지식재산연구
Journal of Intellectual Property

지식재산연구 21권 1호, 2026년 3월, 181-207면
Journal of Intellectual Property Volume 21, Issue 1, March 2026, Pages 181-207.
https://doi.org/10.34122/jip.2026.21.1.181

Print ISSN 1975-5945
Online ISSN 2733-8487

RESEARCH ARTICLE

A Contrastive Learning Patent Document Retrieval Model for Similar Patent Technologies Based on KorPatSTS

Jae-Ok Min[1], Sol-Bin Hwang[2], Young-Hoon Jeon[2], Song-A Chae[2], Bong-Gun Lee[3]

[1]Team Manager, Intelligent Information Strategy Department, Korea Institute of Patent Information, Republic of Korea
[2]Assistant Manager, Intelligent Information Strategy Department, Korea Institute of Patent Information, Republic of Korea
[3]Head of Strategic Planning Department, Korea Institute of Patent Information, Republic of Korea

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

ABSTRACT

This study proposes an advanced deep-learning–based patent retrieval model, NCE-KorPat, and a high-quality training dataset, KorPatSTS (Korean Patent Semantic Textual Similarity), to more precisely assess grounds for rejection arising from technological redundancy during the patent application process. The model recommends cited patent documents based on semantic and technical similarity between an application and prior art patents.

KorPatSTS is a sentence-level dataset of similar patent-technology sentence pairs, drawing on the expertise of the Korea Ministry of Intellectual Property (MOIP) AI Examiner Advisory Group. The dataset aligns the technical constituent elements of claims in an application patent and the corresponding sentences in a detailed description, which explains those elements by considering the matching portions of prior art patents cited as grounds for rejection, thereby forming highly precise sentence-level correspondence pairs.

In this study, the NCE-KorPat model was first developed by fine-tuning KorPatBERT, a patent-domain–specific language model, for CPC subgroup-level classification and then subsequently applying contrastive learning and optimization using the KorPatSTS dataset. When applied to Korean patent retrieval experiments, the proposed model demonstrated superior performance, outperforming both the previously best-performing Korean embedding models and state-of-the-art global embedding models.

To the best of our knowledge, this study represents the first attempt to construct similar-technology sentence pairs systematically by integrating the domain expertise of Korean patent examiners and directly applying them to a practical patent retrieval model. The proposed approach is expected to substantially contribute to improving the accuracy and efficiency of patent examinations in the future.

KEYWORDS

Intellectual Property Rights, Patent, KorPatBERT, KorPatSTS, Contrastive Learning, Patent Search

Received: December 31, 2025
Revised: February 21, 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

KorPatSTS 기반 특허 유사 기술 대조 학습을 활용한 특허 문헌 검색 모델에 관한 연구

민재옥[1], 황솔빈[2], 전영훈[2], 채송아[2], 이봉건[3]

[1]한국특허정보원 지능정보전략실 팀장
[2]한국특허정보원 지능정보전략실 대리
[3]한국특허정보원 전략기획실장

교신저자: 이봉건 (bglee@kipi.or.kr)

차례

국문초록

본 연구는 특허 출원 과정에서 발생하는 기술적 중복으로 인한 거절 사유를 보다 정밀하게 판단하기 위해, 출원 특허와 선행 특허 간의 의미적·기술적 유사성을 기반으로 인용 특허 문헌을 추천하는 고도화된 특허 검색 딥러닝 모델 NCE-KorPat과 고품질의 학습 데이터셋 KorPatSTS(Korean Patent Semantic Textual Similarity)를 제안한다.

KorPatSTS는 한국 지식재산처 「AI 심사관 자문단」의 전문성을 바탕으로 출원 특허의 청구항 기술 구성 요소와 이를 설명하는 발명의 상세한 설명 문장에서 실제 거절 사유로 인용된 선행 특허의 대응되는 내용을 문장 단위 수준으로 정밀하게 매핑하여 구축한 유사 특허 기술 문장쌍 데이터셋이다.

본 연구에서는 특허 분야에 특화된 언어모델인 KorPatBERT 기반으로 CPC 서브그룹 단위 분류 파인튜닝에 이어서 KorPatSTS 데이터셋을 활용한 대조 학습(Contrastive Learning) 및 학습 최적화를 통해 NCE-KorPat 모델을 개발하였으며, 이를 한국어 특허 검색 실험에 적용한 결과, 기존 최고 성능의 한국어 임베딩 모델과 최신 글로벌 임베딩 모델을 모두 상회하는 우수한 성능을 달성하였다.

본 연구는 한국 지식재산처 특허 심사관의 전문성과 도메인 지식을 결합하여 유사 기술 문장쌍을 체계적으로 구축하고 이를 실제 검색 모델에 적용한 최초의 연구로, 향후 특허 심사의 정확성과 효율성을 동시에 높이는 데 실질적으로 기여할 것으로 기대된다.

주제어

지식재산권, 특허, KorPatBERT, KorPatSTS, 대조학습, 특허검색

1. 서론

전 세계 특허 출원량이 기하급수적으로 증가[1]하고 있으며, 지식재산권 경쟁의 무대를 전례 없이 치열하게 만들고 있다. Ali et al.[2]연구에 따르면, 방대한 특허 문헌 속에서 원하는 정보를 얼마나 신속하고 정확하게 찾아내느냐는 단순한 기술적 과제를 넘어, 국가와 기업의 기술 패권을 좌우하는 전략적 핵심 과제로 부상하고 있다고 주장하였다. 이러한 맥락에서 고도화된 특허 검색 시스템은 선택이 아닌 필수가 되었으며, 그 중요성은 과거 어느 때보다도 막중하다. 그러나 Parikh & Dori-Hacohen[3] 연구에서 지금까지의 연구는 주로 AI 기반 검색 모델의 성능 향상에만 집중되어 왔다. 정작 검색 모델의 실질적 성능을 결정짓는 학습 데이터셋, 특히 선행 특허 검색에 활용 가능한 정밀하고 현실적인 데이터셋 구축은 상대적으로 소외되어 왔다는 점에서 한계를 드러낸다고 주장하였다. 창의적이고 모호한 표현이 빈번히 등장하는 특허 문헌의 텍스트[4]는 언어적 유사성에 기반해 구축된 기존 일반 분야의 유사 문장쌍 데이터와는 그 형태가 다르다. 따라서 한국 지식재산처(이하, 지재처) 특허 심사관의 전문성을 반영하여 기술적 유사성에 기초한 고품질의 유사 기술 문장쌍 데이터셋을 구축하는 것은 더 이상 미룰 수 없는 과제가 되었다.

본 연구에서 명시하는 유사 기술은 청구항 전체를 하나의 단위로 비교하는 방식과 구별된다. 이는 특허 심사관이 거절 사유를 판단할 때 수행하는 구성요소별 대비 과정을 모사하여, 출원 발명의 특정 기술 구성 요소와 그에 대응하는 선행 특허의 구체적 기술 내용을 매핑한 요소 단위의 유사성으로 정의한다.

지재처 심사관이 직접 구축한 데이터셋을 모델 학습한 AI 모델은 특허 문헌 간의 복잡한 의미적 관계를 정밀하게 포착하고, 이를 고차원 임베딩 벡터로 변환하여 변별력 높은 유사도 계산을 가능하게 한다. 이는 단순한 검색 정확도의 개선을 넘어, 특허 심사 환경에서의 정보 탐색 방식을 근본적으로 혁신할 수 있는 토대를 제공할 것이다.

본 연구에서는 다음과 같은 구체적인 연구 목표를 설정하였다.

첫째, 특허 산업 분야의 기업·기관에서 실질적으로 활용 가능하도록 전체 한국 특허를 대상으로 검색 데이터셋과 평가 데이터셋의 구축 방법을 제안하며, 특허 심사 특성에 부합한 기술적 유사성 계산을 위해 AI 모델 학습이 가능한 Korean Patent Semantic Textual Similarity(이하, KorPatSTS) 데이터셋을 제안한다.

둘째, KorPatSTS 데이터셋으로 의미적·기술적 유사성 계산을 가능하게 하는 최적의 대조학습(Contrastive Learning)[5] 모델링 기법을 제안한다.

셋째, 특허 검색 실험에서 KorPatSTS 데이터셋의 효과를 검증하고, 다양한 텍스트 임베딩 모델과의 검색 성능을 비교함으로써 가장 우수한 특허 검색 모델을 제안한다.

본 연구는 특허 심사 과정에서 심사관의 전문적 판단을 반영하여 정밀하게 구축된 유사 특허 기술 문장쌍 학습 데이터셋을 토대로, 다양한 임베딩 모델과 대조 학습 기법을 종합적으로 비

1) 지식재산처, “2024 통계로 보는 특허동향”, 지식재산처, <https://www.kipo.go.kr/ko/kpoBultnDetail.do?menuCd=SCD0200640&ntatcSeq=16933&sysCd=SCD02&aprchId=BUT0000048#1>, 검색일: 2025. 8. 20.

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) 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. 7. 16.

4) Su-Jeong Jeong, “Zur Analyse von mehr oder weniger festen Wortverbindungen in Patentschriften im Deutschen und Koreanischen”, *German Literature,* Vol.26 No.3(2016), pp. 360-361.

5) Prannay Khosla et al., “Supervised Contrastive Learning”, Advances in Neural Information Processing Systems 33, 2020, pp. 18661-18673.

교·분석하였다. 이러한 시도는 민재옥, et al.[6] 연구에서 시도된 CPC 서브그룹 분류 기반 특허 검색 모델 연구를 한층 고도화한 연구로, 단순한 성능 개선을 넘어 전문가 데이터셋 설계에서 구축, 모델 학습 전략까지 검색 모델 개발의 전 주기를 아우르는 새로운 접근을 제시한다. 특히, '의견제출통지서'라는 실제 행정 처분 문서를 정답(Answer)으로 활용하여 검색 모델의 실무 재현성을 극대화하였다. 또한 단순 언어적 유사성이 아닌 특허 청구범위의 기술 구성요소 단위로 대응 관계를 설계한 점은 기존 연구와 구별되는 차별점이다.

본 연구는 방대한 특허 문헌 속에서 검색 정확성과 효율성을 획기적으로 향상시키며, 나아가 특허 산업 기관이 지식재산권을 신속하고 정밀하게 확보·관리할 수 있는 실질적 토대를 마련한다. 이러한 성과는 단순한 기술적 진보를 넘어, 치열해지는 글로벌 지식재산 경쟁 속에서 AI 기반 특허 행정 혁신의 방향성을 제시하는 전환점이 될 것으로 기대된다.

2. 이론적 논의

2.1. 대조 학습(Contrastive Learning)

Zhang et al.[7] 연구에 따르면, 대조 학습(Contrastive Learning)5)은 인공지능 모델이 데이터의 유사성과 차이점을 스스로 구분할 수 있도록 훈련하는 방식이다. 다시 말해, 모델이 '비슷한 것은 서로 가깝게', '다른 것은 서로 멀게' 임베딩 공간에 표현하도록 유도하는 학습 방법이다.

