지식재산연구
Journal of Intellectual Property

지식재산연구 21권 1호, 2026년 3월, 151-180면
Journal of Intellectual Property Volume 21, Issue 1, March 2026, Pages 151-180.
https://doi.org/10.34122/jip.2026.21.1.151

Print ISSN 1975-5945
Online ISSN 2733-8487

RESEARCH ARTICLE

# Constructing Korean Patent Retrieval Datasets to Improve Deep Learning-Based Patent Retrieval Performance: An Automated Methodology

**Dong-Uk Lee[1], Woo-Chul Sim[2], Jin-Woo Park[3], Bong-Gun Lee[4]**

[1]Associate of Intelligent Information Strategy Department, Korea Institute of Patent Information, Republic of Korea
[2]Assistant Manager of Intelligent Information Strategy Department, Korea Institute of Patent Information, Republic of Korea
[3]Manager of Intelligent Information Strategy Department, Korea Institute of Patent Information, Republic of Korea
[4]Head of Intelligent Information Strategy Department, Korea Institute of Patent Information, Republic of Korea

Corresponding Author: Bong-Gun Lee (bglee@kipi.or.kr)

## ABSTRACT

Owing to the difficulty of constructing large-scale datasets and the scarcity of Korean-language resources, recent deep learning-based patent retrieval research gaces limitaions in improving model performance. To address these challenges, this study proposes a methodology for automatically building a large-scale patent retrieval dataset from Korean patent documents. The method automatically extracts semantically related pairs of technical components between patent applications and cited prior art using claim comparison tables in office action notices. In addition, the sentences that are most similar to each technical component are extracted from both the patent application and the cited prior art documents. Korean patent XML parsing techniques are combined with a KorPatBERT-based CPC classification model, and a hybrid similarity measure integrating sentence embedding–based semantic similarity with lexical similarity is employed.

Subsequently, a large-scale, high-quality dataset approximately 19 times larger than a manually constructed expert dataset was built and validated through large-scale experiments simulating real-world retrieval environments. Experimental results indicate that models trained on the automatically constructed dataset achieved Top-70 accuracy comparable to or better than those trained on expert-built datasets. Accordingly, this study presents a practical and cost-effective approach for constructing high-quality Korean patent retrieval datasets and demonstrates improved performance and real-world applicability.

### KEYWORDS

Automatic Dataset Construction, CPC Classification, KorPatBERT, Patent Retrieval, Patent Similar Technical Component Dataset, Semantic Similarity



**Received:** December 23, 2025
**Revised:** January 15, 2026
**Accepted:** March 06, 2026
**Published:** March 30, 2026

**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

# 딥러닝 기반 특허 검색 성능 개선을 위한 한국어 특허 검색 데이터셋 자동구축 방법론

**이동욱[1], 심우철[2], 박진우[3], 이봉건[4]**

[1]한국특허정보원 지능정보전략실 주임
[2]한국특허정보원 지능정보전략실 대리
[3]한국특허정보원 지능정보전략실 과장
[4]한국특허정보원 지능정보전략실장

교신저자: 이봉건 (bglee@kipi.or.kr)

## 차례

## 국문초록

최근 딥러닝 기반 특허 검색 연구에서는 대규모 데이터셋 구축의 어려움과 한국어 데이터셋 부족으로 인해 모델 성능 향상에 한계가 존재한다. 본 연구는 이러한 한계를 극복하기 위해, 한국어 특허 문헌을 대상으로 한 대규모 특허 검색 데이터셋을 자동으로 구축하는 방법론을 제안한다. 제안 방법은 의견제출통지서 내 구성대비표 데이터를 활용하여 출원 특허와 인용 선행기술 간 의미적으로 연관된 기술 구성요소 쌍을 자동 추출한다. 또한 기술 구성요소와 가장 유사한 문장을 출원 특허와 인용 선행기술 특허 문헌에서 추출한다. 이를 위해 한국 특허 XML 파싱 기법과 KorPatBERT 기반 CPC 분류 모델을 결합하였으며, 문장 임베딩 기반 의미 유사도와 어휘 유사도를 결합한 하이브리드 유사도 계산 방식을 사용하였다.

본 방법론을 통해 전문가 수작업 데이터셋 대비 약 19배 규모의 대규모 고품질 데이터셋을 구축하였으며, 실제 검색 환경을 모사한 대규모 실험을 통해 품질을 검증하였다. 실험 결과, 제안한 자동 구축 데이터셋을 활용하여 학습한 검색 모델은 전문가 구축 데이터셋 대비 Top-70 정확도가 유사하거나 우수한 검색 성능을 달성하였다. 본 연구는 대규모 고품질 한국어 특허 검색 데이터셋을 비용 효율적으로 구축할 수 있는 실용적인 방법을 제시하며, 한국어 특허 검색 모델의 성능 향상 및 실무적 활용 가능성을 동시에 확보했다는 점에서 의의가 있다.



**주제어**

특허 유사 기술 구성요소 데이터셋, 특허 검색, 의미 유사도, KorPatBERT, CPC 분류, 자동 데이터 구축

# 1. 서론

## 1.1. 개요

전 세계적인 혁신 활동과 기술 경쟁력 평가는 특허출원 건수를 핵심 지표로 활용한다. WIPO(World Intellectual Property Organization)의 World Intellectual Property Indicators 2025 보고서에 따르면, 2024년 전 세계 특허출원 건수는 약 370만 건으로 사상 최고치를 기록하였으며, 전년 대비 4.9% 증가하는 등 지속적인 성장세를 보였다.[1)]

그러나 특허출원 규모의 지속적인 증가는 특허 정보의 탐색과 활용을 점점 더 어려운 과제로 만들고 있다. 매년 수백만 건씩 축적되는 특허문서는 기술적 표현의 복잡성, 법률적 서술 방식, 그리고 중복 및 유사 기술의 증가로 인해 정확한 검색과 선행기술 조사에 제약을 초래한다. 선행 연구들 또한 대규모 특허 데이터 환경에서 정확하고 효율적인 특허 검색이 여전히 도전적인 과제임을 지적하고 있다.[2)]

특허 검색의 효율성을 제고하기 위한 기반 요소로 특허 분류체계의 중요성이 논의되고 있다. 특히 CPC(Cooperative Patent Classification) 분류체계는 유럽특허청(EPO)과 미국특허청(USPTO)이 공동으로 개발한 분류체계로, 수만 개의 세분화된 기술 코드를 통해 기술 영역을 정밀하게 구조화함으로써 특허 문서의 체계적인 분류와 검색 성능을 향상에 기여한다. WIPO 및 EPO의 공식 문서에 따르면 CPC는 기술 동향 분석, 검색 및 심사 과정 전반에서 핵심적인 기준 체계로 활용되고 있으며, 세부 기술 단위의 분석이 필요한 환경에서 그 중요성이 강조되고 있다.[3)4)]

## 1.2. 문제 정의

CPC와 같은 정교한 분류체계의 존재만으로는 특허 검색의 한계를 완전히 해소하기 어렵다. 특허 문서는 동일한 기술 개념을 다양한 표현으로 기술하며, 법적·기술적 서술이 혼합된 구조를 가지기 때문에, 키워드나 분류 정보만으로는 검색자의 의도와 기술적 유사성을 충분히 반영하는 데 한계가 있다.[5)]

이러한 한계를 극복하기 위해 최근에는 딥러닝(Deep Learning) 기반 검색 모델이 대안으로 제시되고 있다. 딥러닝 기반 검색 모델은 문서의 의미적 표현(Semantic Representation)을 학습하여 복잡한 텍스트 간 의미 관계를 포착할 수 있는 잠재력을 가진다.[6)] 그러나 기존 연구들은 주로 제한된 규모의 수작업 데이터셋이나 영어 중심의 데이터셋에 의존하고 있어, 한국어 특허 문서를 대상으로 한 대규모 학습 데이터셋은 거의 존재하지 않는다.[7)8)] 이로 인해 실제 대규

1) WIPO, “World Intellectual Property Indicators 2025”, WIPO, 2025, pp. 1-188.
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), Article No. 91.
3) WIPO, “World Intellectual Property Indicators 2025”, WIPO, 2025, pp. 1-188.
4) EPO & USPTO, “CPC Guide”, Cooperative Patent Classification, <https://www.cooperativepatentclassification.org/home>, 검색일: 2025. 12. 22.
5) Bart Degroote & Pierre Held, “Analysis of the patent documentation coverage of the CPC in comparison with the IPC with a focus on Asian documentation”, *World Patent Information,* Vol.54 Supplement(2018), S78-S84.
6) Liang Chen et al., “A deep learning based method benefiting from characteristics of patents for semantic relation classification”, *Journal of Informetrics,* Vol.16 No.3(2022), Article No. 101312.
7) Julian Risch et al., “PatentMatch: A Dataset for Matching Patent Claims & Prior Art”, arXiv, <https://arxiv.org/abs/2012.13919>, 작성일: 2020. 12. 17.
8) Grigor Aslanyan & Ian Wetherbee, “Patent Phrase to Phrase Semantic Matching Dataset”, arXiv, <https://arxiv.org/abs/2208.01171>, 작성일: 2022. 8. 1.

모 특허 검색 환경, 특히 한국어 특허 검색에 대한 일반화 성능 검증이 제한적이라는 한계가 존재한다.

### 1.3. 연구 목적

본 연구의 목적은 기존 특허 문헌 가운데 유사한 기술 구성요소를 탐색하는 특허 검색 과제를 대상으로, 한국어 특허 검색을 위한 데이터셋을 기계적으로 자동 구축하는 방법론을 제안하는 데 있다. 특허 검색은 방대한 특허 문헌 집합으로부터 질의 기술과 기술적으로 유사한 선행기술 문서를 식별하는 과정으로 정의되며, 선행기술 조사 및 기술 동향 분석의 핵심 단계에 해당한다. 이 과정에서 기술적 유사성을 정확히 반영하는 학습 데이터의 품질은 검색 성능에 직접적인 영향을 미친다.

이를 위해 제안한 방법론으로 구축한 자동구축 데이터셋을 딥러닝 기반 특허 검색 모델의 학습에 활용하고, 특허 전문가가 수작업으로 구축한 데이터셋과 비교 실험을 통해 데이터셋 품질과 방법론의 타당성을 평가한다. 이러한 평가는 데이터셋 구축 방법론의 효과성을 검증하는 과정에 해당한다.

본 연구는 한국어 특허 문헌의 언어적 특성과 비정형적 기술 서술로 인해 고품질 학습 데이터 확보가 어려웠던 기존 특허 검색 연구의 한계를 극복하는 것을 목표로 하며. 자동 구축 데이터셋과 전문가 수작업 데이터셋을 각각 학습한 딥러닝 검색 모델의 성능을 비교·분석하여 제안 방법론의 유효성을 검증한다.

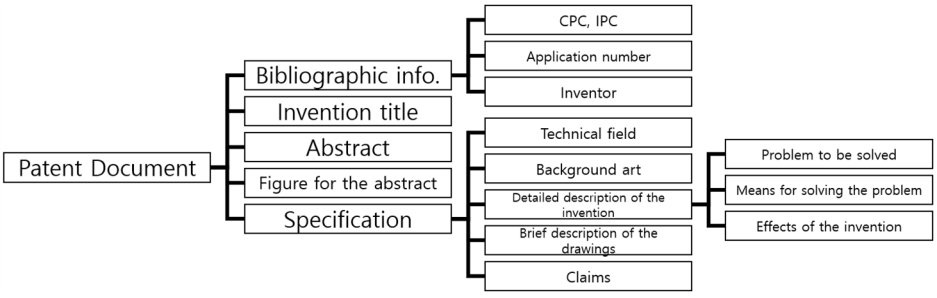
나아가, 본 연구에서 제안한 방법론과 구축된 데이터셋은 한국어 기반 특허 검색 모델의 성능 향상에 기여할 뿐만 아니라, 향후 CPC 자동 분류, 특허 정보 분석 등 다양한 지식재산 분야 연구로의 확장에도 활용될 수 있을 것으로 기대된다.

## 2. 연구 배경

### 2.1. 특허 문서 구조

국내 특허 문서는 그림1과 같은 구조로 구성되어있다.

<그림1 특허 문서의 구조적 구성요소>[9)]




국내 특허 문서는 크게 서지정보(Bibliographic information), 발명의 명칭(Invention title), 초록(Abstract), 대표도(Figure for the abstract), 명세서(Specification)로 구성된

9) 심우철, “CPC 계층적 특성을 고려한 자동 특허 분류 방법”, 충남대학교 일반대학원, 석사, 2025, 5-6면.

다.[10)]

서지정보에는 출원번호(Application number), 발명자(Inventor) 정보, CPC, IPC와 같은 특허 분류 코드(Classification Code) 등 행정적 기본 정보가 제시된다.

발명의 명칭과 초록은 문서의 상단에 위치하여 해당 발명의 핵심 내용을 간략히 파악하는 데 활용된다. 발명의 상세한 설명(Detailed description of the invention)은 명세서를 통해 확인할 수 있는데, 명세서에는 기술 분야(Technical field)와 배경 기술(Background art)이 서술되어 있으며, 이를 통해 발명의 기술적 맥락을 이해할 수 있다.

