹 대상 단어의 전과 후 문맥 정보를 모두 활용하여 해당 단어를 추론하는 양방향 학습을 진행하여, 단방향 학습을 기반으로 하는 기존 모델의 한계를 대폭 개선하였다. 이후 많은 연구들이 BERT 언어모델 기반 전이 학습(Transfer Learning), 즉 BERT를 주어진 분석 과제(Downstream Task)에 적용하여 모델의 가중치를 갱신하는 미세조정(Fine-tuning)을 수행함으로써 여러 분야에서 괄목할 성과를 거두었다. 전이 학습은 학습 데이터가 부족하거나 컴퓨터 자원이 부족할 때, 또한 분석 과제의 학습 모델이 사전학습 모델과 유사한 분포를 갖는 경우 처음부터 새로운 학습을 수행하는 것에 비해 매우 효과적인 것으로 알려져 있다.

하지만 고도로 전문화된 분야의 문서 분석에 BERT 모델을 그대로 사용하는 경우, BERT의 방대한 학습량에도 불구하고 해당 전문 분야에서 높은 빈도로 사용되는 전문어가 충분히 학습되지 못한 경우가 존재할 수 있다. 이러한 경우 해당 분야에서 매우 중요하게 사용되는 전문어임에도 문맥적 의미가 제대로 학습되지 않아, 결과적으로 분석 품질이 저하되는 경우가 발생할 수 있다. 특히, 일부 용어의 경우 BERT가 학습한 일반적 의미와 전문 분야에서 사용하는 전문적 의미가 서로 다를 수 있기 때문에, 해당 전문 분야의 문맥 정보 학습을 통해 BERT 언어모델의 가중치를 갱신하는 작업이 반드시 필요하다. 이에 따라 최근 연구에서는 BERT로 일반적인 단어의 의미를 학습하고, 추가 사전학습(Further Pre-training)을 통해 전문 분야의 지식을 추가로 학습하는 연구들이 다양한 분야에서 시도되고 있다(I. Beltagy et al., 2019; J. Lee et al., 2019).

국가 R&D 분야에서도 위와 같은 문제가 동일하게 발생하며, 이는 <그림 1>을 통해 확인할 수 있다. <그림 1>은 “치매 바이오마커 개발 및 유용성 평가”라는 국가 R&D 과제가 어떤 분류에 속하는지 예측하는 분




〈그림 1〉 전문어 학습 부족으로 인한 분석 과제의 성능 저하

석 과정을 보이며, 여기서 "치매", "바이오마커"와 같은 전문어는 BERT에는 포함되어 있지 않고 비교적 일반적인 용어인 "발굴", "유용성", "평가"는 BERT에 포함되어 있다고 가정한다. 이 경우 해당 분야의 핵심어로 분류에 큰 영향을 미칠 수 있는 "치매"와 "바이오마커"는 그 문맥적 의미가 충분히 학습되지 않은 상태로 분류 분석에 사용되게 된다.

이러한 한계를 극복하기 위해, 본 연구에서는 추가 사전학습 기법을 기반으로 국가 R&D 분야에서 사용하는 전문어를 추가로 학습한 언어모델을 구축하고자 한다. 이는 국가 R&D 분야의 전문 지식을 BERT 언어모델에 전이한 국가 R&D 특화 언어모델이라고 할 수 있다. 또한 본 연구에서 제안한 모델의 우수성과 실무 적용 가능성을 확인하기 위해, 국가 R&D 분야 중 보건의료와 정보통신 분야의 과제 데이터를 대상으로 제안 모델을 적용한 실험을 수행하고 그 결과를 공유하고자 한다.

본 논문의 이후 구성은 다음과 같다. 2장에서는 딥러닝 기반 텍스트 임베딩, 사전학습 언어모델 그리고 국가 R&D 정보 분류 체계에 대한 관련 연구를 소개하고, 3장에서는 본 연구에서 제안한 모델의 전체 프로세스에 대해 설명한다. 4장에서는 제안한 모델의 검증을 위해 수행한 성능 평가 실험에 대한 결과를 요약하고, 마지막 장인 5장에서는 본 연구의 기여와 앞으로의 연구 계획을 제시한다.

## 2. 관련 연구

### 2.1. 딥러닝 기반 텍스트 임베딩

딥러닝이란 인간의 두뇌 작동 원리와 유사한 방식으로 구현된 인공 신경망을 기반으로, 입력층(Input Layer)과 출력층(Output Layer) 사이에 여러 은닉층(Hidden Layer)을 쌓아 층마다 존재하는 가중치(Weight)를 학습하는 기법이다. 딥러닝 기술이 발전함에 따라, 자연어 처리 분야에서도 텍스트를 효율적으로 처리하고 텍스트에 내포된 지식을 효과적으로 표현하기 위해 딥러닝 기술을 적용한 연구가 활발히 이루어지고 있다. 자연어 처리 분야의 전통적인 텍스트 임베딩에는 단어 집합 크기만큼의 차원을 갖는 벡터에 특정 단어의 값을 0과 1로 표현하는 이산 표현(Discrete Representation) 방식인 One-hot Encoding 기법이 주로 사용되었다. 그러나 해

당 기법은 단어 집합이 커질수록 벡터 차원의 크기도 커지게 되어 메모리의 낭비가 심하고 계산 효율성이 저하된다는 한계를 갖는다. 반면, 인공 신경망 기반의 텍스트 임베딩 기법은 각 단어의 관계 정보를 신경망으로 학습한 후, 단어를 다차원 공간에 투영하는 분산 표현(Distributed Representation) 방식을 통해 기존의 한계를 극복하였다. 대표적인 단어 임베딩(Word Embedding) 모델로는 Word2Vec(Mikolov et al., 2013), GloVe(Gloval 그리고 Vectors for Word Representation)(Pennington at al., 2014) 등이 있다.

인공 신경망 기반의 단어 임베딩이 우수한 성과를 나타냄에 따라, 확률에 기반하여 문장 벡터를 추출하는 언어모델(Language Model)에도 딥러닝을 적용하려는 이른바 문장 임베딩(Sequence Embedding)에 대한 연구들이 활발히 수행되었다. 대표적인 문장 임베딩 모델인 RNN(Recurrent Neural Network)(Mikolov et al., 2010)은 이전 단어 벡터가 가지고 있는 정보를 다음 단어 벡터 계산의 입력으로 사용하여, 과거 단어의 지식 정보를 충분히 반영한 문장 벡터를 추출한다. 이후 RNN을 개선한 모델인 LSTM(Long Shot-Term Memory) (Hochreiter & Schmidhuber, 1997)과 두 개의 언어모델을 인코더(Encoder)와 디코더(Decoder)로 연결한 Seq2Seq (Sequence-to-Sequence)(Sutskever at al., 2014) 모델이 등장하였다.