Denize et al.[8] 연구에서 대조 학습의 기본 원리는 데이터에서 두 가지 유형의 쌍(Pair)을 만드는 것에서 시작한다. 하나는 서로 유사한 데이터를 묶은 양성(Positive) 쌍이며, 다른 하나는 서로 유사하지 않은 데이터를 묶은 음성(Negative) 쌍이다. 모델은 학습을 통해 양성 쌍의 임베딩 표현은 서로 비슷하게 만들어 가까워지게 하고, 음성 쌍의 임베딩 표현은 서로 달라져 멀어지도록 한다. 이 과정을 반복하면, 모델은 새로운 데이터를 봤을 때도 자연스럽게 유사성을 효과적으로 판단할 수 있는 능력을 갖추게 된다.

대조 학습 기법은 레이블이 부족한 상황에서도 강력한 표현 학습 성능을 보여주어 최근 자연어처리 분야에서도 널리 주목받고 있다[9].

2.1.1. Triplet Loss

구글의 Schroff, et al.[10] 연구에서 Triplet loss 기반 모델 학습은 FaceNet10)에서 처음 사용되어 서로 다른 각도의 동일 인물 얼굴 이미지 임베딩을 학습하는 데 활용된 바 있다. 하나의

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

7) Tao Zhang & Mingming Hu, "Learning Representation for Clustering via Dual Correlation", 2023 IEEE 6th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), IEEE, 2023, pp. 1579-1583.

8) Julien Denize et al., "Similarity Contrastive Estimation for Image and Video Soft Contrastive Self-Supervised Learning", *Machine Vision and Applications,* Vol.34 No.6(2023), p. 111.

9) Yuanmeng Yan et al., "ConSERT: A Contrastive Framework for Self-Supervised Sentence Representation Transfer", Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 5065-5075.

10) Florian Schroff et al., "FaceNet: A Unified Embedding for Face Recognition and Clustering", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 815-823.

기준이 되는 앵커(Anchor) 샘플 x와 이에 양성 관계인 샘플 x+, 그리고 음성 관계인 샘플 x-의 세 개 샘플로 학습을 진행한다. 여기서 x와 x+는 동일한 클래스에 속하거나 의미적으로 유사한 한 쌍이며, x-는 다른 클래스에 속하거나 무관한 샘플로 선택된다. Triplet loss의 목표는 앵커-양성 쌍의 거리는 가깝게 줄이고 앵커-음성 쌍의 거리는 일정 마진(Margin) 이상으로 벌리도록 만드는 것이다.

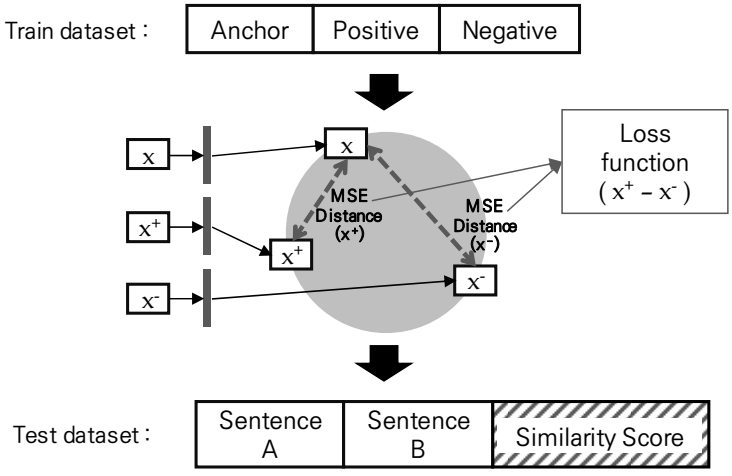
<수식 1> Triplet loss

$$L_{triplet}(x, x^+, x^-) = \max(\|f(x) - f(x^+)\|_2 - \|f(x) - f(x^-)\|_2 + \epsilon, 0)$$

<수식 1>에서 $f(x)$는 입력 샘플을 임베딩 벡터로 변환하는 인코더 함수이고, ϵ는 양성-음성 간 거리 차이에 대한 최소 여유분을 정의하는 마진 하이퍼파라미터이다. 위 손실 함수를 최소화하면 모델은 자연스럽게 x와 x+의 거리를 좁히는 동시에 x와 x-의 거리는 ϵ만큼 더 크게 만들도록 학습된다.

Triplet을 활용한 학습 데이터셋 구성 방식과 동작 원리 및 두 문장을 입력받아 유사도 점수를 출력하는 과정을 <그림1>에 제시하였다.

<그림1 Triplet 모델 학습 동작 원리 및 데이터 입·출력 구조 >



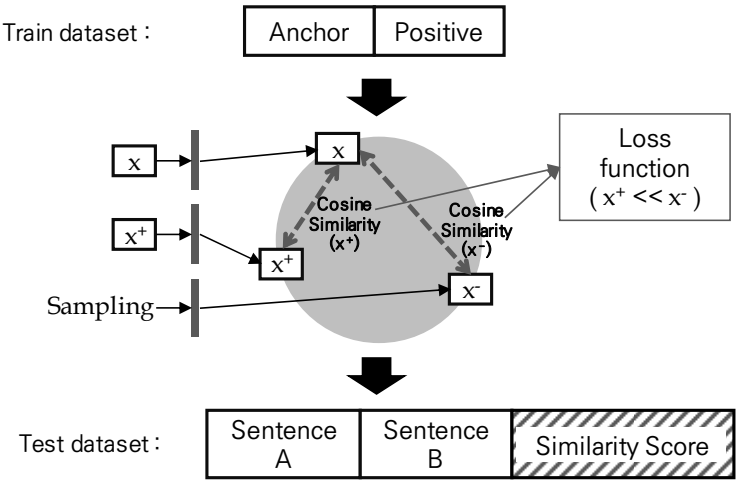
2.1.2. InfoNCE Loss

InfoNCE[11]는 한 앵커가 주어졌을 때 그와 의미적으로 짝을 이루는 양성을 여러 음성 샘플 후보 사이에서 올바르게 골라내도록 소프트맥스 기반 분류 형태로 학습시킨다. 이 손실 함수는 구글 CPC(Contrastive Predictive Coding)[12]에서 발전했으며, 모델이 주어진 앵커 입력 x에 대해 진짜 양성 x+를 여러 후보 중에서 맞추도록 학습시킨다.

주어진 기준 샘플이 있을 때 이 기준 샘플과 짝을 이루는 진짜 양성 샘플 하나를, 배치(batch) 안에 포함된 많은 음성 샘플들 중에서 모델이 정확히 찾아내도록 학습시키는 것이다.

11) Evgenia Rusak et al., "InfoNCE: Identifying the Gap Between Theory and Practice", arXiv, <https://arxiv.org/abs/2407.00143>, 작성일: 2025. 4. 16.
12) Aaron van den Oord et al., "Representation Learning with Contrastive Predictive Coding", arXiv, <https://arxiv.org/abs/1807.03748>, 작성일: 2019. 1. 22.

<그림2 InfoNCE 모델 학습 동작 원리 및 데이터 입·출력 구조>



Li et al.[13], Yuan et al.[14] 연구에 따르면, InfoNCE는 배치(In-Batch) 내 다수의 음성 샘플을 동시에 활용함으로써 한 쌍의 Triplet에만 의존하는 손실 대비 효율적으로 학습을 진행할 수 있다는 장점이 있다고 주장하였다. 배치 전체를 음성으로 활용하여 비교함으로써, 정보 밀도가 높고 수렴 속도도 빠르며, 대규모 데이터셋을 다룰 때에도 매우 효과적인 특성을 보인다. Lu & Lu[15], Xuan et al.[16]연구에서는 InfoNCE loss가 Triplet loss 보다 안정적으로 수렴한다고 주장하였다. 이러한 이유로 InfoNCE loss 기반 모델 학습은 현재 대조 학습의 표준으로 자리매김하고 있으며, 자연어처리 분야 등 다양한 모델에서 핵심적인 학습 방법으로 사용되고 있다.

2.2. 관련 선행 연구

박상언[17] 연구에 따르면, 특정 도메인의 대규모 말뭉치를 기반으로 '사전학습(Pre-training)'된 언어모델은 해당 도메인의 다양한 자연어처리(NLP) 과제에서 범용 모델을 압도하는 성능을 보이며, 이에 대한 학술적·산업적 관심이 급격히 고조되고 있다고 주장하였다. 이와 관련하여 박진우 등[18] 연구에서는 특허 문헌의 대규모 데이터를 활용해 사전학습을 수행하여 KorPatBERT를 개발하였고, 이를 특허 분류 태스크에 적용하여 그 효과를 입증하였다. 이러한 결과는 특허 도메인 전용 언어모델이 실제 심사·분류 업무에서 탁월한 성능을 발휘할 수 있음을 보여준다. 더 나아가 민재옥, et al.[6] 연구에서는 KorPatBERT 기반 CPC 분류 파인튜닝 모델을 특허 검색 실험에 적용함으로써 도메인 특화 파인튜닝 모델이 특허 문헌 검색 성

13) Haochen Li et al., "Rethinking Negative Pairs in Code Search", Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 12760-12774.
14) Ye Yuan et al., "In Defense of the Triplet Loss Again: Learning Robust Person Re-Identification with Fast Approximated Triplet Loss and Label Distillation", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 1454-1463.
15) Zhenyu Lu & Yonggang Lu, "A Balanced Triplet Loss for Person Re-Identification", *International Journal of Pattern Recognition and Artificial Intelligence,* Vol.37 No.1(2023), Article No. 2256022.
16) Hong Xuan et al., "Hard Negative Examples Are Hard, but Useful", European Conference on Computer Vision, Springer International Publishing, 2020, pp. 126-142.
17) 박상언, "딥러닝 기반 사전학습 언어모델에 대한 이해와 현황", 「한국빅데이터학회지」, 제7권 제2호(2022), 11-29면.
18) 박진우 외 4인, "한국어 특허 문장 기반 CPC 자동분류 연구-인공지능 언어모델 KorPatBERT를 활용한 딥러닝 기법 접근", 「지식재산연구」, 제17권 제3호(2022), 209-256면.

능을 유의미하게 향상시킬 수 있음을 입증하였다. 그럼에도 불구하고, 기존 연구는 주로 모델 구조와 학습 기법의 고도화에 집중되어 왔으며, 정작 선행기술·유사기술 검색에 실질적으로 활용 가능한 정밀하고 현실적인 데이터셋 구축 연구는 EPO 해외 사례[19]가 있으나, 국내 연구 사례는 찾아보기 어려웠다.

Alexander V. Giczy, et al.[20] 연구에 따르면, 고품질 데이터셋 구축을 위해서는 도메인 전문가의 검증이 필수적이라고 주장하였다. AIPD(Artificial Intelligence Patent Dataset)는 AI 전문 특허 심사관들의 수동 검토를 통한 평가 지표를 제시하여, 모델 학습 접근법이 기존 문헌의 대안들보다 우수한 성능을 달성함을 보여주었다.

유동건, et al.[21] 연구에서는 유사한 기술 문장쌍을 학습하기 위해 Google Patent에서 출원 특허의 청구항 데이터와 인용 특허의 청구항 데이터를 수집하는 방식으로 연구를 진행하였다. 이 데이터로 대조 학습 모델을 구축하여 검색 성능 향상을 입증하였다. 그러나 보다 더 정확하고 실제 특허 심사관의 의사결정 기준을 충실히 반영하려면, 단순한 '청구항 – 인용 청구항'의 연결을 넘어 거절이유 통지 등 심사 기록에서 근거가 된 문장 수준의 정밀한 매핑, 그리고 특허 텍스트의 기술적 의미 등가성을 반영하는 전문가 주도 데이터셋이 필요하다.

