공공 R&D 가치창출을 위한 다목적 문서추천 자동화 기술 개발

김성진*, 최윤정*, 변정은*, 송경우**
*한국과학기술정보연구원 기술사업화연구센터
**연세대학교 응용통계학과
e-mail:sungjin.kim@kisti.re.kr

Multi-Purpose Document Recommendation for Public R&D Value Creation

Sungjin Kim*, Yunjeong Choi*, Jeongeun Byun*, Kyungwoo Song**
*Technology Commercialization Research Center, Korea Institute of Science and Technology Information
**Dept of Applied Statistics, Yonsei University

요 약

공공 R&D를 통해 논문, 특허, 연구보고서 등 수많은 결과물이 생산되고 있으나 실제 이 결과물을 활용하여 사업화까지 연결시키는 성공률은 선진국 대비 다소 낮은 현실이다. 이러한 한계를 해결하기 위한 방법 중 하나로 본 논문에서는 사용자가 원하는 다양한 목적에 맞는 다목적 문서추천 자동화 기술을 개발하고자 한다. 기존의 BERT 언어모델을 바탕에 기술사업화 특화 문서를 추가로 학습한 Tc-BERT를 구축하고 이를 활용하여 기술 문서를 구성하고 있는 각 문장별 의미를 13개의 태그로 구분하였다. 이를 통해 사용자가 입력하는 문서와 사용자가 원하는 목적에 가장 부합하는 문서들을 추천해 줄 수 있다. 이로써 사용자는 공공 R&D의 결과물에서 사용자가 원하는 문서를 보다 정확하게 추천을 받을 수 있고 이를 기술사업화에 활용함으로써 공공 R&D의 가치를

1. 서론

공공 R&D를 통해 획득되는 결과물은 논문, 특허, 보고서, 연구노트, 이미지 데이터 등 다양한 종류로 나타날 수 있다. 이들 결과물의 상당수는 텍스트, 즉 문서의 형태로 존재하고 관리되고 있다. 국내의 공공 R&D에 대한 정보는 국가과학기술지식정보서비스(NTIS)에서 확인할 수 있으며 현재 수십만 건의 공공 R&D 정보가 저장되고 관리되고 있다. 그러나 공공 R&D가 사업화까지 연결되는 성공률은 선진국에 비해 크게 낮고 공공 R&D의 특허 활용도 역시 현저히 낮은 수준이다. 공공 R&D의 결과를 기술이전 받아 자사 제품의 성능을 개선하거나 신제품을 개발하려고 하는 수요기업 측면에서는 수많은 R&D 정보 혹은 특허 문서 중에서 자사에 필요한 문서를 찾는 것이 매우 중요하다.

따라서 본 논문에서는 이러한 문제를 해결하기 위해 BERT 언어모델을 활용하여 사용자의 목적에 부합하는 문서를 추천해주는 모델을 제시하고자 한다. 본 논문에서 제시하는 연구모델은 인공지능 기반 공공R&D 가치창출 플랫폼에 탑재되어 활용을 목표로 개발중에 있다.

2. 관련 문헌 연구

언어모델(Language Model)은 문서를 이해하고 분류하는 등의 여러 용도에 널리 활용되고 있으며 최근에는 기계번역, 문서분류, 문장생성 등 다양한 영역에서 가능성을 보여주고 있다. 특히 최근 ChatGPT의 등장으로 거대언어모델을 활용한 생성형 AI가 텍스트는 물론 이미지 등 다양한 콘텐츠를 만들어내는 것에 많은 연구가 진행되고 있다.

