합성 매핑 임베딩을 통한 교차 분류체계 매핑

*

Cross-Taxonomy Mapping via Compositional Mapping Embeddings

*

지원기관표기

요 약

서로 다른 목적으로 구축된 분류체계를 연결하는 교차 분류체계 매핑은 두 분류기를 독립 학습해 결합하는 2단계 방식이 일반적이나, 분류 오차가 매핑으로 전이되고 관측되지 않은 새로운 결합을 표현하지 못한다. 본 논문은 매핑 자체를 하나의 라벨로 보아 문서와 정렬·검색하는 문제로 재정의한다. 매핑 라벨을 두 분류 코드 임베딩의 합성으로 표현하고, 양방향 구조적 네거티브 전략을 통해 두 분류 축의 분리를 유도한다. 이 합성 표현은 분류기의 고정 라벨 집합을 넘어 미관측 결합까지 좌표로 나타내므로, 관측 분포로부터의 거리로 신규 융합 매핑을 직접 탐지한다. 과학-특허와 산업-특허 두 데이터셋에서 제안 방법은 2단계 방식보다 상위 검색과 순위 품질에서 우수했으며, 특히 롱테일 구간과 신규 결합 탐지에서 큰 향상을 보였다.

Abstract

Cross-taxonomy mapping, which links classification systems built for different purposes, is commonly addressed by a two-stage pipeline that trains a classifier per taxonomy and combines their predictions; however, classification errors propagate into the mapping and novel unobserved combinations cannot be represented. We reformulate it as an alignment-and-retrieval problem that treats the mapping itself as a single label. Each mapping label is represented as a composition of two classification-code embeddings, and documents are aligned with mappings using bidirectional structural hard negatives that separate the two classification axes. Since this compositional representation assigns coordinates even to combinations beyond a classifier's closed label set, novel fusion mappings are detected directly via distance from the observed distribution. On two datasets science-patent and industry-patent the proposed method outperforms the two-stage baseline in top-k retrieval and ranking quality, with notable gains on the long tail and in novel-combination detection.

Keywords

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

* .

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]. 과학기술분류가 연구 주제와 학문적 지식 구조를 중심으로 한다면[3], 특허분류는 기술의 권리화와 산업적 활용 가능성을 중심으로 하며[4], 산업분류는 산업 활동의 구조를 중심으로 한다[5]. 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[6].

기존 접근 방식은 일반적으로 과학분류기와 특허분류기를 각각 독립적으로 학습한 뒤, 두 분류 결과를 동시출현 기반 매핑 테이블에 연결하는 2단계 구조를 따른다[7]. 이 방식은 각 분류체계를 개별적으로 다룰 수 있다는 점에서 직관적이지만, 교차 분류체계 매핑을 직접 모델링하지 못한다는 한계를 가진다[8].

첫째, 분류 오차가 곧 매핑 오차로 전이된다. 기존 방식에서 최종 매핑은 과학분류 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[9].

둘째, 매핑 자체에 대한 의미 표현이 부재하다. 기존의 2단계 방식은 과학분류와 특허분류를 각각 예측한 뒤, 두 결과를 사후적으로 연결할 뿐이다. 따라서 과학분류와 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[10]. 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[11].

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, 과학분류와 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[12], 기존에 관측된 매핑뿐만 아니라 새로운 융합 매핑의 가능성도 거리 기반으로 분석할 수 있게 한다[13].

본 논문의 주요 기여는 다음과 같다.

첫째, 교차 분류체계매핑을 2단계 분류가 아니라 매핑 자체를 하나의 라벨로 보는 정렬·검색 문제로 재정의한다. 기존 연구가 각 체계의 분류 결과를 사후에 잇는 방식이었던 데 반해, 본 논문은 문서를 매핑 라벨이 놓인 공동 임베딩 공간에 직접 정렬함으로써 앞단 분류의 오차 전파를 원천적으로 제거한다.

둘째, 매핑 라벨을 두 코드 임베딩의 합성으로 표현하여, 분류기의 닫힌 라벨 집합 한계를 넘어 관측되지 않은 과학-특허 결합까지 탐지한다. 이는 신규·융합 매핑을 거리 기반으로 다룰 수 있는 구조적 기반을 제공한다.

셋째, 두 분류 축의 분리를 명시적으로 유도하는 양방향 구조적 하드 네거티브 학습을 제안한다. 구분 난이도를 높이는 일반적 하드 네거티브와 달리, 한쪽 축을 공유하는 네거티브를 양방향으로 구성하여 합성 함수가 두 축을 모두 반영하도록 강제한다.