하지만 LSTM, Seq2Seq와 같은 모델은 문장의 길이가 길어질수록 과거의 정보들을 잊어버리는 문제와 함께, 모든 정보를 하나의 고정된 벡터 공간에 압축하는 과정에서 정보의 손실이 발생한다는 한계를 갖는다. 이를 해결하기 위해 전체 문장 정보와 각 단어에 대한 정보를 모두 반영하여, 각 단어의 예측에 기여한 단어들에 집중하는 어텐션 메커니즘(Attention Mecanism) (Bahdanau at al., 2014)이 고안되었다. 어텐션 메커니즘은 어텐션 함수(Attention Function)를 통해 입력 문장 내의 단어들이 출력 문장 내의 단어들 중 관련성이 높은 단어에 집중할 수 있도록 학습하여 어텐션 가중치 (Attention Weight)를 구하기 때문에, 각 단어들 간의 관계를 시각적으로 확인할 수 있어 학습의 결과를 해석 가능하게 하였다. 이후, 오직 어텐션 메커니즘만을 활용하여 인코더-디코더 구조를 구축한 트랜스포머(Transformer) (Vaswani et al., 2017) 모델이 등장함에 따라 자연어 처리 분야의 획기적인 발전이 가능하게 되었다.

## 2.2. 사전학습 언어모델

사전학습 언어모델은 대규모의 텍스트 데이터에서 일반적인 의미의 텍스트 정보와 특징들을 사전에 학습한 후, 학습된 지식을 여러 자연어 처리 분야의 하위 분석 과제(Downstream Task)에 적용하는 방법이다. 사전학습 언어모델은 많은 양의 학습된 지식 정보를 상대적으로 학습 데이터가 부족한 모델에 전달하여 성능을 높이는 방법인 전이 학습을 통해 소량의 데이터만으로도 분석 모델의 성능을 높일 수 있다는 장점이 있으며, 크게 특징 기반 접근법(Feature-based approach)과 미세 조정 접근법(Fine-tuning approach)을 통해 학습된 지식을 전이할 수 있다. 특징 기반 접근법은 하위 분석 모델의 학습 네트워크에 사전학습 언어모델을 연결하여 학습된 특징 정보를 반영하는 방법으로, 대표적인 모델로는 ELMo(Embeddings from Language Model)(Peter at al., 2018)를 꼽을 수 있다. 반면, 미세 조정 접근법은 사전학습 언어모델의 파라미터(Parameter)를 활용하여 하위 분석 과제 수행 과정에서 가중치를 조정하는 방식으로, 대표적인 모델로는 BERT(Devlin at al., 2018)를 들 수 있다.

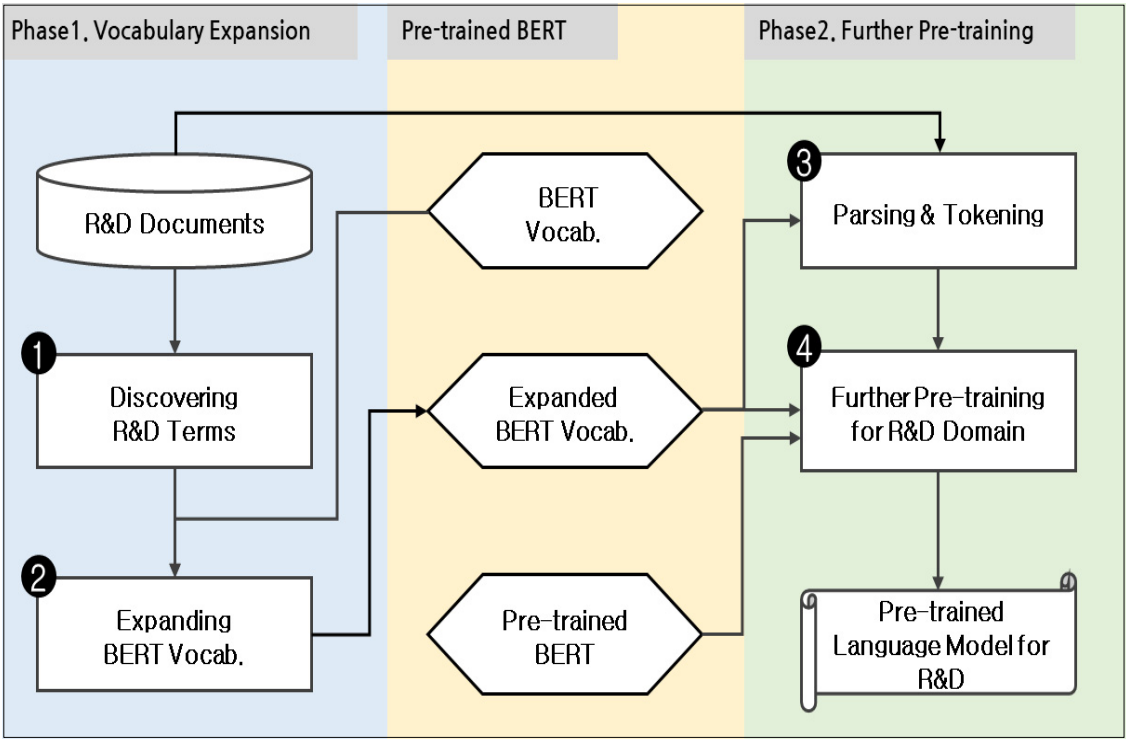
BERT는 트랜스포머 모델의 인코더 부분을 기반으로 하여 구축된 모델로, 기존 사전학습 모델의 한계인 단방향 학습 방식을 양방향으로 개선하여 더욱 풍부한 텍스