명세서 내 발명의 상세한 내용(Detailed description of the invention)은 다시 여러 하위 구성으로 나뉜다. 예를 들어, 과제의 해결 수단(Problem to be solved), 해결하려는 과제(Means for solving the problem), 발명의 효과(Effect of the invention) 등이 포함되며, 발명을 명확하게 이해할 수 있도록 구조화된 설명으로 제시된다. 한편, 청구항(Claim)은 발명의 핵심 기술 구성요소가 법적 표현으로 정의되는 부분으로, 학습데이터 구축에 있어 매우 중요한 영역이다.

본 연구에서 학습데이터 구축에 사용 되는 영역은 주로 발명의 상세한 설명과 청구항이 된다. 먼저 기술 구성요소와 유사한 문장을 발명의 상세한 설명에서 우선적으로 추출하며, 적절한 문장을 찾기 어려운 경우 다른 구성 영역을 추가적으로 검토한다.

## 2.2. 의견 제출 통지서

의견제출통지서는 출원된 발명에 대해 심사관이 심사를 진행한 결과, 특허 요건을 모두 충족하지는 않으나 거절 결정을 내리기 전에 출원인에게 의견을 제출할 기회를 부여하기 위해 발송하는 문서이다[11)]. 인공지능 관점에서는 이러한 통지서가 특허 AI 검색 모델용 학습데이터 구축에 유용한 정보를 포함하는 문서로서 중요한 역할을 한다.

의견제출통지서에는 심사 결과와 관련된 상세한 설명이 포함되며, 특히 출원발명과 대비되는 인용발명의 기술 구성요소가 구체적으로 제시된다. 해당 기술 구성요소를 서술하는 방식은 크게 두 가지로 구분할 수 있다. 하나는 구성대비표를 활용하여 출원발명과 인용발명의 구성요소를 항목별로 비교하는 방식이며, 다른 하나는 구성대비표 없이 서술형으로 기술하는 방식이다. 본 연구에서는 전자를 정형 서술방식, 후자를 비정형 서술방식으로 구분하여 기술하고자 한다.

### 2.2.1. 의견 제출 통지서의 정형 서술방식

지식재산처의 의견제출통지서 표준 문안집[12)]에 따르면, 심사 대상이 출원 발명을 인용발명과 대비하여 제시하기 위해 구성대비표를 활용하도록 안내하고 있다. 구성대비표는 출원발명과 인용발명의 기술 구성요소를 항목별로 비교함으로써 동일점과 차이점을 명확히 판단할 수 있도록 구성된다. 표의 형식은 다소 다양하나, 대부분 일정한 구조를 갖추고 있어 파싱과정을 통해 자동으로 추출하는 것이 가능하다.

10) 지식재산처, “2020년 CPC 매뉴얼”, 지식재산처, 2020, 1-68면.
11) 지식재산처, “통지서 표준 문안집”, 지식재산처, <http://www.moip.go.kr/ko/kpoBultnDetail.do?menuCd=SCD0200640&parntMenuCd2=SCD0200281&aprchId=BUT0000048&pgmSeq=10148&ntatcSeq=10148>, 검색일: 2025. 12. 8.
12) 지식재산처, “통지서 표준 문안집”, 지식재산처, <http://www.moip.go.kr/ko/kpoBultnDetail.do?menuCd=SCD0200640&parntMenuCd2=SCD0200281&aprchId=BUT0000048&pgmSeq=10148&ntatcSeq=10148>, 검색일: 2025. 12. 8.

<그림2 의견제출통지서 표준 문안집[13] 참조: 정형 서술방식 예시>

◈ 제29조 제1항

청구항 1 발명은 아래표에서와 같이 인용발명과 대비될 수 있습니다.

| 청구항 1 발명 | | 인용발명 (식별번호 <8>, 도면4) | 비고 |
|---|---|---|---|
| 구성 1 | A | A | **동일** |
| 구성 2 | B | B | **동일** |
| 구성 3 | C | C′ | **실질적 동일** |

구성 3은 인용발명의 C′과 단순히 표현상의 차이만 있을 뿐 실질적으로 동일한 것입니다.

따라서 청구항 1 발명은 인용발명과 실질적으로 동일합니다.

◈ 제29조 제2항

청구항 1 발명은 아래표에서와 같이 인용발명과 대비될 수 있습니다.

| 청구항 1 발명 | | 인용발명 (식별번호 <8>, 도면4) | 비고 |
|---|---|---|---|
| 구성 1 | A | A | **동일** |
| 구성 2 | B | B | **동일** |
| 구성 3 | C | C′ | **차이** |

구성 3의 차이점에 대해 살펴보면, OOOO하므로, 이 발명이 속하는 기술분야에서 통상의 지식을 가진 자가 C′를 C로 용이하게 설계변경할 수 있는 것입니다.

### 2.2.2. 의견 제출 통지서의 비정형 서술방식

의견제출통지서에서 기술구성요소 대비를 서술하는 방식은 표 형태의 정형 방식뿐만 아니라 비정형 방식도 존재한다. 비정형 방식은 문장형으로 기술되며 구조적 일관성이 부족하여, 자동 추출이나 파싱을 적용하기 어려운 특징을 가진다.

<그림3 의견제출통지서 표준 문안집[14] 참조: 비정형 서술방식 예시>

**청구항 1 발명의 A, B는 인용발명 1에** OOO로 **제시**되어 있습니다. 단지, 청구항 1 발명은 C를 더 구비한 점에서 인용발명 1**과 차이**가 있으나, (△△△라는 점에서) 실질적으로 동일한 c가 인용발명 2에 이미 공지되어 있습니다.
따라서 상기 차이점은 OOO하기 때문에 **통상의 기술자**가 인용발명 1에 인용발명 2의 c를 결합하여 용이하게 도출할 수 있습니다.

◈ 청구항 2 **내지 4 발명에서 각 부가구성**은 인용발명 1**의 C, D, E와 표현만 다를 뿐 실질적으로 동일한 것**입니다.

◈ 청구항 2 **내지 5 발명에서 한정한 사항은 인용발명 2**의 f,g,h 구성을 적용대상에 따라 각각 최적화한 정도에 불과합니다.

◈ 청구항 5 발명은 청구항 1 발명과 **카테고리만 상이한 발명**이므로, 제1항과 **동일한 취지의 거절이유가 적용**됩니다.

### 2.2.3. 인용발명 정보

의견제출통지서에는 인용발명의 공보정보가 포함되어 있다.[15] 의견제출통지서 표준 문안집의 안내에 따르면, 인용발명 정보는 통지서 서두 또는 말미에 일괄적으로 제시하도록 규정되어

13) 지식재산처, “통지서 표준 문안집”, 지식재산처, <http://www.moip.go.kr/ko/kpoBultnDetail.do?menuCd=SCD0200640&parntMenuCd2=SCD0200281&aprchId=BUT0000048&pgmSeq=10148&ntatcSeq=10148>, 검색일: 2025. 12. 8.
14) 지식재산처, “통지서 표준 문안집”, 지식재산처, <http://www.moip.go.kr/ko/kpoBultnDetail.do?menuCd=SCD0200640&parntMenuCd2=SCD0200281&aprchId=BUT0000048&pgmSeq=10148&ntatcSeq=10148>, 검색일: 2025. 12. 8.
15) 지식재산처, “통지서 표준 문안집”, 지식재산처, <http://www.moip.go.kr/ko/kpoBultnDetail.do?menuCd=SCD0200640&parntMenuCd2=SCD0200281&aprchId=BUT0000048&pgmSeq=10148&ntatcSeq=10148>, 검색일: 2025. 12. 8.

있으며, 일반적으로 아래 그림과 같은 서술 패턴을 따른다. 학습데이터 구축 과정에서는 이러한 공보 정보를 활용하여 인용발명에 해당하는 문헌을 확인하고, 필요할 경우 해당 문헌의 상세한 기술 내용을 추가적으로 참조해야한다. 참고로 본 연구에서는 거절이유 중 제29조 제1항 및 제2항에 해당하는 사례만을 대상으로 한다.

<그림4 의견제출통지서 표준 문안집[16] 참조: 인용발명 정보 예시>

◈ 제29조 제1항 내지 제2항

인용발명 : 공개특허공보 제oo-oooo-oooooo호(0000. 00. 00. 공개)

◈ 제29조 제3항

인용발명 : 공개특허공보 제oo-oooo-oooooo호(0000. 00. 00. 공개 : 0000. 00. 00.자 출원서에 최초로 첨부된 명세서 및 도면에 기재된 발명과 동일)

◈ 제36조

인용발명 : 특허출원번호 제10-2008-OOOOOOO호 (0000. 00. 00. 출원)

인용발명 : 특허출원번호 제10-2008-OOOOOOO호의 0000. 00. 00.자 보정서

※ *청구항 3 발명은 인용발명의 청구항 5 발명과 실질적으로 동일합니다.*

◈ 제52조 및 제53조

원출원: 출원번호 제10-2008-OOOOOOO호 (0000. 00. 00. 출원)

※ ***이 출원의 명세서 및 도면*** *(상세한 설명 식별번호 <9>, 청구항 3)****에 기재된 OOO는 원출원의 최초로 첨부된 명세서 또는 도면에 기재되어 있지 않고****....*

## 2.3. 특허 CPC 분류 체계

특허 분류 체계는 검색과도 연관이 있다. 방대한 특허 문헌 속에서 특정 기술 분야를 효율적으로 탐색하기 위해서는 분류 체계가 체계적으로 구축되어 있어야 하며, 이는 검색의 정확성과 속도 향상에 크게 기여한다. 국가별로 특허 분류 방식에는 차이가 있으나, 국제적 호환성을 위해 일반적으로 CPC(Cooperative Patent Classification) 분류 체계를 공통 기준으로 활용한다.

CPC는 미국 특허청(USPTO)과 유럽특허청(EPO)이 공동으로 개발한 분류 체계로, 전 세계에서 가장 세분화된 기술 분류를 제공하는 시스템이다. 약 26만 개 이상의 기술을 단일 코드 체계로 표현할 수 있으며[17], 기술 분야를 매우 미세한 수준까지 구분하여 구조화할 수 있다는 특징을 갖는다.

16) 지식재산처, "통지서 표준 문안집", 지식재산처, <http://www.moip.go.kr/ko/kpoBultnDetail.do?menuCd=SCD0200640&parntMenuCd2=SCD0200281&aprchId=BUT0000048&pgmSeq=10148&ntatcSeq=10148>, 검색일: 2025. 12. 8.

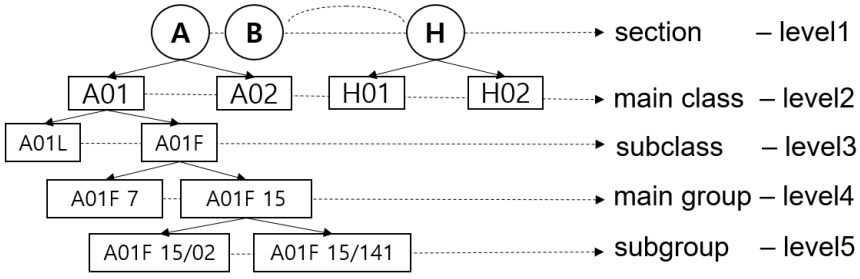
17) 지식재산처, "2020년 CPC 매뉴얼", 지식재산처, 2020, 1-68면.

<표1 CPC 단계별 카테고리 수(ver. 2023.1)>

| section | class | subclass | maingroup | subgroup (CPC) |
|---|---|---|---|---|
| A | 16 | 87 | 1,461 | 29,891 |
| B | 38 | 172 | 2,876 | 57,953 |
| C | 21 | 89 | 1,787 | 38,581 |
| D | 10 | 41 | 411 | 5,793 |
| E | 8 | 32 | 433 | 9,227 |
| F | 19 | 103 | 1,647 | 28,065 |
| G | 15 | 87 | 985 | 38,398 |
| H | 6 | 52 | 806 | 39,608 |
| Y | 3 | 11 | 349 | 18,081 |
| Total | 136 | 674 | 10,757 | 265,597 |

CPC 분류체계는 기술의 세부 수준을 표현하기 위해 계층적 구조를 갖는다. 전체 체계는 총 다섯 단계로 구성되며, 최상위 단계인 섹션(section)의 경우 8개 범주로 이루어져 대부분의 기술 분야를 포괄한다. 이후 클래스(main class), 서브클래스(subclass), 메인그룹(main group), 서브그룹(subgroup)으로 내려갈수록 분류의 수는 증가하고, 기술은 더욱 세밀한 수준으로 구분된다.

<그림5 CPC 단계[18)]>




예를 들어 CPC F16K 1/02를 살펴보면, 섹션 F는 기계공학 분야, 클래스인 F16은 기계요소, 서브클래스인 F16K는 밸브를 의미한다. 또한 메인그룹 F16K 1은 리프트 밸브, 서브그룹 F16K 1/02는 나사 스핀 구조를 나타낸다. 이와 같이 CPC는 기술을 계층적으로 세분화하여 표현함으로써, 특정 기술에 대한 정밀한 분류와 검색을 가능하게 한다.

<표2 CPC 기호 및 설명>

| | section | class | subclass | maingroup | subgroup (CPC) |
|---|---|---|---|---|---|
| 기호 | F | F16 | F16K | F16K 1 | F16K 1/02 |
| 설명 | 기계공학 | 공학적 요소 및 단위 | 밸브 | 리프트 밸브 | 스크류-스핀들 |

18) 지식재산처, “2020년 CPC 매뉴얼”, 지식재산처, 2020, 1-68면.

