Check for updates

Journal of KIIT. Vol. 24, No. 10, pp. 1-14, Oct. 31, 2026. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2026.24.10.1

합성 임베딩을 통한 교차 분류체계 매핑

곽지호*[1], 김민석*[2], 김준호**, 정유철***

Cross-Taxonomy Mapping via Compositional Embedding

Jiho Gwak*[1], Minseok Kim*[2], Junho Kim**, and Yuchul Jung***

본 연구는 한국과학기술정보연구원(KISTI)의 위탁연구개발과제로 수행한 것입니다. 아울러 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-학·석사연계ICT핵심인재양성 (IITP-2026-RS-2022-00156394) 및 2026년도 교육부 및 경상북도의 재원으로 경북앵커센터의 지원을 받아 수행된 지역성장 인재양성체계(앵커)-(지역성장 혁신LAB)의 결과입니다.(2026-anchor-15-105)

요 약

서로 다른 목적의 분류체계를 연결하는 교차 분류체계 매핑은 보통 두 분류기를 독립 학습해 결합하는 2단계 방식을 사용한다. 이 방식은 분류 오차가 최종 매핑으로 전이되고 학습에서 관측되지 않은 코드 결합을 표현하기 어렵다. 본 논문은 매핑 자체를 하나의 라벨로 보고 문서와 직접 정렬·검색하는 문제로 재정의하고, 문서와 직접 정렬·검색하는 방법을 제안한다. 매핑 라벨은 두 분류 코드 임베딩의 합성으로 표현하며, 양방향 구조적 네거티브 전략으로 두 분류 축의 분리를 유도한다. 이를 통해 미관측 결합까지 좌표로 표현하고, 관측 분포와의 거리로 신규 융합 매핑을 탐지할 수 있다. 두 연계 데이터셋 실험에서 제안 방법은 기존 2단계 방식보다 상위 검색·순위 품질, 롱테일 및 신규 결합 탐지 성능에서 우수한 결과를 보였다.

Abstract

Cross taxonomy mapping links taxonomies with different purposes, typically by combining two independently trained classifiers. However, this two stage approach propagates errors to final mappings and struggles with unseen code combinations. This paper treats each mapping as a single label and directly aligns documents with mapping labels for retrieval. Each label combines two classification code embeddings, while bidirectional structural negatives separate the taxonomy axes. This represents unseen combinations as coordinates and detects novel fusion mappings by distance from the observed distribution. Experiments on two linked datasets show superior top k retrieval, ranking quality, long tail mapping, and novel combination detection over the two stage approach.

Keywords

classification mapping, compositional embedding, contrastive learning, dense retrieval, novel-fusion detection

* 국립금오공과대학교 컴퓨터 · AI융합공학과 석사과정
- ORCID[1]: https://orcid.org/0009-0000-7524-0357
- ORCID[2]: https://orcid.org/0009-0007-9180-7983
** 한국과학기술정보연구원 기술원
- ORCID: https://orcid.org/0009-0001-1750-1616
*** 국립금오공과대학교 인공지능공학과 부교수(교신저자)
- ORCID: http://orcid.org/0000-0002-8871-1979

· Received: Jul. 02, 2026 Revised: Aug 19, 2026 Accepted: Aug. 22, 2026
· Corresponding Author: Yuchul Jung
Dept. of AI Engineering, Kumoh National Institute of Technology, 61 Daehak-ro (yangho-dong), Gumi, Gyeongbuk, [39177] Korea
Tel.: +82-54-478-7536, Email: jyc@kumoh.ac.kr

Ⅰ. 서 론

연구 성과와 산업 활동이 실제 기술 가치로 이어지는 과정은 단일한 분류체계만으로 충분히 설명되기 어렵다. 분류체계는 그 목적에 따라 지식을 서로 다르게 조직한다[1][2]. 과학기술표준분류[3]가 연구 주제와 학문적 지식 구조를 중심으로 한다면[4], 특허분류[5]는 기술의 권리화와 산업적 활용 가능성을 중심으로 하며[6], 한국표준산업분류[7]는 산업 활동의 구조를 중심으로 한다. 이처럼 서로 다른 기준으로 구축된 두 체계는 동일한 문서나 기술 내용도 각기 다른 방식으로 표현하므로, 두 체계를 잇는 분석에서는 의미적 간극이 발생한다. 따라서 교차 분류체계 매핑은 단순한 코드 대응표를 만드는 작업을 넘어, 서로 다른 관점의 지식을 구조적으로 연결하기 위한 핵심 기반이 된다. 이는 기술 동향 분석, 연구 성과의 산업적 파급 경로 추적, 유망 융합기술 탐색, 연구개발 전략 수립 등에 중요한 역할을 한다[8].

기존 접근 방식은 일반적으로 과학기술표준분류기와 특허분류기를 각각 독립적으로 학습한 뒤, 두 분류 결과를 동시출현 기반 매핑 테이블에 연결하는 2단계 구조를 따른다[9]. 이 방식은 각 분류체계를 개별적으로 다룰 수 있다는 점에서 직관적이지만, 교차 분류체계 매핑을 직접 모델링하지 못한다는 한계를 가진다[10].

첫째, 분류 오차가 곧 매핑 오차로 전이된다. 기존 방식에서 최종 매핑은 과학기술표준분류 예측값과 특허분류 예측값의 조합으로 결정된다. 따라서 두 예측 중 하나라도 잘못되면, 최종 매핑 역시 잘못된 조합을 기반으로 산출된다. 즉, 매핑 단계가 존재하더라도 문서와 매핑 간의 관계를 직접 판단하는 것이 아니라, 앞단에서 독립적으로 산출된 분류 결과에 의존한다. 이로 인해 최종 매핑 성능은 두 분류기의 개별 성능에 구조적으로 종속되며, 각 단계의 오류가 최종 결과에 누적된다[11].

둘째, 매핑 자체에 대한 의미 표현이 부재하다. 기존의 2단계 방식은 과학기술표준분류와 특허분류를 각각 예측한 뒤, 두 결과를 사후적으로 연결할 뿐이다. 따라서 과학기술표준분류와 특허분류의 결합 자체가 하나의 의미 단위로 학습되거나 표현되지 않는다. 이 경우 특정 문서가 왜 해당 매핑에 대응되는지, 또는 기존에 관측되지 않은 새로운 조합이 어떤 의미적 근거를 갖는지를 설명하기 어렵다[12]. 매핑 테이블은 단순한 연결 규칙으로 작동할 뿐, 문서와 매핑 사이의 거리, 인접 매핑과의 관계, 과학 축과 특허 축에서의 상대적 위치와 같은 해석 가능한 정보를 제공하지 못한다[13].

본 논문은 이러한 한계를 해결하기 위해 교차 분류체계 매핑을 단순한 분류 결과의 사후 연결 문제가 아니라, 매핑 자체를 하나의 의미 단위로 정렬하고 검색하는 문제로 재정의한다. 구체적으로, 과학기술표준분류와 특허분류를 각각 독립적으로 예측한 뒤 연결하는 것이 아니라, 두 분류 항목의 결합을 하나의 매핑 라벨로 정의한다. 이후 문서와 매핑 라벨을 동일한 공동 임베딩 공간에 사상하고, 문서 임베딩과 매핑 라벨 임베딩 간의 거리를 기반으로 적합한 매핑을 검색·판정한다. 이와 같은 방식은 문서가 직접 서로 다른 분류 체계 결합 공간에서 해석되도록 하며[14], 기존에 관측된 매핑뿐만 아니라 새로운 융합 매핑의 가능성도 거리 기반으로 분석할 수 있게 한다[15].