II. 관련 연구

2.1 분류와 분류체계 매핑

특허 및 기술문서의 자동 분류는 오랫동안 연구되어 온 주제로[14], 초기에는 TF-IDF와 같은 통계적 자질에 기반한 기계학습 분류기가 주로 사용되었다 [15]. 이후 단어 임베딩과 RNN(Recursive Neural Network), CNN(Convolution Neural Network) 기반 모델이 도입되며 성능이 향상되었고 [15], 최근에는 BERT 계열의 사전학습 언어모델을 미세조정하는 방식이 표준으로 자리 잡았다 [16]. 특히 특허 도메인에 특화된 사전학습 모델이나 한국어 특허·기술문서를 위한 분류 모델이 제안되며, 단일 분류체계 내에서의 범주 예측 성능은 꾸준히 개선되어 왔다 [17].

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 [18], 각 체계의 분류기를 독립적으로 학습한 뒤 그 예측 결과를 연결하는 2단계 방식을 사용해 왔다. 국제특허분류(IPC)와 경제분류[19], 또는 기술분류와 특허분류를 연계하려는 시도가 이러한 범주에 속한다[20].

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 [21]. 특히 SBERT 계열의 연구는 BERT를 문장 임베딩에 적합하도록 재구성하여 문장 간 유사도를 효율적으로 계산할 수 있게 하였다[22]. 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 [23]. 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 [24]. 다만 기존의 하드 네거티브는 주로 구분 난이도를 높여 표현력을 강화하는 데 목적을 두며, 본 논문은 이를 두 분류 축의 분리를 유도하는 방향으로 확장한다는 점에서 차이가 있다.

2.3 밀집 검색과 거리 기반 탐지

대조학습으로 학습된 임베딩은 Dense Retrieval의 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[25]. 본 논문은 이러한 정렬·검색 패러다임을 문서 - 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 [26]. 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.

III. 제안 프레임워크: 합성 매핑 임베딩을 통한 정렬·검색

3.1 문제 형식화

문서 집합 X의 각 문서 x에는 서로 다른 두 분류체계의 라벨이 연결되어 있다고 가정한다. 두 분류체계의 라벨 집합을 각각 S와 P로 두며, 문서 d에 연결된 라벨을 $(s \in S, p \in P)$로 표기한다. 여

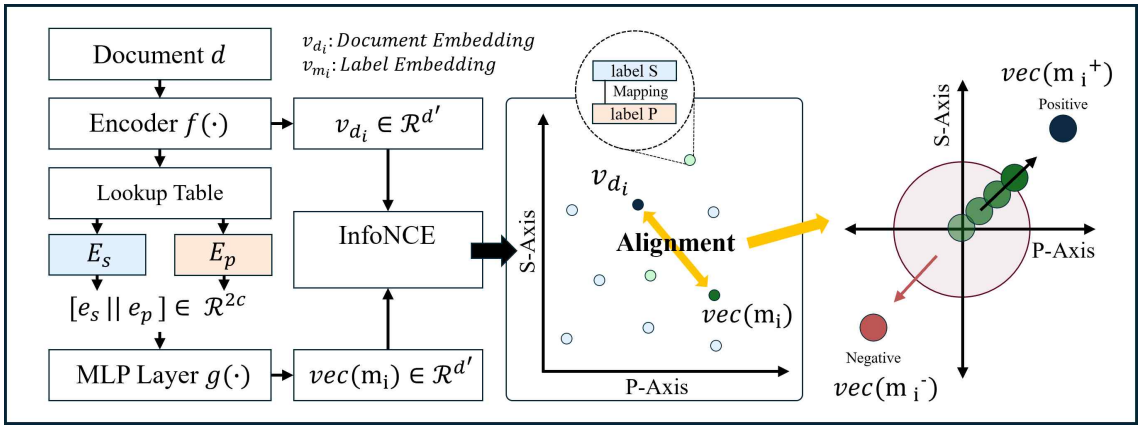


그림 1. 제안하는 합성 매핑 임베딩 프레임워크의 전체 구조. 문서 d는 인코더 $f(\cdot)$를 통해 임베딩 v_{di}로 변환된다. 매핑 라벨은 두 분류 축의 코드 임베딩 E_s, E_p를 결합한 뒤 MLP $g(\cdot)$로 합성하여 매핑점 $\mathrm{vec(m_i)}$로 표현된다. 두 표현은 동일한 공동 임베딩 공간에서 InfoNCE 손실로 정렬되며, 이때 한 축을 공유하는 양방향 구조적 하드 네거티브 $\mathrm{vec(m_i^+)}$, $\mathrm{vec(m_i^-)}$를 사용하여 S축과 P축이 분리되도록 학습한다.