## 3. 관련 연구

본 연구에서 제안하는 한국어 특허 유사 기술구성요소 쌍 데이터셋은 시멘틱 텍스트 유사도(Semantic Text Similarity, STS) 유형에 해당한다. 시멘틱 텍스트 유사도란 자연어 처리에서 두 문장이 의미적으로 얼마나 유사한지를 정량적으로 평가하는 기법으로, 모델 간 성능을 공정하게 비교하기 위해 다양한 벤치마크 데이터셋이 사용된다. 대표적인 STS 데이터셋으로는 STS-B(Semantic Textual Similarity Benchmark)[19], SICK(Sentence Involving Compositional Knowledge)[20], MRPC(Mircosoft Research Paraphrase Corpus)[21], PIT(Paraphraser and Semantic Similarity in Twitter)[22] 등이 있다. 그러나 이러한 데이터셋은 일반 도메인 문장 유사도 평가를 위해 설계된 것으로, '특허'와 같은 기술적 특성과 '한글'이라는 언어적 특성을 고려한 STS 데이터셋은 아직 구축된 사례가 없다. 영문 특허를 대상으로 한 데이터셋 구축 연구는 일부 존재하며, 자동 구축 기법을 적용한 사례도 보고되고 있다. 대표적인 연구를 살펴보면 다음과 같다.

Arav Parikh 외는 ClaimCompare: A Data Pipeline for Evaluation of Novelty Destroying Patent Pairs[23]에서 신규성 거절에 해당하는 특허 쌍 대상으로 미국 특허 검색용 데이터셋 구축 방법론을 제안했다. 출원특허 1건에 대해 신규성 거절에 해당하는 positive와 신규성 거절이 아닌 negative로 자동 추출하는데, positive는 T5모델, negative는 keyBERT를 사용하였다. 최종 데이터셋은 전기, 화학분야에서 약 2만 7천여 개 기술 쌍을 구축하였다.

Grigor Aslanyan 외는 Patents Phrase to Phrase Semantic Matching Dataset[24]에서 기계적 방식과 전문가 수작업 방식을 혼합하여 미국 특허 STS 데이터셋을 구축했다. 특허 문서 내 명사구, 기능적 구 단위로 기술을 파싱하고 100번 이상 출현한 구만 필터링하여 노이즈를 제거한다. 완성된 1천여 개의 구를 anchor 문장의 기준으로 두고 유사한 구를 문헌 내에서 BERT 모델을 사용해 후보 문장을 추출한다. 그 후 전문가가 후보 문장 중 최종 target 문장을 결정한다. 최종 데이터셋은 48,548개 쌍으로 5개 컬럼으로 구성된다. anchor 문장, target 문장, CPC(클래스 레벨), 유사관계, 유사점수가 된다.

Julian Risch 외는 PatentMatch : A Dataset for Matching Patent Claims & Prior Art[25]에서 EPO 유럽 특허 문헌을 XML 파싱해서 특허 검색용 기술 쌍 데이터셋을 구축하는 방법을 제안하였다. AI 모델이나 전문가가 투입되지 않았지만 EPO 심사관이 작성한 공식 검색보고서를 활용하였다. 기계적 파싱 방법으로 출원서의 청구항, 그에 대응되는 선행 기술(신규성, 진보

19) Daniel Cer et al., "SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation", Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Association for Computational Linguistics, 2017, pp. 1-14.

20) Marco Marelli et al., "A SICK Cure for the Evaluation of Compositional Distributional Semantic Models", Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), European Language Resources Association, 2014, pp. 216-223.

21) William B. Dolan & Chris Brockett, "Automatically Constructing a Corpus of Sentential Paraphrases", Proceedings of the Third International Workshop on Paraphrasing (IWP 2005), 2005, pp. 9-16.

22) Wei Xu et al., "SemEval-2015 Task 1: Paraphrase and Semantic Similarity in Twitter (PIT)", Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Association for Computational Linguistics, 2015, pp. 1-11.

23) Arav Parikh & Shiri Dori-Hacohen, "ClaimCompare: A Data Pipeline for Evaluation of Novelty Destroying Patent Pairs", arXiv, <https://arxiv.org/abs/2407.12193>, 작성일: 2024. 6. 16.

24) Grigor Aslanyan & Ian Wetherbee, "Patent Phrase to Phrase Semantic Matching Dataset", arXiv, <https://arxiv.org/abs/2208.01171>, 작성일: 2022. 8. 1.

25) Julian Risch et al., "PatentMatch: A Dataset for Matching Patent Claims & Prior Art", arXiv, <https://arxiv.org/abs/2012.13919>, 작성일: 2020. 12. 17.

성 거절 사유), 문헌 내 문단번호로 구성된다. 최종 데이터셋은 약 625만 기술 쌍을 구축하였다.

Jaewoong Choi 외는 Deep learning-based citation recommendation system for patents[26]에서 Google BigQuery에서 미국 특허 심사관 기반 인용 데이터셋으로 PatentNet을 제안하였다. positive는 인용 쌍, negative는 인용되지 않은 쌍으로 구분하고 문헌 단위 159,157개 쌍을 구축하였다.

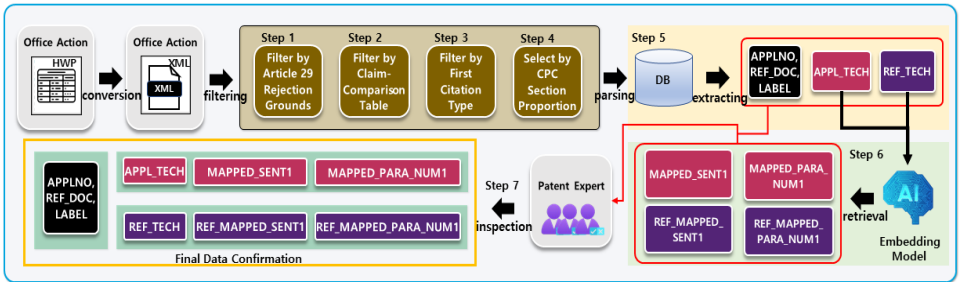
## 4. 한국어 특허 검색 데이터셋 자동구축 방법론 제안

본 절에서는 한국어 특허 검색 데이터셋을 자동으로 구축하기 위한 전체적인 방법론을 제시한다. 또한 자동 구축된 결과의 품질을 검증하기 위해 별도로 구축한 특허 전문가 수작업 데이터셋의 구성 방식도 함께 설명한다.

### 4.1. 한국어 특허 검색 데이터셋 자동구축

제안하는 자동 구축 파이프라인은 7단계로 구성된다. 전체적인 구조는 기존 연구들[27][28][29]과 유사하나 5단계 한국어 특허 전용 XML파싱, 6단계 KorPatBERT[30] CPC 분류 모델과 유사도 알고리즘을 혼합하여 데이터 자동 추출한 작업, 7단계 전문가 검수 절차를 강화한 점에서 차별점이 있다. 1~4단계는 적합한 문헌을 선별하는 단계이며, 5~6단계는 파싱 알고리즘과 AI 모델을 활용해 학습 항목을 자동으로 추출한다. 7단계는 특허 전문가가 추출 결과를 검수·보정하여 최종 데이터셋을 확정하는 과정이다.

<그림6 AI기반 자동구축 데이터셋 구축 절차>




작업 대상은 발송연도 기준 2007~2024년까지의 원시 통지서 데이터 총 298만 건이다. 1단계에서는 특허법 제29조에 해당하는 거절 사유 문헌만을 남기기 위한 1차 필터링을 수행한다. 2단계는 구성대비표의 존재 여부를 판별하고 XML을 파싱하는 과정이다. 표 태그를 탐색하여 헤더 구성과 라벨 존재 여부를 기반으로 구성대비표를 판별하여 정형데이터를 선별된다. 3단계는

26) Jaewoong Choi et al., "Deep learning-based citation recommendation system for patents", arXiv, <https://arxiv.org/abs/2010.10932>, 작성일: 2020. 10. 21.
27) Arav Parikh & Shiri Dori-Hacohen, "ClaimCompare: A Data Pipeline for Evaluation of Novelty Destroying Patent Pairs", arXiv, <https://arxiv.org/abs/2407.12193>, 작성일: 2024. 6. 16.
28) Grigor Aslanyan & Ian Wetherbee, "Patent Phrase-to-Phrase Semantic Matching Dataset", arXiv, <https://arxiv.org/abs/2208.01171>, 작성일: 2022. 8. 1.
29) Julian Risch et al., "PatentMatch: A Dataset for Matching Patent Claims & Prior Art", arXiv, <https://arxiv.org/abs/2012.13919>, 작성일: 2020. 12. 17.
30) 박진우 외 4인, "한국어 특허 문장 기반 CPC 자동분류 연구 - 인공지능 언어모델 KorPatBERT를 활용한 딥러닝 기법 접근", 「지식재산연구」, 제17권 제3호(2022), 209-256면.

구성대비표에 포함된 제1선행발명이 국내 또는 해외 문헌인지 확인하며, 국내 문헌을 우선적으로 선택한다. 4단계는 CPC 섹션별 연간 평균 출원 비중을 고려하여 데이터의 분포를 균형 있게 선별하는 과정이다. 이 과정을 통해 약 70만 건이 선별된다. 5단계에서는 파싱한 구성대비표로부터 기술 구성요소, 인용 정보 등 검색 모델 학습에 필요한 데이터를 자동 추출한다. 이는 한국어 특허의 언어적 특징을 고려하여 문서를 파싱 방법으로 그림 7의 (a), (b)의 알고리즘을 통해 확인할 수 있다.

<그림7 (a): 문서 내 구성대비표 검출 알고리즘, (b): 구성대비표에서 기술적 구성요소를 추출하기 위한 알고리즘>

```
Checking compared-composition Tables
Input : XML file
Output : Set of compared-composition tables

1 : Load XML tree from file
2 : Define compared-composition keyword ← {"청구항","인용발명"}
3:  Define relevance keyword ← {"동일","유사","차이"}
4:  Initialize Results
5:  For each top-level TableBody T do
6:      first row ← first TableRow of T
7:      if any cell in first row contains compared-composition keyword then
8:          table data ← ParseTable Function(T)
9:          relation count ← count of rows whose last column contains relevance keyword
10:         if relation count ≥ (|table data |) − 1)/2 then
11:             Results ← Results ∪ { table data }
12: return Results
```

(a)

```
Mapping column indices based on table header
Input : table header, col data list
Output : claim index, citation index, relevance index

1 : Initialize all indices to -1
2 : Detect relevance index by scanning col data list for any relevance-related keyword
3:  For each head cell :
4:      Normalize text via CleanText Function.
5:      If Text includes any citation-related keyword ( excluding "대비", "비교"):
6:          update citation index; mark conflict if multiple inconsistent matches occur.
7:      Else if text includes any claim-related keyword or matches claim patterns:
8:          Set claim index.
9: Return claim index, citation index, relevance index
```

(b)

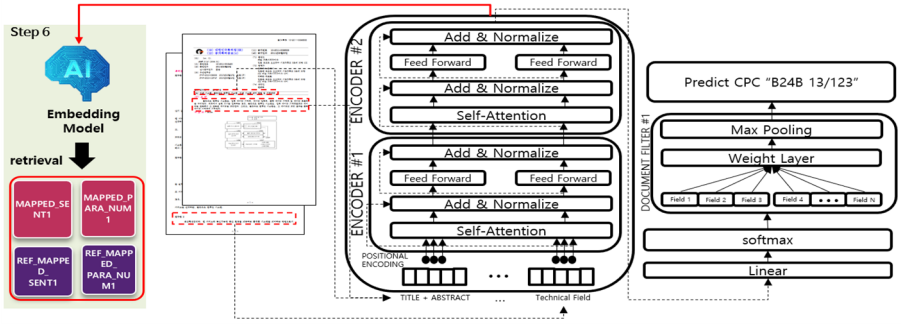
6단계는 특허 AI 모델을 활용해 구성대비표에서 추출된 기술 구성요소와 유사한 문장을 특허 공보에서 자동 추출하는 단계이다. 유사 문장은 발명의 상세한 설명 필드를 우선적으로 탐색하며, 해당 필드가 존재하지 않을 경우 다른 필드에서 검색한다.