3. 연구 방법

본 연구는 <그림 3>에 제시된 절차에 따라 특허 검색 성능 향상 및 비교 실험을 위한 체계적인 절차를 통해 진행되었다.

3.1. 데이터 수집 및 구축 단계에서는 키프리스 플러스(KIPRIS Plus)[22]를 활용하여 전체 한국 특허 데이터를 수집하고 전처리하여 특허 검색을 위한 검색 데이터셋을 구축하였다.

특허 심사관이 출원 특허에 대한 심사를 통해 최종적으로 '거절'하였고, 거절 사유와 인용 특허를 기록한 의견제출통지서에서 객관적인 평가 데이터셋을 구축하였다. 또한, 2024년 지재처에서 위촉한 AI 심사관 자문단[23]에서 우수 심사관[24]을 선별하여 대조 학습을 위한 고품질의 학습 데이터셋인 KorPatSTS 데이터셋을 구축하는 방안을 제안하였다. KorPatSTS는 심사관의 전문성을 바탕으로 발명의 권리 범위를 최소 단위로 쪼개어 분석함으로써, 실질적으로 기술적 대비가 이루어지는 유사 기술 문장만을 선별하여 데이터의 변별력을 확보하였다.

다음으로, 3.2장의 검색 모델 설계 단계에서는 MTEB[25][26]의 한국어 검색 태스크 리더보드

19) Julian Risch et al., "PatentMatch: A Dataset for Matching Patent Claims & Prior Art", arXiv, <https://arxiv.org/abs/2012.13919>, 작성일: 2020. 12. 27.

20) Alexander V. Giczy et al., "Identifying Artificial Intelligence (AI) Invention: A Novel AI Patent Dataset", *The Journal of Technology Transfer*, Vol.47 No.2(2022), pp. 476-505.

21) 유동건·한지헌, "결합발명 진보성 판단의 인용문헌 자동 추천 딥러닝 모델에 관한 연구: BERT-for-patents 및 대조학습 기법을 중심으로", 「지식재산연구」, 제20권 제1호(2025), 119-143면.

22) 한국특허정보원, "특허정보 활용 서비스", 한국특허정보원, <https://plus.kipris.or.kr/portal/main.do>, 검색일: 2025. 8. 20.

23) 지식재산처, "AI 심사관 자문단 위촉장 수여식", 지식재산처, <https://www.kipo.go.kr/ko/kpoBultnDetail.do?menuCd=SCD0200615&ntatcSeq=1338&sysCd=SCD02&aprchId=BUT0000026>, 검색일: 2025. 8. 20.

24) 지식재산처, "심사명장 소개", 지식재산처, <https://www.kipo.go.kr/ko/kpoContentView.do?menuCd=SCD0200461>, 검색일: 2025. 8. 20.

25) Niklas Muennighoff et al., "MTEB: Massive Text Embedding Benchmark", Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, 2023, pp. 2014-2037.

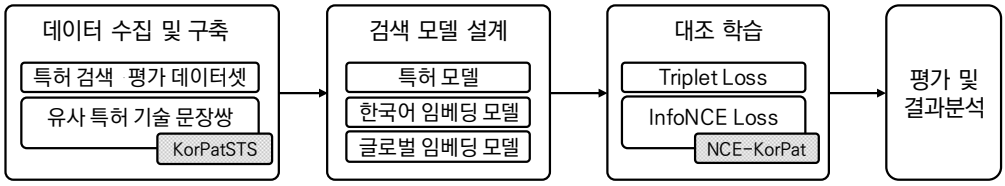
26) MTEB, "MTEB", Embedding Benchmark Github, <https://github.com/embeddings-benchmark/mteb>, 검색일: 2025. 8. 20.

(MTEB-ko-retrieval Leaderboard)에서 우수한 성능을 달성한 한국어 임베딩 모델과 최신 글로벌 임베딩 모델을 선정하였고, 대조 학습 방법에 따라 모델별 학습을 진행하였다.

마지막으로 4장에서 모델별 KorPatSTS 데이터셋 적용 및 검색 비교 실험을 진행하였고, 최종적으로 특허 검색 성능이 가장 우수한 NCE-KorPat 모델을 제안하였다.

모델별 실험 결과를 종합적으로 분석하여 5장에서 연구 목표 달성 여부를 검증하였고, 본 연구의 한계와 연구 방향을 제안하였다.

<그림3 연구 절차>



3.1. 데이터 수집 및 구축

본 연구는 신뢰성 높은 데이터셋의 구축이 필수적이라는 문제에서 출발하였다. 검색 모델의 성능을 객관적으로 검증하기 위해서는, 검색 대상이 되는 방대한 특허 문헌 전체에 대한 데이터와, 이 위에서 모델 성능을 정량적으로 측정할 수 있는 평가 데이터셋이 체계적으로 준비되어야 한다.

이를 위해 본 연구에서는 국내 전체 특허 문헌 데이터를 기반으로 대규모 검색 데이터셋을 구축하였으며, 특허 심사 실무의 실제 결과 문서인 의견제출통지서를 바탕으로 평가 데이터셋을 구축하였다. 아울러, 지재처의 AI 심사관 자문단[23]의 전문 지식을 활용하여 문장 수준에서 정밀한 유사성을 반영한 학습 데이터셋인 KorPatSTS 데이터셋을 구축함으로써, 의미 기반의 고도화된 특허 검색 모델 학습이 가능하도록 하였다.

3.1.1. KorPatSTS : Korean Patent Semantic Textual Similarity

일반적인 AI 학습 데이터 구축과 달리, 특허 심사 분야의 데이터 구축은 출원 발명의 기술적 핵심을 정확히 이해하고, 이에 대응하는 선행기술 문헌의 내용을 정밀하게 연결하는 고도의 전문성을 요구한다. 특히, 특허 심사관 개개인이 축적해온 전문 지식을 AI가 학습 가능한 구조로 전환하는 과정은 도메인 전문가의 참여 없이는 실현이 어렵다.

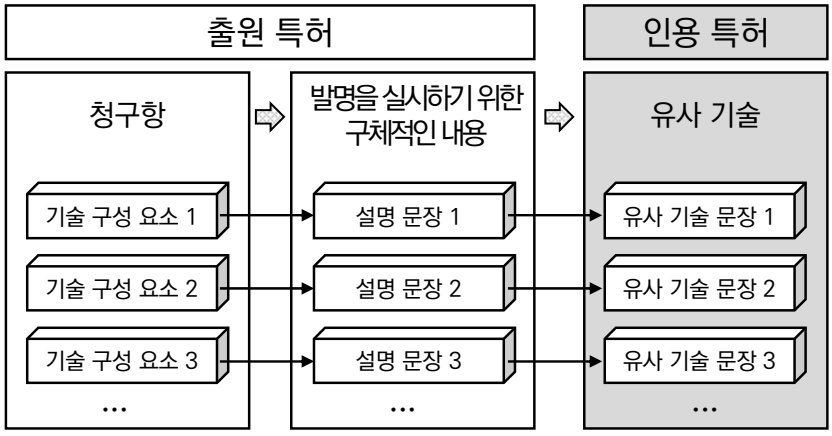
본 연구에서는 특허 심사관과 협업하여 전문적으로 데이터를 구축하였다. 지재처는 AI 기반 특허 심사 지원을 위해 AI 심사관 자문단[23]을 발족하였다. 자문단은 지재처의 공정한 평가를 거쳐 선발된 우수 특허 심사관[24]들로 구성되며, 특허 심사의 전문성과 AI 기술 간의 연결고리로서 핵심적인 역할을 수행한다. 이들은 특허 심사 업무에 대한 실질적 경험과 이해를 바탕으로, AI 학습에 필요한 고품질 데이터 구축과 품질 검증 전반에 걸쳐 중심적인 기여를 하도록 하였다. 데이터 구축의 전 과정에 참여하며, 심사 과정에 반드시 필요한 정보가 학습 데이터로 정교하게 반영되도록 설계하였다.

구체적으로, 실제 특허 심사관의 심사 지침[27]과 전문성에 기반하여 출원 특허의 권리 범위인

27) 지식재산처, "지식재산 심사 기준/매뉴얼", 지식재산처, <https://www.kipo.go.kr/ko/kpoContentView.do?menuCd=SCD0201119>, 검색일: 2025. 8. 20.

청구항에서 기술 구성 요소를 식별하고, 기술 구성 요소별로 대응하는 상세한 기술 설명 내용을 식별한 후, 유사 기술로 판단되는 선행기술 문헌의 해당 문장을 추출하여 문장 쌍(Pairing) 형태로 구성한다. 이러한 유사 기술 문장 쌍의 구축 절차는 <그림 4>에 제시하였다.

<그림4 유사 기술 문장 쌍 데이터 >



데이터의 일관성과 품질을 위해 다음과 같은 가이드 라인을 적용하였다.

- 문장 단위 대응 매핑: 출원 특허의 청구항 문장과 선행기술 문헌의 해당 구절을 일대일로 대응시켜 연결하는 작업을 수행한다.
- 원본 텍스트 유지: 텍스트가 반드시 존재해야 하고, 실제 특허 문헌에 존재하는 원문 그대로 사용해야 한다.
- 데이터 출처 및 식별 정보 명시: 추출한 텍스트의 문헌 정보, 문단 정보, 특허 필드 정보를 정확히 기입하고, 기술 구성 요소별로 한 행(row)씩 작성한다.
- 데이터 검수 및 교차 검토: 매핑된 기술 쌍의 정확성과 타당성 여부를 자문단 심사관이 직접 검토하며, 다수 전문가 간의 교차 검토를 통해 최종 데이터를 확정한다.
- 품질 평가 및 피드백: 정확성, 일관성, 다양성 등 품질 지표를 기준으로 구축한 데이터셋을 지속적으로 평가하고, 개선이 필요한 경우 피드백을 반영하여 수정한다.

작업 대상은 키프리스 플러스(KIPRIS Plus)22)를 통해 수집된 2010년부터 2024년까지 공개된 의견제출통지서로 한정하였다.

또한, KorPatSTS 데이터셋 구축 과정에서 AI 심사관 자문단이 수행하는 문장 페어링 및 검수 작업의 일관성과 효율성을 향상시키기 위해 웹(Web) 기반 데이터 구축 지원 시스템을 개발·적용하였다.

이 시스템은 작업자가 시스템에 접근하여 각 특허 출원 특허의 청구항과 인용 특허 문헌을 동시에 열람하고, 두 특허 문헌의 유사성을 판단하여 출원 특허의 기술 구성 요소별 유사 기술 인용 문장을 짝지어 매핑할 수 있도록 설계되었다. 각 작업자는 시스템 상에서 문장 쌍을 직접 선택하고 저장할 수 있으며, 본인이 수행한 이전 작업 기록을 시각적으로 확인하고 체계적으로 관리할 수 있는 사용자 인터페이스(UI)가 제공된다.

이 시스템은 다음과 같은 주요 기능을 포함한다.

- 문장 매핑 인터페이스: 각 작업자에게 할당한 작업 대상을 불러와서, 의견제출통지서, 출원 특허, 인용 특허 문서를 나란히 표시하여, 유사 문장을 손쉽게 비교하고 선택할 수 있도

록 시각적인 구성을 제공한다.