본 논문의 기술적 공헌은 다음과 같다. 첫째, 교차 분류체계매핑을 두 독립 분류기의 예측을 사후적으로 결합하는 문제가 아니라, 문서와 매핑 라벨 간의 직접적인 정렬 및 검색 문제로 정식화한다. 이를 통해 어느 한 분류기의 오분류가 최종 매핑 오류로 전이되는 2단계 방식의 오차 전파 경로를 제거한다. 둘째, 매핑 라벨을 두 분류 코드 임베딩의 합성으로 표현함으로써, 분류기의 고정된 라벨 집합을 넘어 관측되지 않은 신규 결합에 대한 좌표를 부여할 수 있다. 셋째, 과학-특허 및 산업-특허 데이터셋에서 고정 라벨 검색, 롱테일 매핑 및 신규 결합 탐지를 종합적으로 평가하여 제안 방법의 효과를 검증한다.

본 논문의 구성은 다음과 같다. 제2장에서는 분류체계 매핑, 대조학습 기반 임베딩 및 거리 기반 신규성 탐지에 관한 관련 연구를 살펴본다. 제3장에서는 교차 분류체계 매핑 문제를 형식화하고, 합성 매핑 임베딩과 양방향 구조적 하드 네거티브 학습 방법을 설명한다. 제4장에서는 실험에 사용한 데이

터셋, 비교 방법, 평가지표 및 구현 세부사항을 제시한다. 제5장에서는 고정 라벨 매핑, 롱테일 구간 및 확장 라벨 탐지 결과를 분석하고 신규 융합 매핑 사례를 정성적으로 살펴본다. 마지막으로 제6장에서는 연구 결과를 요약하고 한계와 향후 연구 방향을 논의한다.

II. 관련 연구

2.1 분류와 분류체계 매핑

특허 및 기술문서의 자동 분류는 오랫동안 연구되어 온 주제로[16], 초기에는 TF-IDF와 같은 통계적 자질에 기반한 기계학습 분류기가 주로 사용되었다[17]. 이후 단어 임베딩과 RNN(Recurrent Neural Network), CNN(Convolution Neural Network) 기반 모델이 도입되며 성능이 향상되었고[18], 최근에는 BERT(Bidirectional Encoder Representations from Transformers) 계열의 사전학습 언어모델을 미세조정하는 방식이 표준으로 자리 잡았다[19]. 특히 특허 도메인에 특화된 사전학습 모델이나 한국어 특허·기술문서를 위한 분류 모델이 제안되며, 단일 분류체계 내에서의 범주 예측 성능은 꾸준히 개선되어 왔다[20].

그러나 이러한 연구는 대부분 하나의 분류체계 안에서 문서를 특정 범주로 할당하는 단일 분류 문제에 집중한다. 서로 다른 두 분류체계를 연결하는 교차 분류체계 매핑은 이와 구별되는 문제로, 기존 연구는 주로 두 체계가 함께 부여된 문서의 동시출현 통계를 집계하여 대응표를 구축하거나[21], 각 체계의 분류기를 독립적으로 학습한 뒤 그 예측 결과를 연결하는 2단계 방식을 사용해 왔다. 국제특허분류(IPC)와 경제분류[22], 또는 기술분류와 특허분류를 연계하려는 시도가 이러한 범주에 속한다[23].

2.2 대조학습 기반 문장 임베딩과 밀집 검색

문장이나 문서를 의미 공간의 벡터로 표현하는 임베딩 학습은 의미 유사도 측정과 검색의 기반이 되어 왔다. 초기에는 단어 임베딩의 평균이나 순환신경망 기반 표현이 사용되었으나, 사전학습 언어모델이 등장하면서 문장 단위 표현의 품질이 크게 향상되었다[24]. 특히 SBERT[25] 계열의 연구는 BERT를 문장 임베딩에 적합하도록 재구성하여 문장 간 유사도를 효율적으로 계산할 수 있게 하였다. 이러한 표현 학습에서 핵심적인 역할을 하는 것이 대조학습(Contrastive learning)이다. 대조학습은 의미적으로 관련된 표본을 가깝게, 무관한 표본을 멀게 배치하도록 임베딩 공간을 학습하며, InfoNCE 손실이 널리 사용된다[26]. 이때 어떤 부정 표본(Negative)을 사용하는지가 표현 품질을 좌우하는데, 정답과 유사하지만 다른 하드 네거티브(Hard negative)를 활용하면 더 변별력 있는 표현을 학습할 수 있다고 알려져 있다[27]. 다만 기존의 하드 네거티브는 주로 구분 난이도를 높여 표현력을 강화하는 데 목적을 두며, 본 논문은 이를 두 분류 축의 변별 정보가 각각 반영되도록 유도하는 방향으로 확장한다는 점에서 차이가 있다.

2.3 밀집 검색과 거리 기반 탐지

대조학습으로 학습된 임베딩은 밀집 검색의 토대가 된다. 질의와 문서를 같은 공간에 사상한 뒤 벡터 유사도로 관련 대상을 검색하는 방식은, 어휘 일치에 의존하는 전통적 검색을 넘어 의미 기반 검색을 가능하게 한다[28]. 본 논문은 이러한 정렬·검색 패러다임을 문서와 매핑 라벨 정렬로 확장한다. 일반적인 밀집 검색이 질의와 문서를 정렬하는 데 반해, 본 논문은 검색 대상을 두 분류 코드의 합성으로 구성된 매핑 라벨로 둔다는 점에서 구별된다.

밀집 검색의 틀은 신규성 판정으로도 자연스럽게 확장된다. 질의를 관측된 검색 대상들의 분포와 비교할 때, 어떤 대상과도 충분히 가깝지 않은 질의는 기존 항목으로 설명되지 않는 신규 입력으로 볼 수 있다[29]. 본 논문은 이 관점을 매핑에 적용하여, 문서가 관측된 매핑점들로부터 일정 거리 이상 떨어져 있으면 신규·융합 매핑으로 판정한다. 이는 분류기의 확신도와 같은 간접 신호에 의존하지 않고, 매핑이 놓인 표현 공간에서 거리로 신규성을 직접 측정한다는 점에서 분류 기반 방식과 구별된다.

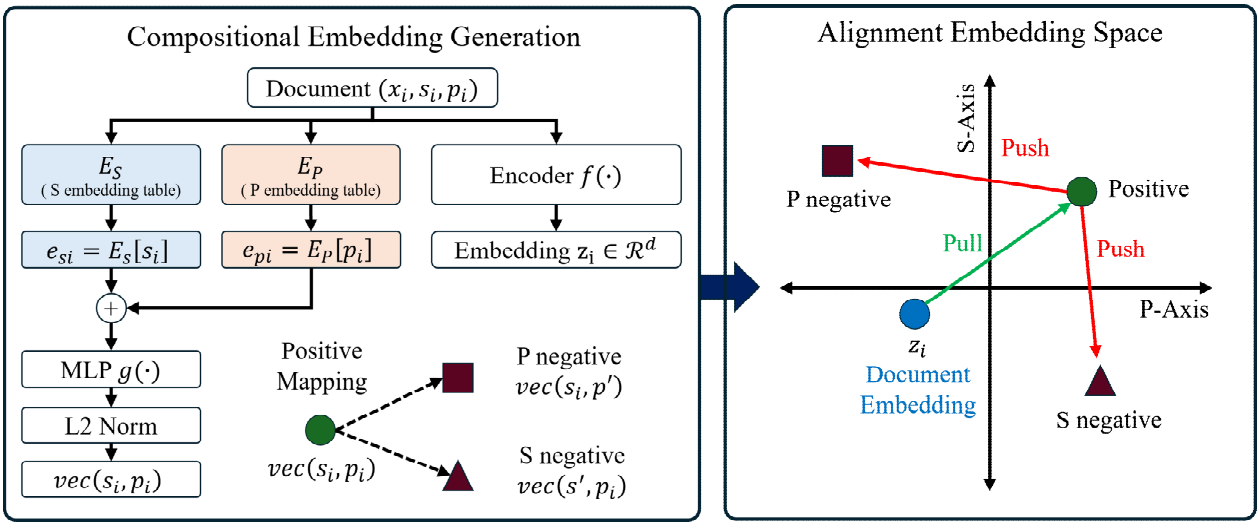


그림 1. 합성 매핑 임베딩 생성 및 정렬 과정
Fig. 1. Compositional mapping embedding generation and alignment

III. 제안 프레임워크: 합성 매핑 임베딩을 통한 정렬·검색

3.1 문제 형식화

제안하는 매핑 정렬 및 학습 프레임워크의 전체 구조는 그림 1과 같다. 문서 집합 X의 각 문서 x_i에는 서로 다른 두 분류체계의 라벨 $s_i \in S$와 $p_i \in P$가 함께 연결되어 있다고 가정한다. 여기서 S와 P는 특정 도메인에 고정된 의미가 아니라, 매핑 대상이 되는 두 분류체계를 일반적으로 나타낸다. 본 연구의 과학-특허 매핑에서는 S가 과학기술표준분류, P가 특허분류를 의미하며, 산업-특허 매핑에서는 S가 한국표준산업분류, P가 특허분류를 의미한다. 이하에서는 표기의 편의를 위해 전자의 매핑을 과학-특허(SCI-PAT), 후자의 매핑을 산업-특허(IND-PAT)로 각각 지칭한다.