사용된 특허 AI 모델은 과거 CPC 자동분류 방법론 연구[31]에서 제안한 CPC 주·부분류 모델을 사용하였다. 간단히 설명하면, 한국어 특허 문헌 전체를 사전 학습한 KorPatBERT 모델을 기반으로 CPC 분류 문제를 미세학습한 모델로, 관련 기존 연구들[32][33][34][35]을 비교해서 가장 성능이 좋은 방법론의 모델이다. 본 연구에서는 2024년까지 출원된 특허 문헌을 추가 학습하여 차용된 연구보다 더 많은 데이터를 학습하였다. CPC 학습데이터는 문서 내 11개 필드 중 '발명의 상세한 설명'을 제외한 10개 필드를 대상으로 CPC 코드를 라벨링하여 구성하였다. 5,088,350건 문헌의 32,374,607건 데이터를 학습에 활용하였으며, 최종 레벨인 서브그룹 기준 117,066종(서브클래스 657종, 메인그룹 8,682종)을 분류하였다. Y섹션의 경우, 분류 당 문서 수가 다른 분류에 비해서 상대적으로 매우 적은 문제로 딥러닝 학습이 진행 불가능한 상황이고, 융합 기술이기 때문에 특정 기술을 찾아야하는 검색 문제에서는 오히려 성능을 하락시키는 요인이 되기 때문에 제외하였다. 모델 성능을 보면, 최근 10년동안 출원한 CPC 비율과 동일한 비율로 구성되고 학습하지 않은 50만건의 평가 데이터셋으로 CPC 분류 평가 시 Recall 기준 서브클래스 Top3 96.59%, 메인그룹 Top5 93.93%, 서브그룹 Top10 79.67% 성능을 보였다. 이는

31) 박진우 외 4인, "한국어 특허 문장 기반 CPC 자동분류 연구 – 인공지능 언어모델 KorPatBERT를 활용한 딥러닝 기법 접근", 「지식재산연구」, 제17권 제3호(2022), 209-256면.
32) 김용일 외 4인, "딥러닝-규칙 기반 병행 모델을 이용한 특허문서의 자동 IPC 분류 방법", 제31회 한글 및 한국어 정보처리 학술대회 논문집, 한국정보과학회 언어공학연구회, 2019, 347-350면.
33) 박진우 외 3인, "Patent Tokenizer: 형태소와 SentencePiece를 활용한 특허문장 토크나이즈 최적화 연구", 제37회 한글 및 한국어 정보처리 학술대회 논문집, 한국정보과학회 언어공학연구회, 2020, 441-445면.
34) 심우철 외 4인, "한국 특허문헌 특성 및 딥러닝 기반 분류모델을 고려한 CPC 자동분류에 관한 연구", 한국정보과학회 2020 한국소프트웨어종합학술대회 논문집, 2020, 406-408면.
35) 곽민학 외 3인, "한국어 특허 언어모델 Scaling-up에 관한 연구", 한국정보과학회 2023 한국소프트웨어종합학술대회 논문집, 2023, 621-623면.

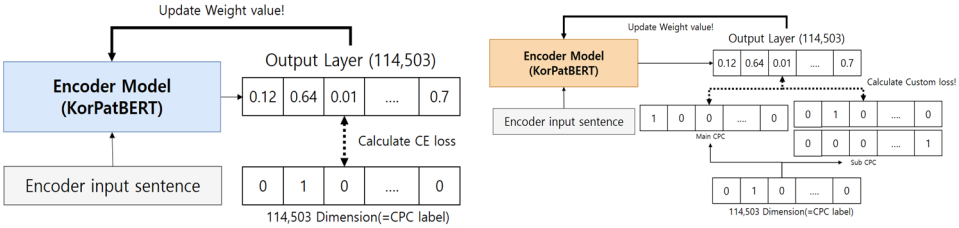
차용한 연구 방법론을 적용한 모델[36]보다 서브클래스 2.74% 상승, 메인그룹 7.36% 상승한 수치이다.

<그림8 CPC 분류 모델 아키텍처>




KorPatBERT CPC 모델은 문서 단위의 CPC를 예측하지만 입력 데이터는 필드 단위이므로, 최종 레이어에 필드 예측을 문서 단위 예측으로 변환하기 위한 문서 필터 모듈(sum layer, weight layer)을 추가하였고, 그림8에서 확인할 수 있다. CPC 필드 조합 최적화 실험 결과, 명칭요약·기술분야·배경기술·청구항·해결하려는 과제·해결수단·발명의 구성 및 작용·도면의 간단한 설명·발명의 효과 등 10개 필드를 사용할 때 가장 우수한 성능을 보여 해당 조합을 자동분류에 사용하였다.

<그림9 (a): CPC 메인 CPC 코드 학습을 최적화하기 위한 손실 함수, (b): 메인 & 서브 CPC>




추가적으로, KorPatBERT 기반 CPC 분류 모델은 단일 기술 분류(멀티 클래스 분류)가 아니라 여러 기술을 동시에 분류(멀티 레이블 분류)하는 모델이다. 한국특허기술진흥원 CPC 분류 실무자의 자문을 바탕으로 분석한 결과, 하나의 문서에는 복수의 CPC 코드가 부여(최대 50개)되며, 이때 CPC 분류 체계에는 주분류 코드(main CPC)와 부분류 코드(sub CPC)가 존재한다. 주분류 코드는 문서 내에서 주된 기술이 되는 코드이며 여러 CPC 코드 중 반드시 하나만 존재해야 하는 특징을 가진다.

주·부분류를 모두 분류하기 위해서 일반적인 멀티 레이블 분류 손실 함수를 그대로 적용할 경우 주분류만 학습한 모델보다 예측 성능이 떨어지는 문제가 있다. 또한 주분류만 학습한 모델을 사용하기에는 부분류를 예측할 수 없는 문제가 있기 때문에 적절하지 않은 상황이다. 이런 현상은 과거 CPC 자동분류 연구[37]에서도 잘 설명하고 있다. 이를 해결하기 위해, 본 연구에서는 주

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

분류 코드를 더 정확하게 학습하도록 설계한 맞춤형 손실 함수(Custom loss)를 적용하였다. 손실 계산 과정에서 주분류 코드에 더 큰 가중치를 부여함으로써 모델이 주분류를 우선적으로 학습하도록 유도하였으며, 이를 통해 전체 학습 효율을 향상시켰다. 표3을 보면, 주분류 모델이 주·부분류 모델보다 성능이 높지만, 맞춤형 손실 함수를 적용하면 주·부분류 모델이 주분류 모델보다 성능이 높아지는 것을 확인할 수 있다. 평가데이터는 최근 10년동안 출원한 문서의 CPC 비율과 동일한 비율로 구성되고 학습하지 않은 24만건이다.

<표3 주분류 모델과 주·부분류 모델의 서브그룹 성능 비교>

| 분류방식 | 손실함수 | 평가(건) | 분류종수 | Top1 | Top3 | Top5 | Top10 |
|---|---|---|---|---|---|---|---|
| 주분류 | CE loss | 241,623 | 64,000 | 39.71 | 61.73 | 70.80 | 80.83 |
| 주·부분류 | CE loss | | 117,066 | 39.53 | 61.01 | 69.80 | 79.70 |
| 주·부분류 | Custom loss | | 117,066 | 40.05 | 62.05 | 71.10 | 81.21 |

참고로, CPC 분류는 본질적으로 계층적 다중 레이블 분류 문제를 내포하고 있다. CPC의 특징인 계층적 제약이나 상·하위 제약을 고려하기 위해 KorPatBERT 서브그룹 분류 방식을 대표적인 계층적 학습 방식의 모델 2종과 비교하였다. KE-T5모델은 생성형 모델로 CPC레벨마다 분류 예측을 학습한 방식이고, mBERT는 CPC레벨 별 분류 예측을 병렬적으로 학습한 방식이다. 표4을 보면, 한국어 특허 문헌의 경우 특이하게도 CPC 계층적 학습 방식이 非계층적 학습 방식보다 더 낮은 성능을 보인 것을 확인할 수 있다.

<표4 한국어 특허 문헌의 CPC 계층적 학습 방식과 非계층적 학습 방식 성능 비교>

| 학습 방식 | 언어모델 | 평가(건) | 분류종수 | Top1 | Top3 | Top5 | Top10 |
|---|---|---|---|---|---|---|---|
| 계층적 | KE-T5 | 241,623 | 117,066 | 26.53 | 42.85 | 49.66 | 57.63 |
| 계층적 | mBERT | | | 38.78 | 59.86 | 68.65 | 78.38 |
| 비계층적 | KorPatBERT | | | 40.05 | 62.05 | 71.10 | 81.21 |

CPC 주·부분류 모델의 맞춤형 손실함수(Custom loss)의 최적화값은 실험을 통해 확인하였다. 부분류 대비 주분류에 1.0~3.0까지 가중치 학습하였을 때 2.0 가중치가 가장 높은 성능을 보였다. 표5을 보면, 주·부분류 모델 손실함수 가중치 범위별 모델 성능을 평가한 결과이다. 2.0 가중치가 가장 높은 성능임을 확인할 수 있다.

<표5 주·부분류 모델 손실함수 가중치 범위별 모델 성능 평가>

| | 가중치 값 | 평가(건) | 분류종수 | Top1 | Top3 | Top5 | Top10 |
|---|---|---|---|---|---|---|---|
| KorPatBERT | 1.5 | 241,623 | 117,066 | 39.63 | 61.59 | 70.45 | 80.67 |
| | 1.75 | | | 40.03 | 62.01 | 71.05 | 81.18 |
| | 2.0 | | | 40.05 | 62.05 | 71.10 | 81.21 |
| | 2.25 | | | 39.97 | 62.81 | 70.87 | 80.81 |
| | 2.5 | | | 39.86 | 62.74 | 70.64 | 80.68 |

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

모델 학습은 GPU 8개 환경에서 수행되었으며, 러닝레이트 0.00003, 에폭 30, 배치 크기 3,328, 입력 길이 256 토큰으로 설정하여 CPC 서브그룹 레벨의 분류를 학습하였다.

이렇게 학습된 KorPatBERT 기반 CPC 분류 모델은 이후 의견제출통지서에서 구성대비표의 기술 구성요소와 가장 유사한 문장을 탐색하기 위해 활용된다. 이를 위해 특허 공보 전체 문장을 분리한 뒤, 각 문장의 임베딩 벡터를 추출한다. 이후 코사인 유사도와 자카드 유사도를 결합하여 유사 문장을 탐색하며, 두 유사도의 가중치 비율은 특허 문헌의 의미적 유사성과 어휘적 중복성을 균형있게 반영하기 위해 사전 실험을 통해 결정하였다. 비율을 변화시키며 Top-K 정확도를 비교하였으며 결과는 표6에 제시하였다.

본 사전 실험에는 2024년에 지식재산처 심사관이 구축한 구성요소-유사문장 데이터셋을 활용하였다. 해당 데이터셋은 C섹션 496건, H섹션 617건으로 총 1,113건의 문헌에서 추출된 10,964개의 정답 쌍으로 구성되어 검증의 객관성을 확보하였다.

사전 실험 결과 의미 정보와 어휘 정보를 8:2 비율로 혼합하였을 때 Top-K 매칭 정확도가 가장 높게 나타났으며, 해당 비율을 최종 가중치로 채택하였다. 계산된 유사도 점수는 문장 단위로 정렬되며, 가장 높은 점수를 가진 문장을 유사 문장으로 선정한다.

<그림10 문장 유사도 검출을 위한 자카드 유사도 알고리즘>

```
JaccardChar(technical element, sentence)
Input : Strings technical element, sentence
Output : Jaccard Similarity score between technical element and sentence

1 : A ← set of characters in technical element
2 : B ← set of characters in sentence
3:  Intersection size  ← A ∩ B
4:  Union size ← A ∪ B
5:  Similarity score  ← |Intersection size | / |Union size |
6: return Similarity score
```

<표6 코사인-자카드 유사도 가중치 비율에 따른 유사 문장 추출 성능>

| 코사인:자카드 | 5:5 | 6:4 | 7:3 | 8:2 | 9:1 |
|---|---|---|---|---|---|
| Top-1 | 54.23 | 54.82 | 55.01 | 55.26 | 54.47 |
| Top-3 | 80.18 | 80.31 | 80.49 | 80.54 | 79.12 |
| Top-5 | 86.59 | 86.84 | 87.02 | 87.31 | 86.03 |

7단계는 특허 전문가가 자동 추출된 결과를 검토하는 단계로, 의견제출통지서에서 및 특허 공보에서 파싱된 문장 및 기술 요소가 적절한지 확인하고 필요한 경우 수정한다. 검토 과정에서 식별된 보완 사항은 파싱 알고리즘에 반영하여 전체 구축 시스템을 재정비하였다. 검수는 전체 구축 데이터 중 10%를 무작위로 샘플링하여 진행하였다.

<그림11 7단계 검수 시스템 UI>

이와 같은 자동 구축 방법론을 통해 문헌 기준 404,721건, 기술 구성요소 기준 1,633,626개의 데이터가 최종 구축되었다.

### 4.2. 한국어 특허 검색 자동구축 데이터셋 항목 상세

자동 구축된 데이터셋은 총 9개 항목으로 구성되며, 각 항목의 의미와 데이터 출처는 표7에 정리하였다. 각 컬럼은 특허 검색 모델 학습에 필수적인 핵심 정보를 담고 있다.

<표7 CPC 분류 모델 기반 자동구축 데이터셋의 컬럼 구성>

| 컬럼 | 설명 | 출처 |
|---|---|---|
| APPLNO | 특허 출원에 대한 고유 식별자 | 의견제출통지서 |
| APPL_TECH | 청구항으로부터 분리·식별된 개별 구성요소 | 의견제출통지서, 특허 출원서 |
| MAPPED_SENT1 | 특허 문서에서 해당 구성요소에 대한 상세 설명을 포함한 문장 | 특허 공보문서, 특허 출원서 |
| MAPPED_PARA_NUM1 | 구성요소 설명이 위치한 문단번호 | 특허 공보문서, 특허 출원서 |
| REF_DOC | 출원발명에 대응되는 인용발명의 공보번호 | 의견제출통지서 |
| REF_TECH | 출원발명 구성요소에 대응되는 인용발명의 구성요소 | 의견제출통지서, 특허 공보문서 |
| REF_MAPPED_SENT1 | 인용발명 구성요소에 대한 상세 설명을 포함한 문장 | 특허 공보문서 |
| REF_MAPPED_PARA_NUM1 | 인용발명 구성요소 설명이 위치한 문단번호 | 특허 공보문서 |
| LABEL | 출원발명 구성요소와 인용발명 구성요소간 유사도 | 의견제출통지서 |

자동구축 학습데이터 구성항목은 총 9개 컬럼으로 구성된다. 출원번호(APPLNO)는 대상 출원발명 문헌의 고유 식별번호를 의미한다. 출원발명 기술 구성요소(APPL_TECH)는 통지서에서 구성대비표 내 출원발명의 기술 구성요소를 의미한다. 이 정보는 5단계에서 추출된다. 기술구성요소와 유사한 문장(MAPPED_SENT1)과 기술구성요소와 유사한 문장의 문단번호(MAPPED_PARA_NUM1)는 6단계에서 특허 AI모델을 활용하여 해당 공보 내 유사한 문장을 추출한 정보이다. 인용발명 공보번호(REF_DOC)는 통지서 내 서두에 있는 인용발명 번호를 의미한다. 인용 발명 기술 구성요소(REF_TECH), 인용발명 기술구성요소와 유사한 문장(REF_MAPPED_SENT1), 인용발명 기술구성요소와 유사한 문장의 문단번호(REF_MAPPED_PARA_NUM1), 관련도(LABEL) 모두 5단계에서 추출된다. 관련도는 출원발명의 기술 구성요소와 인용발명의 기술 구성요소의 유사 관련 정도를 의미한다. 관련도는 '실질적 동일', '동일', '일부차이', '차이' 4개 단계로 유사 정도가 구분된다.

