지식재산연구
Journal of Intellectual Property

지식재산연구 20권 1호, 2025년 3월, 89-117면.
Journal of Intellectual Property Volume 20, Issue 1, March 2025, Pages 89-117.
https://doi.org/10.34122/jip.2025.20.1.4

Print ISSN 1975-5945
Online ISSN 2733-8487

RESEARCH ARTICLE

Improving the Performance of a Korean Patent Document Search Model using KorPatBERT-based CPC Classification Model

Jaeok Min[1,2†], Hansung Noh[1,2†], Minhak Kwak[2], Solbin Hwang[2], Taehoon Kim[3]

[1]Ph.D. Candidate, Dept. of Intellectual Property Convergence, Chungnam National University, Republic of Korea
[2]Intelligent Information Strategy Dept., Korea Institute of Patent Information, Republic of Korea
[3]Professor, Dept. of Electric, Electronic & Communication Engineering Education, College of Education, Chungnam National University, Republic of Korea
[†]These authors contributed equally to this work as first authors.

Corresponding Author: Taehoon Kim (kth0423@cnu.ac.kr)

ABSTRACT

The global competition for technological supremacy is intensifying, prompting every country to focus on securing technological advantages through patent acquisition. In this environment, efficient and accurate patent searching is a key factor for establishing national technological sovereignty and strengthening global competitiveness. However, identifying prior art patents accurately and effectively within vast patent data remains a challenging task. To address this challenge, this study proposes an advanced patent search model that leverages artificial intelligence technology.

This study presents a method for creating models according to the CPC classification model based on the KorPatBERT(Korean Patent BERT) that can deeply understand the detailed technical context of patent documents through pre-training involving vast patent data. Furthermore, this study presents a method for generating high-dimensional document embedding vectors that can effectively reflect the technical subject and context of patent documents and a method for building a search system capable of processing large volumes of patent data in real time. By integrating the proposed patent search model into this system, the study successfully demonstrated improved search performance compared with existing methods in objective performance evaluations.

This study can contribute toward enhancing industrial applicability and practical usability by applying the processes of currently operational patent search data and systems. The current study's findings are expected to provide a foundation for nations and companies to continuously lead innovation and efficiently manage and utilize patents.

KEYWORDS

Intellectual Property Rights, Patent, KorPatBERT, Artificial Intelligence, Prior-art Patent, CPC, Patent Classification, Patent Search, Embedding Vector

Open Access

Received: December 24, 2024
Revised: January 26, 2025
Accepted: February 27, 2025
Published: March 30, 2025

Funding: The author received manuscript fees for this article from Korea Institute of Intellectual Property.

Conflict of interest: No potential conflict of interest relevant to this article was reported.

원저

KorPatBERT 기반 CPC 분류 모델을 활용한 한국어 특허 문헌 검색 모델 성능 향상 연구

민재옥[1,2†], 노한성[1,2†], 곽민학[2], 황솔빈[2], 김태훈[3]

[1]충남대학교 대학원 지식재산융합학과 박사과정
[2]한국특허정보원 지능정보전략실
[3]충남대학교 사범대학 전기 · 전자 · 통신공학교육과 교수
†민재옥과 노한성은 공동 제1저자로 이 논문에 기여하였습니다.

교신저자: 김태훈 (kth0423@cnu.ac.kr)

차례

국문초록

전 세계적으로 기술 패권을 둘러싼 경쟁이 날로 심화되고 있으며, 각국은 특허 확보를 통해 기술 우위를 확보하려는 노력을 강화하고 있다. 이 과정에서 신속하고 정확한 특허 검색은 국가 기술 주권 확립과 글로벌 경쟁력 강화를 위한 핵심 요소이다. 그러나 방대한 특허 데이터 속에서 선행 기술을 효과적이고 정밀하게 찾아내는 일은 여전히 도전적인 과제로 남아 있으며, 이를 해결하기 위해, 본 연구는 인공지능 기술을 활용한 고도화된 특허 검색 모델을 제안하였다.

본 연구는 방대한 특허 데이터를 사전 학습하여 특허 문헌의 세부적인 기술적 맥락을 깊이 이해할 수 있는 KorPatBERT 기반 CPC 분류 체계별 모델을 생성하는 방법을 제시하였다. 아울러, 이 모델을 통해 특허 문헌의 기술적 주제와 맥락을 효과적으로 반영하는 고차원의 문헌 임베딩 벡터를 생성하는 방법을 제시하였고, 대량의 특허 데이터를 실시간으로 처리할 수 있는 검색 시스템 구축 방법을 함께 제시하였다. 이를 통해 제안된 특허 검색 모델은 시스템과 결합하여 객관적인 성능 평가에서 기존 방식보다 개선된 검색 성능을 성공적으로 입증하였다.

본 연구에서는 현재 운영 중인 특허 검색 데이터 및 검색 시스템의 프로세스를 적용함으로써 산업적 활용성과 실질적 유용 가능성을 높이는데 기여하였다. 이러한 연구 결과는 국가와 기업이 지속적으로 혁신을 선도하고, 특허를 효율적으로 관리하고 활용할 수 있는 기반을 제공할 것으로 기대한다.

주제어

지식재산권, 특허, KorPatBERT, 인공지능, 선행기술, CPC, 특허분류, 특허검색, 임베딩벡터

1. 서론

최근에는 특허 출원의 양이 전 세계적으로 기하급수적으로 증가[1]하고 있으며, Ali, Amna, et al.[2] 연구에서 이러한 방대한 특허를 효율적으로 관리하고 필요한 정보를 정확하게 검색할 수 있는 특허 검색 시스템의 중요성이 더욱 커지고 있다고 주장하였다. 특허 검색 시스템은 정밀한 문헌 분석을 기반으로 신속하고 정확한 검색 결과를 제공하는 것이 필수적이다. 단순히 기술 정보를 확보하는 역할을 넘어 적시에 정확한 특허 정보를 제공함으로써, 새로운 발명이 신규성과 진보성을 충족하는지 판단하는 데 중요한 역할을 한다. 이를 통해 기술 개발과 혁신을 촉진하고 경쟁 전략을 지원하며 법적 보호를 강화하는 도구로 기능한다.

Kang, Dylan Myungchul, et al.[3] 연구에서 기존의 키워드 매칭 기반의 검색 모델은 문헌의 세부적인 기술적 특성을 정확히 반영하지 못하고 언어적 차이를 효과적으로 처리하지 못한다는 한계를 지닌다고 주장하며, 이러한 한계를 극복하기 위해 최근 AI 기반의 자연어 처리(NLP) 기술이 특허 문헌 검색 및 분류에 활발히 도입되고 있다고 하였다. 이에 따라, 특허 문헌의 기술적 맥락을 깊이 이해할 수 있는 보다 정교한 AI 기반의 검색 모델이 요구된다.

최근 AI 자연어처리 기술이 발전하면서 검색 분야에서도 BERT(Bidirectional Encoder Representations from Transformers)[4]와 같은 딥러닝 기반 언어 모델이 활발히 도입되고 있다. 임준호, et al.[5] 연구에서는 BERT와 같은 언어 모델이 문헌의 복잡한 언어적 맥락을 이해하고, 기술적 구조를 깊이 이해할 수 있는 고도화된 모델 개발을 가능하게 했다고 주장하였다. 특히, 특허 분야에서는 대량의 특허 문헌 데이터를 사전 학습(Pre-training)한 KorPatBERT[6][7]와 KorPatELECTRA[8]를 발표한 바 있다. 이 모델들은 딥러닝을 통해 특허 문헌의 세부적인 기술적 맥락을 깊이 이해하도록 설계되어, 특허 도메인 지식에 최적화된 학습이 가능하다는 점에서 특허 검색 연구의 활용 가능성을 크게 확장시킬 잠재력을 보여준다. 이러한 배경에서 특허 검색 모델의 성능을 향상시키기 위한 새로운 접근법이 필요하며, 기존의 단순한 키워드 매칭 방식이나 일반적인 언어 모델을 넘어 특허 문헌의 기술적 주제와 의미를 효과적으로 반영할 수 있는 모델 연구가 요구된다.

본 연구에서는 단순한 텍스트 매칭이나 보편적인 언어모델의 활용을 넘어, KorPatBERT를 기반으로 CPC 분류 모델을 생성하여 특허 문헌에 내재된 기술적 주제와 의미를 효과적으로 반영하는 임베딩 벡터를 추출하고, 이를 활용하여 다양한 검색 모델별 성능을 비교·분석함으로써 특허 검색 모델의 성능을 향상시키는 새로운 접근법을 제안하고자 한다. 이를 위해 다음과 같은

---

1) 특허청, “2023 통계로 보는 특허동향”, 특허청, 2023, 19면.
2) Amna Ali et al., “Innovating patent retrieval: a comprehensive review of techniques, trends, and challenges in prior art searches”, *Applied System Innovation*, Vol.7 No.5 (2024), p. 91.
3) Dylan Myungchul Kang et al., “Patent prior art search using deep learning language model”, Proceedings of the 24th Symposium on International Database Engineering & Applications, 2020, pp. 1-5.
4) Jacob Devlin et al., “Bert: Pre-training of deep bidirectional transformers for language understanding”, arXiv preprint arXiv:1810.04805, <https://arxiv.org/abs/1810.04805>, 작성일: 2019. 5. 24.
5) 임준호 외 2인, “딥러닝 사전학습 언어모델 기술 동향”, 「전자통신동향분석」, 제35권 제3호(2020), 9-19면.
6) 박진우 외 4인, “한국어 특허 문장 기반 CPC 자동분류 연구-인공지능 언어모델 KorPatBERT 를 활용한 딥러닝 기법 접근”, 「지식재산연구」, 제17권 제3호 (2022), 209-256면.
7) 한국특허정보원, “kipi-ai/korpatbert”, 한국특허정보원 github, <https://github.com/kipi-ai/korpatbert>, 검색일: 2024. 11. 10.
8) 민재옥 외 3인, “Korean Patent ELECTRA: 한국 특허문헌 자연어처리 연구를 위한 사전 학습된 언어모델(KorPatELECTRA)”, 「한국컴퓨터정보학회 학술발표논문집」, 제29권 제2호(2021): 69-71면.

구체적인 연구 목표를 설정하였다.

첫째, 3.1장에서 특허 분야의 기관 및 기업에서 실질적으로 활용 가능하도록 한국어 특허 문헌 검색 연구를 위한 평가 대상 데이터셋과 검색 대상 데이터셋 구축 방법을 제안한다.

둘째, 3.2장에서 다양한 유형의 CPC 분류 모델을 생성하는 방법을 제시하고, 이를 활용하여 특허 문헌에 내재된 기술적 정보를 효과적으로 추출하는 방법을 제안한다.

셋째, 3.3장에서 효과적인 특허 검색 시스템을 구축하여 검색 모델에서 추출된 특허 문헌 임베딩 벡터의 검색 성능 개선 가능성을 검증한다.

본 연구에서는 현재 실제로 운용 중인 특허 검색 데이터와 검색 시스템의 프로세스를 제시하고 이를 토대로 실험함으로써, 민재옥, et al[9] 특허 검색 연구와 차별화된 기대 효과를 제공하고자 한다. 이러한 시도를 통해 실질적 유용성과 산업적 적용 가능성을 높이고, 나아가 국가와 기업이 지속적으로 혁신을 주도하며 지식재산권을 효과적으로 관리하고 활용할 수 있는 기반을 제공할 것으로 기대한다.

2. 관련 연구

2.1. 지식재산권의 이론적 배경

Chun, Youngsam, et al.[10] 연구에 따르면, 현대 사회에서 기술 발전은 바이오테크놀로지, 반도체, 특히, 인공지능(AI) 기술 분야를 중심으로 그 속도가 전례 없이 가속화되고 있으며, 이러한 첨단 분야의 기술 혁신은 국가와 기업의 경제적 경쟁력을 좌우하는 핵심 요소로 자리 잡고 있음을 주장하였고, Kanwar, Sunil, et al.[11] 연구에는 전 세계적으로 기술 패권을 둘러싼 경쟁이 날로 심화되고 있으며, 이러한 환경 속에서 기술적 우위를 선점하기 위한 각국의 지식재산권 확보 및 보호 전략은 중요한 정책 과제로 부상하고 있다고 주장하였다.

급변하는 글로벌 시장에서 기술의 독점적 권리를 확보하지 못할 경우, Edler, Jakob, et al.[12] 연구에서는 국가적으로 글로벌 기술 패권 경쟁에서 뒤처질 위험이 커지며, 국가 경제 성장의 핵심 동력을 상실하게 하고 기술적 자립과 산업 경쟁력 확보에 부정적인 영향을 미칠 수 있다고 주장하였다. 따라서, 지식재산권을 신속하고 정확하게 확보하는 것은 기업의 생존 문제를 넘어, 국가 차원의 기술 주권 확립과 글로벌 경쟁력 강화를 위한 핵심 조건이라 할 수 있다. 특허는 기술적 혁신의 법적 보호를 위해 발명자의 권리를 보장하고, 이를 기반으로 산업 및 경제발전을 촉진하는 지식재산권의 핵심적인 형태로 정의[13]된다. Narin, Francis[14] 연구에서 특허 문헌은 기술적 진보를 기록하는 동시에, 관련 분야 연구자와 산업계 종사자들에게 기술 동향, 경쟁사 활동, 신규 발명 및 시장 기회와 관련된 유의미하고 실질적인 정보를 제공하는 중요한

9) 민재옥 외 3인, “특허 언어모델 기반 CPC 클러스터링 필터와 토픽 벡터를 활용한 선행기술 특허검색 성능 향상 연구”, 한국정보과학회 학술발표논문집, 2022, 380-382면.