트 정보를 추출할 수 있다는 장점이 있다. BERT는 크게 NSP(Next Sentence Prediction)와 MLM을 통해 학습을 수행하며, 그중 MLM 기법은 주어진 문장 내의 단어를 무작위로 선택하여 [Mask] 토큰으로 대체한 후, 주변 맥락 정보를 통해 마스킹 단어를 예측하는 학습을 수행함으로써 양방향 정보를 모두 반영한 텍스트 표현을 학습할 수 있게 한다. BERT의 성능이 입증됨에 따라 여러 자연어 처리 분야에서 BERT를 활용하여 특정 분야에 특화된 사전학습 언어모델을 구축하려는 시도가 이어지고 있다. 특히 사전학습을 통해 범용적인 의미 표현을 학습한 BERT에 특정 분야에 대한 지식 정보를 추가하여 해당 분야에 특화된 언어모델을 재구축하는 추가 사전학습 방법이 주목받고 있다. 추가 사전학습은 기존에 학습된 BERT에 특정 분야의 데이터를 추가하여 재학습하는 방식으로, 텍스트의 일반적인 의미뿐만 아니라 해당 분야의 전문 지식도 학습할 수 있다는 장점이 있다. 구체적으로 금융 분야 지식을 학습한 FinBERT(Araci 2019), 법률 분야 지식을 학습한 LEGAL-BERT(Chalkidis at al., 2020) 등 여러 분야에서 추가 사전학습을 통해 전문 분야에 특화된 BERT를 구축하려는 시도가 이루어지고 있다.

### 2.3. 국가 R&D 정보 분류 체계

과거 국가 R&D 사업이 부처별로 개별 관리됨에 따라 중복 투자의 문제가 심각하게 발생하였으며, 이를 해결하기 위해 국가 R&D에서 산출된 과학기술 지식정보를 적극적으로 공유하고 활용하는 체계, 즉 지식경영(Knowledge management) 관점에서의 통합 시스템 구축에 대한 필요성이 제기되었다. 지식경영이란 지식이전이 효과적으로 이루어질 수 있도록 지식의 공유, 활용, 학습 창조 프로세스를 관리하는 활동을 말한다(김창식 & 곽기영, 2015; 백윤정 & 김은실, 2008; 최은수 & 이윤철, 2009). 이러한 관점에서 2008년 국가 R&D 사업의 전주기 프로세스를 체계적으로 관리하고, 국가 R&D 사업에서 산출된 모든 정보를 한 곳에서 공유될 수 있도록 하는 국가 과학기술 지식정보 포털(NTIS)이 구축되었다(김재수, 2008). 현재 NTIS는 2021년 8월 기준으로 사업·과제 92만 건, 참여인력 20만 건, 논문 153만 건 그리고 특허 74만 건 등 총 775만 건의 정보를 서비스하고 있다. 방대한 양의 국가 R&D 정보를 체계적으로 관리하고 공유할 수 있도록 국가과학기술표준분류 체계를 사용하여 분류하고 있으며, 구체적으로 과학기술표준분류 체계는 대분류 33개, 중분류 371개 그리고 소분류 2898개로 이루어져 있다(김태현 외, 2019).

이렇듯 국가 R&D 정보 관리에 있어서 분류 체계의 기준 설정 및 적용은 매우 중요한 요소로 다루어지고 있으며, 이에 따라 R&D 정보의 분류 예측, 새로운 분류 체계 구축 그리고 다른 분류 체계와의 매핑과 같은 분류 체계에 대한 다양한 연구가 진행되고 있다. 대표적으로 딥러닝 분야에서 주로 이미지(Image)나 영상(video)을 분류할 때 많이 사용되는 CNN(Convolutional Neural Networks)(LeCun, 1989)을 기반으로 국가 R&D 성과물 정보인 연구 보고서의 연구보고서명과 키워드로부터 과학기술표준분류를 예측하는 자동 분류 모델을 제안(최종윤 외, 2020) 한 것을 들 수 있다. 또한, 연구 트렌드가 빠르게 변화함에 따라 연구 분야의 세분화 및 변화를 즉시 반영되기 어렵다는 한계를 극복하기 위해, 정보에 따라 분류 체계를 자동으로 구축하고(김현종 외, 2020; 김선우 외, 2018; 오효정 외, 2003) 나아가 다양한 분류 체계를 아우를 수 있는 기본 분류 체계를 만들고 이를 통해 분류 체계 간 상호 운용성을 확인할 수 있는 방법(고영만 외, 2006)이 연구되었다. 또한 분류 체계에 대해 설명된 텍스트 데이터를 대상으로 분류 체계 간 유사도를 계산하여 서로 다른 분류




〈그림 2〉 추가 사전학습 기반 국가 R&D 전문 언어모델 구축 개요

체계를 매핑하려는 하는 연구가 수행되었으며(이재성 외, 2018), 특허청 등 관련 기관에서도 과학기술표준분류 체계, 표준산업분류, 그리고 특허청 IPC 분류 간 연계 표를 구축하여 공개하고 있다.

## 3. 제안 모델

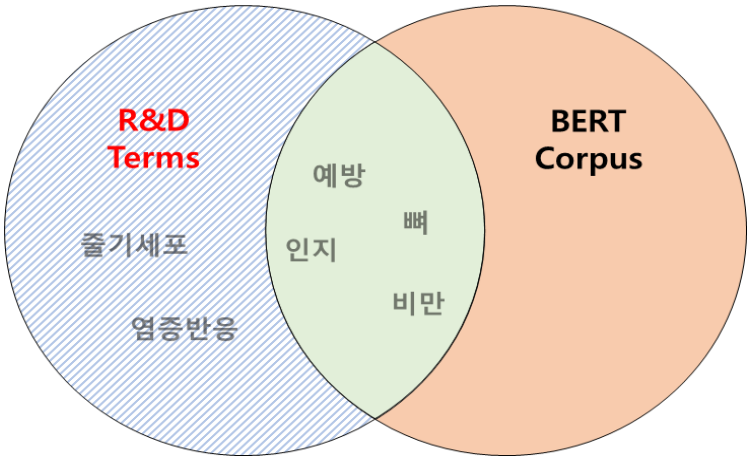
### 3.1. 구축 절차

<그림 2>는 본 연구에서 제안하는 언어모델, 즉 추가 사전학습 기반 국가 R&D 전문 언어모델을 구축하는 전체 과정을 나타낸다.

먼저, 전문어를 국가 R&D 과제 문서에서 추출하고(단계 1), 이를 사전학습된 BERT 말뭉치와 비교하여 BERT에 존재하지 않는 전문어를 식별해 토크나이저(Tokenizer)를 확장한다(단계 2). 이를 통해 국가 R&D 과제 문서를 분절한 후(단계 3), 전문어를 포함한 추가 사전학습을 수행하여 국가 R&D 분야에 특화된 언어모델을 구축한다(단계 4). 각 과정에 대한 구체적인 내용은 이후 절에서 설명한다.