### 4.3. 전문가 수작업 데이터셋 구축

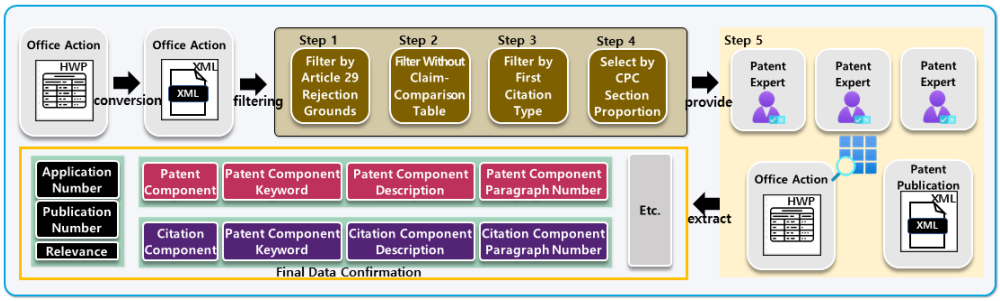
자동 구축된 데이터셋의 품질을 확인하기 위해서 기존 연구들에서는 데이터셋을 직접 전문가가 평가[38]하거나 하고 검색 성능[39]을 통해 비교·평가하였다. 본 연구에서는 검색 성능을 통

38) Arav Parikh & Shiri Dori-Hacohen, "ClaimCompare: A Data Pipeline for Evaluation of Novelty Destroying Patent Pairs", arXiv, <https://arxiv.org/abs/2407.12193>, 작성일: 2024. 6. 16.
39) Julian Risch et al., "PatentMatch: A Dataset for Matching Patent Claims & Prior Art", arXiv, <https://arxiv.org/abs/2012.13919>, 작성일: 2020. 12. 17.

해 평가한다.

전문가 수작업 데이터셋은 의견제출통지서 중 구성대비표가 포함되지 않은 문헌을 대상으로, 검색에 필요한 핵심 정보를 전문가가 직접 추출하여 구축하는 방식으로 마련하였다. 전체 절차는 총 5단계로 구성되며, 1~4단계는 자동 구축 과정과 동일한 방식으로 필터링을 수행하고, 5단계에서 전문가의 수작업 구축이 이루어진다.

<그림12 전문가 수작업 데이터셋 구축 절차>




1~4단계의 기본적인 흐름은 자동 구축 과정과 유사하다. 1단계는 특허법 제29조에 해당하는 거절 사유 문헌만을 남기고 필터링하는 단계로, 자동 구축 방법과 동일하다. 2단계는 자동 구축 과정과 다소 차이가 있다. 전문가 수작업 데이터셋은 구성대비표가 없는 비정형 문헌을 대상으로 구축된다. 3단계는 제1인용발명이 국내 문헌인 경우만을 선별하며, 4단계는 CPC 섹션별 비중을 구분하는 단계로, 자동 구축 방법과 동일하다. 이 과정에서 약 53만 건으로 축소된다.

5단계는 특허 전문가가 직접 통지서 내 기술 구성요소, 인용 정보 등 검색 성능 향상에 필요한 학습 데이터 요소를 수작업으로 추출하는 단계이다. 이 작업에는 선행기술조사·분석 분야의 전문 인력 약 30명이 기술 분야별로 참여하였다.[40)]

정리하면, 전문가 수작업 데이터셋 구축은 구성대비표가 없는 비정형 통지서를 대상으로 전문가가 직접 정보를 추출한다는 점에서 고품질의 데이터셋으로 평가할 수 있다. 다만, 전 과정이 수작업으로 이루어지기 때문에 구축 규모는 상대적으로 제한적이며, 최종적으로 문헌 기준 20,934건, 기술 구성요소 기준 85,272개가 구축되었다.

## 4.4. 전문가 수작업 데이터셋 구축 항목 상세

전문가 수작업 데이터셋은 총 15개 항목으로 구성되며, 각 항목의 정의와 데이터 출처는 표8에 정리하였다.

40) 한국특허진흥원, “특허 선행기술조사”, 한국특허진흥원, <https://www/kipro.or.kr/business/priorArtSearch>, 검색일: 2026. 1. 22.

<표8 전문가 수작업 데이터셋의 컬럼 구성>

| 컬럼 | 설명 | 출처 |
|---|---|---|
| 출원발명_출원번호 | 특허 출원에 대한 고유 식별자 | 의견제출통지서 |
| 출원발명_공보번호 | 출원발명의 공개번호 또는 등록번호 | 특허 공보문서 |
| 출원발명_청구항 | 비교 대상으로 선택된 특정 청구항 번호 | 의견제출통지서, 특허 출원서 |
| 순번 | 출원별 문장 쌍의 순번 | - |
| 출원발명_구성요소 | 청구항으로부터 분리·식별된 개별 구성요소 | 의견제출통지서, 특허 출원서 |
| 출원발명_구성요소_핵심키워드 | 출원발명 구성요소의 기술적 의미를 대표하는 핵심 기술 용어 | 의견제출통지서, 특허 출원서 |
| 출원발명_핵심기술여부 | 해당 구성요소가 특허의 기술적 핵심을 나타내는지 여부 | 의견제출통지서, 특허 출원서 |
| 출원발명_구성요소설명 | 특허 문서에서 해당 구성요소에 대한 상세 설명을 포함한 문장 | 특허 공보문서, 특허 출원서 |
| 출원발명_구성요소설명_문단번호 | 구성요소 설명이 위치한 문단번호 | 특허 공보문서, 특허 출원서 |
| 인용발명_공보번호 | 출원발명에 대응되는 인용발명의 공보번호 | 의견제출통지서 |
| 인용발명_구성요소 | 출원발명 구성요소에 대응되는 인용발명의 구성요소 | 의견제출통지서, 특허 공보문서 |
| 인용발명_구성요소_핵심키워드 | 인용발명 구성요소의 기술적 의미를 대표하는 핵심 기술 용어 | 의견제출통지서, 특허 공보문서 |
| 인용발명_구성요소설명 | 인용발명 구성요소에 대한 상세 설명을 포함한 문장 | 특허 공보문서 |
| 인용발명_구성요소설명_문단번호 | 인용발명 구성요소 설명이 위치한 문단번호 | 특허 공보문서 |
| 관련도 | 출원발명 구성요소와 인용발명 구성요소간 유사도 | 의견제출통지서 |

전문가 수작업 학습데이터 구성항목은 총 15개 컬럼으로 구성된다. 출원발명_출원번호, 출원발명_공보번호, 출원발명_청구항, 순번, 출원발명_구성요소, 출원발명_구성요소_핵심키워드, 출원발명_핵심기술_여부, 출원발명_구성요소설명, 출원발명_구성요소설명_문단번호, 인용발명_공보번호, 인용발명_구성요소, 인용발명_구성요소_핵심키워드, 인용발명_구성요소설명, 인용발명_구성요소설명_문단번호, 관련도가 있다.

전문가 수작업 데이터셋은 사람이 직접 구축한 만큼, 핵심 키워드나 핵심 기술 여부 같은 추가 정보가 포함되어 있어 고품질 데이터셋으로 평가된다. 이러한 요소는 데이터셋의 정밀도를 높일 뿐 아니라, 후속 연구 및 분석에 활용될 수 있다. 반면, 자동 구축 데이터셋은 동일한 정보 구조를 유지하면서도 대규모 데이터를 효율적으로 생성할 수 있는 장점이 있다. 본 연구에서 두 데이터셋의 규모는 약 19배 정도의 차이가 있음을 확인하였다.

### 4.5. 최종 구축된 데이터셋 정리 및 통계

자동 구축 데이터셋이 문헌 기준 총 404,721건, 기술구성요소 기준 1,633,626개, 전문가 수작업 데이터셋은 문헌 기준 20,934건, 기술구성요소 기준 85,272개가 구축되었다.

<표9 전문가 수작업, CPC 모델 기반 자동구축 데이터셋의 섹션별 문헌 수 및 비율>

| | 전문가 수작업 데이터셋 | | CPC 모델 기반 자동구축 데이터셋 | |
|---|---|---|---|---|
| Section | 문헌 수 | 비율 (%) | 문헌 수 | 비율 (%) |
| A | 4,186 | 20.00 | 54,853 | 13.55 |
| B | 4,011 | 19.16 | 66,448 | 16.42 |
| C | 2,614 | 12.49 | 26,661 | 6.59 |
| D | 399 | 1.91 | 5,883 | 1.45 |
| E | 887 | 4.24 | 14,253 | 3.52 |
| F | 1,820 | 8.69 | 26,388 | 6.52 |
| G | 3,098 | 14.80 | 103,110 | 25.48 |
| H | 3,919 | 18.72 | 107,125 | 26.47 |
| Total | 20,934 | 100.0 | 404,721 | 100.0 |

## 5. 실험

본 절에서는 앞 절에서 구축한 자동구축 데이터셋과 전문가 수작업 데이터셋으로 모델을 학습한다. 기본 언어 모델로는 일반 도메인 텍스트를 대상으로 사전학습된 BERT 대신, 특허 문헌을 대상으로 사전학습된 KorPatBERT를 base 모델로 사용한다. 이는 특허 문헌이 일반 텍스트와 비교하여 용어 분포 및 문체적 특성이 상이한 도메인 특화 문서라는 점을 고려할 때, 도메인 특화 사전학습 모델이 검색 성능 향상에 보다 유리할 것으로 판단하였기 때문이다. 또한 base 모델에 CPC 분류 정보를 추가적으로 학습한 모델을 비교군으로 포함하여, 분류 정보 활용 여부가 특허 검색 성능에 미치는 영향을 함께 분석한다. 이러한 설정을 통해 본 실험에서 사용된 검색 모델은 총 8종으로 구성된다. 더불어 검색 평가를 위해 평가 데이터셋도 구축한다.

<표10 실험에 사용된 모델 및 STS 학습 데이터셋>

| 순번 | 모델 | STS 학습 데이터셋 |
|---|---|---|
| 1 | KorPatBERT | - |
| 2 | KorPatBERT + STS 자동구축 | 자동구축 데이터셋 |
| 3 | KorPatBERT + STS 전문가구축 | 전문가구축 데이터셋 |
| 4 | KorPatBERT + STS 혼합 | 자동구축 & 전문가구축 혼합 데이터셋 |
| 5 | KorPatBERT + CPC | - |
| 6 | KorPatBERT + CPC + STS 자동구축 | 자동구축 데이터셋 |
| 7 | KorPatBERT + CPC + STS 전문가구축 | 전문가구축 데이터셋 |
| 8 | KorPatBERT + CPC + STS 혼합 | 자동구축 & 전문가구축 혼합 데이터셋 |

구체적인 실험 목적은 3가지가 있다. 첫 째, 자동구축 데이터셋과 전문가 수작업 데이터셋을 각각 STS 미세학습한 검색 모델 간의 검색 성능을 비교하는 것이다. 둘 째, 자동구축 데이터셋을 STS 미세학습한 모델과 학습하지 않은 모델의 검색 성능을 비교하여 검색 학습용 데이터셋의 유효성을 확인하는 것이다. 셋 째, STS 미세학습 외에 CPC 분류 미세학습[41]을 선행 작업하여도 데이터셋의 성능이 유지되는 지 확인한다.

41) 민재욱 외 4인, "KorPatBERT 기반 CPC 분류 모델을 활용한 한국어 특허 문헌 검색 성능 향상 연구", 「지식재산연구」, 제20권 제1호(2025), 89-117면.

실험을 위해 먼저 최종 구축된 자동구축 데이터셋을 문서 기준 9:1 비율로 학습/검증데이터로 분할하며, 전문가 수작업 데이터셋도 동일하게 분할한다. 여기서 사용되는 검증데이터는 최적 에폭 학습을 찾기 위함이다.

학습 데이터는 라벨이 '동일', '실질적동일'인 구성요소 쌍을 사용하고, 검증 데이터는 출원발명 출원번호, 인용발명 출원번호 쌍을 사용한다. 검증 데이터는 실제 검색 시스템이 활용되는 평가 환경과 동일한 구조로 구성하여, 모델이 학습한 유사도 표현이 문헌 검색 성능으로 얼마나 효과적으로 이어지는지를 확인하기 위한 목적의 데이터셋이다. 모델 학습에 활용될 데이터셋에서 '동일', '실질적동일' 데이터만 사용한 것은 손실함수인 InBatch InfoNCE[42)43)44)]의 특성에 기인한다. 이는 하나의 query, positive로 구성되고, 배치 내 데이터를 negative로 활용하여 대조학습을 수행하는 방식이다.