그림 1의 왼쪽은 문서 임베딩과 합성 매핑 임베딩의 생성 과정 및 두 가지 라벨 구성 방향의 구조적 네거티브 샘플 구성 과정을 나타낸다. 그림 1의 오른쪽은 문서 임베딩, 정답 매핑점, 두 방향에서 구성된 구조적 네거티브를 공동 임베딩 공간에 배치하여 정렬하는 과정을 나타낸다. 매핑 임베딩의 구체적인 합성 방식과 학습 방법은 각각 3.2절, 3.3절, 3.4절에서 설명한다.

3.2 매핑 라벨의 합성 표현

매핑 라벨을 표현하는 가장 단순한 방법은 각 매핑을 서로 독립적인 항목으로 간주하고, 관측된 매핑마다 하나의 임베딩 벡터를 할당하는 것이다. 그러나 이러한 표현은 관측된 매핑 집합에 포함된 매핑에 대해서만 좌표를 가지므로, 학습에서 관측되지 않은 라벨 조합을 표현하기 어렵다.

이러한 한계를 완화하기 위해 매핑 라벨 $m = (s, p)$을 두 분류 코드 임베딩의 합성 표현으로 정의한다. 두 분류체계 S와 P에 대한 코드 임베딩 테이블을 각각 E_S와 E_P로 정의하고, 라벨 s와 p에 대응하는 코드 임베딩을 식 (1)와 같이 조회한다.

$$e_s = E_S[s], e_p = E_P[p] \tag{1}$$

두 코드 임베딩은 연결 연산을 통해 하나의 벡터로 결합되고, 합성 함수 $g(\cdot)$와 L2 정규화를 거쳐 식 (2)와 같이 매핑 임베딩으로 변환한다.

$$\overrightarrow{(m)} = \mathrm{normalize}(g([e_S; e_P])) \tag{2}$$

여기서 $[\cdot\,;\cdot]$는 두 벡터의 연결(Concatenation)을 의미하며, 합성 함수 g는 다층 퍼셉트론으로 구성된다. 해당 합성 함수는 두 코드 임베딩 간의 복잡한

순차적 의존성을 모델링하는 것이 아니라, 두 임베딩을 결합하여 문서 임베딩과 동일한 차원의 표현으로 변환하는 것이 주된 목적이다. 따라서 Transformer 또는 Attention 구조 대신 상대적으로 경량인 다층 퍼셉트론을 사용하였다. normalize는 생성된 매핑 임베딩을 단위 벡터로 변환하는 L2 정규화 연산이다.

문서 x는 문서 인코더 $f(\cdot)$를 통해 d차원의 문서 임베딩으로 변환되며, 매핑 임베딩과 마찬가지로 L2 정규화된다. 문서와 매핑 라벨 간의 정렬 정도는 식 (3)의 코사인 유사도로 측정한다.

$$\sim(\mathrm{x},\mathrm{m}) = \frac{\mathrm{f}(\mathrm{x})^{\mathrm{T}} \cdot \overrightarrow{(\mathrm{m})}}{|\mathrm{f}(\mathrm{x})| \cdot |\overrightarrow{(\mathrm{m})}|} \tag{3}$$

문서 임베딩과 매핑 임베딩이 모두 단위 벡터로 정규화되므로, 식 (3)의 코사인 유사도는 두 벡터의 내적과 동일하다. 이 유사도는 학습 단계에서는 문서와 정답 매핑 및 구조적 네거티브를 정렬하는 기준으로 사용되며, 추론 단계에서는 고정 라벨 매핑 검색과 신규 융합 탐지에 활용된다.

이하에서는 학습 데이터를 M, 실제로 관측된 매핑 라벨의 집합을 M_{obs}로 정의한다. 먼저 고정 라벨 매핑 검색에서는 식 (4)와 같이 관측 매핑 집합에서 입력 문서 x와 가장 높은 유사도를 갖는 매핑을 선택한다.

$$\hat{m}x = \mathrm{argmax}_{\mathrm{m}\in\mathrm{M}_{\mathrm{obs}}} \sim(\mathrm{x},\mathrm{m}) \tag{4}$$

예측 결과는 두 분류 라벨의 결합인 $\hat{m}_x = (\hat{s},\hat{p})$이며, 문서 x의 정답 매핑 라벨이 모두 일치하는 경우에 올바른 예측으로 평가한다.

신규 융합 탐지에서는 입력 문서가 기존 관측 매핑으로 충분히 설명되는지를 판단한다. 이를 위해 식 (5)과 같이 문서와 가장 유사한 관측 매핑점 사이의 코사인 거리를 신규성 점수로 정의한다.

$$\delta_{\min}(x) = 1 - \max_{m\in M_{obs}} \sim(\mathrm{x},\mathrm{m}) \tag{5}$$

신규성 점수 $\delta_{\min}(x)$가 임계(Υ)를 초과하면, 즉 $\delta_{\min}(x)>\Upsilon$ 이면 기존 어떤 대응에도 속하지 않는 신규/융합으로 판정한다.

3.3 네거티브 데이터 구성

합성 표현이 의미를 가지기 위해서는 두 분류체계의 정보가 모두 매핑 임베딩에 반영되어야 한다. 그러나 두 축의 라벨이 모두 다른 일반적인 네거티브를 사용할 경우, 정답 매핑과 네거티브 사이의 차이가 어느 분류 축에서 발생한 것인지 구분하기 어렵다. 따라서 모델이 한쪽 라벨에 공통으로 나타나는 의미 성분과 다른 축의 변화에 따라 달라지는 의미 성분을 분리하여 학습하기 어렵다.

두 분류 축의 의미 성분을 구분하기 위해서는 동일한 라벨이 서로 다른 상대 축의 라벨과 반복적으로 결합하는 구조가 필요하다. 예를 들어 동일한 S가 여러 P와 결합하는 경우, 각 매핑에 공통으로 나타나는 성분은 S축의 정보로, P의 변화에 따라 달라지는 성분은 P축의 정보로 추정할 수 있다. 반대로 동일한 P가 여러 S와 결합하는 경우에는 P축에 공통된 성분과 S축에 따라 달라지는 성분을 구분할 수 있다. 반면 S와 P가 항상 일대일로만 결합한다면 두 라벨의 효과를 분리하기 어려우며, 해당 결합은 사실상 하나의 원자적 매핑 라벨과 같이 작동한다.