Fig. 1. Overall architecture of the proposed framework. A document is encoded by $f(\cdot)$ into v_{di}, and a mapping label is represented as $\mathrm{vec(m_i)}$ by composing the two code embeddings E_s and E_p via $g(\cdot)$; the two are aligned in a joint embedding space with InfoNCE. The positive mapping $\mathrm{vec(m_i^+)}$ and a structural hard negative $\mathrm{vec(m_i^-)}$ induce the separation of the two classification axes between S and P.

기서 S와 P는 특정 도메인에 고정된 것이 아니라, 매핑 대상이 되는 두 분류체계를 일반적으로 나타낸다. 예를 들어 (S, P) 는 {(과학기술분류, 특허분류), (산업분류, 특허분류)} 와 같은 조합을 구성한다.

본 논문은 두 라벨의 결합을 하나의 매핑 라벨 $m_i = (s_i, p_i)$로 정의한다. 실제 데이터에서 가능한 모든 라벨 조합($|S| \times |P|$)이 관측되는 것이 아니므로, 문서에서 실제로 함께 등장한 매핑 라벨의 집합을 M_{obs}로 둔다.

문서 인코더 f는 문서 x를 d차원의 임베딩 표현으로 변환한다. 매핑 라벨 m은 그 구성요소인 두 라벨의 코드 임베딩으로부터 합성되어 동일한 공간의 한 점 $\mathrm{vec(m)}$으로 표현된다. 구체적인 합성 방식은 3.2절에서 설명한다. 문서와 매핑 라벨 간의 정렬 정도는 코사인 유사도로 측정한다.

$$\mathrm{sim(x,m)} = \frac{\mathrm{f(x)^T \cdot vec(m)}}{|\mathrm{f(x)}| \cdot |\mathrm{vec(m)}|} \quad (1)$$

본 프레임워크는 두 과제를 포함한다. 매핑 검색(고정 라벨)은 문서 x를 관측 매핑 라벨 집합에 대해 검색하여 가장 유사한 매핑을 예측한다.

$$\hat{mx} = \mathrm{argmax}_{\mathrm{m} \in \mathrm{M_{obs}}} \mathrm{sim(x,m)} \quad (2)$$

정답은 그 문서의 (s, p)쌍이며, 예측 쌍과의 일치로 측정한다. 신규 융합 탐지(확장 라벨)는 최근접 매핑점까지의 거리

$$\delta_{\mathrm{min}}(x) = 1 - \max_{m \in M_{obs}} \mathrm{sim(x,m)} \quad (3)$$

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

3.2 매핑 라벨의 합성 표현

매핑 라벨을 표현하는 가장 단순한 방식은 각 매핑 라벨을 독립적인 항목으로 간주하는 원자적 표현이다. 이 경우 관측된 매핑 라벨마다 하나의 임베딩 벡터를 할당하는 룩업 테이블을 사용한다. 그러나 원자적 표현은 M_{obs}에 포함된 관측 매핑 라벨

에 대해서만 좌표를 가지므로, 학습에서 관측되지 않은 라벨 조합 $(s,p) \notin M_{obs}$에는 임베딩을 부여할 수 없어 신규 또는 융합 매핑을 표현하지 못한다.

본 논문은 이러한 한계를 완화하기 위해 매핑 라벨을 두 라벨의 합성 표현으로 구성한다. 두 분류체계의 코드 임베딩 테이블을 각각 다음과 같이 정의한다.

$$E_S \in \mathcal{R}^{|S| \times c}, E_P \in \mathcal{R}^{|P| \times c} \quad (4)$$

여기서 c는 코드 임베딩 차원이다. 매핑 라벨 m은 각 코드 임베딩을 결합한 뒤, 합성 함수 g를 통해 d차원의 매핑점으로 다음과 같이 변환된다.

$$\mathrm{vec}(\mathrm{m}) = \mathrm{normalize}(\mathrm{g}([\mathrm{E_S};\mathrm{E_P}])) \quad (5)$$

여기서 $[\bullet ; \bullet]$는 두 벡터의 결합(Concatenation)이고, normalize는 L2 정규화로 매핑점을 단위 벡터로 만들어 3.1에서 계산된 코사인 유사도와 정합되며, 합성 함수 g는 다층 퍼셉트론으로 정의된다.

즉 $\mathrm{vec}(\mathrm{m})$은 별도의 매핑 라벨 인코더가 아니라, 두 코드 임베딩 E_S, E_P와 합성 함수 g가 함께 생성하는 매핑점이다. 이 방식은 관측된 매핑 라벨을 단순히 암기하는 것이 아니라 두 분류체계 라벨의 조합으로 매핑을 표현하므로, 두 코드 임베딩이 학습되어 있으면 M_{obs}에 포함되지 않은 조합에 대해서도 매핑 좌표를 계산할 수 있다.