10) Youngsam Chun et al., “AI technology specialization and national competitiveness”, *Plos one,* Vol.19 No.4(2024), e0301091.

11) Sunil Kanwar & Robert Evenson, “Does intellectual property protection spur technological change?”, *Oxford Economic Papers,* Vol.55 No.2(2003), pp. 235-264.

12) Jakob Edler et al., “Technology sovereignty as an emerging frame for innovation policy. Defining rationales, ends and means”, *Research Policy*, Vol.52 No.6(2023), 104765.

13) Kenneth W. Dam, “The economic underpinnings of patent law”, *The Journal of Legal Studies*, Vol.23 No.1(1994), pp. 247-271.

14) Francis Narin, “Patents as indicators for the evaluation of industrial research output”, *Scientometrics,* Vol.34 No.3(1995), pp. 489-496.

자료로 기능하고 있다고 주장하였다.

2.2. AI 기반의 특허 검색 연구

Devlin, Jacob, et al.[15] 연구에서 발표한 구글(Google) BERT(Bidirectional Encoder Representations from Transformers)은 자연어처리 분야의 혁신적인 전환점을 가져왔다. BERT는 Transformer[16] 구조를 적용하여 양방향 컨텍스트(Bidirectional Context)를 동시에 학습할 수 있도록 설계되었다. 이를 통해 단어가 특정 문맥에서 사용되는 의미를 동적으로 반영할 수 있는 문맥적 임베딩(Contextualized Embedding) 기법이 가능해졌다. 박상언[17] 연구에서는 특정 분야의 대규모 데이터로 '사전학습(Pre-training)'된 모델이 해당 분야의 여러 NLP 태스크에서 탁월한 성능을 보인다는 점을 강조하며, 이에 대한 연구가 활발히 진행되고 있다고 주장하였다. 이러한 접근 방식의 결과로 SciBERT(과학기술)[18], BioBERT(의료)[19], LegalBERT(법률)[20] 등 도메인 특화된 언어모델이 지속적으로 개발되었다. 이와 같은 특화된 언어모델은 특정 산업이나 연구 분야와 밀접한 결합을 가능하게 하며, 본 연구에서도 특허 분야의 데이터를 사전 학습하였고 특허 분류 태스크에서 가능성을 확인[21]한 KorPatBERT를 특허 검색 연구에 활용하기로 하였다.

자연어처리 기술 발전에 따라 특허 문헌 검색에 대한 연구도 활발하게 진행되고 있다. Vaish, Kanishka, et al.[22] 연구에 따르면, 매년 수백만 건의 특허가 출원되면서 수동 검색은 비효율적이고 시간 소모가 크다고 주장하며, 인공지능 기술을 활용하면 검색 시간과 비용을 줄이고 정확도를 향상시킬 수 있다고 주장하였다. Kang, Dylan Myungchul, et al.[23] 연구에서는 BERT 모델을 활용하여 '요약'과 '청구 1항'을 학습하고, 관련성이 높은 특허 문서를 분류하여 노이즈 특허를 제거하는 방법을 제안하였다. Bekamiri, Hamid, et al.[24] 연구에서는 PatentBERT[25],

15) Jacob Devlin et al., "Bert: Pre-training of deep bidirectional transformers for language understanding", arXiv preprint arXiv:1810.04805, <https://arxiv.org/abs/1810.04805>, 작성일: 2019. 5. 24.

16) Ashish Vaswani et al., "Attention is all you need", 31st Conference on Neural Information Processing Systems (NIPS 2017), 2017, pp. 1-11.

17) 박상언, "딥러닝 기반 사전학습 언어모델에 대한 이해와 현황", 「한국빅데이터학회지」, 제7권 제2호(2022), 11-29면.

18) Iz Beltagy et al., "SciBERT: A pretrained language model for scientific text", arXiv preprint arXiv:1903.10676, <https://arxiv.org/abs/1903.10676>, 작성일: 2019. 9. 10.

19) Jinhyuk Lee et al., "BioBERT: a pre-trained biomedical language representation model for biomedical text mining", *Bioinformatics*, Vol.36 No.4(2020), pp. 1234-1240.

20) Ilias Chalkidis et al., "LEGAL-BERT: The muppets straight out of law school", arXiv preprint arXiv:2010.02559, <https://arxiv.org/abs/2010.02559>, 작성일: 2020. 10. 6.

21) 박진우 외 4인, "한국어 특허 문장 기반 CPC 자동분류 연구-인공지능 언어모델 KorPatBERT 를 활용한 딥러닝 기법 접근", 「지식재산연구」, 제17권 제3호(2022), 209-256면.

22) Kanishka Vaish et al., "Artificial Intelligence Reducing the Intricacies of Patent Prior Art Search", 2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), IEEE, 2023, pp. 978-982.

23) Dylan Myungchul Kang et al., "Patent prior art search using deep learning language model", Proceedings of the 24th Symposium on International Database Engineering & Applications, 2020, pp. 1-5.

24) Hamidet Bekamiri al., "Patentsberta: A deep nlp based hybrid model for patent distance and classification using augmented sbert", *Technological Forecasting and Social Change*, Vol.206(2024), 123536.

25) Jieh-Sheng Lee & Jieh Hsiang, "Patentbert: Patent classification with fine-tuning a pre-trained bert model", arXiv preprint arXiv:1906.02124, <https://arxiv.org/abs/1906.02124>, 작성일: 2019.

DeepPatent[26] 모델보다 높은 성능으로 CPC를 예측하고, 범주 안에 있는 특허를 AugsSBERT[27] 기반으로 특허 문헌의 텍스트 임베딩 간 유사성을 계산하는 방법을 제안하였다. 하지만, 특허 언어모델을 파인 튜닝을 하고, 이를 통해 생성된 특허 임베딩 벡터를 활용한 특허 검색에 대한 연구는 찾아보기 어려웠다.

최근에는 생성(Generation) 능력을 갖춘 대규모 언어모델(Large Language Model)이 각광받고 있다. OpenAI의 GPT[28] 시리즈 및 구글의 PaLM[29] 시리즈 등은 방대한 규모의 매개변수(Parameter)와 데이터셋을 바탕으로 자연스러운 문장 생성, 요약, 번역은 물론 코드 생성이나 창의적인 글쓰기 등 폭넓은 작업을 수행한다. 또한, 생성 AI를 활용한 ChatGPT[30] 및 생성 AI와 RAG(Retrieval-Augmented Generation)[31]를 결합한 검색 시스템은 방대한 정보를 바탕으로 다양한 언어적 표현과 사용자 친화적인 대화형 응답 방식을 제공하여 사용자의 높은 만족도를 제공하였다. 그러나, 생성 AI는 잘못된 정보를 생성하여 정답처럼 답변하는 환각 증상(Hallucination)[32] 및 막대한 경제적 비용 부담에 대한 문제점이 존재한다. 또한, 사용자 친화적인 응답 방식은 최종 표현 단계(Generator)에 해당하기 때문에 결과적으로 검색 단계(Retrieval)에서의 검색 모델 품질이 시스템 전체 성능을 좌우하는 핵심 요소라 할 수 있다.

2.3. 특허 문헌과 CPC

2.3.1. 특허 문헌

특허 문헌은 기술 산업의 발전과 혁신의 중요한 산물로, 발명의 기술적 내용을 기록하고 이를 명확히 전달하기 위해 체계적으로 작성되어 있다. 임소라, et al.[33] 연구에 따르면, 한국 특허 문헌은 문헌을 식별하기 위한 행정적 식별 정보와 발명의 기술적 내용을 설명하는 기술적 서술 부분으로 구성되며, 기술 내용을 전달하기 위해 여러 필드로 구성된 체계를 따르는 구조화된 문서라고 주장하였다. 이러한 구조적인 체계는 특허 정보를 명확히 전달하고 효율적인 관리와 검색을 가능하게 하는 데 필수적이다.

Suzgun, Mirac, et al.[34] 연구에 따르면, 특허 문헌의 본문에는 발명의 기술적 내용의 상세

---

7. 1.

26) Shaobo Li et al., “DeepPatent: patent classification with convolutional neural networks and word embedding”, *Scientometrics,* Vol.117 No.2(2018), pp. 721-744.

27) Nandan Thakur et al., “Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks”, arXiv preprint arXiv:2010.08240, <https://arxiv.org/abs/2010.08240>, 작성일: 2021. 4. 12.

28) Alec Radford et al., “Improving language understanding by generative pre-training”, OpenAI, <https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf>, 검색일: 2024. 10. 30.

29) Rohan Anil et al., “Palm 2 technical report”, arXiv preprint arXiv:2305.10403, <https://arxiv.org/abs/2305.10403>, 작성일: 2023. 9. 13.

30) Long Ouyang et al., “Training language models to follow instructions with human feedback”, Advances in neural information processing systems 35, 2022, pp. 27730-27744.

31) Patrick Lewis et al., “Retrieval-augmented generation for knowledge-intensive nlp tasks”, Advances in Neural Information Processing Systems 33, 2020, pp. 9459-9474.

32) Ziwei Ji et al., “Survey of hallucination in natural language generation”, *ACM Computing Surveys,* Vol.55 No.12(2023), pp. 1-38.

33) 임소라·권용진, “특허문서 필드의 기능적 특성을 활용한 IPC 다중 레이블 분류”, 「인터넷정보학회논문지」, 제18권 제1호(2017), 77-88면.

34) Mirac Suzgun et al., “The harvard uspto patent dataset: A large-scale, well-structured, and multi-purpose corpus of patent applications”, Advances in neural information processing

한 설명을 포함하는 발명의 설명부가 위치한다. 이 설명부에서는 발명의 명칭, 배경, 기술 분야, 과제의 해결 수단, 발명의 효과 등을 구체적으로 다룬다. 이는 발명이 기존 기술과 어떻게 차별화되는지, 그리고 어떤 문제를 해결하는지 명확하게 기술되어 있다. 발명의 독창성과 실용성을 뒷받침하는 중요한 역할을 하며, 이를 통해 특허 심사관이 발명의 신규성과 진보성을 평가하는 기준을 제공한다.

또한, 특허 문헌에는 발명 보호 범위를 구체적으로 정의하는 청구항이 포함된다. IIncarbone, Stefano.[35] 연구에서 청구항은 발명의 법적 보호 범위를 명확히 하기 위해 사용되는 중요한 부분으로, 유사 기술에 대한 방어 전략을 수립하는 데 필수적이라고 주장하였다. 발명의 설명부와 청구항의 핵심 내용을 활용하면 유사 발명과의 비교 분석이 용이해지며, 신규성 및 진보성 검토를 체계적으로 수행할 수 있다.

특허 문헌 검색의 정확성과 효율성을 높이기 위해서는 특허 문헌의 구조적 특성을 이해하고 적절한 필드 조합을 선택하여 검색을 진행할 수 있다.

2.3.2. CPC

특허 문헌의 식별 정보에는 출원인, 대리인, 발명자 등 인적 정보와 고유한 출원번호가 포함되며, 기술 범위를 나타내는 CPC도 포함된다. CPC(Cooperative Patent Classification)[36]는 미국 특허청(USPTO)과 유럽 특허청(EPO)이 공동으로 개발한 특허 분류 체계로, 전 세계적으로 특허 출원이 급증하고 기술 발전 속도가 가속화되는 상황에서 보다 정밀하고 체계적인 특허 관리를 위해 도입되었다. CPC는 특허 문헌의 기술적 범주를 세분화하여 체계적으로 정리함으로써, 기술 연관성을 명확히 파악하고 효율적인 검색과 관리가 가능하도록 지원한다. IP5 Statistics Report 2022[37] 연구에 따르면, 현재 CPC는 특허 출원의 약 80% 이상을 차지하는 선진 5개 특허청을 중심으로 점차 확대되고 있으며, 국제적인 표준으로 자리 잡고 있다고 보고하였다.

<표1 CPC 구조 예시>

전체	섹션	서브 클래스	메인 그룹	서브 그룹
H01L 21/28	H	H01L	H01L 21	H01L 21/28

CPC는 A부터 H까지의 8개 섹션과 Y섹션으로 구성되어 기술 분야별로 특허를 체계적으로 분류한다. <표1>과 같이 세부적으로는 클래스(Class), 서브 클래스(Sub Class), 메인 그룹(Main Group), 서브 그룹(Sub Group)으로 단계적으로 나뉘어 특허 문헌의 기술적 특성대로 구분할 수 있다[38]. 이는 효율적인 특허 정보 관리에 도움이 되며, 연관된 기술 특징을 그룹화

systems 36, 2023, pp. 57908-57946.