- 작업 이력 관리: 작업자가 매핑한 문장 쌍과 유사성 평가 결과를 저장 및 시각화하고, 이전 작업 내역에 대한 조회 및 수정이 가능하다.
- 다중 사용자 협업 평가: 여러 작업자가 동일한 데이터에 대해 동시에 평가하거나 상호 검토할 수 있는 협업 기능을 포함함으로써, 데이터의 신뢰도를 높인다.
- 피드백 및 의견 공유 기능: 작업 결과에 대한 피드백을 입력하거나 동료 작업자와 의견을 교환하여, 평가 기준에 대한 해석 차이를 조율할 수 있도록 지원한다.

<그림 5> 데이터 구축 지원 시스템

이 시스템은 KorPatSTS 데이터셋 구축을 위해 전용 개발된 도구이지만, 시스템 도입 초기부터 수렴된 현장 피드백을 바탕으로 UI와 기능을 지속적으로 개선해 왔다. 그 결과 사용자 경험(UX)과 작업 효율성이 유의미하게 향상되었으며, 향후 유사한 AI 학습 데이터 구축 과제 전반에 적용 가능한 공공 플랫폼으로의 확장 가능성을 가진다.

KorPatSTS 데이터셋은 <표1>과 같다.

<표1 KorPatSTS 데이터셋 일부>

출원 특허			인용 특허
번호	구성요소	발명의 상세한 설명	인용내용
1	탄산염 및 인산염을 포함하는 무기물	상기 무기물은 여러 종류의 무기물 중에서도 본 발명에 따라서 탄산염 및 인산염으로부터 선택된 적어도 하나 이상을 포함한다.	일 구현 예는 탄산칼륨 함유 제1 소화성분; 인산암모늄 함유 제2 소화성분; 요소 함유 제3 소화성분; 빙점 강하제; 및 물을 포함하는 리튬이온배터리용 강화액 소화약제를 제공한다.
2	플러그를 꽂을 수 있는 콘센트(30)	콘센트(30);은 다른 가전제품에 전기를 전달하기 위해 연결하는 역할을 한다. 플러그-전선-콘센트로 연결된다.	본 고안에 따른 전선(100)은 전선(100)의 일측에 부착되어 콘센트에 끼워 넣어지는 플러그(10)와, 플러그(10)에 결합되어 플러그(10)의 핀 단자(2)를 통해 입력받은 전원을 전달하는 제1 전선(20) 및 제2 전선(30)과, 전선(100)의 타측에서 상기 제2 전선(30)과 연결되며, 상기 제1 전선(20) 및 제2 전선(30)을 통해 전달된 전원을 개별 전자제품에 공급하는 콘센트(40)를 포함한다.
3	상기 유기박막을 광학적으로 처리하는 단계	다음에, 제1전극(10)이 형성된 기판(1) 상에 유기박막들의 전부 또는 일부가 형성되면, 이러한 유기박막을 광학적으로 처리한다.	패터닝된 ITO 기판을 준비하는 단계, 정공주입층과 수송층의 발광층을 적층하는 단계, 발광층을 레이저로 조사하는 단계, 음극을 증착하는 단계로 이루어지는 유기발광소자 제조방법
4	(b) 상기 발신 이동통신 단말기가 콜백(Call Back) URL을 포함하는 상기 메시지를 SMSC에 전송 요청하는 단계	발신 이동통신 단말기(110)는 현재 접속하고 있는 웹페이지의 주소를 콜 백 URL(Call Back URL)로 포함하는 메시지를 착신 이동통신 단말기(150)에 전송하도록 SMSC(140)에 요청한다(S240).	무선 인터넷 서버에 접속한 이동단말기가 웹페이지의 URL과 수신 전화번호를 무선 인터넷 서버로 전송하고, 상기 서버는 수신 단말기로 메시지 및 콜백 URL을 전송하여 메시지를 수신한 이동단말기가 상기 URL에 해당되는 웹페이지에 접속
5	제 2 고정홈의 입구측으로부터 돌출 형성되어 제 2 고정홈 내에 삽입되어 있는 상태의 와이어에 지지되는 이탈방지돌기	또한, 본 발명에 따른 드럼 세탁기는 다이어프램(15)에 힘이 가해지더라도 제 2 고정부(15a)가 제 1 고정홈(10c)으로부터 이탈하는 것을 방지할 수 있도록 제 2 고정홈(15b)의 입구측 단부에는 이탈방지돌기(15c)가 돌출 형성되어 있다.	상기한 환형요홈(5)의 내측으로는 환형요홈(5)의 간극(G)에 개재되는 돌기(6)를 등간격으로 연속형성함으로써 립(4)의 배면으로부터 정면방향으로 가압력이 작용할 경우에도 상기한 돌기(6)의 작용에 따라 립(4)이 전방으로 과도하게 탄성변형되는 것을 차단한다.

KorPatSTS 데이터셋은 문장 단위 매핑을 통해 특허 청구항 문장과 선행기술 문헌 문장 간의 직접적인 의미적 연결 관계를 확보하였다. 이로써 동일한 기술 내용을 다루지만 서로 다른 특허 문헌에 포함된 문장들을 하나의 쌍으로 구성할 수 있었다. 특히, 동일한 기술 분야 내에서도 문장의 표현 방식이나 구조가 크게 상이할 수 있다는 점을 고려하여, 단순한 문장 형태나 키워드의 일치 여부에 의존하지 않고 기술적 의미의 유사성을 기반으로 문장 쌍을 선정하였다. 이러한 원칙에 따라, KorPatSTS 데이터셋은 표면적인 유사성을 넘어 심층적인 의미 대응 관계를 충실히 반영하도록 구축되었다. 심사관의 교차 검토 과정에서 3.4%의 낮은 오류 수정률로 수행되어 품질을 극대화하였다.

구축 대상으로는 데이터가 충분(<표 4>)하고 기술적 중요성과 산업 파급력을 고려하여 전기

전자분야의 H섹션과 화학 야금 분야인 C섹션을 우선적으로 구축하였다. 구축 현황은 <표2>와 같다.

<표2 KorPatSTS 데이터셋 현황>

섹션	문헌 수	데이터 수
C	497	4,406
H	616	6,558
합계	1,112	10,964

KorPatSTS 데이터셋은 특허 검색 모델을 위한 대조 학습에 활용한다. 대조 학습을 통해 출원 특허의 기술 구성 요소와 선행 기술 문장 사이의 의미적 관계를 더욱 정밀하게 포착할 수 있게 된다. 결과적으로 KorPatSTS 데이터셋은 특허 문헌 특성이 반영된 의미 기반의 특허 검색 정확도 제고에 핵심적인 학습 자원으로 기능한다.

KorPatSTS 데이터셋은 대조 학습을 위한 학습 데이터셋과 검증 데이터셋으로 사용하였다. 그 결과는 <표3>에 제시하였으며, 평가 데이터셋은 3.1.3장에서 구축한 C섹션과 H섹션의 전체를 대상으로 하여 실질적인 산업적 유용성 조건을 갖추도록 하였다.

<표3 학습 데이터셋>

	학습(Train)	검증(Validation)	평가(Test)
	9,868	1,096	8,806 (C섹션), 16,967 (H섹션)
합계	10,964 건		25,773 문헌

3.1.2. 검색 데이터셋

검색 데이터셋은 실제 검색 모델이 탐색하게 될 특허 문헌의 전체 풀(pool)을 의미하며, 전체 한국어 특허 문헌을 사용함으로써, 모델의 학습 및 평가 환경을 현실에 가깝게 구성하였다. 본 연구에서는 키프리스 플러스(KIPRIS Plus)[22]에서 제공하는 국내 공개 특허 문헌 전체 데이터를 수집하여 검색 코퍼스로 활용하였다.

수집된 데이터는 1946년부터 2024년까지 발행된 한국어 특허 문헌 총 5,470,177건이며, 세계 특허정보 표준인 ST.96 기반의 XML 형식으로 제공된다. 방대한 원문 데이터를 효율적으로 처리하기 위해, 본 연구에서는 선행 연구[18] 연구에서 정의한 특허 필드 구조와 전처리 규칙을 적용하여 설계하였다. XML 태그 구조에서 불필요한 기호 및 텍스트를 정제하고, 식별 정보와 본문 텍스트를 체계적으로 분리함으로써 분석과 검색에 최적화된 형태로 데이터를 가공하였다.

정제된 데이터는 데이터베이스에 저장하여 대용량 검색 실험 시 신속한 색인(indexing)과 검색이 가능하도록 구성하였다. 최종 구축된 검색 데이터셋은 한국 특허 전 분야를 포괄하며, 총 5,470,177건, 약 210GB 규모로 집계되었다.

<표4>에서 CPC 분야별 데이터셋 현황을 제시하였다.

<표4 검색 데이터셋 현황(1946년 ~ 2024년)>

CPC 섹션	특허 문헌 수(건)	비율(%)	합계(건)	데이터 크기
A (생활 필수품)	813,830	14.88	5,470,177	210GB
B (처리조작, 수송)	916,754	16.76		
C (화학, 야금)	574,289	10.50		
D (섬유, 종이)	87,379	1.60		
E (고정 구조물)	250,326	4.58		
F (기계공학, 조명, 가열, 무기)	463,368	8.47		
G (물리)	1,080,692	19.76		
H (전기)	1,283,539	23.46		

3.1.3. 평가 데이터셋

평가 데이터셋은 특허 심사관이 특정 출원 특허를 심사하면서 해당 발명이 등록 거절된 사유와 인용한 선행 특허들을 상세히 기록한 문서인 의견제출통지서를 활용하여 구축하였다. 의견제출통지서 데이터는 키프리스 플러스(KIPRIS Plus)[22]에서 추출하였으며, 2024년까지 행정처분이 최종 종료된 건 중에서 선별된 113,219건의 출원 특허에 대한 기록을 수집하였다. 이 데이터셋은 실제 심사 과정에서 사용된 출원 특허-인용 특허의 문헌 쌍 정보를 포함하고 있어, 모델 개발 방향을 설정하고 성능 개선의 기준을 제공하는 동시에 모델 성능을 검증하는 중요한 척도로 활용될 수 있다[6]. 이에 따라, AI 심사관 자문단에서 수작업으로 구축할 대상으로 평가 데이터셋을 제공하였다. 평가 데이터셋 현황은 <표5>에 제시하였다.

<표5 평가 데이터셋 현황(1946년 ~ 2024년)>

CPC 섹션	문헌 쌍(건)	비율(%)	합계(건)
A (생활 필수품)	22,404	19.79	113,219
B (처리조작, 수송)	16,828	14.86	
C (화학, 야금)	8,806	7.78	
D (섬유, 종이)	1,067	0.94	
E (고정 구조물)	7,005	6.19	
F (기계공학, 조명, 가열, 무기)	7,864	6.95	
G (물리)	32,278	28.51	
H (전기)	16,967	14.99	

하나의 출원 특허가 심사를 받을 때, 해당 분야뿐만 아니라 여러 기술 분야의 선행 특허가 인용될 수 있다. 즉, 출원 발명이 다수의 기술 요소로 구성된 경우 각 요소별로 관련된 선행기술이 비교되어 한 출원에 여러 개의 인용 특허가 포함될 수 있다. 다만 모든 인용 특허를 평가에 동일하게 활용할 경우 한 Query(질의)에 다수의 문헌 Answer(정답)가 존재하게 되어 일관성 있는 평가가 어렵다. 따라서 본 연구에서는 의견제출통지서에 가장 먼저 기록된 인용 특허를 출원 특허의 신규성 및 진보성 부정의 핵심 근거가 되는 주된 인용 특허로 간주하여 대표 Answer로 설정하였다. 이는 평가의 일관성을 확보하기 위함이나, 다중 인용이 빈번한 실무의 맥락에서는 검색 모델이 정답 외의 타당한 문헌을 도출하더라도 누락되는 보수적 평가 결과로 나타날 수 있다.