이에 본 논문은 정답 매핑 (s_i,p_i)을 기준으로 한쪽 축의 라벨을 유지하고 다른 축의 라벨만 변경하는 방식으로 네거티브를 구성한다. 먼저 동일한 S축 라벨을 공유하지만 P축 라벨이 다른 매핑을 P축 네거티브로 정의한다. 이 네거티브는 s가 동일하므로, 정답 매핑을 구분하려면 p의 차이를 반영해야 한다. 반대의 경우에도 동일하게 정의된다. 따라서 P축 및 S축 구조적 하드 네거티브 집합은 각각 식 (6)과 식 (7)로 정의한다.

$$N_P(s,p) = (\mathrm{s},\mathrm{p}')\in\mathrm{M}_{\mathrm{obs}}|\mathrm{p}'\neq\mathrm{p} \tag{6}$$

$$N_S(s,p) = (\mathrm{s}',\mathrm{p})\in\mathrm{M}_{\mathrm{obs}}|\mathrm{s}'\neq\mathrm{s} \tag{7}$$

이러한 구성은 한쪽 분류 축을 동일하게 유지한 상태에서 다른 축의 차이를 비교하므로, 모델이 어

느 한 축의 정보에만 의존하는 것을 줄인다. 기존 하드 네거티브 방식은 문서와 유사한 오답을 선택하여 정답과의 구분 난이도를 높이는 데 초점을 둔다. 그러나 교차 분류체계 매핑에서는 선택된 네거티브의 두 축이 모두 달라질 수 있어, 정답과의 차이가 어느 축에서 비롯된 것인지 구분하기 어렵다. 반면 제안 방식은 S축을 고정하고 P축만 변경하는 경우와 P축을 고정하고 S축만 변경하는 경우를 함께 구성함으로써, 각 비교에서 변화한 축의 차이가 매핑 표현에 반영되도록 한다. 결과적으로 양방향 구조적 네거티브는 단순히 구분 난이도를 높이는 것을 넘어, 합성 함수가 두 분류체계의 정보를 모두 활용하도록 유도한다. 각 방향의 구조적 네거티브 후보 중 문서 임베딩과의 유사도가 높은 상위 k개를 하드 네거티브로 선택하며, 본 실험에서는 방향별 최대 10개를 사용하였다.

3.4 InfoNCE 기반 문서-매핑 임베딩 정렬

학습의 목표는 문서 임베딩 $f(x_i)$를 정답 매핑점에 가깝게 정렬하고, 3.3절에서 구성한 네거티브 매핑 임베딩과는 멀어지도록 하는 것이다. 이를 위해 문서와 매핑 간의 코사인 유사도를 기반으로 InfoNCE 손실을 적용한다.

문서 x_i의 정답 매핑을 $m^+ = (s_i, p_i)$라 하고, 네거티브 매핑 집합을 N_i라 할 때 손실 함수는 식 (8)과 같이 정의한다.

$$L(x) = -\log\left[\frac{\exp(\sim(\mathrm{x},\mathrm{m}^{+)/}\tau)}{\exp(\sim(\mathrm{x_i},\mathrm{m_i^+}/\tau)) + \Sigma_{\mathrm{m}^-\in N_i}\exp(\sim(\mathrm{x},(\mathrm{m}^-)/\tau)}\right] \quad (8)$$

여기서 τ는 유사도 분포의 집중도를 조절하는 온도 파라미터이며, N_i는 3.3절에서 정의한 P축 및 S축 네거티브로 구성된다. 식 (8)을 최소화하면 문서 임베딩과 정답 매핑 임베딩 간의 유사도는 증가하고, 네거티브 매핑 임베딩과의 유사도는 감소한다. 이에 따라 그림 1의 오른쪽과 같이 문서 임베딩은 정답 매핑점에 가까워지고 네거티브 매핑점과는 멀어지도록 공동 임베딩 공간에 정렬된다.

Ⅳ. 실험 설계

4.1 데이터셋

본 논문은 하나의 문서에 두 분류체계의 라벨이 함께 부여된 데이터셋을 활용한다. 이러한 데이터는 한 문서에 과학기술표준분류와 특허분류가 함께 부여되거나, 한국표준산업분류와 특허분류가 함께 부여된 형태이다. 제안 프레임워크는 이러한 두 종류의 데이터셋에서 평가한다.

데이터셋은 AI-Hub에서 제공하는 과학기술표준분류 대응 특허데이터[30]와 특허 분야 자동분류 데이터[31]를 사용한다. 해당 데이터셋은 각 문서의 제목, 초록, 청구항 등으로 구성되어 있으며, 하나의 문서에 두 개의 라벨이 모두 부여되어 있다. 이 정보를 매핑의 목적에 맞게 정제하여 활용한다.

학습 검증 분할은 기존 데이터셋과 동일하게 유지하였으며, 구축된 데이터셋의 상세 통계는 표 1과 같다. 각 분류체계는 본래 다층적인 계층 구조를 지니고 있으나, 최하위 계층으로 갈수록 가능한 매핑쌍의 수가 기하급수적으로 증가하여 라벨당 샘플 희소성 문제가 심화된다. 반면 최상위 계층은 지나치게 추상적이어서 교차 매핑의 실익이 낮다. 따라서 본 논문은 정보의 손실을 최소화하면서도 모델의 학습 안정성을 확보할 수 있는 최적의 해상도로서, 과학기술표준분류 및 한국표준산업분류는 중분류, 특허분류는 서브클래스 단위로 라벨 공간을 정의한다. 입력 텍스트는 제목, 초록, 청구항을 결합하여 하나의 입력문으로 구성한다.

표 1. 데이터셋 통계
Table 1. Dataset statistics

Domain	Train	Validation	Paired mapping
SCI-PAT	240,187	30,024	15,281
IND-PAT	443,488	56,392	6,130

한편 검증 집합에는 학습 집합에서 관측되지 않은 매핑쌍을 가진 문서가 존재하며, 이는 확장 라벨 평가의 신규(Novel) 대상이 된다. 이러한 신규 문서는 과학-특허 데이터에서 809건(검증의 약 2.7%), 산업-특허 데이터에서 496건(약 0.88%)이다. 주목할 점은 이들이 미관측 개별 라벨이 아니라 미관측 결

합이라는 것이다. 두 데이터셋 모두 검증 집합의 개별 라벨은 전부 학습 집합에 존재하므로, 신규 문서는 두 라벨이 모두 관측되었으나 그 결합만 새롭게 나타난 경우에 해당한다.

4.2 비교 방법

제안 방법을 다음과 같이 세 가지 접근 방식으로 비교한다.

첫째, 2단계 분류 방식은 두 분류체계 각각에 대해 분류기를 독립적으로 학습한 뒤 그 결과를 결합하는 기존 방식을 대표한다. 본 논문에서는 두 분류체계에 대해 각각 KoBERT[24] 분류기를 학습하고, 각 분류기의 최상위 예측을 결합하여 한 쌍의 매핑을 출력한다. 이 방식은 매핑쌍에 대해 단일 후보만 산출하므로, 상위 후보 집합을 활용하는 순위 기반 평가에서는 최상위 예측으로 측정된다.

둘째, 사전학습(Pre-trained) 임베딩 방식은 별도의 미세조정 없이 사전학습 임베딩만으로 구성한 베이스라인이다. 각 매핑 라벨에 속한 학습 문서들의 임베딩 평균을 그 매핑의 프로토타입으로 두고, 검증 문서를 가장 가까운 프로토타입으로 검색한다. 이는 합성 학습이 가져오는 효과를 분리해 확인하기 위한 대조군이다.

셋째, 제안 방법은 3장에서 기술한 합성 매핑 임베딩으로, 매핑 라벨을 두 코드 임베딩의 합성으로 표현하고 양방향 구조적 하드 네거티브와 InfoNCE 손실로 문서와 매핑 라벨을 정렬한다.

사전학습 임베딩 방식과 제안 방법은 문서 인코더 백본에 따라 성능이 달라질 수 있으므로, 한국어 특화·다국어 임베딩 모델을 아우르도록 Ko-SRoBERTa[32], Qwen3-Embedding-0.6B[33] 두 가지 백본에 대해 각각 평가하고, 사전학습 비교에는 KoBERT, Qwen3-8B를 추가로 포함한다. 2단계 분류는 분류기 기반이므로 KoBERT를 파인튜닝하여 사용한다.

4.3 평가지표

평가는 매핑 검색 성능을 보는 고정 라벨 방식과 신규 융합 탐지 성능을 보는 확장 라벨 방식으로 구성한다.