35) Stefano IIncarbone, “Claim construction: an international convergence in striking the balance between patent protection and legal certainty”, *Journal of Intellectual Property Law and Practice,* Vol.17 No.10(2022), pp. 878-888.

36) 특허청, “CPC 및 IPC 분류코드”, 특허청, <https://www.kipo.go.kr/ko/kpoContentView.do?menuCd=SCD0200269>, 검색일: 2024. 10. 30.

37) Fiveipoffices, “IP5 Statistics Report 2022”, Fiveipoffices, <https://www.fiveipoffices.org/statistics/statisticsreports/2022edition>, 검색일: 2024. 10. 30.

38) 한국특허기술진흥원, “특허분류 조회 서비스”, 한국특허기술진흥원, <https://www.pipc.or.kr/business/cpcService>, 검색일: 2024. 10. 30.

(Clustering)하여 다양한 기술적 연관성을 파악해 볼 수 있다.

민재옥, et al[39] 연구에서는 CPC의 중요성을 인식하고, CPC 분류 모델을 활용하여 특허 문헌의 임베딩 벡터를 추출한 뒤, 이를 검색 실험에 적용하여 그 효과성을 확인하였다. 본 연구에는 CPC 분류 모델의 유효성을 검증하는 동시에 보다 다양한 분류 모델을 활용하여 특허 문헌 검색 실험을 확장하여 진행하였다.

3. 연구 방법

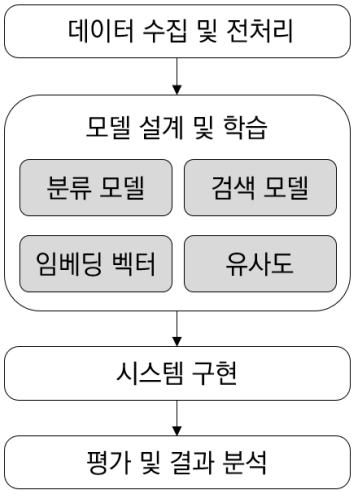
본 연구의 절차는 <그림1>에 제시된 흐름도에 따라 단계별로 진행하였다. 먼저, 3.1장 데이터 수집 및 전처리 단계에서는 검색 연구에 필요한 데이터를 확보하기 위해 키프리스 플러스(KIPRISPlus)[40]를 활용하였다. 수집된 데이터는 연구 목적에 맞게 가공 및 전처리하여 검색 연구에 적합한 형태로 변환하였다.

다음으로, 3.2장 모델 설계 및 학습 단계에서는 특허 검색 모델을 생성하기 위한 핵심 요소로 CPC 분류 체계별로 분류 모델을 생성하였고, 특허 임베딩 벡터를 생성하였다. 또한, 특허 검색 유사도 방법을 설계하였고, 검색 시스템의 기반을 마련하였다.

이어지는 3.3장 시스템 구현 단계에서는 실질적인 검색 시스템을 설계하고 구축하였다.

마지막으로, 4장 평가 및 결과 분석 단계에서는 성능 지표를 통해 모델별 실험 결과를 종합적으로 분석하였고, 이어서 5장에서 연구 목표 달성 여부를 검증하였다. 아울러, 본 연구의 한계를 논의하고 향후 연구 방향을 제안하였다.

<그림1 연구 절차>



3.1. 데이터 수집 및 전처리

본 연구는 한국어 특허 문헌을 대상으로 특허 문헌 검색 실험을 수행하는 것이며, 효과적인

39) 민재옥 외 3인, “특허 언어모델 기반 CPC 클러스터링 필터와 토픽 벡터를 활용한 선행기술 특허검색 성능 향상 연구”, 한국정보과학회 학술발표논문집, 2022, 380-382면.

40) 한국특허정보원, “특허정보 활용 서비스”, 한국특허정보원, <https://plus.kipris.or.kr/portal/main.do>, 검색일: 2024. 10. 30.

특허 검색 모델의 실험을 위해서는 신뢰성 높은 데이터셋의 준비가 필수적이다. 검색 모델의 성능을 정확히 평가하려면 검색 대상 데이터셋과 이를 기반으로 제안된 모델의 성능을 검증할 수 있는 평가 대상 데이터셋이 체계적으로 구축되어야 한다.

이를 위해 본 연구에서는 실제 사용되는 전체 특허 데이터를 기반으로 데이터셋을 구축하여 활용하였다. 실험에 사용된 데이터셋 유형은 <표2>에 상세히 설명하였으며, 이후 본 논문에서는 데이터셋의 영문 명칭을 사용하여 설명을 이어가도록 하겠다.

<표2 특허 검색 데이터셋 유형>

구분	설명	국문 명칭	영문 명칭
평가 대상	특허 심사관이 특허 심사에서 최종적으로 등록을 거절한 특허 문헌	출원 특허	Query
	특허 심사관이 특허 심사에서 특허의 등록을 거절하는 근거로 인용한 특허 문헌	선행 기술 특허	Answer
검색 대상	검색 모델에 입력으로 들어오는 특허 문헌으로, 검색 실험에서는 Query와 동일한 데이터를 의미	입력 데이터	Input source
	검색 모델이 Input Source와 비교하는 특허 문헌 집합	특허 문헌 집합	Documents

3.1.1. 평가 대상 데이터셋

평가 대상 데이터셋은 의견제출통지서를 기반으로 구축하게 된다. 의견제출통지서는 특허 심사관이 출원한 특허에 대해서 인용한 특허와 등록을 거절한 이유를 상세히 기록하여 통보한 문헌으로, 키프리스 플러스(KIPRISPlus)[41]에서 추출하였다. 2010년부터 2023년까지 최근 10개년도를 기준으로 행정처분이 최종적으로 종료된 데이터로 선별한 결과 103,976건의 데이터셋을 수집하였다. 이 데이터셋은 모델 개발의 방향성을 설정하고 성능 개선의 기준을 제공하는 동시에, 모델 성능을 검증하는 중요한 척도로 활용될 수 있을 것으로 기대된다.

하나의 출원 특허가 특허 심사를 받을 때 해당 분야뿐만 아니라 해당 기술이 응용될 수 있는 다양한 분야에서도 기존 특허를 인용할 수 있다. 즉, 출원 특허가 여러 기술적 요소로 구성된 경우, 각 요소가 이미 등록된 특허 기술들과 비교되어 하나의 출원 특허에 여러 개의 인용 특허가 생길 수 있다. 또한, 기존의 출원한 특허를 기준으로 심사하기 때문에 출원 특허의 출원일 이전에 출원된 특허만 인용 대상으로 포함된다. 수집한 데이터셋에서 출원 특허에 대한 인용 특허의 현황은 <표3>에 제시하였다.

<표3 평가 대상 데이터셋의 Query에 대한 인용 특허 수 현황>

인용 특허 수	문헌 수(건)	비율(%)
1건	32,272	31.04
2건	38,886	37.40
3건	19,659	18.91
4건 ~ 10건	13,037	12.54
10건 초과	122	0.12
합계	103,976	100.00

41) 한국특허정보원, “특허정보 활용 서비스”, 한국특허정보원, <https://plus.kipris.or.kr/portal/main.do>, 검색일: 2024. 10. 30.

인용 특허 수가 1건에서 2건인 경우가 전체 103,976건 중 68.44%로 대부분을 차지하는 것으로 확인하였다.

평가 대상 데이터셋에서 하나의 출원 특허에 여러 인용 특허가 모두 Answer로 사용될 수 있지만, 의견제출통지서를 기준으로 첫 번째로 기록된 인용 특허는 주(Main) 선행 기술 특허로 지칭하며, 출원 특허의 핵심 기술과 가장 유사하다고 판단된 것이다. 따라서 유일한 Query로 구성하고 다른 Query들과의 비교에서 일관성을 유지하기 위해 첫 번째 인용 특허를 Answer로 하여 하나의 Query와 Answer로 이루어진 1대 1의 문헌 쌍(Pair) 데이터로 가공하였다.

3.1.2. 검색 대상 데이터셋

검색 대상 데이터셋은 키프리스 플러스(KIPRISPlus)[42]에서 제공하는 한국 특허 공보 데이터를 수집하여 사용하였다. 1946년부터 2023년 5월까지의 아카이빙(Archiving)된 특허 문헌으로 구성되어 있으며, 수집된 한국 특허 문헌은 세계 특허정보 표준인 ST.96 기반의 XML 형식으로 제공된다. 이를 효율적으로 전처리하기 위해 박진우, et al.[43] 연구에서 정의한 특허 필드 구조를 참고하였다. 이를 바탕으로 XML 태그와 문자열 정제 규칙을 적용하여 특허 문헌의 식별 정보와 특허 텍스트 데이터를 체계적으로 구분 및 추출하였다.

추출된 데이터는 효율적인 관리와 검색 실험에 활용되도록 데이터베이스에 저장하였다. 최종적으로 수집된 검색 대상 데이터셋은 5,145,909건이며, 이 데이터셋을 CPC 섹션별로 분류한 결과는 <표4>에 제시하였다.

<표4 검색 대상 데이터셋 현황(1946년 ~ 2023년 5월)>

CPC 섹션	특허 문헌 수(건)	비율(%)	합계(건)	데이터 크기
A (생활 필수품)	758,816	14.75	5,145,909	197GB
B (처리조작, 수송)	869,352	16.89		
C (화학, 야금)	533,144	10.36		
D (섬유, 종이)	84,373	1.64		
E (고정 구조물)	238,020	4.63		
F (기계공학, 조명, 가열, 무기)	446,720	8.68		
G (물리)	1,000,774	19.45		
H (전기)	1,214,043	23.59		
Y (새로운 기술)	667	0.01		

3.1.3. 최종 실험 데이터셋

Liu, Kuo-tsan, et al.[44] 연구에 따르면, H섹션에 해당하는 반도체 관련 기술 특허는 기업의

42) 한국특허정보원, “특허정보 활용 서비스”, 한국특허정보원, <https://plus.kipris.or.kr/portal/main.do>, 검색일: 2024. 10. 30.

43) 박진우 외 4인, “한국어 특허 문장 기반 CPC 자동분류 연구-인공지능 언어모델 KorPatBERT 를 활용한 딥러닝 기법 접근”, 「지식재산연구」, 제17권 제3호(2022), 209-256면.

44) Kuo-tsan Liu & Chia-Ho Chen, “Formulation of research and development strategy by analysing patent portfolios of key players the semiconductor industry according to patent strength and technical function”, *World Patent Information*, Vol.70(2022), 102125.

경쟁력을 좌우하는 중요한 요소로 평가되며, 많은 기업이 독창적인 기술을 보호하고 시장에서 우위를 차지하기 위해 지속적으로 특허 출원에 집중하고 있다고 주장하였다. <표4>에서 검색 대상 데이터셋 5,145,909건 중 H섹션은 전기·전자공학, 통신, 컴퓨터, 반도체 기술 분야 등의 분야로 다른 섹션 보다 가장 많은 23.59%의 특허 문헌을 가지고 있다. 이에 따라 H섹션이 거의 모든 산업에서 응용될 수 있는 핵심 기술 분야인 중요성을 고려하여 본 연구는 H섹션 1,214,043건으로 한정하였다.

마찬가지로, 평가 대상 데이터셋 103,976건(<표3>) 중 H섹션 및 첫 번째 인용 특허만 선별하여, Query와 Answer가 1대 1로 매칭되는 15,669건의 쌍을 구성하였다. 하나의 Input source에 H섹션에 속하는 1,214,043건 대상으로 검색 실험을 진행하였다.

최종적으로 본 연구에서 사용하는 실험 데이터셋(평가 대상 데이터셋, 검색 대상 데이터셋)은 <표5>에 제시하였다.

<표5 최종 실험 데이터셋 현황>

구분	명칭	특허 문헌 수(건)
평가 대상 데이터셋	Query	15,669
	Answer	15,669
검색 대상 데이터셋	Input source	15,669
	Documents	1,214,043

본 연구에서는 특허 심사관의 특허 심사 결과를 바탕으로 인용 관계를 반영한 평가 대상 데이터를 구축하였고, 특허 산업 분야의 기업·기관에서 활용되는 전체 문헌을 대상으로 구성하여 특허 문헌 검색 모델의 성능을 객관적으로 평가하고 개선하는 데 필요한 기초 데이터셋을 마련하였다. 또한, 핵심 기술 분야인 H섹션을 다룸으로써 검색 모델의 실질적 유용성과 산업적 적용 가능성을 높이고자 하였다.

3.2. 모델 설계 및 학습

3.2.1. CPC 분류 모델

Kang, Dylan Myungchul, et al.[45] 연구에 따르면, 기존 키워드 매칭 기반의 검색 방식에 비해 AI 기반 임베딩 모델을 사용했을 때 검색 성능이 향상된다고 주장하였다. 본 연구에서는 CPC 분류 모델을 활용하여 검색 모델을 개선하고자 한다.