<표11 전문가 수작업, 자동구축 데이터셋의 학습/검증 분할별 문헌(구성요소) 수 및 비율>

| | 전문가 수작업 데이터셋 | | CPC 모델 기반 자동구축 데이터셋 | |
|---|---|---|---|---|
| 데이터셋 | 문헌(구성요소) 수 | 비율 (%) | 문헌(구성요소) 수 | 비율 (%) |
| Train | 364,253(1,475,235) | 90 | 18,846(76,389) | 90 |
| Dev | 40,468(158,391) | 10 | 2,088(8,883) | 10 |
| Total | 404,721(1,633,626) | 100 | 20,934(85,272) | 100 |

학습에 사용된 하이퍼 파라미터는 다음과 같다. 토크나이저의 최대 길이는 512, 배치 크기는 128, 에폭은 30으로 설정한다. 학습 자동 스케줄링을 적용해서 러닝레이트 최고값은 9e-6, 최소값은 최고값의 30%로 설정하고 웜업 스텝을 200, 전체 스탭 20,000으로 설정한다. 손실함수는 InBatch InfoNCE를 사용했고 temperature값은 0.07로 설정한다.

<표12 모델 학습에 사용된 하이퍼 파라미터 및 값>

| 하이퍼 파라미터 | 값 | 하이퍼 파라미터 | 값 |
|---|---|---|---|
| Epoch | 30 | Total Steps | 20,000 |
| Max Length | 512 | Optimaizer | AdamW |
| Batch Size | 128 | Weight Decay | 1e-3 |
| Learning Rate Max | 9e-6 | loss Function | InBatchInfoNCE Matron |
| Learning Rate Min | 9e-6*0.3 | InfoNCE temperature | 0.07 |
| Warmup Steps | 200 | InfoNCE Dim list | [256, 512, 768] |

최대 30 에폭으로 학습하며, 매 에폭마다 검증 데이터셋을 활용해 검색 성능 평가를 수행하여 최적의 체크포인트를 선정하였다.

검색 성능 평가 지표는 기존 연구들[45)46)47)]에서 가장 많이 사용되는 Top-k Accuracy[48)]를

42) Aditya Kusupati et al., "Matryoshka Representation Learning", arXiv, <https://arxiv.org/abs/2205.13147>, 작성일: 2024. 2. 8.
43) Xianming Li et al., "2D Matryoshka Sentence Embeddings", arXiv, <https://arxiv.org/abs/2402.14776>, 작성일: 2024. 11. 30.
44) Aaron van den Oord et al., "Representation Learning with Contrastive Predictive Coding", arXiv, <https://arxiv.org/abs/1807.03748>, 작성일: 2019. 1. 22.
45) Arav Parikh & Shiri Dori-Hacohen, "ClaimCompare: A Data Pipeline for Evaluation of Novelty Destroying Patent Pairs", arXiv, <https://arxiv.org/abs/2407.12193>, 작성일: 2024. 6. 16.

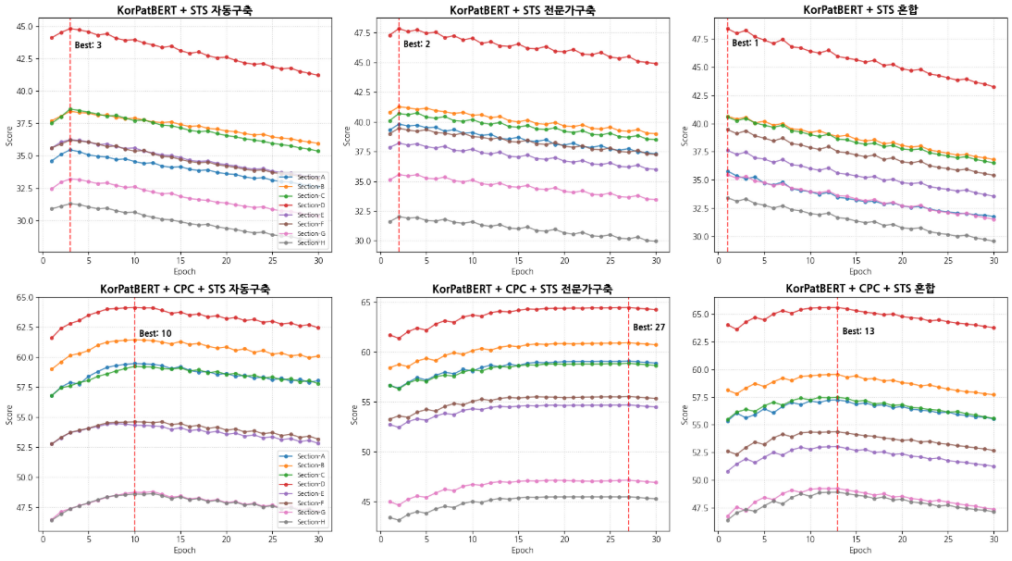
적용하였다. Top-k Accuracy(%) 공식은 다음과 같이 정의된다.

$$\text{Top-K } Accuracy = \frac{1}{N}\sum_{i=1}^{N} y_i \in \hat{Y}_i^K \quad (1)$$

여기서 변수 $N$는 전체 쿼리 수로 즉, 평가데이터 문헌 수가 된다. $y_i$는 i번째 쿼리로 평가데이터의 정답이 된다. $\hat{Y}_i^K$는 모델이 반환한 상위 K개의 검색 결과이다.

검색 실험 결과는 CPC 섹션 별로 구분하였고 Top-70의 평가 결과를 대표적인 지표로 사용하였다. 이는 실제 특허 검색 및 검토 과정에서 상위 수십 건의 문헌이 주로 검토 대상이 된다는 점을 반영한 것으로, 지식재산처 심사관 인터뷰를 통해 확인된 실무적 검색·검토 관행에 근거한다.[49] 따라서 본 지표는 모델의 실질적인 검색 효용성을 평가하기 위한 적절한 기준으로 판단된다. 에폭 별 검색 성능 평가 추이는 그림13과 같다.

<그림13 모델-데이터셋 조합별 학습 에폭에 따른 검색 성능 평가 점수>




검색 실험을 준비를 위해 검색 문헌의 전체 모수 풀을 구축한다. 정답을 찾아야하는 검색 모수는 1946년부터 2024년까지 전체 특허 문서로 5,470,177건이다. 이 전체 검색 모수에는 학습 데이터셋과 이후 구축할 평가데이터셋이 모두 포함되어 있다. 전체 특허 문서의 CPC 섹션 별 건수는 표13을 보면 확인할 수 있다.

다음은 평가데이터를 구축한다. 평가데이터를 구축할 때는 학습데이터가 되는 자동구축 데이터셋과 전문가 수작업 데이터셋을 포함시키지 않는다. 또한 전체 실험 데이터의 섹션 별 비율

46) Julian Risch et al., “PatentMatch: A Dataset for Matching Patent Claims & Prior Art”, arXiv, <https://arxiv.org/abs/2012.13919>, 작성일: 2020. 12. 17.
47) Jaewoong Choi et al., “Deep learning-based citation recommendation system for patents”, arXiv, <https://arxiv.org/abs/2010.10932>, 작성일: 2020. 10. 21.
48) Vladimir Karpukhin et al., “Dense Passage Retrieval for Open-Domain Question Answering”, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 6769-6781.
49) 지식재산처, “지식재산처 홈페이지”, 지식재산처, <https://www.moip.go.kr/ko>, 검색일: 2026. 1. 22.

과 유사한 경향으로 구성한다. 문서 한건 당 검색 정답 한건의 쌍으로 총 92,845건을 구축하였다.

<표13 전체 실험 데이터, 평가 데이터셋의 섹션별 분포 및 비율>

| | 전체 실험 데이터 | | 평가 데이터셋 | |
|---|---|---|---|---|
| Section | 문헌 수 | 비율 (%) | 문헌 수 | 비율 (%) |
| A | 813,830 | 14.9 | 18,427 | 19.85 |
| B | 916,754 | 16.76 | 13,907 | 14.98 |
| C | 574,289 | 10.50 | 7,396 | 7.97 |
| D | 87,379 | 1.60 | 790 | 0.85 |
| E | 250,326 | 4.58 | 6,140 | 6.61 |
| F | 463,368 | 8.47 | 6,607 | 7.12 |
| G | 1,080,692 | 19.76 | 26,287 | 28.31 |
| H | 1,283,539 | 23.46 | 13,291 | 14.32 |
| Total | 5,470,177 | 100.0 | 92,845 | 100.0 |

마지막은 검색 실험을 수행한다. 학습이 완료된 8종의 모델을 사용하여 전체 검색 모수 문헌에 대한 벡터를 생성한다. 각 문서는 11개 필드로 구성되어 있으며, 문서 단위 표현은 각 필드 벡터값의 평균으로 계산한다. 즉, 평가 문서가 입력되면 해당 문서의 모든 필드에 대해 임베딩을 추출한 후 이를 평균하여 단일 문서 벡터를 생성한다. 동일한 방식으로 전체 검색 모수 문헌에 대해서도 사전에 문서 벡터를 구축한다.

이후 입력 문서의 벡터와 검색 모수 문헌 벡터간의 코사인 유사도를 계산하고, 유사도 기준 내림차순으로 정렬하여 검색 결과를 도출한다. 최종적으로 상위 Top-K 검색 결과 내에 정답 문서가 포함되는지를 기준으로 검색 성능을 평가한다.

## 6. 평가 및 분석

본 절에서는 8종의 모델에 대한 검색 실험 결과를 확인하고 분석한다. 검색 실험 결과는 아래 표11을 통해 확인할 수 있다. 이때 검색 성능 평가 지표는 앞선 검증 단계에서 사용한 평가지표와 동일한 기준을 적용하였다. 각 모델 별 Top-1~1000까지의 자세한 평가 결과는 부록(Appendix)을 통해서 확인할 수 있다.

8종 모델 중 자동구축 데이터셋을 학습한 모델 6이 가장 높은 성능의 모델임을 확인할 수 있다. 또한 앞 절에서 구체적인 실험 목적 3가지에 대해서 확인해보면 다음과 같다.

<표14 전체 실험 결과 : Top-70 정확도(%) 기준 검색 성능>

| 순번 | 모델 | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|---|
| 1 | KorPatBERT | 19.72 | 23.08 | 26.35 | 30.38 | 21.55 | 22.01 | 19.93 | 18.33 |
| 2 | KorPatBERT + STS 자동구축 | 34.45 | 39.41 | 39.60 | 43.80 | 35.23 | 35.70 | 32.70 | 30.29 |
| 3 | KorPatBERT + STS 전문가구축 | 38.81 | 42.27 | 42.70 | 48.35 | 37.20 | 38.96 | 34.06 | 31.01 |
| 4 | KorPatBERT + STS 혼합 | 36.78 | 42.14 | 41.56 | 47.72 | 37.31 | 38.46 | 34.25 | 31.71 |
| 5 | KorPatBERT + CPC | 54.25 | 53.44 | 55.42 | 61.27 | 47.33 | 49.25 | 40.09 | 38.70 |
| 6 | KorPatBERT + CPC + STS 자동구축 | 57.77 | 60.24 | 58.73 | 64.43 | 54.46 | 54.94 | 48.47 | 47.08 |
| 7 | KorPatBERT + CPC + STS 전문가구축 | 57.55 | 59.71 | 59.09 | 65.19 | 53.09 | 54.41 | 45.03 | 45.19 |
| 8 | KorPatBERT + CPC + STS 혼합 | 57.05 | 59.88 | 58.23 | 64.43 | 53.99 | 54.67 | 48.12 | 46.69 |

첫째는, 자동구축 데이터셋을 STS 미세학습한 검색 모델과 전문가 수작업 데이터셋을 학습한 모델의 검색 성능을 비교하는 것이다. 이는 모델 2,3을 보면, 전 섹션에서 전문가 수작업 데이터셋을 학습한 모델이 자동구축 데이터셋을 학습한 모델보다 우위에 있음을 알 수 있다. 그러나 그 차이가 크지않다. 뿐만 아니라, 모델 6,7을 보면 C섹션과 D섹션을 제외하고 전 섹션에서 자동구축 데이터셋을 학습한 모델이 전문가 수작업 데이터셋을 학습한 모델보다 뛰어난 것을 볼 수 있다.

둘째, 자동구축 데이터셋을 STS 미세학습한 검색 모델과 STS 미세학습하지 않은 언어모델의 검색 성능을 비교하여 검색 학습용 데이터셋의 유효성을 확인하는 것이다. 이는 모델 1,2를 보고, 모델 5,6을 봐도 STS 데이터셋의 유효성을 확인할 수 있다. 모델 1,2의 경우 2배 가까운 상승을 보인다.

셋째, STS 미세학습 외에 CPC 분류 미세학습을 선행 작업하여도 데이터셋의 성능이 유지되는 지 확인한다. 이는 모델 1,2,3,4,5,6,7,8을 보면 알 수 있다. KorPatBERT에 CPC 분류 미세학습을 선행학습하고 STS 미세학습을 진행하면 오히려 검색 성능이 비약적으로 상승되는 것을 알 수 있다.