### 3.2. 전문 용어 식별 및 전문어 확장 토크나이저 구축

본 절에서는 <그림 2>의 Phase 1에 대한 과정을 상세히 소개한다. 국가 R&D 전문어를 대상으로 추가 사전학습을 수행하기 위해서는 사전학습된 BERT 말뭉치에 전문어를 추가하는 과정이 필요하다. BERT는 사전에 구축된 토크나이저를 통해 단어를 분절하기 때문에, 전문어가 BERT 말뭉치에 존재하지 않는다면 이를 여러 개의 하위 단어로 분절하여 전문어가 가진 고유 의미가 손실된다. 이러한 문제를 해결하기 위해

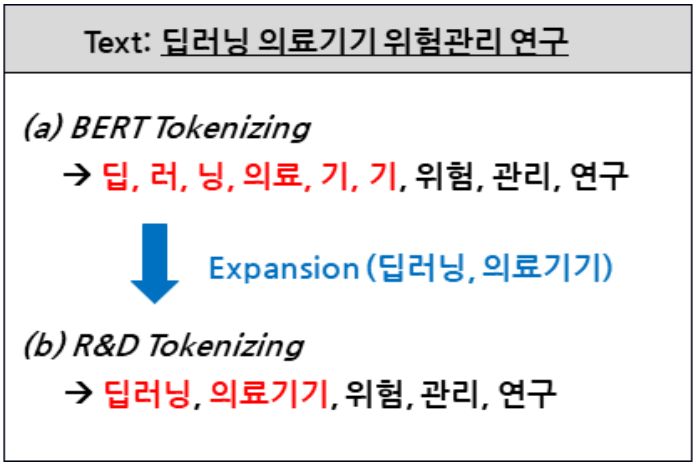



〈그림 3〉 R&D 용어 확장을 위한 전문어 식별 예

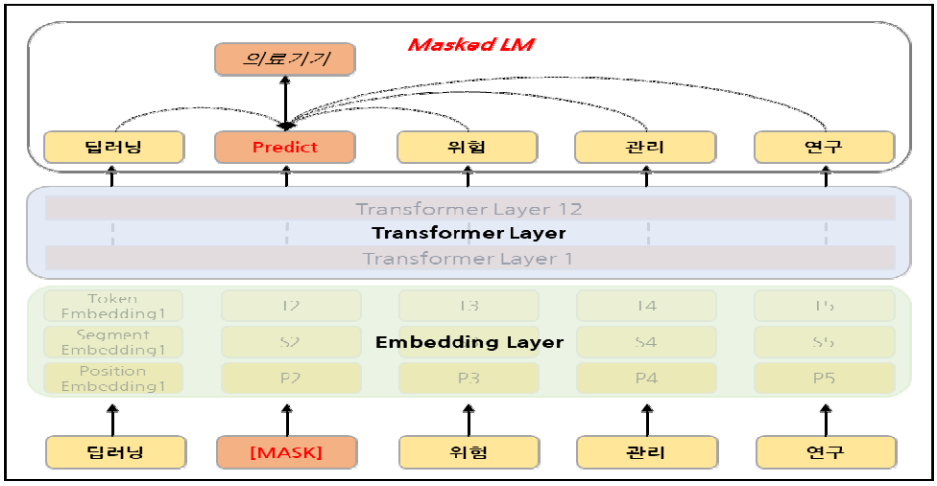
국가 R&D 과제 문서에서 추출한 전문어를 사전에 학습된 BERT 말뭉치에 추가하는 과정을 수행한다. 구체적으로 국가 R&D 과제 문서에서 특정 빈도 이상 사용된 고빈도 전문어를 추출하고(단계 1), 이를 BERT 말뭉치와 비교하여 말뭉치에 존재하지 않는 고빈도 전문어를 추가하여 토크나이저를 확장한다(단계 2). <그림 3>은 이러한 과정을 통해 “줄기세포”와 “염증반응”을 R&D 용어 확장을 위한 전문어로 식별한 예시를 보여준다.

### 3.3. MLM 기반 추가 사전학습

본 절에서는 Phase 1에서 구축한 전문어 토크나이저를 활용하여 국가 R&D 과제 문서에 대한 분절을 수행하고(단계 3), BERT의 학습 방식 중 하나인 MLM을 활용하여 추가 사전학습을 수행하는(단계 4) Phase 2의 과정을 소개한다. <그림 4>는 주어진 R&D 과제 문서를 확장 전 토크나이저와 확장 후 토크나이저를 통해 분절한 결과를 비교한 가상의 예시이다. <그림 4(a)>에서 일반 BERT 토크나이저는 “딥러닝”과 “의료기기”를 토큰으로 인식하지 못하여 더 작은 단위로 분절하였지만, 제안 방법론을 통해 확장된 전문어 토크나이저인 <그림 4(b)>는 해당 용어들을 토큰으로 잘




〈그림 4〉 확장 토크나이저 기반 전문어 보존 분절




〈그림 5〉 R&D 문장 추가 사전학습 예

식별해 냄을 확인할 수 있다.

<그림 4>와 같이 식별된 토큰들은 다음 단계인 추가 사전학습의 입력으로 사용된다(단계 4). 일반적인 BERT의 내부 구조는 임베딩 층(Embedding Layer)에서 단어, 문장 그리고 위치 정보를 결합하여 벡터를 구성하고, 12개의 트랜스포머 인코더로 이루어진 트랜스포머 층(Transformer Layer)을 통해 가중치를 학습하게 된다. <그림 5>는 BERT의 내부 학습 구조를 시각화한 것으로, 앞에서 소개한 전문어 토크나이저를 통해 과제 문서를 분절한 결과를 BERT 학습에 사용한 예를 보여준다.

BERT의 MLM 학습 기법은 무작위로 단어를 선정하여 [MASK] 토큰으로 대체하고 마스킹된 단어를 예측하는 학습을 수행하는데, 본 예에서는 “의료기기”가 [MASK] 토큰으로 대체된 경우를 보이고 있다. 추가 사전학습은 일반 사전학습과 동일한 방식으로 수행되며, 질의, 키 그리고 값에 대한 가중치인 $W_Q$, $W_K$, $W_V$ 행렬을 임의의 값이 아닌 BERT를 통해 사전학습된 값을 사용한다는 점, 방대한 양의 일반 문서가 아닌 상대적으로 소량의 R&D 과제 문서를 학습에 사용한다는 점에서 차이가 있다. 이처럼 R&D 과제 문서에 대한 추가 사전학습을 통해 최종적으로 전문어의 의미 정보와 국가 R&D 분야에 대한 지식을 충분히 정확하게 표현할 수 있다.