CPC 분류 모델은 검색 모델의 기반이 되며 기존 키워드 매칭 기반 검색 방식의 한계를 극복하고 AI 기반 임베딩 모델을 활용하여 검색 성능을 향상시키는 데 중점을 두고 있다.

3.2.1.1. 데이터셋 구축

CPC 분류 모델을 생성하기 위해 Documents를 가공하여 학습 데이터셋을 구축하였다. 심우철, et al.[46] 연구에서는 특허 필드 조합에 따른 분류 성능 비교 결과, '발명의 명칭', '요약', '배

45) Dylan Myungchul Kang et al., "Patent prior art search using deep learning language model", Proceedings of the 24th Symposium on International Database Engineering & Applications, 2020, pp. 1-5.

46) 심우철 외 4인, "한국 특허문헌 특성 및 딥러닝 기반 분류모델을 고려한 CPC 자동분류에 관한 연구", 한

경기술', '기술분야', '청구 1항' 필드를 사용할 때 상대적으로 우수한 성능을 보인다고 주장하였고, 박진우, et al.[47] 연구에서 서브 클래스 분류는 '발명의 명칭', '요약', '배경기술', '기술분야' 필드를 사용하고, 메인 그룹 분류는 서브 클래스 대비 더욱 세분화 되고 좁은 범위의 특성을 고려해서 '청구 1항'을 추가하여 5개의 필드로 학습 데이터셋을 구성했을 때 효과적임을 주장하였다. 본 연구에서도 해당 필드의 텍스트를 사용하여 학습 데이터셋을 구축하였으며, 서브 클래스 분류 모델과 메인 그룹 분류 모델에 더하여 추가적으로 서브 그룹 분류 모델을 생성하기로 하였다. 이를 통해 다양한 검색 모델을 실험하고 그 성능을 비교 분석하고자 하였다.

서브 클래스 분류의 Context는 각 특허 문헌의 '발명의 명칭', '요약', '배경기술', '기술분야'의 4개 특허 필드를 기반으로 구성하였다. 이 중 텍스트의 길이 짧은 '발명의 명칭'과 '요약' 필드는 결합하여 단일 필드로 처리하여 하나의 특허 문헌당 총 3개의 행(Row)으로 구성하였다. 한편, 메인 그룹 분류와 서브 그룹 분류의 Context를 구성할 때에는 '청구 1항'만 사용하는 것보다 '전체 청구항'을 결합하여 단일 필드로 처리한 경우 성능이 향상됨을 사전에 확인하였다. 따라서, 최종적으로 '전체 청구항'으로 사용하였다. 위와 동일한 방식으로 총 5개 특허 필드를 사용하여 4개의 행(Row)으로 구성하였다. 학습 데이터셋의 유형별로 사용한 특허 필드는 <표6>에 제시하였다.

<표6 학습 데이터셋 유형별 사용한 특허 필드>

분류 모델	특허 필드
서브 클래스	발명의 명칭·요약, 배경기술, 기술분야
메인 그룹	발명의 명칭·요약, 배경기술, 기술분야, 전체 청구항
서브 그룹	발명의 명칭·요약, 배경기술, 기술분야, 전체 청구항

CPC 분류 모델의 학습 데이터셋은 최신 데이터를 대상으로 2024년부터 순차적으로 사용하였으며, 특허 문헌의 텍스트를 Context로 사용하고, 해당 문헌의 CPC를 정답 클래스(Class)로 구성하였다. 클래스 수의 선택은 분류 모델의 유형을 결정하는 데 있어 중요한 요소로 작용한다. 클래스 수가 많아질수록 세밀한 군집화가 가능하다는 장점이 있으나, 분류 모델의 정확도가 상대적으로 낮아질 수 있으며, 그 결과로 특허 검색에서 일부 문헌이 검색 시작 단계부터 제외될 위험이 존재한다. 반대로, 클래스 수가 적을 경우 분류 모델의 정확도는 높아질 수 있으나 검색 대상이 지나치게 넓어져 노이즈 데이터가 다수 포함될 가능성이 커진다. 이러한 이유로, 분류 모델의 정확도와 클래스 수를 균형 있게 고려하여 모델을 선택하는 것이 중요하다. 또한, Akbani, Rehan, et al.[48], Oreski, Goran, et al.[49] 연구에 따르면, 클래스 간 데이터가 불균형할 경우, 모델은 빈도수가 높은 클래스에 치우쳐 학습될 가능성이 높아지고, 그 결과 희소 클래스에 대한 예측 성능이 저하될 수 있다고 주장하였다. 따라서, 학습 데이터셋에서 클래스 간 데이터 균형을 유지하는 것은 학습 최적화를 위해 필수적이며, 이는 모델이 특정 클래스에 편중

국정보과학회 학술발표논문집, 2020, 406-408면.

47) 박진우 외 4인, "한국어 특허 문장 기반 CPC 자동분류 연구-인공지능 언어모델 KorPatBERT를 활용한 딥러닝 기법 접근", 「지식재산연구」, 제17권 제3호(2022), 209-256면.

48) Rehan Akbani et al., "Applying support vector machines to imbalanced datasets", Machine Learning: ECML 2004: 15th European Conference on Machine Learning, Pisa, Italy, September 20-24, 2004. Proceedings 15, Springer Berlin Heidelberg, 2004, pp. 39-50.

49) Goran Oreski & Stjepan Oreski, "An experimental comparison of classification algorithm performances for highly imbalanced datasets", Central European Conference on Information and Intelligent Systems. Faculty of Organization and Informatics Varazdin, 2014, p. 4.

되지 않고 모든 클래스에서 균형 잡힌 성능을 발휘할 수 있도록 하기 위함이다.

본 연구에서는 충분한 규모의 균형 잡힌 데이터를 확보함으로써 모델은 다양한 기술 분야에서 안정적이고 일반화된 성능을 발휘할 수 있도록 하였다. 이를 위해 클래스별 데이터 분포를 면밀히 분석하고 이를 반영하여 균형 잡힌 클래스를 가진 데이터셋으로 구축하였다.

서브 클래스에서 서브 그룹으로 갈수록 더욱 세분화되며, 서브 그룹은 서브 클래스에 비해 약 100배 이상의 클래스 수를 가진다. 분류할 수 있는 클래스의 수가 많을수록 클래스 간의 세분화된 미세한 차이점을 효과적으로 구분할 수 있다. 이러한 특성을 고려하여 각 세부 분류에 대한 충분한 데이터를 확보하기 위해 메인 그룹 분류와 서브 그룹 분류의 학습 데이터셋을 기존 대비 2배 이상 확장하여 구축하였다. 최종 구축한 학습 데이터셋은 <표7>과 같다.

<표7 학습 데이터셋 현황>

분류 모델	클래스(분류) 수	데이터셋 수(문헌 수)	
		학습용	평가용
서브 클래스	632	2,052,584	123,388
메인 그룹	5,503	4,832,805	255,545
서브 그룹	64,000		

3.2.1.2. 모델 학습

구축한 학습 데이터셋을 대상으로 KorPatBERT 기반의 파인 튜닝을 진행하였다. 모든 모델 학습은 NVIDIA Tesla A100 80GB GPU 8개가 장착된 Linux 기반 GPU 서버에서 동일하게 수행하였다.

CPC 분류 모델 학습은 Devlin, Jacob, et al.[50]연구에서 공개한 BERT 아키텍처를 기반으로 파인 튜닝(Fine-tuning)[51]하는 소스를 참고하여 진행하였고, 학습 파라미터는 Mosbach, Marius, et al.[52] 연구에서 제안한 BERT 파인 튜닝 최적화 전략을 바탕으로 설정하였다. 학습 파라미터에서 옵티마이저(Optimizer)[53]는 AdamW[54]를 사용하여 손실(Loss)을 최소화하면서 과적합(Overfitting)의 위험을 효과적으로 줄였고, 활성화 함수(Activation function)로는 Softmax(<수식 1>)를 적용하였다.

50) Jacob Devlin et al., “Bert: Pre-training of deep bidirectional transformers for language understanding”, arXiv preprint arXiv:1810.04805, <https://arxiv.org/abs/1810.04805>, 작성일: 2019. 5. 24.

51) Google, “google-research/bert”, Google github, <https://github.com/google-research/bert>, 검색일: 2024. 11. 12.

52) Marius Mosbach et al., “On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines”, arXiv preprint arXiv:2006.04884, <https://arxiv.org/abs/2006.04884>, 작성일: 2021. 3. 25.

53) 딥러닝에서 모델이 손실(loss)을 최소화하도록 파라미터를 조정해주는 알고리즘.

54) Ilya Loshchilov & Frank Hutter, “Decoupled Weight Decay Regularization”, arXiv preprint arXiv:1711.05101, <https://arxiv.org/abs/1711.05101>, 작성일: 2019. 1. 4.

<수식1 Softmax Activation Function>

$$P(y_i) = \frac{e^{zj}}{\sum_{j=1}^{K} e^{zj}}$$

Softmax는 출력 값을 0에서 1 사이의 값으로 변환하며, 모든 클래스 출력 값들의 합이 1이 되도록 정규화하는 특성을 지닌다. 이 과정에서 각 클래스의 출력 값이 다른 클래스의 출력 값에 의존하여 계산되는 것으로, 이러한 클래스 간 상호의존적 특성은 다중 클래스 분류(Multi-Class Classification) 문제를 해결하는 데 적합한 함수로 널리 사용된다.

반복 학습 시 가중치 조정의 크기를 결정하는 Learning rate는 3×10^{-5}로 설정하였으며, 해당 값이 학습할 때 가장 안정적으로 수렴함을 사전에 확인하였다. 또한, 모델이 한 번에 학습하는 데이터의 크기를 결정하는 Batch size는 가용 자원에서의 최대 값인 400으로 설정하였다.

분류 학습을 진행하기에 앞서, 학습 데이터셋의 Context를 모델이 이해할 수 있는 형태로 변환하기 위해 인코딩의 과정이 필요하다. 이 과정에서 모델의 허용 가능한 최대 토큰 수(Max sequence length)는 256으로 설정하였고, 특허 텍스트를 효과적으로 토큰화하기 위해 특허 텍스트에 최적화된 사전(Vocabulary)과 토크나이저(Tokenizer)를 사용하였다.

3.2.1.3. 평가

분류 모델의 평가는 Fall, Caspar J., et al.[55] 연구에서 제안하는 CPC 분류 평가지표를 기반으로 평가 데이터셋을 평가하였다. 모델은 해당 지표에서 우수한 점수를 달성할 때까지 반복 학습(Epochs)을 진행하였으며, 이 과정에서 학습 손실(Loss)이 지속적으로 0으로 수렴하여 더 이상 유의미하게 낮아지지 않는 수준에서 학습을 종료하였다. 최종적으로 3종의 CPC 분류 학습 모델을 생성하였고, 학습한 분류 모델의 현황 및 평가 결과를 <표8>에 제시하였다.

<표8 CPC 분류 모델 생성 현황 및 평가 점수>

분류 모델명*	CPC 분류 기준	클래스 수	평가 점수(%)		
			Top Prediction	Three Guesses	Five Guesses
CLS-SC	서브 클래스	632	78.63	93.85	96.95
CLS-MG	메인 그룹	5,503	72.36	90.44	94.35
CLS-SG	서브 그룹	64,000	39.63	61.33	70.24

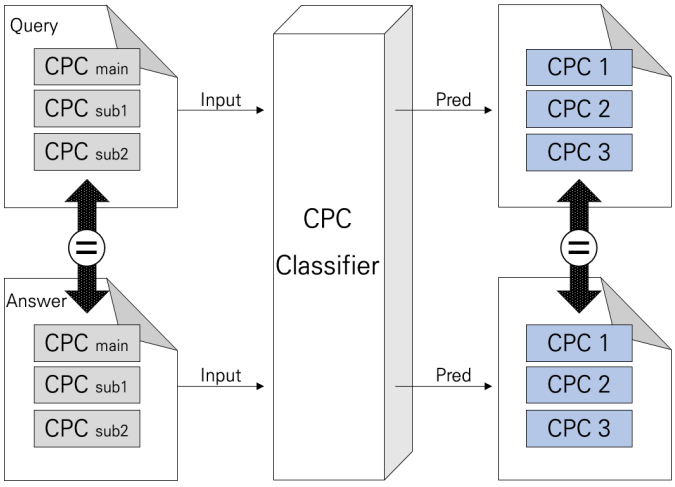
* 예시) CLS(Classification), SC(Sub Class), 256(최대 토큰 길이)

생성한 CPC 분류 모델의 실효성을 검증하기 위해 대표적으로 CLS-SC 모델을 기준으로 평가 대상 데이터셋에 대한 CPC 서브 클래스를 비교 실험하였다. 이때 평가 대상 데이터셋은 검색 실험에서 사용된 H섹션의 15,669건이 아닌, 전체 섹션에 해당하는 103,976건으로 설정하여 비교 실험을 진행하였다. 비교 실험에서는 최대 3개의 CPC를 대상으로, Query의 CPC와 Answer의 CPC 중 하나라도 일치하는 경우 이를 일치로 간주하여 평가하였다.