고정 라벨 방식에서는 검증 문서를 관측된 매핑 라벨 집합에 대해 검색하고, 그 문서의 정답 (s, p) 쌍과 일치하는지를 쌍 단위로 측정한다. 한쪽 축만 맞은 경우는 오답으로 처리하여 두 축을 모두 맞춘 경우만 정답으로 인정한다. 상위 k개 후보 안에 정답 매핑이 포함되는 비율을 쌍 단위 Recall@k(Pair-R@k, 이하 R@k)로 정의하며, 식 (9)와 같이 계산한다.

$$R@k = (1/N)\Sigma_x[rank(x) \le k] \tag{9}$$

또한, 전체 평균인 MRR과 함께 매핑쌍 단위로 평균을 산출한 macro-MRR를 함께 보고한다. macro는 희소한 매핑쌍에서도 동일한 가중치를 부여하므로 롱테일 구간의 성능을 반영하며 MRR 계산은 식 (10)과 같이 계산한다.

$$MRR = (1/N)\Sigma_x 1/rank(x) \tag{10}$$

또한 매핑쌍의 학습 빈도에 따라 head·mid·tail 세 구간으로 나누어 구간별 MRR을 제시하고, 최상위로 예측한 매핑에서 각 축을 맞춘 비율을 참고 지표로 보고한다.

확장 라벨에서는 검증 집합 중 학습에서 관측되지 않은 매핑쌍을 가진 문서에 대해 평가한다. 각 문서에 대해 최근접 매핑점까지의 거리를 신규성 점수로 사용하여, 분류 기반 2단계 방식은 연속적인 거리 점수가 없으므로 두 분류기의 최대 소프트맥스 확신도(MSP, Maximum Softmax Probability)의 곱을 신규성 점수로 사용한다. 지표로는 AUROC와 FPR@95TPR(참 양성률 95%에서의 거짓 양성률)을 사용한다. 신규 문서 집합을 X_{nov}, 기존 문서 집합을 X_{seen}이라 하고, 각각에서 선택한 문서를 $x^+ \in X_{nov}, x^- \in X_{seen}$으로 정의한다. 이때 AUROC는 신규 문서의 점수가 기존 문서의 점수보다 높을 확률로 계산되며 식 (11)과 같이 계산한다.

$$AUROC = (1/(N_{nov} \cdot N_{seen})) \times \Sigma_{x^+ \in X_{nov}} \Sigma_{x^- \in X_{seen}} (score(x^+) > score(x^-)) \tag{11}$$

4.4 구현 세부

문서 인코더 백본은 Ko-SRoBERTa, Qwen3-Embedding-0.6B를 사용한다. 코드 임베딩 차원 c는 256으로 두고, 매핑점과 문서 임베딩은 백본의 은닉 차원 d로 정렬한다. 합성 함수 g는 결합된 코드 임베딩(차원 2c)을 입력받아 d차원으로 변환하는, GELU 활성을 포함한 다층 퍼셉트론이다.

학습은 InfoNCE 온도 $\tau = 0.05$, 방향별 구조적 하드 네거티브 최대 10개, AdamW 옵티마이저(가중치 감쇠 0.01), 학습률 2×10^{-5}, 배치 크기 32, 10 에폭으로 수행하며, 워밍업 비율은 0.1로 둔다. 디코더 계열 백본은 학습 안정성을 위해 학습률을 절반으로 낮춘다. 문서 임베딩과 매핑점은 모두 L2 정규화하여 코사인 유사도로 정렬한다.

2단계 분류의 KoBERT 분류기는 각 분류체계에 대해 학습률 2×10^{-5}, 최대 입력 길이 512로 미세조정한다. 사전학습 임베딩 베이스라인은 별도 학습 없이 각 백본의 임베딩으로 매핑별 프로토타입을 구성한다.

모든 학습은 RTX 4090 2장 환경에서 수행하며, 문서 인코더는 데이터 병렬로 처리한다. 코드 임베딩 테이블과 매핑점은 32비트 부동소수점으로 유지하여 정규화·거리 계산의 수치 안정성을 확보한다.

V. 결과 및 분석

5.1 단일 분류기 성능

표 2는 2단계 방식의 기반이 되는 KoBERT 분류기 성능이다. 과학기술표준분류는 정확도 0.684, 산업분류는 0.818, 특허분류는 0.718로 단일 분류체계 내에서는 비교적 안정적인 성능을 보인다.

표 2. 각 분류체계에 대한 KoBERT 분류기의 단일 분류 성능
Table 2. Single-classification performance of the KoBERT classifier for each taxonomy

Domain	Acc	Macro-F1	Weighted-F1
Science	0.684	0.679	0.682
Industry	0.818	0.595	0.811
Patent	0.718	0.301	0.699

다만 특허분류는 정확도 0.718에 비해 Macro-F1이 0.301로 크게 낮고 산업분류도 Macro-F1이 0.595에 그쳐, 빈도가 높은 라벨은 잘 분류하지만 희소 라벨은 거의 분류하지 못하는 롱테일 편향을 드러낸다.

더 근본적인 문제는 이러한 분류 성능이 매핑으로 결합될 때 그대로 유지되지 못한다는 점이다. 2단계 방식에서 매핑이 정답이 되려면 두 분류 축이 동시에 맞아야 하므로, 매핑 정확도는 두 분류 정확도의 곱에 가깝게 하락한다. 표 3의 2단계 분류 결과가 이를 직접 보여준다. 과학-특허에서 두 분류 정확도의 곱($0.684 \times 0.718 \approx 0.491$)은 표 3의 2단계 매핑 R@1(0.496)과 거의 일치하며, 산업-특허에서도 곱($0.818 \times 0.718 \approx 0.587$)이 R@1(0.578)에 근접한다. 즉 각 분류기가 개별적으로는 0.68~0.82의 정확도를 보이더라도, 두 결과를 결합하는 순간 한쪽의 오차가 다른 쪽의 정답을 무효화하여 매핑 정확도가 0.5~0.58 수준으로 떨어진다. 이는 분류 결과를 사후에 결합하는 2단계 구조에서 두 분류기의 오차가 곱해져 누적되는 오차 전파를 정량적으로 보여준다. 제안 방법은 문서를 매핑 라벨에 직접 정렬함으로써 이러한 단계적 오차 누적을 거치지 않는다.

5.2 고정 라벨(Closed-set) 매핑 성능

표 3은 두 도메인의 고정 라벨 매핑 검색 성능을 보여준다. 학습을 포함하는 2단계 방식과 제안 방법은 서로 다른 세 시드로 실행한 평균과 표준편차를 보고하며, 별도의 미세조정을 거치지 않는 사전학습 방식은 단일 값으로 제시한다. 학습 기반 방법의 표준편차는 모두 0.006 이하로 나타나 반복 실험에서도 안정적인 결과를 보였다.

최상위 정확도(R@1)에서 과학-특허 도메인은 2단계 분류(0.496)와 제안 방법(Ko-SRoBERTa 0.497, Qwen3-0.6B 0.503)이 유사한 수준을 보였다. 반면 산업-특허 도메인에서는 제안 방법이 0.614~0.623으로 2단계 방식의 0.578을 상회하였다. 상위 후보를 함께 고려하면 차이는 더욱 분명해진다. 제안 방법은 R@5에서 과학-특허 0.837~0.859, 산업-특허 0.903~0.912를 기록하였으며, MRR에서도 2단계 방식보다 우수하였다. 이는 문서를 단일 매핑으로 즉시 결정하는 대신, 매핑 후보를 순위화하여 상위권

에 배치할 수 있다는 제안 방식의 특성을 보여준다. 사전학습 임베딩만 사용한 경우에는 모든 백본에서 제안 학습보다 낮은 성능을 보였다. 특히 KoBERT 기반 프로토타입의 R@1은 과학-특허 0.105, 산업-특허 0.171로 가장 낮았다. 사전학습 상태에서는 Qwen3 계열이 Ko-SRoBERTa보다 높은 검색 성능을 보였으나, 제안 학습 이후에는 Ko-SRoBERTa와 Qwen3-0.6B 간의 성능 차이가 크게 줄어들었다. 이는 초기 임베딩 품질뿐 아니라 문서와 합성 매핑을 직접 정렬하는 학습 과정이 최종 성능에 중요한 영향을 미침을 보여준다.