추가로, 자동구축 데이터셋과 전문가 수작업 데이터 셋을 혼합하여 학습한 모델의 경우, 모델 4,8에서 볼 수 있듯이 크게 튀는 수치가 나오지 않는 것을 확인할 수 있다. 전문가 수작업 데이터 학습한 모델과 자동구축 데이터셋을 학습한 모델의 성능 경계 안에서 점수가 나오기 때문에 두 데이터의 변별력이 크지 않음을 간접적으로 확인할 수 있다.

## 7. 결론

본 연구에서는 한국어 특허 검색 성능 향상을 위해 특허 검색 데이터셋 자동구축 방법론을 제안하고, 이를 통해 자동구축 데이터셋을 생성한 후 특허 검색 실험을 통해 제안 방법론의 효과를 검증하였다. 실험 결과, 표14에서 확인할 수 있듯이 자동구축 데이터셋으로 학습한 KorPatBERT 기반 CPC 미세학습 모델 및 STS 미세학습 모델은 전반적으로 검색 성능이 유의미하게 향상됨을 보였다. 이는 CPC 분류 정보가 1차적인 필터링 단계로 작용하여 검색 공간을

효과적으로 축소함으로써 성능 향상에 기여한 결과로 해석된다. 더불어 자동구축 데이터셋은 전문가가 작성한 의견제출통지서를 기반으로 생성되어, 전문가 구축 데이터셋과 높은 내용적 유사성을 가지면서도 상대적으로 충분한 데이터 규모를 확보할 수 있다. 이러한 특성으로 인해 고품질의 대용량 데이터가 구성되었으며, 그 결과 모델 성능 향상에 긍정적인 영향을 미친 것으로 판단된다.

반면, C섹션과 D섹션의 경우에는 자동구축 데이터셋을 활용한 모델보다 전문가 수작업 데이터셋으로 학습한 모델이 여전히 더 높은 성능을 유지하는 경향을 나타냈다. C섹션은 화학 반응, 화학 물질, 금속 및 재료 자체에 관한 기술을 다루는 영역으로, 특허 문서 내에 화학식과 기호 중심의 표현이 다수 포함되어 있다. 이러한 정보는 일반적인 언어모델이 효과적으로 인식하거나 의미를 학습하기 어려운 형태로 나타나는 경우가 많다. 또한 D섹션은 섬유, 직물, 종이의 가공·제조·처리 기술과 같이 산업 공정 및 가공 방법을 중심으로 한 기술 분야로, C섹션과 유사하게 전문적인 화학 용어와 공정 중심 표현이 빈번하게 등장한다.

이와 같은 특성으로 인해 언어모델이 문맥적 의미를 충분히 파악하는 데 한계가 존재하는 것으로 판단되며, 향후 연구에서는 화학식 및 공정 중심 표현을 효과적으로 반영할 수 있는 학습 전략이나 표현 방식 개선을 통해 해당 섹션의 검색 성능을 추가적으로 향상시키는 방향으로 연구를 확장할 계획이다.

## 참고문헌


### 학술지(국내 및 동양)

민재옥 외 4인, “KorPatBERT 기반 CPC 분류 모델을 활용한 한국어 특허 문헌 검색 성능 향상 연구”, 「지식재산연구」, 제20권 제1호(2025).

박진우 외 4인, “한국어 특허 문장 기반 CPC 자동분류 연구 – 인공지능 언어모델 KorPatBERT를 활용한 딥러닝 기법 접근”, 「지식재산연구」, 제17권 제3호(2022).

### 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).

Bart Degroote & Pierre Held, “Analysis of the patent documentation coverage of the CPC in comparis on with the IPC with a focus on Asian documentation”, *World Patent Information*, Vol.54 Supple ment(2018).

Liang Chen et al., “A deep learning based method benefiting from characteristics of patents for semantic relation classification”, *Journal of Informetrics*, Vol.16 No.3(2022).

### 학위논문(국내 및 동양)

심우철, “CPC 계층적 특성을 고려한 자동 특허 분류 방법”, 충남대학교 일반대학원, 석사, 2025.

### 인터넷 자료

지식재산처, “지식재산처 홈페이지”, 지식재산처, <https://www.moip.go.kr/ko>, 검색일: 2026. 1. 22

지식재산처, “통지서 표준 문안집”, 지식재산처, <http://www.moip.go.kr/ko/kpoBultnDetail.do?menuCd=SCD0200640&parntMenuCd2=SCD0200281&aprchId=BUT0000048&pgmSeq=10148&ntatcSeq=10148>, 검색일: 2025. 12. 8.

한국특허진흥원, “특허 선행기술조사”, 한국특허진흥원, <https://www/kipro.or.kr/business/priorArtSearch>, 검색일: 2026. 1. 22.

Aaron van den Oord et al., “Representation Learning with Contrastive Predictive Coding”, arXiv, <https://arxiv.org/abs/1807.03748>, 작성일: 2019. 1. 22.

Aditya Kusupati et al., “Matryoshka Representation Learning”, arXiv, <https://arxiv.org/abs/2205.13147>, 작성일: 2024. 2. 8.

Arav Parikh & Shiri Dori-Hacohen, “ClaimCompare: A Data Pipeline for Evaluation of Novelty Destroying Patent Pairs”, arXiv, <https://arxiv.org/abs/2407.12193>, 작성일: 2024. 6. 16.

EPO & USPTO, “CPC Guide”, Cooperative Patent Classification, <https://www.cooperativepatentclassification.org/home>, 검색일: 2025. 12. 22.

Grigor Aslanyan & Ian Wetherbee, “Patent Phrase to Phrase Semantic Matching Dataset”, arXiv, <https://arxiv.org/abs/2208.01171>, 작성일: 2022. 8. 1.

Jaewoong Choi et al., “Deep learning-based citation recommendation system for patents”, arXiv, <https://arxiv.org/abs/2010.10932>, 작성일: 2020. 10. 21.

Julian Risch et al., “PatentMatch: A Dataset for Matching Patent Claims & Prior Art”, arXiv, <https://arxiv.org/abs/2012.13919>, 작성일: 2020. 12. 17.

Xianming Li et al., “2D Matryoshka Sentence Embeddings”, arXiv, <https://arxiv.org/abs/2402.14776>, 작성일: 2024. 11. 30.

### 기타 자료

곽민학 외 3인, “한국어 특허 언어모델 Scaling-up에 관한 연구”, 한국정보과학회 2023 한국소프트웨어종합

학술대회 논문집, 2023.

김용일 외 4인, “딥러닝-규칙 기반 병행 모델을 이용한 특허문서의 자동 IPC 분류 방법”, 제31회 한글 및 한국어 정보처리 학술대회 논문집, 한국정보과학회 언어공학연구회, 2019.

박진우 외 3인, “Patent Tokenizer: 형태소와 SentencePiece를 활용한 특허문장 토크나이즈 최적화 연구”, 제37회 한글 및 한국어 정보처리 학술대회 논문집, 한국정보과학회 언어공학연구회, 2020.

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

지식재산처, “2020년 CPC 매뉴얼”, 지식재산처, 2020.

Daniel Cer et al., “SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation”, Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Association for Computational Linguistics, 2017.

Marco Marelli et al., “A SICK Cure for the Evaluation of Compositional Distributional Semantic Models”, Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), European Language Resources Association, 2014.

Vladimir Karpukhin et al., “Dense Passage Retrieval for Open-Domain Question Answering”, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020.

Wei Xu et al., “SemEval-2015 Task 1: Paraphrase and Semantic Similarity in Twitter (PIT)”, Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Association for Computational Linguistics, 2015.

William B. Dolan & Chris Brockett, “Automatically Constructing a Corpus of Sentential Paraphrases”, Proceedings of the Third International Workshop on Paraphrasing (IWP 2005), 2005.

WIPO, “World Intellectual Property Indicators 2025”, WIPO, 2025.

## 부록(Appendix)

<표15 KorPatBERT 모델의 TOP@K 기준 섹션별 검색 성능>

| Top@K | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | 4.28 | 4.62 | 4.77 | 8.10 | 4.32 | 4.16 | 3.11 | 2.73 |
| 2 | 6.17 | 6.85 | 7.75 | 11.14 | 6.74 | 6.31 | 4.76 | 4.30 |
| 3 | 7.58 | 8.25 | 9.49 | 13.92 | 7.83 | 7.76 | 5.93 | 5.41 |
| 5 | 9.11 | 10.25 | 11.87 | 16.71 | 9.84 | 9.82 | 7.56 | 6.82 |
| 10 | 11.52 | 13.08 | 15.56 | 19.87 | 12.49 | 12.34 | 10.09 | 8.89 |
| 20 | 14.07 | 16.25 | 18.94 | 22.53 | 15.10 | 15.57 | 12.96 | 11.85 |
| 30 | 15.80 | 18.30 | 21.19 | 25.19 | 16.92 | 17.60 | 14.89 | 13.55 |
| 40 | 17.14 | 19.95 | 23.00 | 26.58 | 18.26 | 18.80 | 16.61 | 15.05 |
| 50 | 18.17 | 21.10 | 24.19 | 27.72 | 19.74 | 19.96 | 17.92 | 16.30 |
| 70 | 19.72 | 23.08 | 26.35 | 30.38 | 21.55 | 22.01 | 19.93 | 18.33 |
| 100 | 21.57 | 25.50 | 28.58 | 32.66 | 23.68 | 24.10 | 22.16 | 20.50 |
| 150 | 23.85 | 28.16 | 31.22 | 34.81 | 26.24 | 26.50 | 24.78 | 23.16 |
| 200 | 25.63 | 30.20 | 33.53 | 36.58 | 28.14 | 28.56 | 26.83 | 25.24 |
| 300 | 28.34 | 32.96 | 36.84 | 39.24 | 31.06 | 31.38 | 29.90 | 28.21 |
| 400 | 30.11 | 34.85 | 39.16 | 42.78 | 33.32 | 33.36 | 31.98 | 30.66 |
| 500 | 31.49 | 36.74 | 40.94 | 45.19 | 35.34 | 35.19 | 33.62 | 32.55 |
| 1000 | 37.04 | 42.04 | 46.74 | 51.39 | 40.60 | 40.50 | 39.15 | 38.52 |

<표16 KorPatBERT + STS Matryoshka 자동 구축 모델의 TOP@K 기준 섹션별 검색 성능>

| Top@K | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | 5.61 | 6.65 | 6.42 | 10.51 | 5.47 | 5.48 | 4.21 | 3.87 |
| 2 | 8.81 | 10.36 | 10.15 | 14.05 | 8.55 | 8.51 | 6.79 | 6.37 |
| 3 | 11.15 | 12.80 | 13.05 | 17.34 | 10.85 | 10.91 | 8.68 | 8.01 |
| 5 | 14.10 | 16.23 | 16.97 | 21.14 | 14.14 | 14.14 | 11.29 | 10.35 |
| 10 | 18.28 | 21.71 | 22.40 | 24.81 | 18.86 | 19.09 | 15.74 | 14.38 |
| 20 | 23.26 | 27.53 | 28.06 | 32.28 | 23.99 | 24.17 | 21.18 | 19.19 |
| 30 | 26.82 | 31.38 | 31.73 | 35.19 | 27.26 | 27.52 | 24.68 | 22.49 |
| 40 | 29.25 | 33.93 | 34.26 | 37.72 | 29.56 | 30.18 | 27.20 | 25.06 |
| 50 | 31.13 | 35.97 | 36.33 | 39.87 | 31.76 | 32.41 | 29.39 | 26.93 |
| 70 | 34.45 | 39.41 | 39.60 | 43.80 | 35.23 | 35.70 | 32.70 | 30.29 |
| 100 | 38.01 | 43.30 | 43.33 | 48.99 | 38.73 | 39.43 | 36.35 | 33.90 |
| 150 | 42.05 | 47.65 | 47.55 | 52.41 | 42.98 | 43.64 | 40.97 | 38.05 |
| 200 | 44.91 | 50.78 | 50.46 | 55.70 | 45.99 | 46.57 | 44.41 | 41.40 |
| 300 | 49.79 | 55.09 | 55.03 | 59.62 | 50.99 | 51.60 | 49.06 | 46.47 |
| 400 | 52.99 | 57.99 | 58.86 | 64.81 | 54.63 | 54.84 | 52.55 | 50.18 |
| 500 | 55.49 | 60.57 | 61.28 | 66.58 | 57.13 | 57.29 | 55.22 | 53.04 |
| 1000 | 62.93 | 67.98 | 69.65 | 74.56 | 64.97 | 64.87 | 63.58 | 61.46 |