55) Caspar J. Fall et al., "Automated categorization in the international patent classification", *Acm Sigir Forum,* Vol.37 No.1(2003), pp. 10-25.

<그림2>에서 제시된 진행 방법에 따라, 첫 번째는 Query와 Answer에 기재된 문헌의 CPC 간 일치 여부를 비교하였고, 두 번째로는 CLS-SC를 사용하여 Query와 Answer의 CPC를 각각 Top3(상위 3개)까지 예측하고 그 일치 여부를 비교하였다.

<그림2 Query와 Answer의 CPC 비교 실험>



실험 결과, 특허 문헌에 기재된 CPC를 비교한 결과는 94,067건이 일치하여 90.47% 일치율을 보였으며, 분류 모델에서 예측한 CPC를 비교한 결과는 100,041건이 일치하여 96.22%의 일치율을 보였다. 두 실험에서 모두 90% 이상의 일치율을 기록하였으나, <표8>에서 CLS-SC 모델의 Three guesses 93.85%를 기준으로 비교해 보았을 때, 특허 문헌에 기재된 CPC를 비교한 실험은 더 낮은 일치율을 보였고, 반면, 분류 모델에서 예측한 CPC는 2.37%p 더 높은 일치율을 보였다.

실험 결과를 통해 Query와 Answer의 CPC가 매우 높은 상관성을 가지고 있음을 확인 할 수 있었다. 또한, 분류 모델이 특허 문헌의 기술적 내용을 일관성 있게 학습하고 효과적으로 내재화했음을 입증하였다. 이는 분류 모델이 특허 문헌에 기재된 CPC 보다 더욱 효과적으로 CPC를 예측할 수 있음을 시사하며, 제안된 분류 모델의 실효성을 뒷받침하였다.

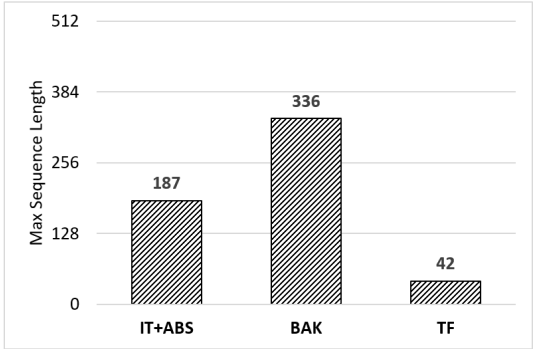
3.2.2. 검색 모델

검색 모델은 앞서 생성한 분류 모델을 기반으로 Input source가 입력되고 최종적으로 특허 문헌 임베딩 벡터로 생성되는 과정을 제시한다. Input source의 토큰 길이가 설정한 최대 토큰 길이보다 짧은 경우 남은 공간은 패딩(Padding)으로 채워진다. 이러한 패딩은 불필요한 계산 비용을 증가시키며 임베딩 벡터의 품질 향상에는 기여하지 않는다. 반대로, 최대 토큰 길이를 작게 설정하여 Input source의 토큰 길이가 이를 초과할 경우 초과된 토큰은 잘라내어 처리하므로 문맥 정보가 손실될 위험이 크다. 따라서, 검색 모델에서 적절한 토큰 길이를 설정하기 위해 평가 대상 데이터셋의 특허 필드별 토큰 길이를 분석하였다. 검색 대상 데이터셋 중에서 검색 실험에서 사용된 H섹션의 1,214,043건에 한정하지 않고, 전체 섹션에 해당하는 5,145,909건을 대상으로 분석을 진행하였다. 분석은 '발명의 명칭+요약', '배경 기술', '기술 분야'의 3가지 항목을 대상으로 진행하였다. <그림3>은 각 항목별 평균 토큰 길이를 보여주며, <그림4>는 256 토큰 길이 기준으로 Input source 토큰의 포함 여부를 비율로 표현하였다. <그림5>는 512

토큰 길이 기준으로 Input source 토큰의 포함 여부를 비율로 나타낸 결과를 제시하였다.

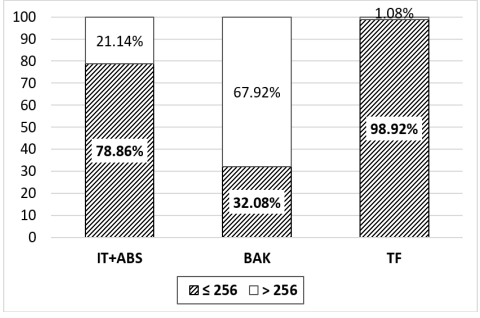
데이터셋의 청구항 항목 중 대표 청구항은 512개 토큰 내에 포함되었다. 그러나 불필요한 패딩을 최소화하고 청구항의 문맥 정보를 최대한 반영하기 위해 남은 토큰 공간을 나머지 청구항으로 결합하였다. 따라서 청구항 항목은 토큰 길이 통계에서 제외하였다.

<그림3 특허 필드별 평균 토큰 길이>

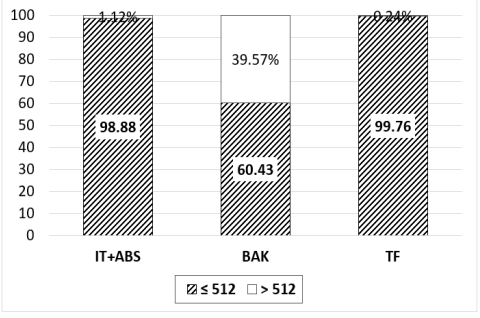


* IT(Invention Title): 발명의 명칭, ABS(Abstract): 요약, TF(Technical Field): 기술분야, BAK(Background): 배경기술

<그림4 256 토큰 기준 특허 문헌 범위(%)>



<그림5 512 토큰 기준 특허 문헌 범위(%)>



*IT(Invention Title): 발명의 명칭, ABS (Abstract): 요약, TF (Technical Field): 기술분야, BAK (Background): 배경기술

<그림3> 특허 필드별 평균 토큰 길이 현황에서는 '배경기술' 필드를 제외한 모든 필드에서 256 토큰 이하로 나타났다. 그러나, '배경기술' 필드는 특허 문헌 건수를 기준으로 분석한 결과, <그림4> 256 토큰 기준에서는 67.92%가 초과되어 256 토큰 길이가 충분하지 않는 것을 확인하였고, <그림5> 최대 토큰 길이인 512 토큰 기준으로 분석했을 때 39.57% 이상 되는 것으로 나타나 여전히 문맥 정보 손실 가능성을 내포하고 있음을 보여주었다.

이에 따라, 본 연구에서는 <표8>의 모델을 기반으로 분류 모델별로 256 토큰 길이와 최대 허용 가능한 512 토큰 길이를 사용하는 두 가지 형태를 가진 6종의 검색 모델을 정의하여 비교 실험하기로 하였다. 최종적으로 본 연구에서 검색 실험할 모델의 유형은 <표9>에 제시하였다.

<표9 검색 모델 유형>

검색 모델명*	CPC 분류 기준	클래스 수	최대 토큰 길이
IR-SC256	서브 클래스	632	256
IR-SC512			512
IR-MG256	메인 그룹	5,503	256
IR-MG512			512
IR-SG256	서브 그룹	64,000	256
IR-SG512			512

* 예시) IR(Information Retrieval), SC(Sub Class), 256(최대 토큰 길이)

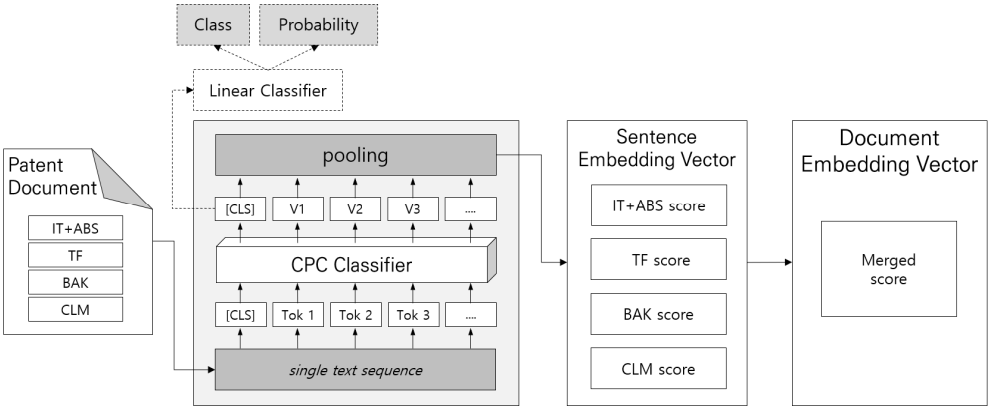
3.2.3. 특허 임베딩 벡터

분류 모델에서는 클래스(Class)와 클래스를 예측한 확률 값을 추출할 수 있을 뿐만 아니라, 입력 데이터의 문맥 정보와 의미를 포함한 고차원의 임베딩 벡터를 추출할 수 있다. Vargas, Max, et al.[56] 연구에서는 AI 모델에서 추출한 임베딩 벡터는 학습 데이터셋에 대한 내재된 의미를 효과적으로 포착하며, 언어적 차이를 정교하게 반영하는 수치화된 정보를 제공한다고 주장하였다. CPC 분류 모델에서 추출한 임베딩 벡터는 입력된 특허 문장의 내재된 주제 및 기술적인 특징을 반영하는 토픽(Topic) 정보[57]를 효과적으로 표현한다. 이러한 특성은 CPC 분류 모델의 임베딩 벡터가 검색 모델 성능에 직접적인 영향을 미칠 수 있음을 시사한다. 이를 통해 임베딩 벡터 간의 유사성을 정밀하게 평가할 수 있는 기반을 마련하였다.

모델에서 임베딩 벡터를 추출할 때, 설정한 최대 토큰 길이까지 인코딩이 가능하다. 그러나 하나의 특허 문헌인 Input source는 여러 개의 특허 필드로 구성되어 있어 최대 토큰 길이를 초과할 수 있다. 이를 해결하기 위해, 각 특허 필드별로 임베딩 벡터를 개별적으로 추출한 뒤, 이를 통합하여 최종적으로 Input source 문헌 임베딩 벡터를 생성하기로 하였다.

특허 문헌의 문맥 정보를 내재화한 문헌 임베딩 벡터 생성 과정은 <그림6>과 같다.

<그림6 CPC 분류 모델에서 클래스 추출 및 특허 문헌 임베딩 벡터 생성 과정>



* IT(Invention Title): 발명의 명칭, ABS(Abstract): 요약, TF(Technical Field): 기술분야, BAK(Background): 배경기술, CLM(Claim): 청구항

56) Max Vargas et al., "Understanding Generative AI Content with Embedding Models", arXiv preprint arXiv:2408.10437, <https://arxiv.org/abs/2408.10437>, 작성일: 2024. 8. 19.

57) 민재욱 외 3인, "특허 언어모델 기반 CPC 클러스터링 필터와 토픽 벡터를 활용한 선행기술 특허검색 성능 향상 연구", 한국정보과학회 학술발표논문집, 2022, 380-382면.

첫 번째로, 여러 특허 필드로 구성된 Input source를 각 필드별로 분리하여 모델에 개별적으로 입력한 뒤, 각 필드로부터 문장 임베딩 벡터를 추출한다. 예를 들어, 입력 문장으로 '에피택셜 성장면에 지지 기판을 가접합하여 제1의 화합물 반도체 접합기판으로 하는 공정' 이 주어진 경우, 해당 문장은 KorPatBERT의 토크나이저와 사전(Vocabulary)를 통해 입력 토큰으로 토큰화 된다. 문장의 시작을 나타내는 [CLS] 토큰과 끝을 나타내는 [SEP] 토큰이 추가되어 '[CLS], 에피, ##택, ##셜, 성장, ##면, ##에, 지지, 기판, ##을, 가, ##접, ##합, ##하여, 제, ##1, ##의, 화합물, 반도체, 접합, ##기, ##판, ##으로, 하,##는, 공정, [SEP]' 순서로 변환된다. 이후 입력 토큰은 학습한 분류 모델의 임베딩 레이어를 통과하여 각 토큰별로 토큰 벡터로 변환된다. 변환된 토큰 벡터 중 [CLS] 토큰에 해당하는 벡터를 선택할 수 있는데, 이는 [CLS] 토큰 벡터가 입력 문장의 전체 의미를 함축한 대표 문맥 벡터로 사용되기 때문이다. 이때, 클래스를 예측하는 태스크에서는 [CLS] 토큰 벡터를 추가된 분류 층(Fully Connected Layer)을 통과하여 분류 대상 클래스의 개수에 해당하는 확률 값(Logits)을 출력한다(예: 클래스가 3종인 경우, [2.1, -1.5, 0.8]로 출력). 마지막으로 softmax 활성화 함수(<수식1>)가 적용되어 모든 예측 확률 값의 합이 1이 되도록 정규화한다(예: 클래스가 3종인 경우, [0.75, 0.05, 0.20]로 출력). IR-SC256을 사용한 경우, 632종 클래스에 대한 예측 확률 값이 출력되고, IR-MG256을 사용한 경우에는 5,503종 클래스에 대한 예측 확률 값이 출력된다. 이 중 가장 높은 출력 값을 가진 클래스를 최종 선택할 수 있으며, 필요에 따라서 출력 값을 정렬하여 후보 클래스의 범위를 확대할 수 있다. 이러한 과정을 통해 특허 문헌 군집화를 진행할 수 있다.