또한 표 4의 절제 연구에서는 제안 방법의 주요 구성에 따른 성능 변화를 분석하였다. 원자적 매핑 임베딩과 일반 하드 네거티브를 사용한 기준 조건에 비해, 합성 표현과 구조적 네거티브를 적용하면서 검색 성능이 전반적으로 향상되었다. 특히 동일한 합성 표현을 사용하는 조건에서 단방향 구조적 네거티브를 양방향으로 확장하면 R@1은 0.463에서 0.497, MRR은 0.515에서 0.540, macro-MRR는 0.403에서 0.437로 향상되었다. 이는 두 분류 축을 모두 고려한 네거티브 구성이 매핑 표현의 변별력을 높이는 데 기여함을 보여준다.

이러한 경향은 희소 매핑을 동일한 가중치로 평가하는 macro-MRR에서 더욱 뚜렷하다. 이는 빈도가 높은 매핑뿐 아니라 상대적으로 학습 사례가 적은 매핑에서도 제안 방법이 보다 안정적인 순위 품질을 유지함을 의미한다.

그림 2는 매핑쌍의 빈도를 3단계로 균등하게 분배하여 구간별 성능을 보여준다. tail로 갈수록 학습 데이터가 적은 것을 의미한다. 모든 방법에서 head에서 tail로 갈수록 성능이 전반적으로 감소한다. 기존 방식은 mid, tail 구간에서 성능 저하가 크게 나타나는 반면, 제안 방법은 상대적으로 완만한 감소를 보인다. 이는 제안 방법이 특정 매핑쌍의 관측 빈도가 낮더라도, 이를 구성하는 개별 라벨이 다른 조합에서 충분히 관측되었다면 해당 정보를 추가로 활용하여 매핑 표현을 형성할 수 있음을 의미한다.

표 3. 도메인별(과학기술표준분류-특허분류, 한국표준산업분류-특허분류)의 매핑 검색 성능
Table 3. Mapping retrieval performance by domain (SCI - PAT, IND - PAT)

Domain	Method	Backbone	R@1	R@3	R@5	MRR	macro-MRR
SCI-PAT	2-stage	KoBERT	0.496±0.004	-	-	0.496±0.004	0.285±0.006
	Pre-trained	KoBERT	0.105	0.210	0.272	0.190	0.174
		Ko-SRoBERTa	0.213	0.362	0.478	0.337	0.299
		Qwen3-0.6B	0.303	0.481	0.637	0.453	0.358
		Qwen3-8B	0.306	0.549	0.653	0.462	0.410
	Proposal	Ko-SRoBERTa	0.497±0.003	0.756±0.002	0.837±0.002	0.540±0.003	0.437±0.005
		Qwen3-0.6B	**0.503±0.004**	**0.777±0.003**	**0.859±0.002**	**0.657±0.003**	**0.480±0.005**
IND-PAT	2-stage	KoBERT	0.578±0.004	-	-	0.578±0.004	0.321±0.006
	Pre-trained	KoBERT	0.171	0.332	0.414	0.287	0.227
		Ko-SRoBERTa	0.384	0.647	0.718	0.5143	0.479
		Qwen3-0.6B	0.402	0.665	0.758	0.5591	0.483
		Qwen3-8B	0.426	0.711	0.804	0.597	0.491
	Proposal	Ko-SRoBERTa	0.614±0.003	0.853±0.002	0.903±0.001	0.741±0.002	0.515±0.005
		Qwen3-0.6B	**0.623±0.004**	**0.867±0.003**	**0.912±0.002**	**0.748±0.003**	**0.523±0.005**

표 4. Ko-SRoBERTa 기반 과학-특허 매핑 절제 연구
Table 4. Ablation Study of Ko-SRoBERTa-Based SCI-PAT Mapping

Method		R@1	R@3	R@5	MRR	macro-MRR
Atomic	General	0.256±0.005	0.633±0.003	0.711±0.005	0.455±0.003	0.312±0.002
Compositional	Unidirectional	0.463±0.003	0.747±0.005	0.817±0.003	0.515±0.002	0.403±0.004
	Bidirectional	0.497±0.003	0.756±0.002	0.837±0.002	0.540±0.003	0.437±0.005

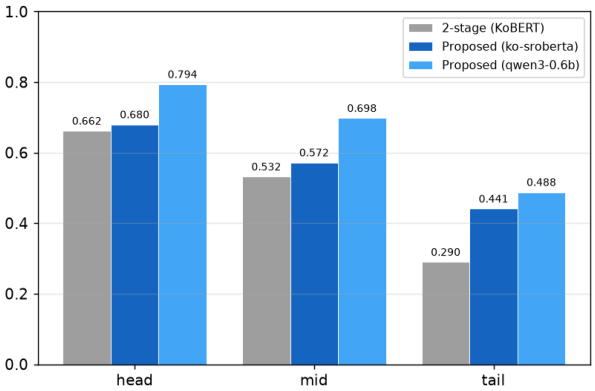


그림 2. 빈도 구간별 매핑 성능
Fig. 2. Mapping performance by frequency interval

5.3 확장 라벨(Open-set) 매핑 성능

표 5는 학습에서 관측되지 않은 매핑쌍을 탐지하는 확장 라벨 성능이다. 두 도메인 모두 제안 방법이 2단계 분류의 MSP 기반 탐지보다 우수하다. 과학-특허에서 제안 방법은 AUROC 0.823(Ko-SRoBERTa)으로 2단계(0.735)를 앞서고, 산업-특허에서도 제안 0.783 대 2단계 0.712이다. FPR@95TPR 또한 두 도메인 모두 제안 방법이 더 낮아, 동일한 탐지율에서 오탐이 적다(과학 0.592 대 0.778, 산업 0.689 대 0.818).

표 5. 신규 매핑 탐지 성능
Table 5. Novel mapping detection performance

Domain	Backbone	AUROC	FPR@95TPR
SCI-PAT	2-stage	0.735	0.778
	Ko-SRoBERTa	**0.823**	**0.592**
	Qwen3-0.6B	0.822	0.605
IND-PAT	2-stage	0.712	0.818
	Ko-SRoBERTa	**0.783**	**0.689**
	Qwen3-0.6B	0.781	0.792

주목할 점은 두 방식의 탐지 원리가 근본적으로 다르다는 것이다. 4.1에서 확인하였듯 신규 문서의 개별 라벨은 모두 학습에 존재하며 그 결합만이 새롭다. 2단계 분류는 각 축의 라벨을 예측할 수 있으나 그 결합이 신규라는 사실 자체를 표현하지 못하고, 분류 확신도가 낮아지는 간접 신호로만 신규성을 포착한다. 반면 제안 방법은 매핑을 단일 좌표로 표현하므로 관측 분포로부터의 거리로 신규성을 직접 측정하며, 이를 통해 더 높은 AUROC와 더 낮은 오탐을 달성한다. 다만 확장 라벨 성능의 절대 수준은 고정 라벨에 비해 낮은 편이며, 이는 신규 융합 탐지가 본질적으로 어려운 과제임을 보여준다. 또한 확장 라벨에서는 Ko-SRoBERTa가 Qwen3-0.6B와 같거나 다소 앞서, 고정 라벨과 달리 백본 규모가 신규 탐지 성능으로 직결되지는 않음을 보인다.