최종적으로 각 출원 특허 Query에 대해 1대1 문헌 쌍(출원 특허 - 주요 인용 특허)으로 구성

된 평가 데이터셋을 확보하여 일관된 성능 평가가 가능하도록 정제하였다.

<표6 검색 실험 데이터셋 현황>

섹션	평가 데이터셋(Query-Answer)	검색 데이터셋(Documents)
C	8,806	574,289
H	16,967	1,283,539
합 계	25,773	1,857,828

3.2. 모델 설계 및 학습

기존 선행 연구[6]에서는 CPC 분류 수준이 세분화될수록 검색 성능이 향상된다고 주장하며, KorPatBERT 기반의 64,000종 CPC 서브그룹 분류 모델(이하, CLS-KorPat-6.4)을 제안하였다. 이후 해당 모델을 활용하여 특허 문헌의 임베딩 벡터를 추출하고 이를 검색 실험에 적용함으로써 그 효과성을 입증하였다.

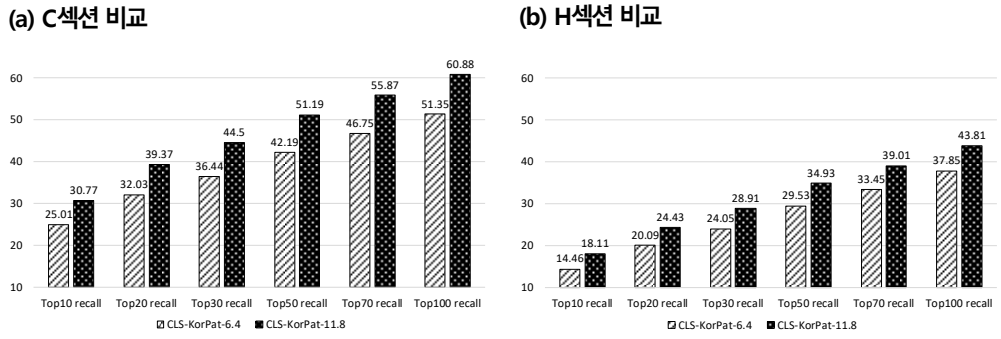
본 연구에서는 이를 확장하여 117,158종의 서브그룹 분류가 가능한 학습 데이터셋을 구축하고 학습하여, 더 세분화된 분류 모델(이하, CLS-KorPat-11.8)을 개발하였다.

본 연구에서는 사전에 두 모델을 비교 실험한 결과, 선행 연구의 주장과 동일하게 분류 수준이 세분화될수록 더 우수한 성능을 보였다.

모든 평가 구간에서 CLS-KorPat-11.8이 높은 성능을 보였고, 이후 실험에서 특허 모델은 CLS-KorPat-11.8로 진행하였다.

그 결과는 <그림6>에 제시하였다.

<그림6 CLS-KorPat-6.4와 CLS-KorPat-11.8 검색 비교 실험 >



다음으로 임베딩 모델 간 비교 실험을 진행하였다.

한국어 비교 모델로는 글로벌 텍스트 임베딩 벤치마크 플랫폼인 MTEB(Massive Text Embedding Benchmark)[28][29]의 한국어 검색 태스크 리더보드(MTEB-ko-retrival Leaderboard)에서 가장 우수한 성능을 기록한 고려대학교 AI 자연어처리 연구실의

28) Niklas Muennighoff et al., "MTEB: Massive Text Embedding Benchmark", Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, 2023, pp. 2014-2037.

29) MTEB, "MTEB", Embedding Benchmark Github, <https://github.com/embeddings-benchmark/mteb>, 검색일: 2025. 8. 20.

KURE-v1[30][31][32](이하, KURE) 모델(2024년 12월 21일 공개)을 선정하였다.

또한, 한국어뿐만 아니라 다국어 임베딩이 가능하며 MTEB에서 우수한 성능을 보인 최신 글로벌 임베딩 모델로 SnowFlake-arctic-embed-l-v2.0[33](이하, SnowFlake) 모델을 선정하였다. 두 모델은 Hugging Face[34]에 공개되어 있는 버전으로 사용하였다.

본 연구에서 비교 실험을 위한 임베딩 모델은 <표7>에 제시하였다.

<표7 검색 비교 실험을 위한 임베딩 모델>

모델명	지원 언어	모델 설명
CLS-KorPat-6.4	한국어	KorPatBERT 기반 CPC 서브그룹 분류(64,000종) 파인튜닝 모델
CLS-KorPat-11.8	한국어	KorPatBERT 기반 CPC 서브그룹 분류(117,158종) 파인튜닝 모델
KURE-v1	한국어	BAAI/BGE-M3[35]을 기반으로 구축된 모델로 고려대 자연어처리 연구실에서 개발 및 공개
SnowFlake-arctic-embed-l-v2.0	다국어	BAAI/BGE-M3을 기반으로 구축된 다국어 모델로 SnowFlake에서 개발 및 공개

다음 실험을 위해, KorPatSTS 데이터셋(<표 3>)을 활용하여 대조 학습을 진행하였다.

먼저, Triplet loss 기반 대조 학습을 진행하였다. Triplet은 앵커(Anchor)와 양성(Positive) 데이터 간의 임베딩 거리를 최소화하고, 동시에 앵커와 음성(Negative) 데이터 간의 거리를 일정 마진(Margin) 이상으로 벌리도록 학습시키는 방식(<그림 1>)이다. KorPatSTS 데이터셋에서 앵커는 출원 특허의 청구항 구성요소와 이에 대응하는 상세한 기술 설명 내용 문장을 연결하여 구성하였으며, 선행 특허 문헌에서 기술적으로 유사하다고 판단하여 인용한 문장을 양성 데이터로 하였다. 음성 데이터는 학습 과정에서 양성 데이터로 지정되지 않은 다른 샘플을 자동으로 선택하도록 하였다.

하이퍼파라미터 설정에서는 Margin 값을 1.0으로 고정하였고, Adam 옵티마이저를 적용하였으며, 학습률(Learning rate)은 2e-5로 설정하였다.

다음으로 InfoNCE loss 기반 대조 학습을 진행하였다. InfoNCE는 앵커와 양성 데이터 쌍을 올바르게 식별하도록 모델을 학습시키는 방식(<그림 2>)으로, 배치(In-Batch) 내에 포함된 다수의 음성 샘플을 동시에 활용함으로써 학습 효율성을 높이는데, 동일 배치 내에서 앵커와 짝지어지지 않은 나머지 문장들을 음성 샘플로 간주하였다.

하이퍼파라미터 설정에는 temperature는 0.07로 고정하였고, AdamW 옵티마이저를 적용하였으며, 학습률은 Min=9e-7, Max=9e-6으로 하여 선형적으로 감소하도록 설정하였다.

모든 실험은 동일한 자원 환경으로 리눅서 GPU서버 기반의 Nvidia A100(80GB) 1대에서 학

30) 고려대학교, “KURE”, 고려대 NLP & AI 연구실, <https://hiai.korea.ac.kr/kure/>, 검색일: 2025. 8. 20.
31) 고려대학교, “KURE”, 고려대학교 NLP & AI 연구실 Github, <https://github.com/nlpai-lab/KURE?tab=readme-ov-file#mteb-ko-retrieval-leaderboard>, 검색일: 2025. 8. 20.
32) 윤진석, “고려대 임희석 교수님 한국어 특화 임베딩 모델 ‘KURE’ 공개”, 아시아타임즈, <https://www.asiatime.co.kr/article/20241202500146#_mobwcvr>, 검색일: 2025. 8. 20.
33) Puxuan Yu et al., “Arctic-Embed 2.0: Multilingual Retrieval Without Compromise”, arXiv, <https://arxiv.org/abs/2412.04506>, 작성일: 2024. 12. 14.
34) Hugging Face, “Hugging Face”, Hugging Face, <https://huggingface.co/>, 검색일: 2025. 8. 20.
35) Jianlv Chen et al., “M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation”, Findings of the Association for Computational Linguistics: ACL, 2024, pp. 2318-2335.

습을 진행하였다.

학습을 진행하면서 검증(Validation) 데이터셋의 정확도가 일정 수준 이상 유지되고, 학습 손실(Loss)이 점차 감소하다가 더 이상 개선되지 않는 시점에서 학습을 종료하고 모델 파일을 저장하였다. 학습을 진행하면서 최적의 학습률을 보이는 지점의 하이퍼파라미터 값으로 최종 선정하였기 때문에 Triplet, InfoNCE 학습 모델의 하이퍼파라미터 값은 각각 상이하다.

최종적으로 만들어진 실험 모델 명칭은 <표8>에서 정의하였다.

<표8 검색 실험 모델 및 명칭>

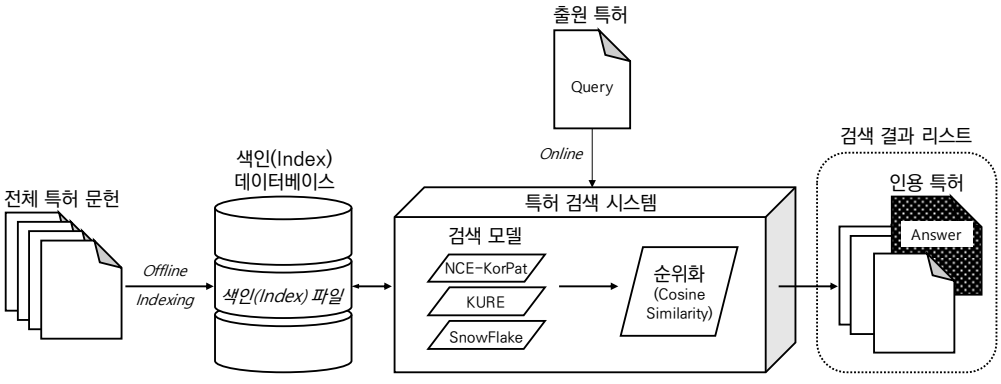
기본 모델	Triplet Loss 학습 모델	InfoNCE Loss 학습 모델
CLS-KorPat-11.8	Triplet-KorPat	NCE-KorPat
KURE-v1	Triplet-KURE	NCE-KURE
SnowFlake-arctic-embed-l-v2.0	Triplet-SnowFlake	NCE-SnowFlake

특허 문헌 벡터를 추출할 때에는 발명의 명칭+요약, 배경 기술, 기술 분야, 전체 청구항에서 각각 임베딩 벡터를 추출한 후, 각 필드별로 생성된 문장 임베딩 벡터의 평균을 계산하여 최종적인 특허 문헌 임베딩 벡터로 하였다.

3.3. 검색 시스템

본 연구에서는 실질적인 유용성과 산업적 적용 가능성을 높이고자 실제 검색 시스템의 프로세스와 동일한 구조로 구축하였다. 특허 검색 시스템의 구조는 <그림7>에 제시하였다.