한편, 문장 임베딩 벡터를 추출하기 위해서는 일반적으로 대표 문맥 벡터인 [CLS] 토큰 벡터가 사용되지만, 임베딩 벡터 간 유사도 실험에서 [CLS] 토큰 벡터를 사용하는 것보다 입력 문장의 시작과 끝을 나타내는 [CLS]와 [SEP] 토큰을 제외한 각 토큰 벡터의 평균을 사용하는 것이 더 효과적임을 사전에 확인하였다. 이에 따라 각 토큰 벡터의 평균 값으로 계산하여 문장 임베딩 벡터(<수식2>)를 생성하였다.

<수식2 문장 임베딩 벡터>

$$E_{field} = \frac{1}{n}\sum_{i=1}^{n} V_i$$

* n: [CLS]와 [SEP]를 제외한 토큰 수

<수식3 문헌 임베딩 벡터>

$$E_{doc} = \frac{1}{n}\sum_{i=1}^{n} E_{field_i}$$

* n: 특허 필드 수

두 번째로, 각 필드별로 생성된 문장 임베딩 벡터들은 다시 평균 값을 계산하여 특허 문헌 임베딩 벡터(<수식3>)로 생성하였다. 최종적으로 하나의 특허 문헌에는 5개의 특허 필드에서 추출된 문장 임베딩 벡터와 이를 통합한 하나의 문헌 임베딩 벡터로 표현된다. 이러한 과정을 통해 생성된 문헌 임베딩 벡터는 문헌 전체의 문맥 정보를 효과적으로 내재하며, 이후 유사도 계산에서 사용하였다.

3.2.4. 특허 문헌 유사도

3.2.4.1. 유사도 알고리즘

KorPatBERT의 기본 아키텍처는 각 토큰을 768 차원(Dimension)의 실수 벡터로 임베딩하여 표현하도록 설계되었다. 이를 기반으로 본 연구 모델에서 추출한 특허 문장 임베딩 벡터와

생성한 특허 문헌 임베딩 벡터 역시 768 차원(Dimension)의 고정된 길이를 가진다. 이러한 고차원의 임베딩 벡터는 기존의 키워드(특정 의미를 가지는 단어 또는 구)의 집합에 비해, 비해 문장의 맥락과 구조를 더욱 풍부하게 반영하여 특허 문헌의 의미를 보다 정교하게 표현할 수 있다는 점에서 차별성을 지닌다. Thada, Vikas, et al.[58] 연구에서는 텍스트 객체 간 유사도를 계산하는 방법으로 코사인 유사도(Cosine similarity)를 적용했을 때 가장 높은 적합도를 나타냈다고 주장하였다. 유사한 특허 문헌을 검색하기 위해 본 연구모델에서 생성한 임베딩 벡터 간의 유사도를 측정하는 방법으로 코사인 유사도를 활용하였다.

코사인 유사도는 두 벡터 간의 방향적 유사도를 측정하는 지표로, 벡터의 크기보다는 방향에 초점을 맞추어 두 벡터가 이루는 각도의 코사인 값을 계산한다. 이를 통해 문헌 간의 문맥적 유사도를 알 수 있다. 우선 비교 대상인 두 벡터 A와 B를 선택한 후, 두 벡터의 내적(dot product)과 각각의 벡터 크기(L2 norm)를 계산한다. 이후 내적 값을 벡터 크기의 곱으로 나누어 코사인 유사도를 산출한다. 코사인 유사도의 값이 1에 가까울수록 두 문헌은 유사한 문맥 정보를 공유한다고 해석할 수 있으며, 값이 0 또는 음수에 가까울수록 두 문헌 간의 문맥적 차이가 크거나 상반된 의미를 가지는 것으로 판단할 수 있다.

반면, 기존 키워드 매칭 기반의 유사도 측정 방법으로 Okapi BM25[59]가 널리 사용되고 있다. Whissell, John S., et al.[60]연구에 따르면, BM25는 유사도를 계산하는 과정이 간단하고 효율적이며, 계산 속도가 빠르다는 점에서 높은 효율성을 제공한다. 또한 문서의 길이가 너무 짧거나 긴 경우를 보정할 수 있는 메커니즘을 포함하고 있어, 다양한 길이를 가진 문서에 대해 안정적인 성능을 발휘한다고 주장하였다. 이러한 특징으로 인해 대규모 데이터셋에서도 빠른 키워드 기반 검색이 가능하여 검색 효율성 측면에서 강점을 가진다. 그러나 키워드 매칭에 의존하는 방식으로 작동하기 때문에 문서의 문맥적 의미나 연관성을 충분히 반영하지 못하는 문맥적 정보 부족의 한계를 지닌다. 동의어나 문장 구조가 다르더라도 동일한 의미를 가지는 문서에 대해 낮은 유사도를 부여할 가능성이 있어 어휘 다양성을 처리하는 데 한계를 보인다[61].

이러한 BM25의 한계를 보완하기 위해, 본 연구에서는 두 문헌에 대해서 토큰 간 BM25 유사도와 임베딩 벡터 간 코사인 유사도를 결합한 접근법을 제안한다. <그림7>에 제시되어 있는 검색 모델의 문헌 유사도 동작 구조에 따라 본 연구에서는 Query와 Documents 간의 BM25 점수와 코사인 유사도 점수를 통합적으로 활용함으로써 검색 성능을 극대화하였다.

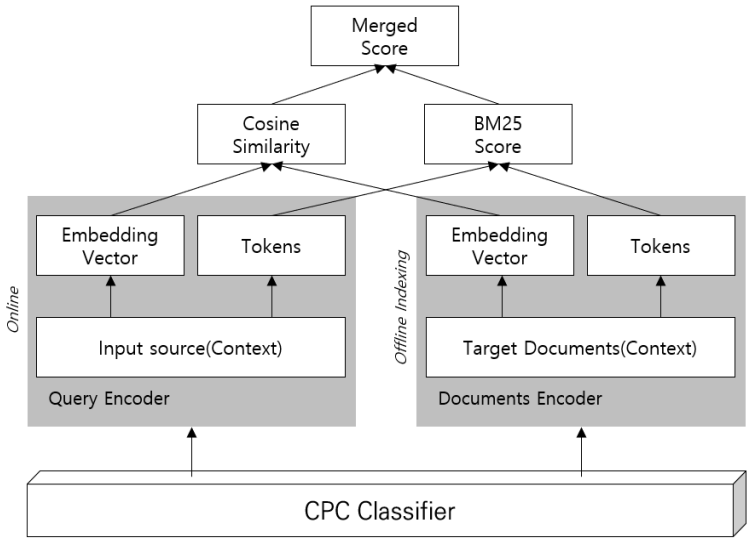
58) Vikas Thada & Vivek Jaglan, “Comparison of jaccard, dice, cosine similarity coefficient to find best fitness value for web retrieved documents using genetic algorithm”, *International Journal of Innovations in Engineering and Technology,* Vol.2 No.4(2013), pp. 202-205.

59) Stephen E. Robertson & Sparck K Jones, “Relevance weighting of search terms”, *Journal of the American Society for Information science,* Vol.27 No.3(1976), pp. 129-146.

60) John S. Whissell & Charles L.A. Clarke, “Improving document clustering using Okapi BM25 feature weighting”, *Information retrieval*, Vol.14(2011), pp. 466-487.

61) Stefan Büttcher et al., “Term proximity scoring for ad-hoc retrieval on very large text collections”, Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, 2006, pp. 621-622.

<그림7 문헌 유사도 동작 구조>



이와 같은 결합 방식은 BM25가 제공하는 중요한 워드에 가중치를 부여하는 것과 임베딩 벡터 유사도를 통한 문맥적 정밀성을 동시에 반영하고자 하였고, 개별 점수만으로 검색을 수행할 때보다 더 정교하고 신뢰도 높은 검색 결과를 제공한다. 특히, 대규모 데이터셋과 복잡한 문맥적 연관성을 동시에 처리해야 하는 특허 문헌 검색 작업에서, 단순 키워드 매칭 기반 방법의 한계를 효과적으로 보완하며 문헌 간 문맥적 의미를 정교하게 반영할 수 있다.

본 연구에서는 BM25만을 이용한 검색 결과를 Baseline으로 설정하였으며, BM25 점수와 벡터간 유사도를 결합한 방법의 성능을 정량적으로 비교 평가하였다.

3.2.4.2. 검색 성능 평가 지표

특허 검색 모델의 성능 평가를 위해 검색 전문가 관점의 평가 지표인 Top-K recall을 제안한다. Top-K recall은 Fall, Caspar J., et al.[62] 연구에서 제안하는 특허 분류 평가 지표와 동일한 맥락을 가진다. Top-K Recall은 사용자가 실제 검색 과정에서 가장 먼저 참고하는 상위 K 구간의 검색 결과의 정확도를 측정하여, 모델의 실질적인 유용성을 정량적으로 평가하는 데 초점을 맞춘다. 기존의 정밀도(Precision)[63] 지표가 전체 검색 결과의 정확도를 평가하는 데 중점을 두었던 반면, Top-K recall은 사용자의 정보 탐색 행동을 반영하여 상위 K 구간의 검색 결과 품질에 집중한다. 이러한 접근은 특허 심사와 같이 사용자가 상위 구간별 검색 결과에 의존하는 실제적인 상황에서 모델의 성능을 더욱 정확하게 평가할 수 있도록 한다. 또한, 검색 결과를 상위 특정 구간에 노출되도록 모델 개선 방향을 설정하는 데 있어 중요한 기준으로 활용될 수 있을 것으로 기대된다.

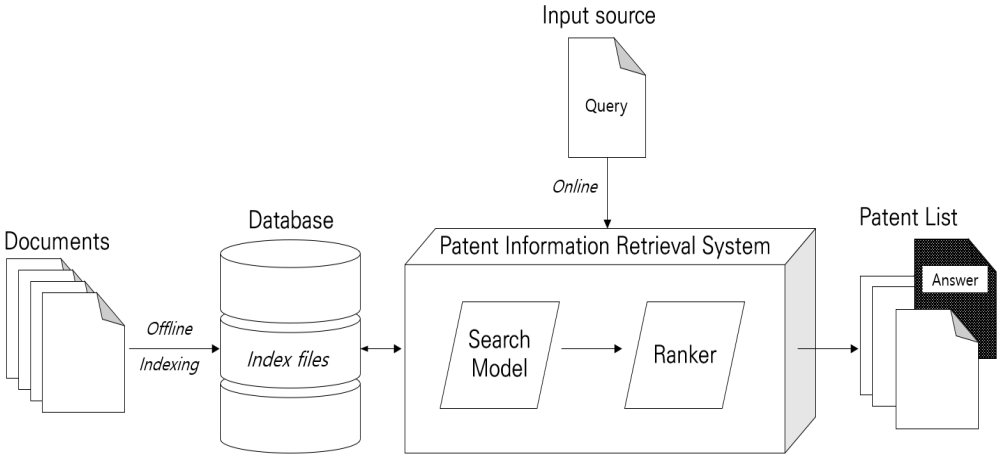
3.3. 검색 시스템 구축

본 연구에서는 현재 운영 중인 검색 시스템의 프로세스를 적용하여 구축하고 검색 모델의 성능을 평가하여 실질적인 유용성을 입증하고 산업적 적용 가능성을 높이고자 한다.

62) Caspar J. Fall et al., "Automated categorization in the international patent classification", *Acm Sigir Forum,* Vol.37 No.1(2003), pp. 10-25.
63) 검색된 문서 중에서 실제로 관련성 있는 문서의 비율

검색 시스템은 사용자에게 실시간 서비스를 제공하기 위해 설계되며, 방대한 데이터를 신속하고 효율적으로 처리할 수 있는 시스템 구조가 요구된다. 본 연구에서는 Query 15,669건을 Documents 1,214,043건과 비교하여 각 문헌 간의 유사도를 계산하고, 검색 결과를 출력하는 실험을 진행하였다. 이러한 과정은 <표9>에 제시한 6종의 검색 모델 각각에 대해 개별 실험을 진행해야 하므로 높은 연산 부담이 요구되었다. 이를 해결하기 위해 대량의 Documents를 색인화(Indexing)[64]하는 작업을 진행하였다.

<그림8 특허 검색 시스템 동작 구조>



<그림8>은 특허 검색 시스템의 동작 구조를 제시하였으며, 데이터베이스의 구체적인 동작 원리는 <그림7>에서 제시된 Documents Encoder로 설명된다.

Documents Encoder는 검색 시스템이 서비스를 시작하기 전에 필요한 데이터를 사전에 처리하여, Input source와의 유사도 계산을 신속하게 수행할 수 있도록 설계되었다. 반면, 사용자가 실시간으로 입력하는 Input source는 Query Encoder에서 처리되며, 사전 작업 없이 즉각적으로 동작 한다.