<표17 KorPatBERT + STS Matryoshka 전문가 구축 모델의 TOP@K 기준 섹션별 검색 성능>

| Top@K | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | 6.46 | 7.20 | 7.02 | 10.76 | 6.01 | 6.31 | 4.49 | 3.96 |
| 2 | 10.00 | 11.04 | 11.24 | 14.94 | 9.63 | 9.79 | 7.10 | 6.67 |
| 3 | 12.53 | 13.73 | 14.45 | 18.61 | 12.08 | 12.15 | 8.97 | 8.29 |
| 5 | 15.93 | 17.39 | 18.54 | 22.03 | 15.16 | 15.39 | 11.88 | 10.68 |
| 10 | 20.79 | 23.17 | 24.51 | 28.10 | 19.95 | 20.40 | 16.46 | 14.81 |
| 20 | 26.74 | 29.51 | 30.46 | 34.18 | 25.68 | 26.43 | 22.13 | 19.68 |
| 30 | 30.54 | 33.35 | 34.52 | 37.85 | 28.75 | 30.20 | 25.72 | 23.18 |
| 40 | 33.26 | 36.25 | 37.18 | 41.39 | 31.47 | 33.48 | 28.52 | 25.99 |
| 50 | 35.50 | 38.53 | 39.16 | 44.18 | 33.71 | 35.93 | 30.62 | 27.88 |
| 70 | 38.81 | 42.27 | 42.70 | 48.35 | 37.20 | 38.96 | 34.06 | 31.01 |
| 100 | 42.71 | 46.34 | 46.35 | 52.91 | 41.34 | 42.65 | 38.22 | 34.67 |
| 150 | 47.28 | 50.75 | 50.88 | 58.23 | 46.06 | 47.15 | 42.69 | 38.91 |
| 200 | 50.46 | 54.31 | 54.25 | 60.63 | 49.54 | 50.34 | 45.95 | 42.30 |
| 300 | 55.25 | 59.13 | 59.23 | 64.30 | 54.71 | 54.61 | 51.03 | 47.47 |
| 400 | 58.77 | 62.10 | 62.56 | 67.47 | 58.03 | 58.17 | 54.46 | 51.01 |
| 500 | 61.50 | 64.38 | 65.18 | 69.37 | 60.85 | 60.78 | 57.08 | 53.90 |
| 1000 | 69.02 | 71.72 | 73.24 | 77.09 | 69.20 | 68.97 | 65.32 | 62.65 |

<표18 KorPatBERT + STS Matryoshka 자동/전문가 혼합 구축 모델의 TOP@K 기준 섹션별 검색 성능>

| Top@K | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | 5.93 | 7.02 | 6.61 | 11.14 | 6.12 | 5.77 | 4.38 | 4.09 |
| 2 | 9.44 | 11.10 | 10.82 | 14.68 | 9.33 | 9.34 | 7.11 | 6.68 |
| 3 | 11.70 | 13.72 | 13.86 | 17.72 | 11.56 | 11.99 | 9.19 | 8.45 |
| 5 | 15.11 | 17.39 | 18.14 | 20.76 | 15.03 | 15.30 | 11.87 | 11.01 |
| 10 | 19.59 | 23.10 | 23.89 | 26.71 | 20.33 | 20.37 | 16.71 | 15.12 |
| 20 | 25.00 | 29.30 | 29.69 | 34.05 | 25.52 | 25.78 | 22.21 | 20.13 |
| 30 | 28.56 | 33.28 | 33.79 | 37.22 | 28.88 | 29.68 | 25.96 | 23.50 |
| 40 | 31.19 | 36.07 | 36.37 | 40.63 | 31.73 | 32.59 | 28.66 | 26.39 |
| 50 | 33.32 | 38.45 | 38.28 | 43.16 | 34.10 | 35.11 | 30.80 | 28.59 |
| 70 | 36.78 | 42.14 | 41.56 | 47.72 | 37.31 | 38.46 | 34.25 | 31.71 |
| 100 | 40.54 | 46.16 | 45.02 | 50.76 | 41.38 | 42.56 | 38.29 | 35.45 |
| 150 | 44.99 | 50.57 | 49.73 | 55.57 | 45.64 | 46.83 | 43.04 | 40.06 |
| 200 | 48.27 | 53.84 | 53.29 | 58.73 | 49.25 | 50.11 | 46.39 | 43.40 |
| 300 | 52.75 | 58.48 | 57.91 | 63.29 | 54.58 | 54.85 | 51.42 | 48.75 |
| 400 | 56.12 | 61.47 | 61.49 | 67.47 | 58.32 | 58.20 | 54.97 | 52.50 |
| 500 | 58.78 | 63.87 | 64.18 | 69.87 | 60.99 | 60.50 | 57.65 | 55.47 |
| 1000 | 66.43 | 71.31 | 72.47 | 76.84 | 69.35 | 68.53 | 66.08 | 63.89 |

<표19 KorPatBERT + CPC 분류 학습 모델의 TOP@K 기준 섹션별 검색 성능>

| Top@K | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | 8.01 | 7.72 | 8.34 | 11.39 | 7.23 | 7.33 | 4.71 | 4.44 |
| 2 | 12.99 | 12.17 | 13.89 | 17.34 | 11.24 | 11.52 | 7.81 | 7.52 |
| 3 | 16.50 | 15.22 | 17.83 | 20.89 | 14.12 | 14.51 | 10.01 | 9.56 |
| 5 | 21.56 | 20.10 | 22.89 | 26.46 | 18.09 | 17.98 | 13.42 | 12.72 |
| 10 | 28.65 | 27.34 | 30.42 | 34.94 | 24.15 | 24.34 | 18.73 | 18.11 |
| 20 | 37.18 | 35.93 | 38.91 | 43.80 | 30.90 | 32.51 | 25.26 | 24.29 |
| 30 | 42.26 | 41.27 | 44.08 | 49.62 | 35.80 | 37.61 | 29.67 | 28.70 |
| 40 | 46.38 | 44.97 | 47.77 | 53.80 | 39.50 | 41.77 | 33.21 | 32.07 |
| 50 | 49.50 | 48.45 | 50.78 | 56.08 | 42.74 | 44.63 | 35.90 | 34.76 |
| 70 | 54.25 | 53.44 | 55.42 | 61.27 | 47.33 | 49.25 | 40.09 | 38.70 |
| 100 | 59.24 | 58.60 | 60.47 | 66.20 | 52.46 | 54.62 | 44.89 | 43.56 |
| 150 | 64.96 | 64.22 | 65.59 | 71.01 | 59.25 | 60.78 | 50.43 | 49.86 |
| 200 | 68.90 | 67.94 | 69.15 | 74.18 | 63.84 | 65.13 | 54.51 | 54.30 |
| 300 | 74.23 | 73.16 | 74.24 | 79.62 | 69.64 | 70.70 | 60.01 | 59.94 |
| 400 | 77.65 | 76.36 | 77.97 | 83.04 | 74.04 | 74.72 | 64.16 | 64.25 |
| 500 | 80.18 | 78.89 | 80.37 | 85.57 | 77.13 | 77.34 | 67.12 | 67.33 |
| 1000 | 86.94 | 85.91 | 87.40 | 91.14 | 85.15 | 85.18 | 76.29 | 76.44 |

<표20 KorPatBERT + CPC 분류 학습+ STS Matryoshka 자동 구축 모델의 TOP@K 기준 섹션별 검색 성능>

| Top@K | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | 8.99 | 9.63 | 9.56 | 12.03 | 8.14 | 8.31 | 5.76 | 5.36 |
| 2 | 14.36 | 14.86 | 15.62 | 18.35 | 13.49 | 13.35 | 9.44 | 9.19 |
| 3 | 18.04 | 18.89 | 19.69 | 20.76 | 16.48 | 16.65 | 12.33 | 11.85 |
| 5 | 23.35 | 24.65 | 25.07 | 27.47 | 20.83 | 21.01 | 16.75 | 16.10 |
| 10 | 31.34 | 33.23 | 33.38 | 36.96 | 28.08 | 28.98 | 23.46 | 22.57 |
| 20 | 40.17 | 42.59 | 42.37 | 47.34 | 36.03 | 38.22 | 31.35 | 30.25 |
| 30 | 45.68 | 48.42 | 47.71 | 53.04 | 41.43 | 43.05 | 36.45 | 35.23 |
| 40 | 49.64 | 52.73 | 51.53 | 56.46 | 45.68 | 46.94 | 40.50 | 39.09 |
| 50 | 52.99 | 55.86 | 54.33 | 59.87 | 49.19 | 49.90 | 43.65 | 42.30 |
| 70 | 57.77 | 60.24 | 58.73 | 64.43 | 54.46 | 54.94 | 48.47 | 47.08 |
| 100 | 62.87 | 65.20 | 63.20 | 70.76 | 59.84 | 59.94 | 53.48 | 52.23 |
| 150 | 68.45 | 71.00 | 68.43 | 75.57 | 65.94 | 65.72 | 59.46 | 58.20 |
| 200 | 72.02 | 74.64 | 71.88 | 79.24 | 69.89 | 69.73 | 63.73 | 62.37 |
| 300 | 76.57 | 79.33 | 76.97 | 84.05 | 75.46 | 75.00 | 69.16 | 67.83 |
| 400 | 79.99 | 82.43 | 80.19 | 86.46 | 78.97 | 78.11 | 72.86 | 71.68 |
| 500 | 82.40 | 84.56 | 82.61 | 87.97 | 81.68 | 80.41 | 75.51 | 74.50 |
| 1000 | 88.66 | 89.99 | 88.85 | 92.91 | 88.50 | 87.18 | 83.28 | 82.62 |

<표21 KorPatBERT + CPC 분류 학습+ STS Matryoshka 전문가 구축 모델의 TOP@K 기준 섹션별 검색 성능>

| Top@K | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | 8.91 | 9.17 | 9.59 | 12.41 | 8.05 | 8.35 | 5.41 | 5.12 |
| 2 | 14.21 | 14.37 | 15.72 | 18.73 | 12.80 | 13.26 | 8.89 | 8.84 |
| 3 | 17.91 | 18.29 | 19.74 | 21.65 | 15.86 | 16.51 | 11.55 | 11.10 |
| 5 | 22.87 | 23.48 | 25.54 | 27.59 | 20.13 | 20.93 | 15.46 | 15.12 |
| 10 | 30.83 | 32.20 | 33.60 | 36.96 | 27.04 | 28.79 | 21.80 | 21.20 |
| 20 | 39.67 | 41.58 | 42.82 | 47.09 | 34.77 | 37.51 | 29.10 | 28.49 |
| 30 | 45.31 | 47.48 | 47.73 | 53.04 | 39.98 | 42.58 | 33.92 | 33.59 |
| 40 | 49.47 | 51.64 | 51.61 | 57.59 | 44.28 | 46.78 | 37.52 | 37.42 |
| 50 | 52.71 | 54.91 | 54.77 | 61.01 | 47.77 | 49.87 | 40.35 | 40.61 |
| 70 | 57.55 | 59.71 | 59.09 | 65.19 | 53.09 | 54.41 | 45.03 | 45.19 |
| 100 | 62.79 | 64.72 | 63.64 | 69.37 | 58.60 | 59.86 | 50.20 | 50.60 |
| 150 | 68.46 | 69.94 | 69.20 | 75.19 | 64.98 | 65.54 | 55.92 | 56.53 |
| 200 | 72.15 | 73.77 | 72.62 | 78.48 | 69.33 | 69.55 | 60.18 | 60.64 |
| 300 | 77.05 | 78.72 | 77.03 | 83.29 | 75.02 | 74.62 | 65.94 | 66.20 |
| 400 | 80.14 | 82.02 | 80.14 | 87.09 | 78.47 | 78.14 | 69.95 | 69.99 |
| 500 | 82.50 | 84.02 | 82.50 | 89.11 | 80.99 | 80.64 | 72.57 | 73.38 |
| 1000 | 88.64 | 89.89 | 89.20 | 93.29 | 88.39 | 87.62 | 80.70 | 81.45 |

<표22 KorPatBERT + CPC 분류 학습+ STS Matryoshka 자동/전문가 혼합 구축 모델의 TOP@K 기준 섹션별 검색 성능>

| Top@K | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | 8.75 | 9.61 | 9.57 | 11.52 | 8.34 | 8.11 | 5.62 | 5.30 |
| 2 | 14.16 | 14.87 | 15.48 | 17.97 | 13.47 | 13.27 | 9.27 | 9.06 |
| 3 | 17.66 | 18.66 | 19.29 | 20.13 | 16.30 | 16.79 | 12.14 | 11.70 |
| 5 | 22.74 | 24.48 | 24.96 | 26.58 | 20.73 | 21.19 | 16.62 | 15.79 |
| 10 | 30.66 | 33.08 | 33.14 | 37.09 | 27.82 | 28.76 | 23.17 | 22.35 |
| 20 | 39.60 | 42.28 | 42.16 | 46.46 | 35.80 | 37.90 | 31.11 | 29.79 |
| 30 | 45.02 | 48.26 | 47.31 | 52.41 | 41.01 | 42.95 | 36.12 | 34.92 |
| 40 | 49.25 | 52.29 | 50.88 | 55.70 | 45.29 | 46.87 | 40.03 | 38.78 |
| 50 | 52.51 | 55.29 | 53.91 | 58.86 | 49.10 | 49.64 | 43.20 | 42.03 |
| 70 | 57.05 | 59.88 | 58.23 | 64.43 | 53.99 | 54.67 | 48.12 | 46.69 |
| 100 | 62.31 | 64.93 | 62.98 | 69.87 | 59.85 | 59.71 | 53.11 | 51.89 |
| 150 | 67.68 | 70.63 | 67.87 | 74.81 | 65.70 | 65.60 | 59.03 | 58.06 |
| 200 | 71.44 | 74.23 | 71.43 | 78.99 | 69.79 | 69.32 | 63.27 | 61.85 |
| 300 | 76.16 | 79.05 | 76.42 | 83.04 | 75.03 | 74.57 | 68.75 | 67.50 |
| 400 | 79.34 | 82.12 | 79.77 | 86.08 | 78.53 | 78.17 | 72.43 | 71.39 |
| 500 | 81.72 | 84.45 | 82.18 | 87.59 | 81.42 | 80.46 | 75.11 | 74.28 |
| 1000 | 88.21 | 89.78 | 88.70 | 92.28 | 88.50 | 86.86 | 82.85 | 82.29 |