## 4. 실험

### 4.1. 실험 개요

본 장에서는 3장에서 제안한 모델을 실제 국가 R&D 데이터에 적용한 실험의 수행 과정 및 결과를 소개한다. 본 연구에서 제안한 R&D KoBERT 모델의 학습 및 이를 검증하기 위한 분류 실험에 사용한 데이터는 국가 R&D 분야 중 수행 과제 건수가 충분히 많으면서 전문성이 두드러지는 “보건의료”와 “정보통신” 분야의 R&D 과제 데이터이다. 2011년부터 2020년까지 최근 10년 동안 수행된 과제 데이터를 분야별로 각각 약 7.5만 건과 4.1만 건 사용하였으며, 구체적으로 R&D KoBERT 모델 구축을 위한 추가 사전학습에 각각 4.5만 건과 3만 건, 제안 모델을 검증하기 위한 분류 실험에 각각 나머지 3만 건과 1만 건의 데이터를

사용하였다. 보건의료, 정보통신 분야 모두 15개의 하위분류를 분류 타겟(Target)으로 하였으며, 입력 데이터로는 “과제명”과 “연구목표”를 연결(Concatenate)하여 사용하였다.

## 4.2. 전문 용어 선정 및 전문어 토크나이저 구축 결과

본 절에서는 전문어를 선정하고 사전학습된 KoBERT 말뭉치에 추가하여 전문어 토크나이저를 구축한 결과를 제시한다. <그림 2>의 Phase 1 과정을 통해 보건의료와 정보통신 분야 각각의 데이터에서 10회 이상 출현한 전문어를 추가 전문어 후보로 추출하였으며, 이들 중 KoBERT의 토크나이저에 포함되지 않은 용어를 추가 전문어로 최종 선정하였다. <표 1>에 나타난 용어들은 추가 후보로 추출된 전문어를 의미하며, 이들 중 밑줄로 표시된 용어들은 추가 전문어로 최종 선정된 용어를 의미한다.

〈표 1〉 추가 전문어 선정 결과 (일부)

| | |
|---|---|
| 보건의료 분야 | 뇌, **미세먼지**, **바이오마커**, 암, 예방, **줄기세포**, **파킨슨병**, 프로그램, ... |
| 정보통신 분야 | 네트워크, 머신러닝, **블록체인**, 스마트, **스마트센서**, 안드로이드, **증강현실**, 클라우드, ... |

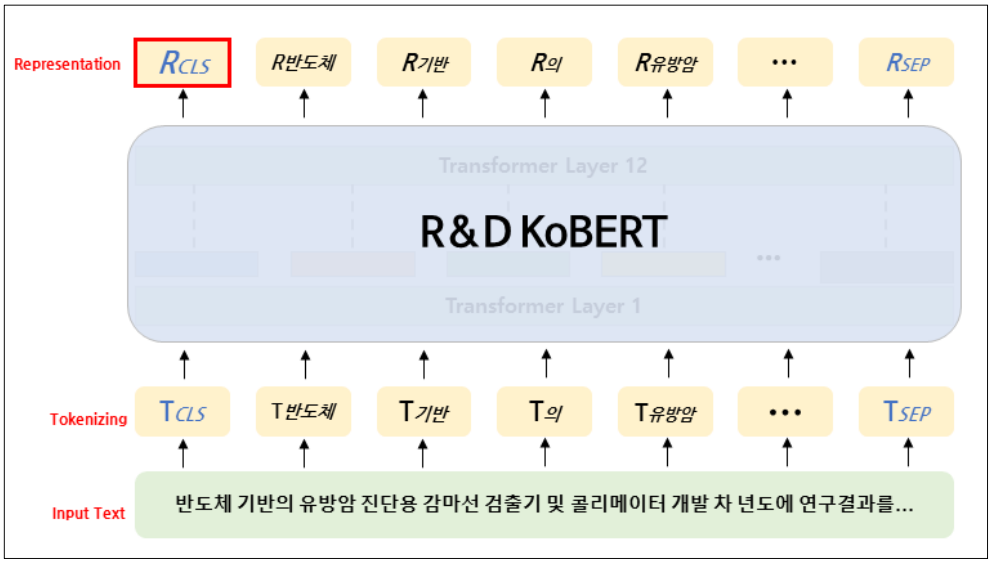
사전학습된 BERT는 SKTBrain에서 개발한 한국어 BERT 언어모델인 KoBERT-base를 사용하였다. KoBERT 토크나이저는 8,002개의 학습된 말뭉치를 가지고 있으며, 본 실험에서는 추가 후보로 추출된 전문어 중 KoBERT 말뭉치에 이미 포함된 단어를 제외하고 최종적으로 분야별로 각각 907개, 487개의 추가 전문어를 선정하였다. 이후, 선정된 전문어를 KoBERT 말뭉치에 추가하여 각각 8,909개, 8,489개의 확장된 어휘를 갖는 전문어 토크나이저를 구축하였다. <표 2>는 확장된 전문어 토크나이저를 통해 각 과제 데이터를 분절한 결과의 일부를 보여준다.

〈표 2〉 확장된 토크나이저를 통한 과제 데이터 분절 (일부)

| | Tokens by Expanded Tokenizer |
|---|---|
| 보건의료 | 간경화', '간암', '_치료', '용', '약물', '_개발', '_및', '전임상', '_시험' |
| | 폐암', '_', '의', '_맞', '춤', '진단', '_및', '표적치료', '_법', '_개발' |
| 정보통신 | 차세대', '_', '바이오', '인식', '_응', '용', '기술', '_표준', '개발' |
| | _모바일', '증강현실', '_', '을', '_위한', '_거리', '인식', '_시스템', '_개발' |

## 4.3. MLM 기반 추가 사전학습 결과

본 절에서는 확장된 전문어 토크나이저를 통해 분절된 과제 데이터에 대해 MLM 기반 추가 사전학습을 수행한 결과를 소개한다. 추가 사전학습을 위한 모델은 트랜스포머 블록 12개, 어텐션 헤드 12개, 그리고 은닉층 768개로 구성되어 있으며, 문장의 최대 길이는 512로 설정하였다. 또한 추가 사전학습을 위한 모델의 가중치 초기값은 KoBERT-base 모델의 가중치를 사용하였다. 이후 전문어 토크나이저를 통해 분절된 과제 데이터에 대해 전체 토큰의 15%를 무작위로 선정하고, 그중 80%를 [MASK] 토큰으로 대체하는 마스킹을 수행하여 최종적으로 MLM 기반의 추가 사전학습을 수행한 R&D KoBERT를 구축하였다. <그림 6>은 R&D KoBERT를 통해 주어진 문장에 대한 학습을 수행하는 과정을 나타내며, 실제 데이터의 동일 문장에 대해 KoBERT-base에 의한 학습과 R&D KoBERT에 의한 학습 결과가 다르게 나타남은 <표 3>을 통해 확인할 수 있다. <표 3>은 <그림 6>에서 나타난 문장 표현 벡터, 즉 첫 토큰인 CLS의 표현 벡터인 $R_{CLS}$를 나타낸다.