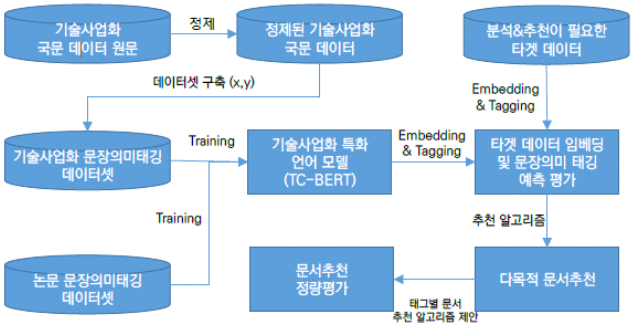
임준호 외(2020)에서는 딥러닝을 활용한 언어모델의 기술동향을 정리하였으며 특히 딥러닝 언어모델을 한국어에 적용한 KorBERT 모델에 대해 소개하였다. KorBERT는 ETRI가 개발한 한국어 전용 딥러닝 언어모델로 백과사전류 텍스트와 신문기사를 대상으로 약 47억 개의 형태소를 학습한 모델이다.

국가 R&D 분야에 BERT모델을 적용하여 국가 R&D 분야에 특화된 모델을 제안한 연구에서는 추가 사전학습 기법을 통해 기존 언어모델에 추가로 국가 R&D 모델을 학습한 언어모델을 제안하였다(유은지 외 2021).

또한 김성진 외(2021)에서는 BERT 모델을 활용하여 기술사업화 분야에 특화된 Tc-BERT를 제안하였고 이를 바탕으로 기술사업화 핵심용어 추출방안을 연구한 바 있다.

3. 연구모형

본 논문에서의 공공 R&D 가치창출을 위한 다목적 문서 추천 자동화 기술의 연구 방법론은 아래 [그림 1]과 같다.



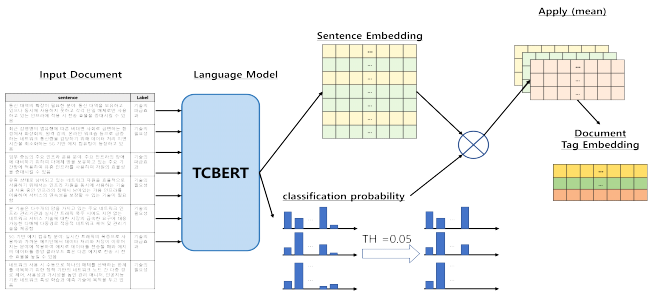
[그림 1] 기술사업화 다목적 문서 추천 프로세스

기술사업화에 특화된 언어모델을 만들기 위해 일반적으로 공개되어 있는 국문 데이터 및 NTIS, NTB, 중소기업기술로드맵 등 기술사업화 특화 문서를 수집 정제하여 데이터셋을 구축하였고 이를 학습하여 BERT 모델을 구축하였다. 또한 문서를 구성하는 각 문장이 어떤 의미를 갖고 있는지를 분류할 수 있는 Task 수행을 위해 논문 및 기술보고서 등으로부터 대량의 문장을 추출하고 각 문장의 의미에 따라 하나의 태그(Tag)를 부여하여 모델 학습 시의 레이블로 활용하였다. 본 연구에서 정의한 태그는 총 13개로 태그별 의미 및 문장 수 등의 통계량은 아래 [표 1]과 같다.