Documents Encoder에서 핵심적인 역할을 하는 작업은 색인(Index)이다. 색인은 검색 쿼리 수행 시 데이터를 신속하게 탐색할 수 있도록 데이터 구조를 최적화하는 작업으로 시스템의 처리 속도와 응답 성능을 크게 향상시킨다. 특히, 역색인(Inverted Index)은 키워드와 해당 키워드가 포함된 데이터의 위치를 매핑하는 구조로, 방대한 데이터에서 관련 정보를 빠르게 탐색할 수 있도록 한다. 본 연구에서는 대규모 BM25 연산을 위해 역색인 기술을 사용하였으며, 대규모 임베딩 벡터 연산은 Faiss(Facebook AI Similarity Search)[65] 오픈 소스를 사용하였다. Faiss는 GPU 가속을 사용하여 대규모 임베딩 벡터 유사도 연산을 신속히 수행할 수 있도록 지원하며, 색인 속도와 시스템 안정성 측면에서도 우수한 성능을 발휘한다.

<그림8> 데이터베이스에는 Documents의 각 문헌별로 식별정보와 함께 “발명의 명칭+요약”, “배경기술”, “기술분야”, “청구항” 특허 필드로 구성된 Context를 포함하고, 분류 모델을 통해 추출 및 생성된 문헌 임베딩 벡터를 저장하여 Faiss로 통합 색인함으로써, BM25 연산과

64) Alfonso F. Cardenas, “Analysis and performance of inverted data base structures”, *Communications of the ACM*, Vol.18 No.5(1975), pp. 253-263.

65) facebookresearch, “facebookresearch/faiss”, facebookresearch github, <https://github.com/facebookresearch/faiss>, 검색일: 2024. 12. 10

코사인 유사도 연산을 신속히 수행할 수 있도록 하였다.

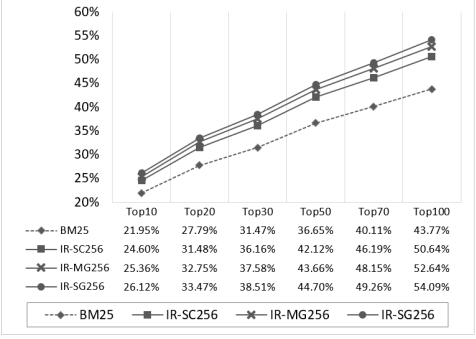
<표9>에 제시된 각 검색 모델에 대해 동일한 색인 과정을 거쳐 데이터베이스에 저장하고 검색 실험을 진행하였다.

4. 실험 및 평가

구축한 특허 문헌 검색 시스템을 활용해서 사전에 정의된 평가 대상 데이터셋을 대상으로 검색 실험을 수행하고 Top-K recall 지표를 통해 그 결과를 분석하였다.

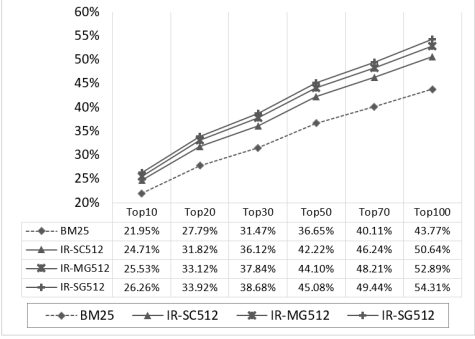
첫 번째, Baseline인 BM25와 각 검색 모델의 검색 결과를 비교하였다. <그림9>, <그림10>의 결과에 따르면, 모든 검색 모델이 Top-K 구간에서 BM25 보다 높은 성능을 보였다. 구체적으로, IR-SC256은 BM25 보다 평균 4.91%p, IR-MG256은 6.40%p, IR-SG256은 7.40%p 더 높은 성능을 기록하였다. 또한, IR-SC512은 5%p, IR-MG512은 6.66%p, IR-SG512은 7.66%p의 성능 향상을 보였다. 이는 BM25의 유사도와 문헌 임베딩 벡터 간의 유사도가 유사한 경향을 보인다고 점을 뒷받침하며, 두 유사도 점수를 결합하여 검색 순위를 결정하는 방식이 검색 성능을 효과적으로 개선했음을 시사한다.

<그림9 BM25와 검색 모델별 성능 비교 (256 토큰 길이)>



	Top10	Top20	Top30	Top50	Top70	Top100
BM25	21.95%	27.79%	31.47%	36.65%	40.11%	43.77%
IR-SC256	24.60%	31.48%	36.16%	42.12%	46.19%	50.64%
IR-MG256	25.36%	32.75%	37.58%	43.66%	48.15%	52.64%
IR-SG256	26.12%	33.47%	38.51%	44.70%	49.26%	54.09%

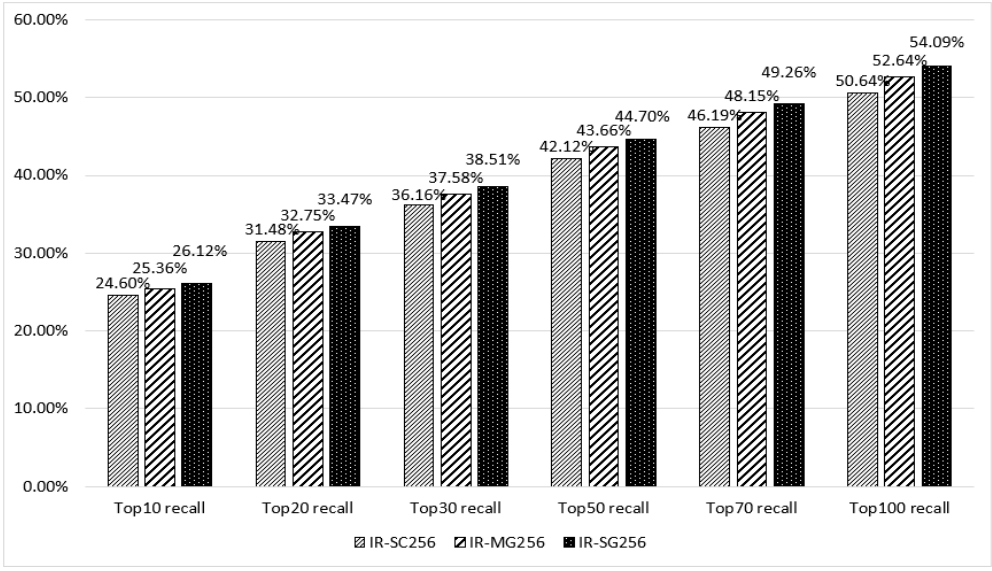
<그림10 BM25와 검색 모델별 성능 비교 (512 토큰 길이)>



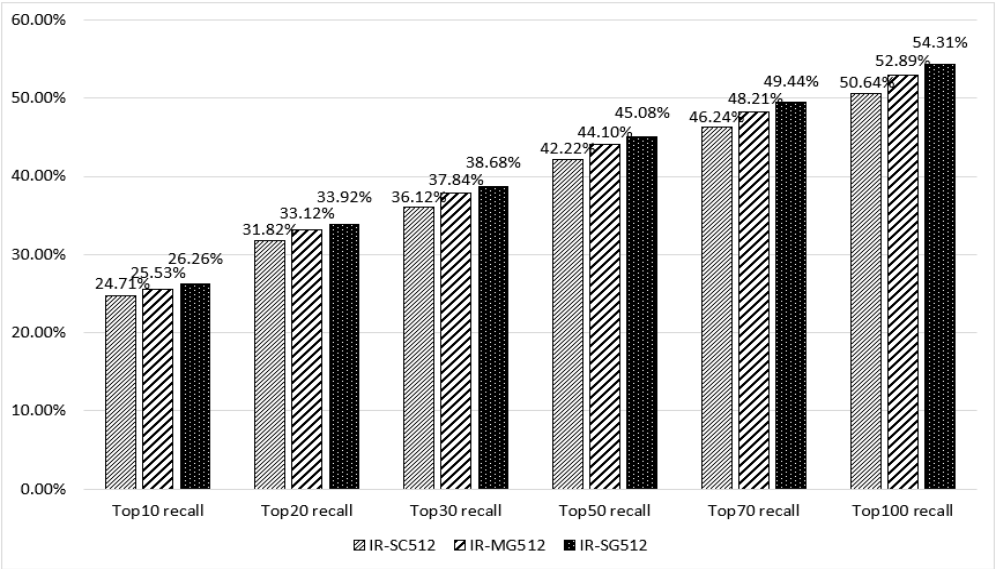
	Top10	Top20	Top30	Top50	Top70	Top100
BM25	21.95%	27.79%	31.47%	36.65%	40.11%	43.77%
IR-SC512	24.71%	31.82%	36.12%	42.22%	46.24%	50.64%
IR-MG512	25.53%	33.12%	37.84%	44.10%	48.21%	52.89%
IR-SG512	26.26%	33.92%	38.68%	45.08%	49.44%	54.31%

두 번째, CPC 분류 체계별 검색 모델의 성능을 비교하였다. <그림11>, <그림12>의 결과에 따르면, 모든 Top-K 구간에서 IR-SG → IR-MG → IR-SC 순으로 성능이 우수한 것을 확인하였다. 특히, 가장 높은 성능 보인 IR-SG512와 가장 낮은 성능 보인 IR-SC256를 비교했을 때, Top10에서 1.66%p, Top20에서 2.44%p, Top30에서 2.52%p, Top50에서 2.96%p, Top70에서 3.25%p, Top100에서 3.36%p의 성능 차이가 나타났다. 그러나, IR-SG512 가 IR-SC256 보다 2배 많은 학습 데이터셋과 100배 더 많은 분류 능력을 갖춘 모델임을 고려할 때, 이 성능 차이는 미흡하다고 판단된다.

<그림11 검색 모델별 성능 비교(256 토큰 길이)>

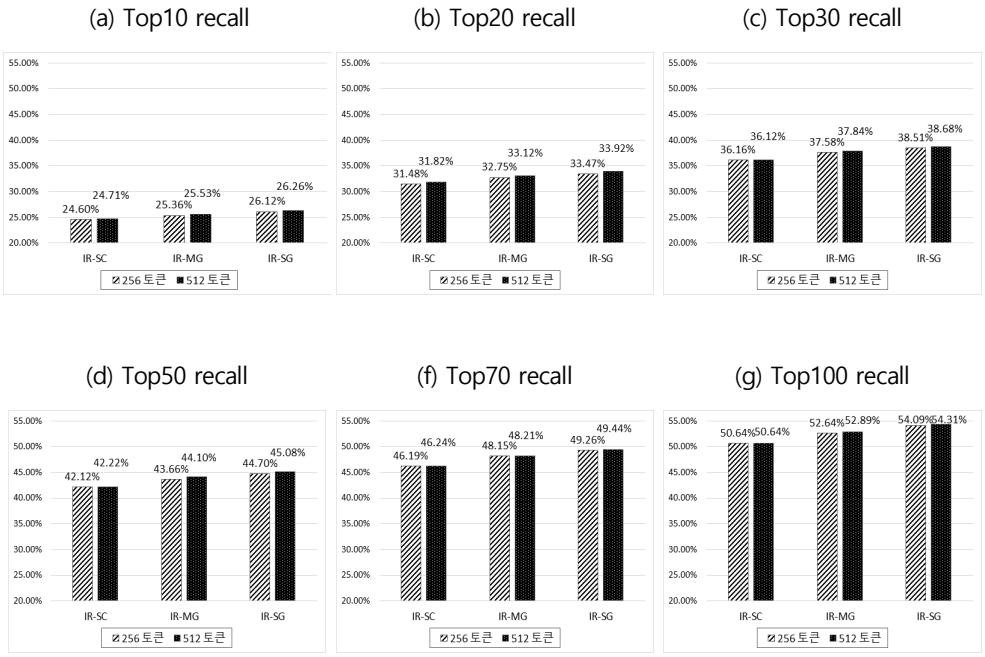


<그림12 검색 모델별 성능 비교(512 토큰 길이)>



세 번째, 토큰 길이별 모델의 성능을 비교하였다. <그림13>의 결과에 따르면, 대부분의 경우 512 토큰 길이의 모델이 256 토큰 길이 모델보다 성능이 우수했지만 그 차이는 실질적으로 크지 않은 것으로 판단되었다. 구체적으로, IR-SC는 평균 0.09%p, IR-MG은 0.26%p, IR-SG은 0.22%p의 성능 차이를 보였다. 이는 문장의 중요한 정보가 주로 앞부분에 집중되어 있어서, 토큰 길이를 길게 입력 받더라도 임베딩 벡터 간 차이가 크지 않다는 것으로 판단해 볼 수 있다. 다만, 이러한 현상에 대해서, 본 연구에서는 구체적인 검증을 수행하지 않았으며, 추가적인 연구를 통해 구체적으로 검증을 하고 모델 성능을 향상시키기 위한 방안은 향후 연구 과제로 남겨둘 예정이다.