5.4 신규 융합 매핑의 정성 분석

제안 방법의 합성 표현은 관측되지 않은 매핑 조합에 대해서도 좌표를 산출할 수 있으므로, 학습에서 나타나지 않은 새로운 매핑을 제안하는 데 활용할 수 있다. 이를 정성적으로 확인하기 위해, 과학-특허 도메인에서 학습에 관측되지 않은 모든 (s, p) 조합을 합성하여 좌표를 산출하고, 각 조합에서 가장 가까운 관측 매핑까지의 거리 δ를 측정하였다. 표 6은 그 거리에 따라 후보를 가까운 경우(A), 중간(B), 먼 경우(C)로 나누어 예시한 것이다.

표 6. 신규 매핑 후보 사례
Table 6. New mapping candidates

Type	Novel mapping	Nearest mapping	δ
(A)	LC01↔B81C	LC01↔B81B	0.211
	EG10↔H04W	EG10↔H04B	0.212
(B)	EG02↔D06C	EG02↔G21B	0.367
	ND09↔A42B	ND01↔A42B	0.368
(C)	EC03↔G06E	EC03↔C08L	0.554
	EC02↔B61H	EH10↔B61H	0.549

가까운 경우(A)는 동일한 과학기술표준분류를 공유하면서 인접한 특허분류와 새롭게 결합되는 후보다. 예를 들어 LC01(의생명과학)↔B81C(미세구조 장치)는 학습에 존재하지 않는 결합이지만, 동일 분야가 인접 특허 B81B(초소형 구조 장치)와 이미 결합된 적이 있어 제안 방법은 이를 가까운 좌표로 추정한다. 주목할 점은 최근접 관측 매핑이 대부분 특허 코드의 한 단계 인접 분류(B81B↔B81C 등)라는 것이다. 이는 해당 사례에서 합성 표현이 특허 축의 인접 분류 관계와 일관된 위치를 형성했음을 보여준다.

반면 먼 경우(C)는 기존의 어떤 관측 매핑과도

충분히 가깝지 않은 결합으로, 거리 δ가 크게 나타난다. 이처럼 거리는 신규 결합이 기존 매핑 공간에 잘 포섭되는지를 정량적으로 나타내는 척도가 된다. 중요한 점은, 미관측 결합이라 하더라도 합성 좌표가 무작위로 흩어지지 않고 두 축의 임베딩에 따라 일관된 위치에 형성되므로, 그 위치와 거리를 근거로 결합의 타당성을 정량적으로 평가할 수 있다는 것이다.

이러한 성질은 제안 방법을 단순한 매핑 검색을 넘어 활용할 수 있게 한다. 거리 점수는 기존 라벨 중 가장 가까운 후보를 선택하는 데 그치지 않고, 기존 라벨 공간으로 충분히 설명되지 않는 영역을 판단하는 기준으로도 쓰인다. 즉 어떤 라벨 조합이 기존 매핑들과 낮은 적합도를 보이면서도 임베딩 공간에서 일관된 위치를 형성한다면, 이를 새로운 매핑 결합의 등장 가능성으로 해석할 수 있다. 따라서 제안 방법은 기존 분류체계 내의 매핑뿐 아니라, 분류체계 개정이나 신규 라벨 추가가 필요한 상황에서 초기 후보를 제안하는 데에도 활용될 수 있다.

Ⅵ. 결 론

본 논문은 교차 분류체계 매핑을 두 분류기의 예측을 사후에 결합하는 문제가 아니라, 매핑 자체를 하나의 라벨로 보아 문서와 정렬·검색하는 문제로 재정의하였다. 제안 방법은 매핑 라벨을 두 분류 축의 코드 임베딩을 합성하여 표현하고, 한 축을 공유하는 양방향 구조적 하드 네거티브와 InfoNCE 손실을 활용하여 문서와 매핑을 공동 임베딩 공간에 정렬한다. 이를 통해 앞단 분류의 오차가 매핑으로 전이되는 2단계 방식의 구조적 한계를 완화하고, 두 분류 축의 정보가 함께 반영된 합성 표현으로 매핑을 구성한다.

과학-특허와 산업-특허 두 데이터셋에서의 실험은 제안 방법의 효과를 일관되게 보여준다. 제안 방법은 상위 후보 검색과 순위 품질에서 2단계 방식을 앞섰으며, 특히 희소한 매핑쌍이 많은 롱테일 구간에서 높은 성능을 보였다. 또한 절제 연구를 통해 일반 하드 네거티브를 사용한 기준 조건에 비해 합성 표현과 구조적 네거티브를 적용한 구성에서 성능이 향상됨을 확인하였다. 특히 동일한 합성 표현 조건에서는 단방향보다 양방향 구조적 네거티브를 적용한 경우 성능 향상이 나타나, 양방향 네거티브 구성이 추가적인 성능 향상에 기여함을 확인하였다. 이러한 결과는 제안 방법이 두 분류 결과를 사후적으로 결합하는 대신, 두 분류체계의 정보를 하나의 매핑 표현으로 학습하고 문서와 직접 정렬하는 방식이 교차 분류체계 매핑에 효과적임을 보여준다. 또한 학습에서 관측되지 않은 신규 결합을 탐지하는 개방형 평가에서도, 매핑 좌표로부터의 거리를 직접 측정하는 제안 방법이 분류 확신도에 의존하는 2단계 방식보다 우수한 성능을 보였다.

본 논문은 다음과 같은 한계와 향후 과제를 가진다. 첫째, 본 연구에서 다룬 신규 결합은 개별 라벨이 모두 학습에 존재하되 그 조합만 새로운 경우에 한정된다. 합성 표현은 구조적으로 미관측 개별 라벨의 조합에도 좌표를 부여할 수 있으나, 해당 라벨의 코드 임베딩이 학습되지 않아 좌표의 신뢰도가 낮으며 본 데이터셋에는 평가 표본이 부족하여 이를 검증하지 못하였다. 코드 임베딩을 라벨의 텍스트 설명으로 초기화하는 등의 방법으로 미관측 라벨로의 일반화를 확장하는 것은 향후 과제로 남긴다. 둘째, 본 논문은 두 분류체계의 결합만을 다루었으나, 셋 이상의 분류체계로 매핑을 확장하는 것은 합성 표현의 자연스러운 확장 방향이 될 수 있다. 제안 방법이 서로 다른 분류체계를 의미 기반으로 연결하는 일반적 틀을 제공하는 만큼, 향후 기술융합 발굴이나 과학-산업-특허 연계 분석 등 다양한 응용으로 이어질 수 있을 것으로 기대한다.

References

[1] G. C. Bowker and S. L. Star, "Sorting things out: Classification and its consequences", MIT Press, 1999. https://doi.org/10.7551/mitpress/6352.001.0001.

[2] E. Jacob, "Classification and categorization: A difference that makes a difference", Library Trends, Vol. 52, No. 3, pp. 515-540, Dec. 2004.

[3] KISTEP, "National Science and Technology Standard Classification System", https://www.kistep.

re.kr/board.es?mid=a10305080000&bid=0002&act=view&list_no=43328. [accessed: Jun. 24, 2026]

[4] OECD, "Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development", OECD Publishing, Oct. 2015. https://doi.org/10.1787/9789264239012-en.

[5] KIPO, "IPC-CPC Patent Classification", https://www.kipo.go.kr/ko/kpoContentView.do?menuCd=SCD0200269. [accessed: Jun. 24, 2026]

[6] WIPO, "Guide to the International Patent Classification", Geneva, Switzerland : World Intellectual Property Organization, 2026. https://doi.org/10.34667/tind.60169.

[7] KOSIS, "Korean Standard Industrial Classification", https://kssc.mods.go.kr:8443/ksscNew_web/index.jsp#. [accessed: Jun. 24, 2026]

[8] B. Hammarfelt, "Linking science to technology: The patent paper citation and the rise of patentometrics in the 1980s", Journal of Documentation, Vol. 77, No. 6, pp. 1413-1429, Oct. 2021. https://doi.org/10.1108/JD-12-2020-0218.

[9] J. H. Gwak and Y. C. Jung, "Enhancing the Performance of Low-Resource Text Classification through KoBERT-based Category Mapping and LLM Integration", JKIIT, Vol. 23, No. 1, pp. 1-11, Jan. 2025. http://doi.org/10.14801/jkiit.2025.23.1.1.