〈그림 6〉 R&D KoBERT를 통한 학습 (일부)

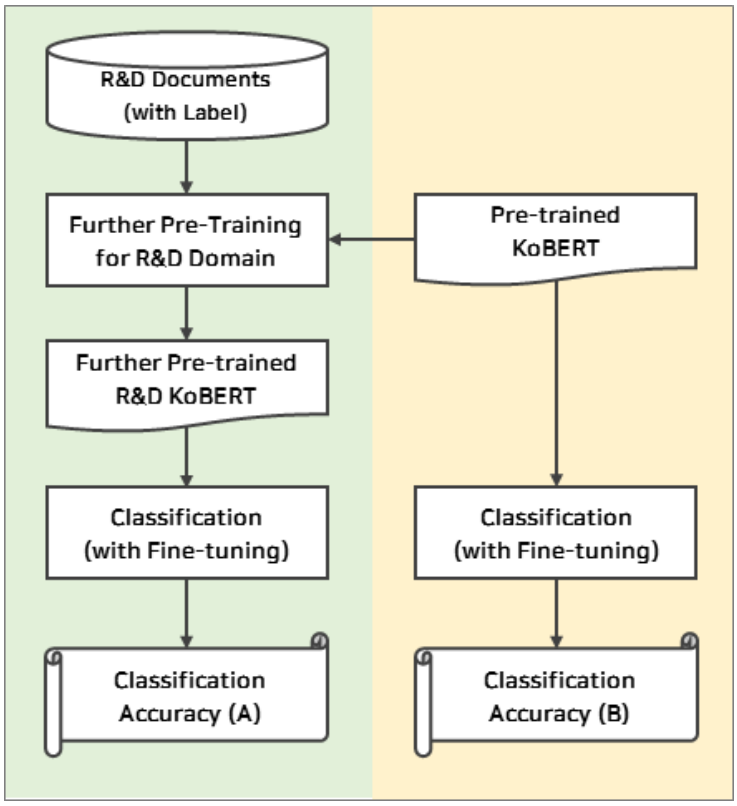
〈표 3〉 KoBERT-base와 R&D KoBERT의 문장 표현 학습 결과 비교

| | | Dim1 | Dim2 | Dim3 | ... | Dim768 |
|---|---|---|---|---|---|---|
| 보건의료 | KoBERT-base | -0.0544 | -0.0315 | 0.0553 | ... | 0.0553 |
| | R&D KoBERT | 0.0135 | -0.0530 | 0.3319 | ... | 0.3319 |
| 정보통신 | KoBERT-base | -0.0874 | -0.0438 | 0.1738 | ... | 0.1738 |
| | R&D KoBERT | 0.0680 | -0.0362 | 0.1869 | ... | 0.0422 |

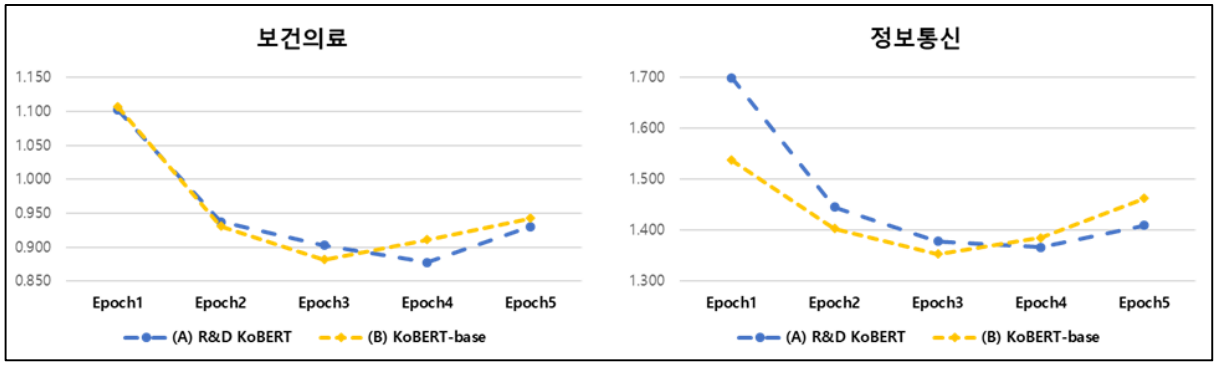
### 4.4. 미세 조정 및 성능 평가

본 절에서는 본 연구에서 제안한 모델인 R&D KoBERT와 순수한 KoBERT 모델의 성능을 비교하기 위해 진행한 성능 평가 실험의 결과를 소개한다. 성능 평가를 위해 보건의료와 정보통신 과제 데이터에서 추가 사전학습에 사용하지 않은 데이터를 각각 3만, 1만 건을 사용하여 각 과제 데이터의 15개 하위분류를 예측하는 다중 분류(Multiclass Classification) 실험을 수행하였다. <그림 7>은 성능 평가 실험의 전반적인 과정을 나타낸다.