<그림 7 특허 검색 시스템 구조 >



방대한 데이터를 효율적으로 처리하기 위해 검색 데이터셋에 대해 실험 모델별 Encoder로부터 추출한 임베딩 벡터를 사전에 색인[36]화(Offline Indexing)하여 데이터베이스에 저장하였고, 실시간 검색에 활용하였다.

먼저, 평가 데이터셋의 Query를 입력으로 받으면 실험 모델의 Encoder에서 실시간으로 임베딩 벡터를 추출한 후, 사전에 색인화된 데이터베이스에서 코사인 유사도(Cosine similarit

36) Alfonso F. Cardenas, "Analysis and Performance of Inverted Data Base Structures", *Communications of the ACM,* Vol.18 No.5(1975), pp. 253-263.

y)[37]를 통해 순위화를 수행한다.

이때, 평가 데이터셋 25,773건(<표6>)에 대해 검색 데이터셋 1,857,828건(<표6>)에서 탐색해야 하므로 총 47,881,801,044회의 대규모 임베딩 벡터 유사도 연산이 요구된다. 이러한 방대한 연산량을 처리하기 위해, GPU 가속을 지원하는 Facebook의 Faiss(Facebook AI Similarity Search)[38]를 사용하였다.

순위화 평가에서는 특허 심사관의 정보 탐색 행동을 반영하여 상위 K 구간의 검색 결과 품질에 집중하는 Top-K recall 지표를 적용하였다. 초기 결과 구간의 재현율을 중심으로 모델의 실무 적합성을 평가하였다.

순위화 평가에서는 특허 심사관의 정보 탐색 행동을 반영하여 Top-K recall 지표를 적용하였다. 이는 심사관이 수백만 건의 문헌 중 단 하나의 정답을 찾는 것이 아니라, 일정 범위(K) 내에 존재하는 유효 문헌들을 검토하여 거절 근거를 확보하는 실무 프로세스를 모사하기 위함이다. 따라서 누락 없는 검색(Recall)은 심사 효율성과 직결되는 핵심 지표로 기능한다.

4. 실험 및 평가

특허 문헌 검색 시스템에서 평가 데이터셋으로 실험 모델별 검색 실험을 수행하였고, Top-K recall 지표를 통해 그 결과를 분석하였다.

첫 번째, KorPatSTS 데이터셋의 효과를 대조 학습 결과로 확인해 보았다. 기본 모델과 InfoNCE loss 기반 학습 모델을 비교하였다. <그림8>, <그림9> 결과에 따르면 CLS-KorPat-11.8과 KURE는 모든 Top-K 구간에서 더 향상된 결과를 보였고, SnowFlake는 기본 모델의 성능이 더 높은 결과를 보였다.

구체적으로, C섹션에서 NCE-KorPat은 CLS-KorPat-11.8 대비 평균 3.03%p, NCE-KURE는 기본 모델 대비 평균 2.16%p 향상되었으며, NCE-SnowFlake는 기본 모델 대비 평균 2.07%p 성능이 낮았다.

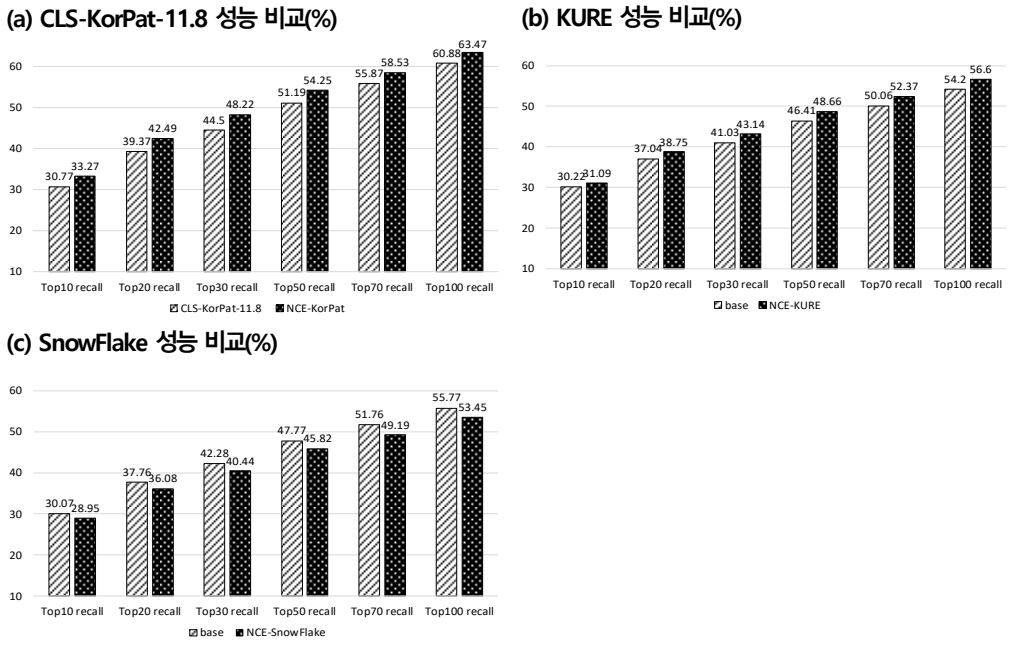
H섹션에서는 NCE-KorPat은 CLS-KorPat-11.8 대비 평균 4.26%p, NCE-KURE는 기본 모델 대비 평균 3.73%p 성능이 향상되었고, NCE-SnowFlake는 기본 모델 대비 평균 0.69%p 낮은 결과를 보였다.

검색 데이터셋의 14.99% 차지하는 H섹션에서 상대적으로 더 큰 개선폭이 관찰된 점은 의미 있는 결과로 판단된다. 그러나 SnowFlake 모델에서는 KorPatSTS 데이터셋을 학습한 결과 더 낮은 성능 결과를 보였다.

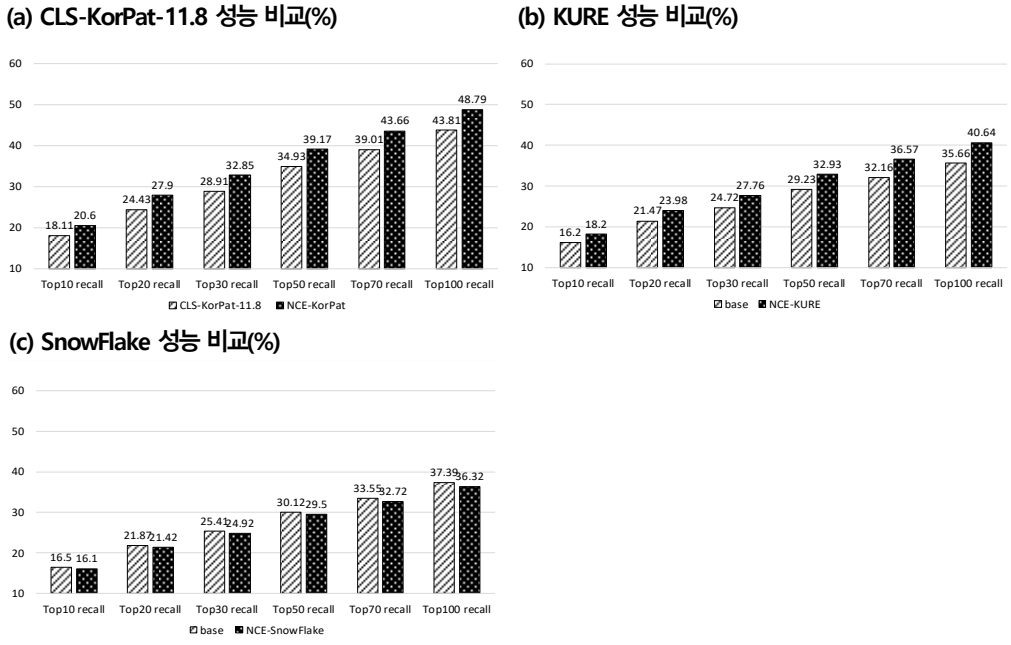
37) Vikas Thada & Vivek Jaglan, “Comparison of Jaccard, Dice, Cosine Similarity Coefficient to Find Best Fitness Value for Web Retrieved Documents Using Genetic Algorithm”, *International Journal of Innovations in Engineering and Technology,* Vol.2 No.4(2013), pp. 202-205.

38) facebookresearch, “facebookresearch/faiss”, facebookresearch Github, <https://github.com/facebookresearch/faiss>, 검색일: 2025. 8. 20.

<그림 8 KorPatSTS 데이터셋 모델 학습 적용에 따른 검색 성능 비교(C섹션)>



<그림 9 KorPatSTS 데이터셋 모델 학습 적용에 따른 검색 성능 비교(H섹션)>



SnowFlake와 KURE는 모두 한국어를 포함하는 BGE-M3[39] 모델을 기반으로 학습되었고, KURE는 국내 기관에서 한국어 중심으로 더욱 특화되도록 학습되었다. SnowFlake는 별도의 파인튜닝 없이도 특허 검색 태스크에 적용 가능한 준수한 성능을 보였다. 하지만 InfoNCE loss

39) Jianlv Chen et al., "M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation", Findings of the Association for Computational Linguistics: ACL, 2024, pp. 2318-2335.

기반 대조 학습을 적용한 경우 오히려 검색 성능이 저하되는 현상이 관찰되었다. SnowFlake 모델의 성능 저하 원인을 모델의 구조적 특성과 학습 역학적 관점에서 분석하였다. 이는 다국어 범용 임베딩 모델이 고도로 특화된 특허 도메인 데이터셋과 결합하는 과정에서 발생한 도메인 부적응(Domain Mismatch) 및 파라미터 민감도 문제로 풀이된다. 구체적인 요인으로는 첫째, 토크나이저의 한계로 인해 특허 특유의 전문 용어와 장문 구조를 효과적으로 처리하지 못해 발생한 정보 손실을 들 수 있다. 둘째, 소규모 KorPatSTS 데이터셋으로 미세 조정(Fine-tuning)을 진행하며 발생한 치명적 망각(Catastrophic Forgetting) 및 과적합 현상이 사전 학습된 가중치 구조를 훼손했을 가능성이 크다. 이에 대한 상세한 분석은 향후 연구 과제로 제언한다.

두 번째, 대조 학습 방법에 따른 검색 성능을 비교하였다. KorPatSTS 데이터셋으로 InfoNCE loss 기반 대조 학습으로 하였을 때, CLS-KorPat-11.8과 KURE는 성능 향상을 확인하였지만, SnowFlake의 결과를 고려하여 추가적인 비교 실험을 진행하였다.