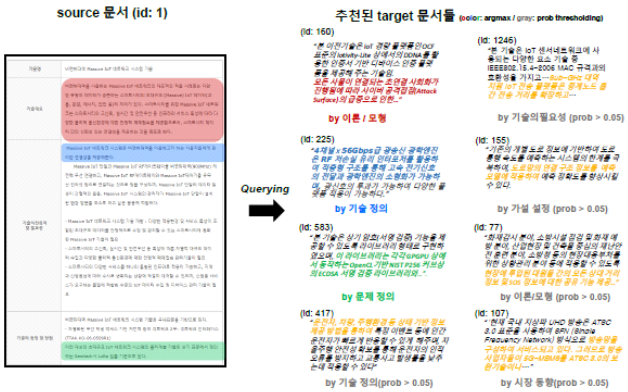
[표 1] 문장 태그별 의미 및 문장 수

태그 명	태그별 의미	태그별 문장 수 (개)
문제 정의	해결 하고자 하는 문제를 표현하는 문장	14,887
가설 설정	방법을 제안하기 위해 가정한 제약이나 현상 등을 표현한 문장	3,072
기술 정의	기술적인 용어의 정의를 내리는 문장	12,301
제안 방법	서론 또는 연구 배경에서 다른 연구와 다른 특징적으로 제시한 문장	24,157
대상 데이터	데이터의 수집 방법이나 출처를 설명한 문장	20,139
데이터 처리	대상 데이터에 대해 논문의 연구 방법을 적용한 결과를 통계적으로 해석하거나 분석하기 위한 처리 방법을 정의한 문장	13,533
이론/모형	연구에서 설정한 주된 문제나 목표를 해결하기 위해 알려진 이론이나 다른 연구자의 결과를 활용하는 내용을 포괄하는 문장	11,130
성능/효과	실험 결과를 수치적으로 보여주고 해석하는 문장	35,397
후속연구	연구의 한계와 확장 및 활용에 대해 설명하는 문장	17,671
기술의 필요성	기술이 활용될 수 있는 분야 및 과업과 관련하여 중요성과 필요성을 설명하는 문장.	7,330
기술의 파급효과	기술이 우리 사회와 과학 분야에 미칠 수 있는 경제적 및 기술적 영향에 대해 설명하는 문장	5,549
기술 동향	기술에 대한 국내 및 국외 동향 및 관련 기술들 소개에 관한 문장	4,987
시장 동향	기술의 국내 및 국외 시장의 규모변화 추세 변화 등에 관한 문장	16,432

구축된 Tc-BERT를 활용하여 문서를 구성하는 문장을 Tc-BERT의 입력으로 사용하여 문장의 임베딩과 각 태그별로 해당 문장이 분류될 확률을 구하고 이를 문서를 구성하는 전체 문장으로 확장하여 문서 하나의 태그에 대한 임베딩을 도출할 수 있다. 이러한 과정을 거쳐 문서 추천 풀 전체와 문서 추천에 활용할 소스 문서에 대해 태그별 임베딩을 구하고 소스 문서에 대해 태그별로 코사인 유사도가 가장 높은 문서를 추천할 수 있고[그림 2], 실제 본 연구모델을 활용한 문서 추천의 예시중 하나는 [그림 3]과 같다.



[그림 2] 문서의 태그별 임베딩 획득 프로세스



[그림 3] 문서추천 결과 예시

4. 결론

기존의 문서추천 방법론은 키워드 혹은 전체적인 문서의 유사도를 기반으로 수행된 연구들이 다수를 이루고 있다. 본 논문에서는 “기술동향” 관점에서의 문서 추천, “시장 동향” 관점에서의 문서 추천 등 사용자의 문서를 찾는 목적에 맞게 문서를 추천해 주는 방법론을 제시하였다. 이를 통해 수많은 기술 문서중에서 사업화를 위해 사용자의 목적과 일치하는 문서를 다양한 의미 관점에서 추천해 줄 수 있는 것에 본 논문의 의의가 있다고 할 수 있다. 향후 본 연구모델을 실제 공공 R&D 가치창출 플랫폼에 탑재함으로써 사용자의 공공 R&D 결과물 활용에 많은 도움이 될 수 있을 것으로 기대한다.

참고문헌

[1] 김성진, 송경우, 최윤정, “기계학습 기반의 공공R&D 기술사업화 핵심용어 추출방안 연구”, 한국지능정보시스템학회 학술대회 논문집, Vol.2021 No. 12(2021), pp. 19-20

[2] 유은지, 서수민, 김남규, “추가 사전학습 기반 지식 전이를 통한 국가 R&D 전문 언어모델 구축”, 지식경영연구, Vol. 22, No. 3(2021), 91~106.

[3] 임준호, 김현기, 김영길, “딥러닝 사전학습 언어모델 기술동향” 전자통신동향분석, Vol. 35, No.3(2020), 9~19.

이 논문은 2022년도 한국과학기술정보연구원(KISTI)의 기본사업으로 수행된 연구입니다.(과제번호: (KISTI)K-22-L03-C03, (NTIS)1711175875)