<그림 7>의 좌측은 제안 모델인 R&D KoBERT를 활용하여 R&D 과제 데이터에 대한 다중 분류 학습과 미세 조정을 수행하는 흐름을 나타낸다. 한편 <그림 7>의 우측은 KoBERT-base 모델을 사용하여 다중 분류 학습을 수행한 것이다. 과제 데이터에 대한 다중 분류 학습을 수행하기 위해 보건의료와 정보통신 데이터를 각각 6:2:2로 분할하여 훈련용(Training), 검증용(Validation) 그리고 평가용(Test) 데이터 집합을 구축하였고, 평가 지표로는 정확도(Accuracy)와 F1 점수를 측정하였다. 각 모델의 성능 비교 결과는 <그림 8>과 <그림 9>에서 확인할 수 있다. <그림 8>은 제안 모델인 R&D KoBERT와 KoBERT-base 모델에 대한 다중 분류 학습의 에폭(Epoch)에 따른 검증용 데이터의 손실(Loss)을 나타낸다. 최종 평가 모델은 손실 값이 가장 낮은 에폭(Epoch)을 채택하여 평가용 데이터에 대한



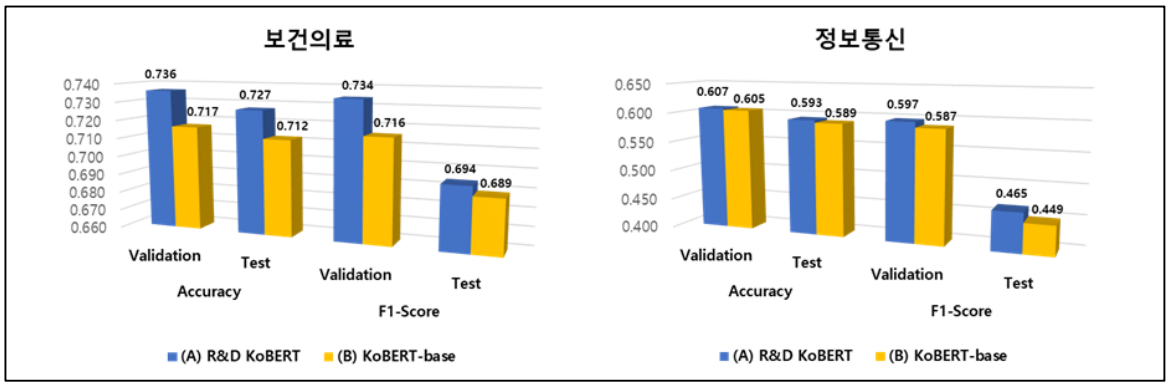

〈그림 7〉 성능 평가 실험 개요




〈그림 8〉 에폭별 손실 값 비교

추론을 수행하였으며, 그 결과는 <그림 9>와 같다.

<그림 9>는 본 연구에서 제안한 R&D KoBERT 모델과 비교 모델인 KoBERT-base 모델에 대한 검증용 데이터와 평가용 데이터의 성능 평가 결과이다. 실험 결과, 보건의료와 정보통신 분야 모두에서, 추가 사전학습을 수행한 제안 모델(A)이 비교 모델(B)에 비해 분류 정확도와 F1점수 측면에서 모두 우수한 성능을 나타냈다. 실험을 통해 본 연구에서 제안한 방법, 즉 사전학습 언어모델에 대해 전문 분야에 대한 추가 사전학습을 수행하여 해당 분야에 특화된 지식을 전이하는 방법이 전문어를 다수 포함하고 있는 문서의 분석에 효과적으로 적용될 수 있음을 확인하였다.




〈그림 9〉 R&D KoBERT와 KoBERT-base 모델의 분류 성능 비교

## 5. 결론

최근 딥러닝 기술을 활용한 모델이 기존의 통계 기반 모델 대비 우수한 성능을 보임에 따라, 다양한 분야에서 딥러닝을 도입하여 활용하고자 하는 수요가 증가하였다. 이러한 흐름에 따라 국가 R&D 분야에서도 과제, 논문 그리고 특허와 같은 방대한 양의 R&D 문서를 분석하기 위한 딥러닝 기반의 자연어처리 기술에 주목하고 있으며, 특히 방대한 텍스트 데이터에 대해 사전학습을 수행한 BERT 언어모델의 활용과 개선에 대한 관심이 높아지고 있다. 그러나 고도로 전문화된 R&D 분야에서 기본 BERT 모델을 그대로 사용하는 경우, BERT의 방대한 학습량에도 불구하고 해당 분야에서 높은 빈도로 사용되는 전문어가 충분히 학습되지 못한 경우가 발생할 수 있으며, 이는 분석 모델의 성능에도 영향을 미칠 수 있다.

이에 본 연구에서는 추가 사전학습 기법을 기반으로 국가 R&D 분야에서 사용하는 전문어를 추가로 학습한 언어모델, 즉 국가 R&D 분야의 전문 지식을 KoBERT 언어모델에 추가한 국가 R&D 특화 언어모델을 제안하였다. 또한, 제안한 모델의 성능 평가를 위해 최근 10년 동안의 보건의료, 정보통신 분야의 과제 데이터를 대상으로 분류 분석을 수행한 결과, 제안 모델이 기본 KoBERT 모델보다 정확도 측면에서 더욱 우수한 성능을 나타내는 것을 확인하였다. 향후 상기 두 분야에 주어진 분석 과제를 해결하고자 하는 경우 본 연구에서 제안한 언어모델이 유용하게 활용될 수 있을 것으로 기대한다. 또한, 본 연구에서 제안한 모델은 사전학습 및 추가 사전학습을 통해 재사용 가능한 방식으로 지식을 전이할 수 있는 방안을 다룬다는 점에서 지식경영 분야에서의 다양한 문제 해결에 직간접적으로 기여할 수 있을 것으로 기대한다.

하지만 본 연구에서는 전체 33개 국가 R&D 분야 중 과제 건수가 상위이면서 전문성이 두드러지는 보건의료, 정보통신의 2개 분야만을 선정하여 추가 사전학습과 성능 평가를 진행하였다. 향후 후속 연구에서는 상기 두 분야 이외에도 매우 활발하게 R&D 과제 수행이 이루어지고 있는 분야인 농림수산식품, 기계, 그리고 생명과학 분야 등에 대한 추가 실험을 수행하여, 각 전문 분야별 데이터의 특징에 따른 제안 방법론의 효과성을 엄밀하게 분석할 필요가 있다.