[10] N. Li, B. Kang, and T. De Bie, "Your Next State-of-the-Art Could Come from Another Domain: A Cross-Domain Analysis of Hierarchical Text Classification", Mach Learn, Vol. 115, No. 80, Mar. 2026. https://doi.org/10.1007/s10994-025-06993-w.

[11] M. Miwa and M. Bansal, "End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures", Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, pp. 1105-1116, Aug. 2016. https://doi.org/10.18653/v1/P16-1105.

[12] Y. Luo, S. Ma, M. Nie, C. Peng, Z. Lin, J. Shao, and Q. Xu, "Domain-Aware Cross-Attention for Cross-domain Recommendation", arXiv preprint, arXiv:2401.11705 Jan. 2024. https://doi.org/10.48550/arXiv.2401.11705.

[13] P. Dorta-González, A. Rodríguez-Caro, and M. I. Dorta-González, "Linking Science and Industry: Influence of Scientific Research on Technological Innovation through Patent Citations", arXiv preprint, arXiv:2502.14570, Feb. 2025. https://doi.org/10.48550/arXiv.2502.14570.

[14] G. Wang, C. Li, W. Wang, Y. Zhang, D. Shen, X. Zhang, R. Henao, and L. Carin, "Joint Embedding of Words and Labels for Text Classification", Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia pp. 2321-2331, Jul. 2018. https://doi.org/10.18653/v1/P18-1216.

[15] D. Li, V. Yadav, Z. L. Zhu, M. M. Fard, Z. Afzal, and G. Tsatsaronis, "Scalable Patent Classification with Aggregated Multi-View Ranking", Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, Torino, Italia, pp. 14336-14346, May 2024. https://aclanthology.org/2024.lrec-main.1249.

[16] C. J. Fall, A. Törcsvári, K. Benzineb, and G. Karetka, "Automated Categorization in the International Patent Classification", ACM SIGIR Forum, Vol. 37, No. 1, pp. 10-25, Apr. 2003. https://doi.org/10.1145/945546.945547.

[17] C. J. Im, D. W. Kim, and T. Mandl, "Text Classification for Patents: Experiments with Unigrams, Bigrams and Different Weighting Methods", International Journal of Contents, Vol. 13, No. 2, pp. 66-74, Jun. 2017. https://doi.org/10.5392/IJoC.2017.13.2.066.

[18] S. Li, J. Hu, Y. Cui, and J. Hu, "DeepPatent: patent classification with convolutional neural networks and word embedding", Scientometrics, Vol. 117, No. 2, pp. 721-744, Sep. 2018.

https://doi.org/10.1007/s11192-018-2905-5.

[19] J. S. Lee and J. Hsiang, "PatentBERT: Patent Classification with Fine-Tuning a Pre-trained BERT Model", World Patent Information, Vol. 61, pp. 101965, Dec. 2020. https://doi.org/10.1016/j.wpi.2020.101965.

[20] J. W. Park, W. C. Sim, S. H. Lee, B. S. Ko, and H. S. Noh, "A Study on Automatic CPC Classification based on Korean Patent Sentence - A Deep Learning Approach using Artificial Intelligence Language Model KorPatBERT -", Journal of Intellectual Property, Vol. 17, No. 3, pp. 209-256, Mar. 2022. https://doi.org/10.34122/jip.2022.17.3.209.

[21] KIPO, "KSIC-IPC Mapping Table", https://www.moip.go.kr/ko/kpoContentView.do?menuCd=SCD0200272. [accessed: Jun. 24, 2026]

[22] T. J. Lybbert, "Getting Patents and Economic Data to Speak to Each Other: An “Algorithmic Links with Probabilities” Approach for Joint Analyses of Patenting and Economic Activity", Economic Research Working Paper, No. 5, 2012. https://doi.org/10.34667/TIND.28868.

[23] KIPO, "National Science and Technology Standard Classification and IPC-CPC Mapping Table", https://www.kipo.go.kr/ko/kpoContentView.do?menuCd=SCD0200273. [accessed: Jun. 24, 2026]

[24] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", Proceedings of NAACL-HLT, Minneapolis, Minnesota, USA, pp. 4171-4186, Jun. 2019. https://doi.org/10.18653/v1/N19-1423.

[25] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks", Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Hong Kong, pp. 3982-3992, Nov. 2019. https://doi.org/10.18653/v1/D19-1410.

[26] A. van den Oord, Y. Li, and O. Vinyals, "Representation Learning with Contrastive Predictive Coding", arXiv preprint, arXiv:1807.03748, Jul. 2018. https://doi.org/10.48550/arXiv.1807.03748.

[27] J. D. Robinson, C. Y. Chuang, S. Sra, and S. Jegelka, "Contrastive Learning with Hard Negative Samples", arXiv preprint, arXiv:2010.04592, Oct. 2020. https://doi.org/10.48550/arXiv.2010.04592.

[28] L. Xiong, C. Xiong, Y. Li, K. F. Tang, J. Liu, P. Bennett, J. Ahmed, and A. Overwijk, "Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval", arXiv preprint, arXiv:2007.00808, Jul. 2021. https://doi.org/10.48550/arXiv.2007.00808.

[29] Y. Sun, Y. Ming, X. Zhu, and Y. Li, "Out-of-Distribution Detection with Deep Nearest Neighbors", Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, Vol. 162, pp. 20827-20840, Jul. 2022. https://proceedings.mlr.press/v162/sun22d.html.

[30] AI-Hub, "Patent Data Mapped to National Science and Technology Standard Classification", https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=71531. [accessed: Jun. 24, 2026]

[31] AI-Hub, "Patent Field Automatic Classification Data", https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=547. [accessed: Jun. 24, 2026]

[32] jhgan, "ko-sroberta-multitask", Hugging Face Model Repository, https://huggingface.co/jhgan/ko-sroberta-multitask. [accessed: Jun. 24, 2026]

[33] Y. Zhang, M. Li, D. Long, X. Zhang, H. Lin, B. Yang, P. Xie, A. Yang, D. Liu, J. Lin, F. Huang, and J. Zhou, "Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models", arXiv preprint, /arXiv.2506.05176, Jun. 2025. https://doi.org/10.48550/arXiv.2506.05176.

저자소개

곽 지 호 (Jiho Gwak)



2025년 2월 : 국립금오공과대학교 컴퓨터공학과(공학사)
2025년 2월 ~ 현재 : 국립금오공과대학교 컴퓨터 · AI융합공학과 석사과정
관심분야 : 텍스트 분류, 임베딩, 강화학습, 자연어처리

김 민 석 (Minseok Kim)



2025년 2월 : 국립금오공과대학교 컴퓨터공학과(공학사)
2025년 7월 ~ 현재 : 국립금오공과대학교 컴퓨터 · AI융합공학과 석사과정
관심분야 : 문서 검색, RAG, 에이전트, 데이터베이스

김 준 호(Junho Kim)



2020년 8월 : 광주과학기술원(GIST) 전기전자컴퓨터공학(석사)
2020년 10월 : 한국한의학연구원 연구원
2023년 11월 : 한국전력공사 전력연구원 연구원
2024년 8월 ~ 현재 : 한국과학기술정보연구원 기술원
관심분야 : 자연어처리 / 지식그래프 / RAG

정 유 철 (Yuchul Jung)



2011년 2월 : 한국과학기술원 (KAIST) 전산학과(공학박사)
2009년 1월 ~ 2013년 7월 : 한국전자통신연구원 (ETRI) 선임연구원
2013년 8월 ~ 2017년 8월 : 한국과학기술정보 연구원(KISTI) 선임연구원
2017년 8월 ~ 2022년 2월 : 국립금오공과대학교 컴퓨터공학과 조교수
2022년 2월 ~ 현재 : 국립금오공과대학교 인공지능공학과 부교수
관심분야 : 거대언어모델, 자연어처리, 지식그래프, 한국어 음석 인식/합성, AI응용