<그림13 토큰 길이별 모델 성능 비교(%)>



5. 결론

본 연구에서는 AI 기술 기반 특허 분야 특화된 모델을 활용하여 특허 검색 성능을 향상시키기 위해 다양한 검색 모델을 제안하고, 이를 정량적으로 평가하였다.

첫째, Baseline인 BM25와 비교한 결과, 제안된 모든 검색 모델이 Top-K 구간에서 BM25 보다 높은 성능을 보였으며, 특허 문헌 임베딩 벡터 유사도와 BM25 유사도를 결합한 방식이 BM25의 표현 한계를 효과적으로 보완함을 확인하였다.

둘째, CPC 분류 체계별 검색 모델 성능 비교에서는 IR-SG → IR-MG → IR-SC 순으로 우수한 성능을 나타냈다. 특히, 분류 모델이 세분화될수록 생성된 임베딩 벡터의 효과가 증가함을 확인하였다. 그러나 학습 데이터의 양과 모델의 복잡도를 고려했을 때 일부 모델은 기대에 미치지 못하는 성능 차이를 보였으며, 이는 추가적인 분석이 필요함을 시사한다.

셋째, 토큰 길이별 비교에서는 더 긴 토큰 길이를 적용한 모델이 약간 더 나은 성능을 보였으나, 그 차이는 미미하였다. 이는 문서의 주요 정보가 주로 앞부분에 집중되어 있음을 시사하며, 토큰 길이를 단순히 증가시키는 것이 특허 검색 성능 향상에 큰 영향을 미치지 않음을 보여준다.

본 연구에서는 1. 서론에서 제시한 연구 목표를 성공적으로 달성하였다.

첫째, 다양한 산업 분야에 널리 활용되는 핵심 기술 분야인 H섹션을 대상으로 실제 특허 데이터를 기반으로 한 고품질의 평가 및 검색 대상 데이터셋을 구축하였다. 이를 통해 실질적이고 객관적인 특허 검색 실험을 가능하게 하였다.

둘째, 한국어 특허 데이터에 특화된 KorPatBERT를 활용하여 CPC 분류 모델을 개발하고, 특허 문헌의 내재된 주제 및 기술적인 특징을 반영한 임베딩 벡터를 활용하여 특허 문헌 검색 모델에서 효과적임을 확인하였다. 특히, 세분화된 분류 모델일수록 검색 모델에 효과적임을 확인

할 수 있었다.

셋째, BM25 기반의 전통적인 검색 방식과 딥러닝 기반의 임베딩 벡터 유사도를 결합한 실시간 검색 시스템을 구축하고, 이를 구현하는 방법을 제시함으로써, 실질적인 특허 검색 환경을 모사하였다.

본 연구에서는 R&D와 산업 현장에서의 차이를 최소화하기 위한 환경을 구축함으로써 실질적 유용성과 적용 가능성을 높이는데 기여하였다. 이는 높은 전문성이 요구되는 선행 기술 조사관과 특허 심사관의 업무 효율성을 향상시키고 의사 결정의 정확성을 높이는 데 긍정적인 영향을 미칠 것으로 기대된다. 아울러, 특허 검색 분야의 학문적 발전에 기여함과 동시에, 국내 기업들이 경쟁력이 있는 핵심 기술을 확보하고, 지속적으로 혁신을 주도할 수 있도록 돕는 선행 연구로서 모범적인 사례를 제공할 것으로 기대된다.

본 연구에서 개발된 특허 문헌 검색 모델은 특허 문헌 간 유사도를 기반하여 검색을 수행하지만, 특허 문헌 간 유사도가 낮더라도 실제 권리 범위를 침해하는 특정 기술 요소가 존재할 수 있다. 따라서 단순히 특허 문헌 검색만으로는 특허에 포함된 실질적인 기술 요소를 완벽히 포착하기에는 한계가 있다.

향후 연구에는 본 연구에서 축적된 노하우와 결과를 바탕으로 더욱 정교하고 세부적인 검색 방법을 연구하고자 한다. 또한, 발전된 인공지능 모델을 연구하고 적용하여 특허 분야에서 직면한 다양한 문제를 해결하는 데 도전하고자 한다.

참고 문헌(References)

학술지(국내 및 동양)

민재옥 외 3인, "Korean Patent ELECTRA: 한국 특허문헌 자연어처리 연구를 위한 사전 학습된 언어모델(KorPatELECTRA)", 「한국컴퓨터정보학회 학술발표논문집」, 제29권 제2호(2021).

박상언, "딥러닝 기반 사전학습 언어모델에 대한 이해와 현황", 「한국빅데이터학회지」, 제7권 제2호(2022).

박진우 외 4인, "한국어 특허 문장 기반 CPC 자동분류 연구-인공지능 언어모델 KorPatBERT 를 활용한 딥러닝 기법 접근", 「지식재산연구」, 제17권 제3호 (2022).

임소라·권용진, "특허문서 필드의 기능적 특성을 활용한 IPC 다중 레이블 분류", 「인터넷정보학회논문지」, 제18권 제1호 (2017).

임준호 외 2인, "딥러닝 사전학습 언어모델 기술 동향", 「전자통신동향분석」, 제35권 제3호(2020).

학술지(서양)

Amna Ali et al., "Innovating patent retrieval: a comprehensive review of techniques, trends, and challenges in prior art searches", *Applied System Innovation*, Vol.7 No.5 (2024).

Jakob Edler et al., "Technology sovereignty as an emerging frame for innovation policy. Defining rationales, ends and means", *Research Policy*, Vol.52 No.6(2023).

Alfonso F. Cardenas, "Analysis and performance of inverted data base structures", *Communications of the ACM*, Vol.18 No.5(1975).

Caspar J. Fall et al., "Automated categorization in the international patent classification", *Acm Sigir Forum,* Vol.37 No.1(2003).

Francis Narin, "Patents as indicators for the evaluation of industrial research output", *Scientometrics*, Vol.34 No.3(1995).

Hamidet Bekamiri al., "Patentsberta: A deep nlp based hybrid model for patent distance and classification using augmented sbert", *Technological Forecasting and Social Change*, Vol.206(2024).

Jinhyuk Lee et al., "BioBERT: a pre-trained biomedical language representation model for biomedical text mining", *Bioinformatics*, Vol36 No.4(2020).

John S. Whissell & Charles L.A. Clarke, "Improving document clustering using Okapi BM25 feature weighting", *Information retrieval*, Vol.14(2011).

Kenneth W. Dam, "The economic underpinnings of patent law", *The Journal of Legal Studies* , Vol.23 No.1(1994).

Kuo-tsan Liu & Chia-Ho Chen, "Formulation of research and development strategy by analysing patent portfolios of key players the semiconductor industry according to patent strength and technical function", *World Patent Information,* Vol.70(2022).

Shaobo Li et al., "DeepPatent: patent classification with convolutional neural networks and word embedding", *Scientometrics,* Vol.117 No.2(2018).

Stefano IIncarbone, "Claim construction: an international convergence in striking the balance between patent protection and legal certainty", *Journal of Intellectual Property Law and Practice,* Vol.17 No.10(2022).

Stephen E. Robertson & Sparck K. Jones, "Relevance weighting of search terms", *Journal of the American Society for Information science,* Vol.27 No.3(1976).

Sunil Kanwar & Robert Evenson, "Does intellectual property protection spur technological change?", *Oxford Economic Papers,* Vol.55 No.2(2003).

Vikas Thada & Vivek Jaglan, "Comparison of jaccard, dice, cosine similarity coefficient to find best fitness value for web retrieved documents using genetic algorithm", *International Journal of Innovations in Engineering and Technology,* Vol.2 No.4(2013).

Youngsam Chun et al., "AI technology specialization and national competitiveness", *Plos one,* Vol.19 No.4(2024).

Ziwei Ji et al., “Survey of hallucination in natural language generation”, *ACM Computing Surveys,* Vol.55 No.12(2023).

인터넷 자료

특허청, “CPC 및 IPC 분류코드”, 특허청, <https://www.kipo.go.kr/ko/kpoContentView.do?menuCd=SCD0200269>, 검색일: 2024. 10. 30

한국특허기술진흥원, “특허분류 조회 서비스”, 한국특허기술진흥원, <https://www.pipc.or.kr/business/cpcService>, 검색일: 2024. 10. 30.

한국특허정보원, “특허정보 활용 서비스”, 한국특허정보원, <https://plus.kipris.or.kr/portal/main.do>, 검색일: 2024. 10. 30.

한국특허정보원, “kipi-ai/korpatbert”, 한국특허정보원 github, <https://github.com/kipi-ai/korpatbert>, 검색일: 2024. 11. 10.

Alec Radford et al., “Improving language understanding by generative pre-training”, OpenAI, <https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf>, 검색일: 2024. 10. 30.

facebookresearch, “facebookresearch/faiss”, facebookresearch github, <https://github.com/facebookresearch/faiss>, 검색일: 2024. 12. 10.

Fiveipoffices, “IP5 Statistics Report 2022”, Fiveipoffices, <https://www.fiveipoffices.org/statistics/statisticsreports/2022edition>, 검색일: 2024. 10. 30.

Google, "google-research/bert", Google github, <https://github.com/google-research/bert>, 검색일: 2024. 11. 12

Ilias Chalkidis et al., “LEGAL-BERT: The muppets straight out of law school”, arXiv preprint arXiv:2010.02559, <https://arxiv.org/abs/2010.02559>, 작성일: 2020. 10. 6.

Ilya Loshchilov & Frank Hutter, “Decoupled Weight Decay Regulariztion”, arXiv preprint arXiv:1711.05101, <https://arxiv.org/abs/1711.05101>, 작성일: 2019. 1. 4.

Iz Beltagy et al., “SciBERT: A pretrained language model for scientific text”, arXiv preprint arXiv:1903.10676, <https://arxiv.org/abs/1903.10676>, 작성일: 2019. 9. 10.

Jacob Devlin et al., “Bert: Pre-training of deep bidirectional transformers for language understanding”, arXiv preprint arXiv:1810.04805, <https://arxiv.org/abs/1810.04805>, 작성일: 2019. 5. 24.

Jieh-Sheng Lee & Jieh Hsiang, “Patentbert: Patent classification with fine-tuning a pre-trained bert model”, arXiv preprint arXiv:1906.02124, <https://arxiv.org/abs/1906.02124>, 작성일: 2019. 7. 1.

Marius Mosbach et al., “On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines”, arXiv preprint arXiv:2006.04884, <https://arxiv.org/abs/2006.04884>, 작성일: 2021. 3. 25.

Max Vargas et al., “Understanding Generative AI Content with Embedding Models”, arXiv preprint arXiv:2408.10437, <https://arxiv.org/abs/2408.10437>, 작성일: 2024. 8. 19.

Nandan Thakur et al., “Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks”, arXiv preprint arXiv:2010.08240, <https://arxiv.org/abs/2010.08240>, 작성일: 2021. 4. 12.

Rohan Anil et al., “Palm 2 technical report”, arXiv preprint arXiv:2305.10403, <https://arxiv.org/abs/2305.10403>, 작성일: 2023. 9. 13.

기타 자료

민재옥 외 3인, “특허 언어모델 기반 CPC 클러스터링 필터와 토픽 벡터를 활용한 선행기술 특허검색 성능 향상 연구”, 한국정보과학회 학술발표논문집, 2022.

심우철 외 4인, “한국 특허문헌 특성 및 딥러닝 기반 분류모델을 고려한 CPC 자동분류에 관한 연구”, 한국정보과학회 학술발표논문집, 2020.

특허청, “2023 통계로 보는 특허동향”, 특허청, 2023.

Ashish Vaswani et al., “Attention is all you need”, 31st Conference on Neural Information Processing Systems (NIPS 2017), 2017.

Dylan Myungchul Kang et al., “Patent prior art search using deep learning language model”, Proceedings of the 24th Symposium on International Database Engineering & Applications. 2020.

Goran Oreski & Stjepan Oreski, “An experimental comparison of classification algorithm performances for highly imbalanced datasets”, Central European Conference on Information and Intelligent Systems. Faculty of Organization and Informatics Varazdin, 2014.

Kanishka Vaish et al., “Artificial Intelligence Reducing the Intricacies of Patent Prior Art Search”, 2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), IEEE, 2023.

Long Ouyang et al., “Training language models to follow instructions with human feedback”, Advances in neural information processing systems 35, 2022.

Mirac Suzgun et al., “The harvard uspto patent dataset: A large-scale, well-structured, and multi-purpose corpus of patent applications”, Advances in neural information processing systems 36, 2023.

Patrick Lewis et al., “Retrieval-augmented generation for knowledge-intensive nlp tasks”, Advances in Neural Information Processing Systems 33, 2020.

Rehan Akbani et al., “Applying support vector machines to imbalanced datasets”, Machine Learning: ECML 2004: 15th European Conference on Machine Learning, Pisa, Italy, September 20-24, 2004. Proceedings 15, Springer Berlin Heidelberg, 2004.

Stefan Büttcher et al., “Term proximity scoring for ad-hoc retrieval on very large text collections”, Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, 2006.