## 〈참고문헌〉

### [국내 문헌]


1. 고영만, 서태설, 조순영 (2006). 국가지식정보 자원 분류 체계 표준화 연구. **한국문헌정보학회지, 40**(3), 151-173.
2. 김선우, 고건우, 최원준, 정희석, 윤화묵, 최성필 (2018). 기술과학 분야 학술문헌에 대한 학습집합 반자동 구축 및 자동분류 통합 연구. **정보관리학회지, 35**(4), 141-164.
3. 김재수 (2008). 국가과학기술종합정보서비스(NTIS)-NTIS 구축사업 개요. **지식정보인프라, 30**, 31-34.
4. 김창식, 곽기영 (2015). 조직구성원의 네트워크 위치가 지식공유에 미치는 영향. **지식경영연구, 16**(2), 67-89.
5. 김태현, 양명석, 최광남 (2019). 국가R&D정보 활용을 위한 전문용어사전 구축. **한국콘텐츠학회 논문지, 19**(10), 217-225.
6. 김현종, 이강배, 류승우, 홍순구 (2020). A study on classification scheme generation for automatic classification of unlabeled documents. **디지털콘텐츠학회 논문지, 21**(12), 2211-2219.
7. 백윤정, 김은실 (2008). 실행공동체(CoP)내 지식공유의 영향요인: 구조적 특성과 관계적 특성의 조절효과를 중심으로. **지식경영연구, 9**(2), 63-86.
8. 오효정, 장문수, 장명길 (2006). 정답문서집합 자동 구축을 위한 속성 기반 분류 방법. **정보과학회논문지, 30**(7/8), 764-772.
9. 이재성, 전승표, 유형선 (2018) 한국표준산업분류를 기준으로 한 문서의 자동 분류 모델에 관한 연구. **지능정보연구, 24**(3), 221-241.
10. 최은수, 이윤철 (2009). 정보기술이 지식경영활동과 성과에 미치는 효과에 대한 실증분석. **지식경영연구, 10**(3), 51-80.
11. 최종윤, 한혁, 정유철 (2020). 국가 과학기술 표준분류 체계 기반 연구보고서 문서의 자동 분류 연구. **한국산학기술학회 논문지, 21**(1), 169-177.


### [국외 문헌]


12. Araci, D. (2019). ***FinBERT: Financial sentiment analysis with pre-trained language models***. arXiv preprint arXiv: 1908:10063.
13. Bahdanau, D., Cho, K., & Bengio, Y. (2014). ***Neural machine translation by jointly learning to align and translate***. arXiv preprint arXiv: 1409.0473.
14. Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. ***Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing(EMNLP-IJCNLP)***, 3615-3620.
15. Chalkids, L., Fergadiotis, M., Malakasioits, P., Aletras, N., & Androutsopoulos, I. (2020). ***LEGAL-BERT: The muppets straight out of law school***. arXiv preprint arXiv: 2010:02559.
16. Devlin, J., Chang, W., Lee, K., & Toutanava, K. (2018). ***BERT: Pre-training of deep bidirectional transformers for language understanding***. arXiv preprint arXiv: 1810.04805.
17. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. ***Journal of Neural Computation, 9***(8), 1735-1780.
18. Le Cun, Y. (1989). Backpropagation applied to handwritten zip code recognition. ***Neural Computation, 1***(4), 541-551.
19. Lee, J. H., Yoon, W. J., Kim, S. D., Kim, D. H., Kim, S. K., So, C. H., & Kang, J. W. (2019). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. ***Bioinformatics, 36***(4), 1234-1240.
20. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. ***In Proceedings of International Conference on Learning Representations(ICLR)***.
21. Mikolov, T., Karafiát, M., Burget, L., & Cernocký, J. (2010). Recurrent neural network based language model. ***In 11th Annual Conference of the International Speech Communication Association(INTERSPEECH)***, 1045-1048.
22. Penningion, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. ***In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing(EMNLP)***, 1532-1543.
23. Peter, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. ***In Proceedings of the 2018 Conference of the North American Chapter of the***

*Association for Computational Linguistics(NAACL), 1*, 2227-2237.

24. Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *In Proceedings, of the 27th International Conference on Neural Information Processing Systems(NIPS), 2*, 3104-3112.

25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *In Proceedings of the 31st International Conference on Neural Information Processing Systems(NIPS)*, 6000-6010.

## [URL]

26. SKTBrain, KoBERT, GitHub Repository. https://github.com/SKTBrain/KoBERT

27. 국가과학기술지식정보서비스(NTIS). www.ntis.go.kr

28. 과학기술정책지원서비스. https://www.k2base.re.kr/clInfo/aboutClInfo.do

29. 특허청(KIPO). www.kipo.go.kr

## 저 자 소 개


### 유 은 지 (Eunji Yu)

현재 국민대학교 비즈니스IT전문대학원 박사과정에 재학 중이며, 한국과학기술정보연구원(KISTI) NTIS센터 연구원으로 재직 중이다. 국민대학교 비즈니스IT전문대학원에서 경영정보학 석사 학위를 취득하였다. 주요 관심분야는 Text Mining, Deep Learning, SNA, National R&D Data Analysis 등이다.

### 서 수 민 (Sumin Seo)

현재 국민대학교 비즈니스IT전문대학원 석사과정에 재학 중이며, 아주대학교에서 사회학 학사 학위를 취득하였다. 주요 관심분야는 Natural Language Processing, Deep Learning, Text Mining 등이다.

### 김 남 규 (Namgyu Kim)

현재 국민대학교 경영대학 경영정보학부 및 비즈니스IT전문대학원 교수로 재직 중이다. 서울대학교 컴퓨터공학과에서 학사 학위를 취득하고, KAIST 테크노경영대학원에서 경영공학 석사 및 박사 학위를 취득하였다. 한국지능정보시스템학회 부회장, 한국정보기술응용학회 부회장, 한국경영학회 상임이사, 한국경영정보학회 이사, 한국인터넷정보학회 이사를 역임하였다. 주요 관심 분야는 Text Mining, Data Mining, Deep Learning, Data Modeling 등이다.

## 〈 Abstract 〉

# Building Specialized Language Model for National R&D through Knowledge Transfer Based on Further Pre-training

Eunji Yu*, Sumin Seo**, Namgyu Kim***

With the recent rapid development of deep learning technology, the demand for analyzing huge text documents in the national R&D field from various perspectives is rapidly increasing. In particular, interest in the application of a BERT(Bidirectional Encoder Representations from Transformers) language model that has pre-trained a large corpus is growing. However, the terminology used frequently in highly specialized fields such as national R&D are often not sufficiently learned in basic BERT. This is pointed out as a limitation of understanding documents in specialized fields through BERT. Therefore, this study proposes a method to build an R&D KoBERT language model that transfers national R&D field knowledge to basic BERT using further pre-training. In addition, in order to evaluate the performance of the proposed model, we performed classification analysis on about 116,000 R&D reports in the health care and information and communication fields. Experimental results showed that our proposed model showed higher performance in terms of accuracy compared to the pure KoBERT model.

Key Words: National R&D, Knowledge Transfer, Pre-trained Language Model, BERT, Further Pre-training

* Graduate School of Business IT, Kookmin University
** Graduate School of Business IT, Kookmin University
*** Graduate School of Business IT, Kookmin University