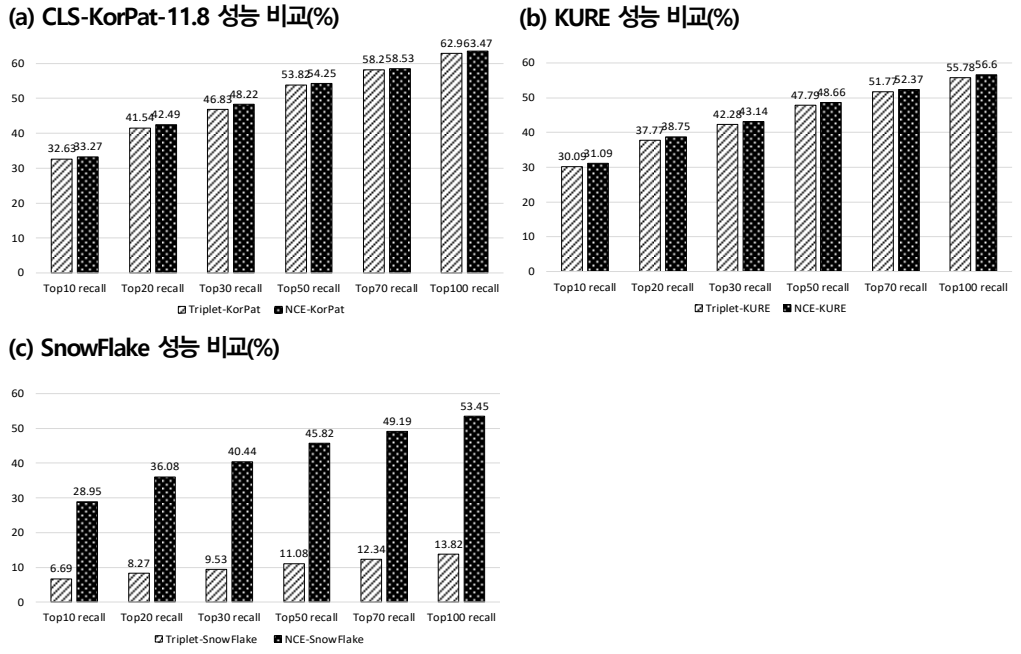
<그림10>, <그림11> 결과에 따르면, 모든 모델에서 Triplet loss 기반 학습 모델 보다 InfoNCE loss 기반 학습 모델의 성능이 우수하였다.

구체적으로, C섹션에서 NCE-KorPat은 Triplet-KorPat 대비 평균 0.73%p, NCE-KURE는 Triplet-KURE 대비 평균 0.83%p 향상되었으며, NCE-SnowFlake는 Triplet-SnowFlake 대비 평균 33.99%p로 큰 폭으로 향상된 결과를 보였다.

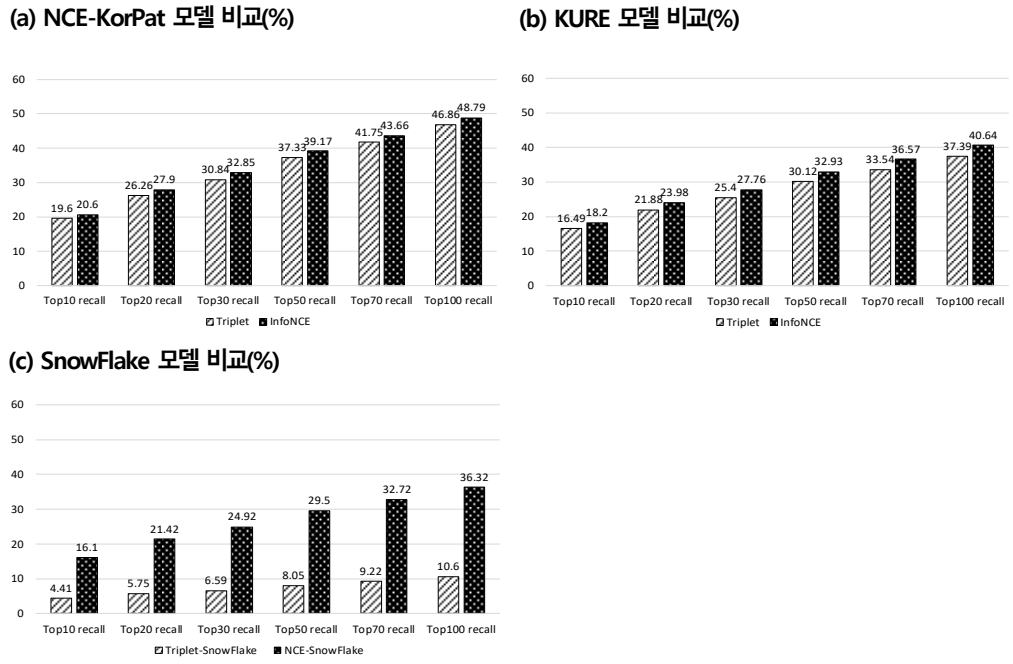
H섹션에서도 NCE-KorPat은 Triplet-KorPat 대비 평균 1.87%p, NCE-KURE는 Triplet-KURE 대비 평균 2.71%p, NCE-SnowFlake는 Triplet-SnowFlake 대비 평균 20.93%p의 성능 향상을 보였다.

NCE-SnowFlake는 기본 모델 대비 성능이 낮게 나타났으나, Triplet 보다 InfoNCE loss 기반 대조 학습 모델이 월등히 우수하였고, 다른 비교 모델에서도 InfoNCE가 Triplet 보다 일관되게 우수하여 특허 검색에 보다 적합한 학습 방식임을 확인하였다.

<그림 10 대조 학습 방법에 따른 검색 성능 비교(C섹션)>



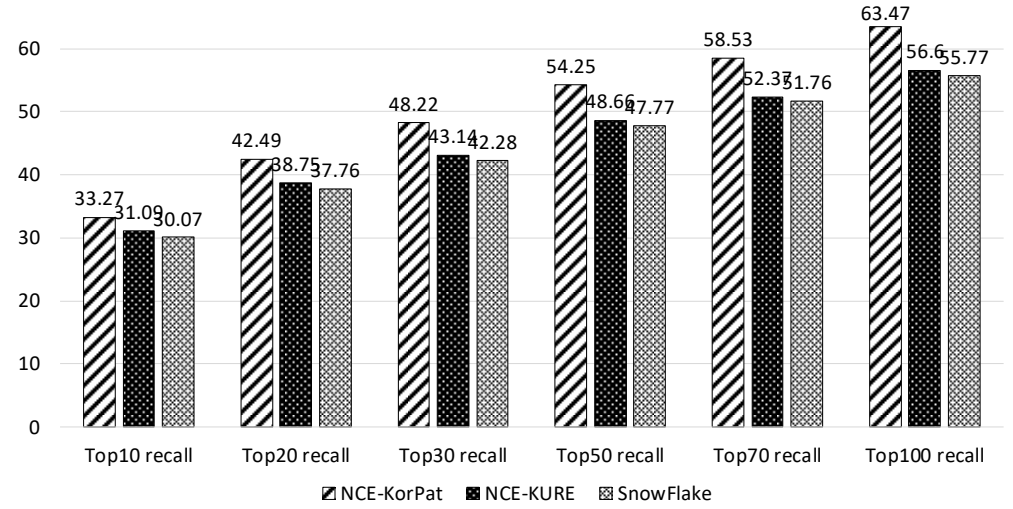
<그림 11 대조 학습 방법에 따른 검색 성능 비교(H섹션)>



세 번째, 모델별 최고 성능을 보인 구성으로 성능을 비교하였다. CLS-KorPat-11.8에서는 InfoNCE 기반으로 학습한 NCE-KorPat과 KURE에서도 동일하게 InfoNCE 기반으로 학습한 NCE-KURE가 최고 성능을 보였다. 반면 SnowFlake는 대조 학습을 적용하지 않은 기본 모델이 가장 우수하였다.

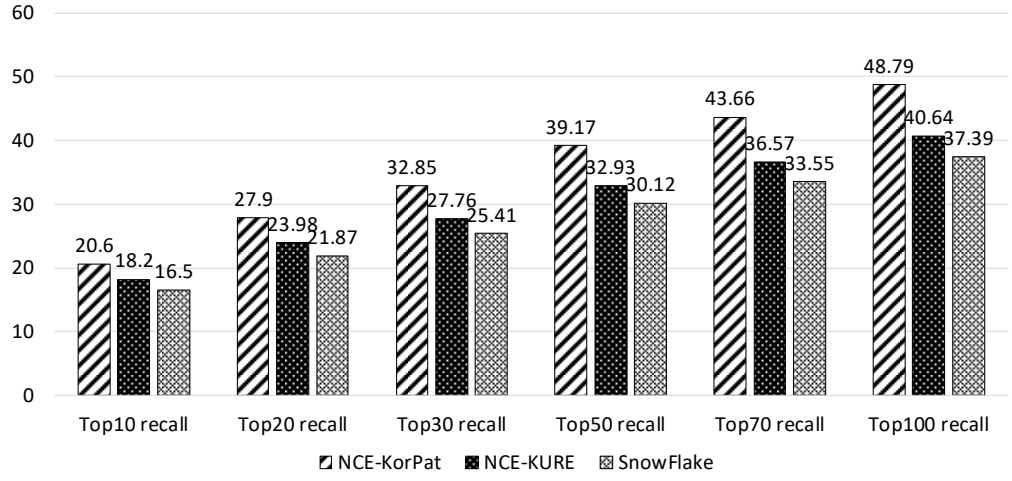
<그림12>, <그림13> 결과에 따르면, KURE와 SnowFlake는 비교적 비슷한 성능 결과를 보였고, 가장 성능이 높은 모델은 NCE-KorPat임을 확인하였다.[40]

<그림12 C섹션 모델별 비교(%)>



40) 전체 결과는 <부록>에 제시하였다.

<그림13 H섹션 모델별 비교(%)>



5. 결론

본 연구는 실제 특허 문헌 검색 시스템을 기반으로 다양한 모델과 대조 학습 방법을 비교하고, Top-K recall 지표로 성능을 정량 평가하였다.

첫째, KorPatSTS 데이터셋을 활용한 결과, CLS-KorPat-11.8과 KURE에서는 InfoNCE loss 기반 대조 학습 모델이 전 구간 Top-K에서 일관된 성능 향상을 보였다. 특히, H섹션에서는 NCE-KorPat이 4.26%p, NCE-KURE가 3.73%p로 개선폭이 C섹션 보다 더 컸다. 검색 데이터셋의 14.99%를 차지하는 H섹션에서 상대적으로 큰 개선이 관찰된 점은 고무적이다.

반면 SnowFlake는 InfoNCE loss 기반 대조 학습 모델이 기본 모델 대비 성능이 낮은 결과를 보였으나, Triplet loss 기반 대조 학습 모델 대비 높은 성능 결과를 보였다.

둘째, 대조 학습의 손실 함수 비교에서 모든 모델에서 InfoNCE가 Triplet보다 우수했다. 이는 텍스트 길이가 길고 언어 다양성이 큰 도메인에서, 배치(In-Batch) 내 다수의 음성을 동시에 활용하는 InfoNCE의 학습 방식이 더 효과적임을 뒷받침한다.

셋째, 모델별 최고 성능의 구성을 기준으로 비교한 결과, NCE-KorPat, NCE-KURE가 각각 최고 성능을 기록하였다. SnowFlake는 대조 학습을 적용하지 않은 기본 모델이 가장 우수했다. KURE와 SnowFlake의 성능은 유사한 수준이었으며, NCE-KorPat이 전체 비교 모델 중 가장 높은 성능을 보였다.

본 연구에서는 1. 서론에서 제시한 연구 목표를 성공적으로 달성하였다.

첫째, 본 연구에서는 특허 산업 분야에서 실질적으로 활용 가능한 검색 시스템 구축을 목표로, 전체 한국 특허 문헌을 대상으로 한 검색 데이터셋과 평가 데이터셋의 체계적인 구축 방법론을 제시하였다. 특히, 실제 심사관의 정보 탐색 과정을 모사할 수 있도록 R&D 및 산업 현장에서의 검색 요구와 차이를 최소화한 실제 검색 환경을 구현함으로써, 연구 결과의 실질적 유용성과 현장 적용 가능성을 크게 제고하였다. 이는 기존 연구들이 주로 실험적 수준에 머물렀던 것과 달리, 실제 활용 환경에 적합한 데이터셋과 시스템 설계 방안을 제시하였다는 점에서 학문적·실무적 의미가 크다.

둘째, 특허 심사의 특수성을 반영하여, 단순한 언어적 유사성 연산을 넘어 기술적 유사성을 보다 정밀하게 포착할 수 있도록 InfoNCE 기반의 대조 학습 방법을 적용하였다. 제안한

NCE-KorPat 모델은 모든 Top-K 구간에서 가장 우수한 성능을 보였으며, 특허 문헌의 내재적 기술 구성을 효과적으로 반영함으로써 실질적인 특허 검색 정확도 향상을 가능하게 하였다. 이는 특허 심사관의 선행기술 조사 효율성을 높이고, 거절 사유 판단의 신뢰성을 강화하는 데 기여할 수 있음을 실험적으로 입증하였다.

셋째, 본 연구에서는 지재처 AI 심사관 자문단의 전문성을 기반으로, 유사 특허 기술의 구조적·언어적 특성에 최적화된 KorPatSTS 데이터셋을 제안하였다. 해당 데이터셋은 청구항 구성요소와 발명의 상세한 설명, 그리고 실제 심사에서 인용된 선행 특허 간의 대응 관계를 정밀하게 구축함으로써, 대조 학습에 적합한 고품질 학습 자원으로서의 활용 가능성을 입증하였다. 나아가, KorPatBERT 기반 CPC 서브그룹 분류 모델에서 KorPatSTS 기반 InfoNCE 대조 학습을 수행한 NCE-KorPat 모델의 우수한 성능을 통해, 해당 데이터셋이 특허 검색 모델의 정밀도 향상에 실질적이고 유의미한 기여를 확인하였다.

본 연구는 특허 심사관의 전문성과 도메인 지식을 결합해 데이터셋을 체계적으로 구축하고 실제 검색 모델에 적용한 최초의 연구로, 특허 분야에서 전문가 주도 데이터셋 구축과 특화 모델의 결합이 검색 성능 향상의 핵심 전략이 될 수 있음을 시사한다. 본 연구를 통해 검증된 검색 성능 향상 효과는 데이터 확보가 충분하고 기술적 중요성이 높은 C(화학 야금) 및 H(전기·전자) 섹션 내에서 특허 심사의 정확성과 효율성을 제고하는 데 실질적으로 기여할 것으로 기대된다. 다만, 기술적 특성이 상이한 타 기술 섹션으로의 범용적 확장을 위해서는 해당 도메인의 특수성을 반영한 추가적인 데이터 구축과 후속 검증이 필요하다.

향후 연구에서는 본 연구에서 축적된 노하우를 기반으로, 특허 심사관이 인용한 문장을 정밀하게 탐색할 수 있는 보다 정교하고 세분화된 검색 기법으로 고도화하고자 한다. 특히 현재의 1:1 대응 평가 모델을 확장하여, 다중 인용 특허 전체를 정답으로 구성하고 검색 모델의 다각적 검색 능력을 정량화하는 연구를 병행하여 평가의 편향성을 해소할 계획이다. 또한, 최신 AI 모델을 적극적으로 적용하여 특허 분야에서 직면한 다양한 문제 해결에 도전할 계획이다.

참고문헌

학술지(국내 및 동양)

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

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

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

유동건·한지헌, “결합발명 진보성 판단의 인용문헌 자동 추천 딥러닝 모델에 관한 연구: BERT-for-patents 및 대조학습 기법을 중심으로”, 「지식재산연구」, 제20권 제1호(2025).

학술지(서양)

Alexander V. Giczy et al., “Identifying Artificial Intelligence (AI) Invention: A Novel AI Patent Dataset”, *The Journal of Technology Transfer,* Vol.47 No.2(2022).

Alfonso F. Cardenas, “Analysis and Performance of Inverted Data Base Structures”, *Communications of the ACM,* Vol.18 No.5(1975).

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).

Julien Denize et al., “Similarity Contrastive Estimation for Image and Video Soft Contrastive Self-Supervised Learning”, *Machine Vision and Applications,* Vol.34 No.6(2023).

Su-Jeong Jeong, “Zur Analyse von mehr oder weniger festen Wortverbindungen in Patentschriften im Deutschen und Koreanischen”, *German Literature,* Vol.26 No.3(2016).

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).

Zhenyu Lu & Yonggang Lu, “A Balanced Triplet Loss for Person Re-Identification”, *International Journal of Pattern Recognition and Artificial Intelligence,* Vol.37 No.1(2023).

인터넷 자료

고려대학교, “KURE”, 고려대 NLP & AI 연구실, <https://hiai.korea.ac.kr/kure/>, 검색일: 2025. 8. 20.

고려대학교, “KURE”, 고려대학교 NLP & AI 연구실 github, <https://github.com/nlpai-lab/KURE?tab=readme-ov-file#mteb-ko-retrieval-leaderboard>, 검색일: 2025. 8. 20.

윤진석, “고려대 임희석 교수님 한국어 특화 임베딩 모델 ‘KURE’ 공개”, 아시아타임즈, <https://www.asiatime.co.kr/article/20241202500146#_mobwcvr>, 검색일: 2025. 8. 20.

지식재산처, “심사명장 소개”, 지식재산처, <https://www.kipo.go.kr/ko/kpoContentView.do?menuCd=SCD0200461>, 검색일: 2025. 8. 20.

지식재산처, “지식재산 심사 기준/매뉴얼”, 지식재산처, <https://www.kipo.go.kr/ko/kpoContentView.do?menuCd=SCD0201119>, 검색일: 2025. 8. 20.

지식재산처, “AI 심사관 자문단 위촉장 수여식”, 지식재산처, <https://www.kipo.go.kr/ko/kpoBultnDetail.do?menuCd=SCD0200615&ntatcSeq=1338&sysCd=SCD02&aprchId=BUT0000026>, 검색일: 2025. 8. 20.

지식재산처, “2024 통계로 보는 특허동향”, 지식재산처 기타 간행물(상세), <https://www.kipo.go.kr/ko/kpoBultnDetail.do?menuCd=SCD0200640&ntatcSeq=16933&sysCd=SCD02&aprchId=BUT0000048#1>, 검색일: 2025. 8. 20.

한국특허정보원, “특허정보 활용 서비스”, 한국특허정보원, <https://plus.kipris.or.kr/portal/main.do>, 검색일: 2025. 8. 20.

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

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. 7. 16.

Evgenia Rusak et al., “InfoNCE: Identifying the Gap Between Theory and Practice”, arXiv, <https://arxiv.org/abs/2407.00143>, 작성일: 2025. 4. 16.

facebookresearch, “facebookresearch/faiss”, facebookresearch Github, <https://github.com/facebookresearch/faiss>, 검색일: 2025. 8. 20.

Hugging Face, “Hugging Face”, Hugging Face, <https://huggingface.co/>, 검색일: 2025. 8. 20.

Julian Risch et al., “PatentMatch: A Dataset for Matching Patent Claims & Prior Art”, arXiv, <https://arxiv.org/abs/2012.13919>, 작성일: 2020. 12. 27.

MTEB, “MTEB”, Embedding Benchmark Github, <https://github.com/embeddings-benchmark/mteb>, 검색일: 2025. 8. 20.

Puxuan Yu et al., “Arctic-Embed 2.0: Multilingual Retrieval Without Compromise”, arXiv, <https://arxiv.org/abs/2412.04506>, 작성일: 2024. 12. 14.

기타자료

Florian Schroff et al., “FaceNet: A Unified Embedding for Face Recognition and Clustering”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015.

Haochen Li et al., “Rethinking Negative Pairs in Code Search”, Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023.

Hong Xuan et al., “Hard Negative Examples Are Hard, but Useful”, European Conference on Computer Vision, Springer International Publishing, 2020.

Jianlv Chen et al., “M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation”, Findings of the Association for Computational Linguistics: ACL, 2024.

Niklas Muennighoff et al., “MTEB: Massive Text Embedding Benchmark”, Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, 2023.

Prannay Khosla et al., “Supervised Contrastive Learning”, Advances in Neural Information Processing Systems 33, 2020.

Tao Zhang & Mingming Hu, “Learning Representation for Clustering via Dual Correlation”, 2023 IEEE 6th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), IEEE, 2023.

Ye Yuan et al., “In Defense of the Triplet Loss Again: Learning Robust Person Re-Identification with Fast Approximated Triplet Loss and Label Distillation”, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020.

Yuanmeng Yan et al., “ConSERT: A Contrastive Framework for Self-Supervised Sentence Representation Transfer”, Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021.

부록

<표9 CLS-KorPat-6.4 검색 실험 결과>

섹션	Top-K recall	CLS-KorPat-6.4	Triplet-KorPat-6.4	NCE-KorPat-6.4
C	5	18.64	23.20	22.29
	10	25.01	29.34	28.97
	20	32.03	36.78	37.32
	30	36.44	41.77	42.26
	50	42.19	47.88	47.99
	70	46.75	51.87	52.40
	100	51.35	56.60	57.32
H	5	10.23	12.58	12.67
	10	14.46	17.56	17.59
	20	20.09	23.91	23.92
	30	24.05	28.21	28.10
	50	29.53	34.19	34.27
	70	33.45	38.41	38.30
	100	37.85	43.08	42.91

<표10 CLS-KorPat-11.8 검색 실험 결과>

섹션	Top-K recall	CLS-KorPat-11.8	Triplet-KorPat-11.8	NCE-KorPat-11.8
C	5	23.43	24.81	25.24
	10	30.77	32.63	33.27
	20	39.37	41.54	42.49
	30	44.50	46.83	48.22
	50	51.19	53.82	54.25
	70	55.87	58.20	58.53
	100	60.88	62.90	63.47
H	5	12.82	13.70	14.38
	10	18.11	19.60	20.60
	20	24.43	26.26	27.90
	30	28.91	30.84	32.85
	50	34.93	37.33	39.17
	70	39.01	41.75	43.66
	100	43.81	46.86	48.79

<표11 KURE 검색 실험 결과>

섹션	Top-K recall	KURE	Triplet-KURE	NCE-KURE
C	5	23.17	23.35	23.87
	10	30.22	30.09	31.09
	20	37.04	37.77	38.75
	30	41.03	42.28	43.14
	50	46.41	47.79	48.66
	70	50.06	51.77	52.37
	100	54.20	55.78	56.60
H	5	12.06	12.26	13.45
	10	16.20	16.49	18.20
	20	21.47	21.88	23.98
	30	24.72	25.40	27.76
	50	29.23	30.12	32.93
	70	32.16	33.54	36.57
	100	35.66	37.39	40.64

<표12 SnowFlake 검색 실험 결과>

섹션	Top-K recall	SnowFlake	Triplet-SnowFlake	NCE-SnowFlake
C	5	23.34	5.03	22.53
	10	30.07	6.69	28.95
	20	37.76	8.27	36.08
	30	42.28	9.53	40.44
	50	47.77	11.08	45.82
	70	51.76	12.34	49.19
	100	55.77	13.82	53.45
H	5	12.27	3.28	11.89
	10	16.50	4.41	16.10
	20	21.87	5.75	21.42
	30	25.41	6.59	24.92
	50	30.12	8.05	29.50
	70	33.55	9.22	32.72
	100	37.39	10.60	36.32