이 성질은 분류 기반 방식과의 근본적 차이를 만든다. 분류 모델은 출력이 학습된 라벨 집합으로 제한되어, 관측되지 않은 라벨 조합은 표현 자체가 불가능하다. 반면 합성 표현은 라벨 조합 공간 전체에 좌표를 부여할 수 있으므로, 분류기의 커버리지 경계 밖에 있는 신규·융합 매핑까지 거리 기반으로 다룰 수 있다.

3.3 두 분류 축의 분리

합성 표현이 의미를 가지기 위해서는 두 분류체계의 라벨 축이 서로 구분되게 학습되어야 한다. 그러나 단일 문서만으로는 문서의 의미가 어느 라벨 축에서 비롯되었는지 분리하기 어렵다. 이에 여러 문서에 걸쳐 반복되는 교차 결합 패턴을 통해 두 축의 분리를 가능하게 설계한다.

예를 들어 동일한 S가 여러 P와 결합하거나 동일한 P가 여러 S와 결합한다면, 모델은 한쪽 라벨에 공통으로 나타나는 의미 성분과 다른 축에 의해 달라지는 성분을 추정할 수 있다. 반대로 S가 하나의 P와만 결합한다면 두 라벨의 효과는 분리되기 어렵고, 해당 결합은 사실상 하나의 원자적 매핑 라벨처럼 작동한다.

따라서 합성 표현의 핵심은 두 코드 임베딩을 결합하는 것이 아니라, 두 라벨 축이 서로 다른 의미 성분을 담당하도록 학습하는 데 있다. 이를 위해 다음 절에서는 한쪽 라벨을 공유하면서 다른 쪽 라벨만 달라지는 구조·의미적 하드 네거티브 전략을 적용한다.

3.4 양방향 구조적 하드 네거티브 학습

학습의 목표는 문서 임베딩 $f(x)$를 정답 매핑점 $\mathrm{vec}(\mathrm{s},\mathrm{p})$에 가깝게 정렬하고 다른 매핑점과는 멀어지도록 하는 것이다. 이를 위해 InfoNCE 손실을 사용하며 다음과 같다.

$$L(x) = -\log\left[\frac{\exp(\mathrm{sim}(\mathrm{x},\mathrm{m})/\tau)}{\Sigma(m')\exp(\mathrm{sim}(\mathrm{x},(\mathrm{m}'))/\tau)}\right] \quad (6)$$

여기서 τ는 온도 파라미터이며, m'은 네거티브를 의미한다. 3.3에서 논의한 두 축의 분리는 이 네거티브 집합을 어떻게 구성하느냐에 달려 있으며, 본 논문에서는 다음과 같이 양방향 구조적 하드 네거티브로 구성한다.

먼저 동일한 S축 라벨을 공유하지만 P축 라벨이 다른 매핑을 P축 네거티브로 정의한다. 이 네거티브는 s가 동일하므로, 정답 매핑을 구분하려면 p의 차이를 반영해야한다. 반대의 경우에도 동일하게 정의된다. 따라서 각 축에 대한 네거티브 구축(N)은 다음과 같이 구성된다.

$$N_P(s,p) = (\mathrm{s},\mathrm{p}') \in \mathrm{M_{obs}} | \mathrm{p}' \neq \mathrm{p} \quad (7)$$

$$N_S(s,p) = (s', p) \in M_{obs} | s' \neq s \quad (8)$$

이러한 구성은 모델이 한쪽 분류 축에만 의존하는 것을 방지한다. 일반적인 네거티브가 구분 난이도를 높여 표현력을 강화하는 데 목적이 있다면, 양방향 구성은 두 분류 축의 분리를 명시적으로 유도한다는 점에서 목적이 다르다. 결과적으로 양방향 네거티브는 합성 함수가 두 코드 임베딩을 단순히 결합하는데 그치지 않고, 두 축이 서로 다른 의미 성분을 담당하도록 학습하게 된다.

Ⅳ. 실험 설계

4.1 데이터셋

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

데이터셋은 AI-Hub에서 제공하는 과학기술표준분류 대응 특허데이터[27]와 특허 분야 자동분류 데이터[28]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.

데이터셋은 AI-Hub에서 제공하는 학습·검증 분할을 유지한다. 분리 후 과학-특허 데이터는 학습 240,187건, 검증 30,024건으로 과학기술표준분류는 188개, 특허분류는 603개의 라벨을 가지며 관측된 매핑쌍은 15,281개로 구성된다. 산업-특허 데이터는 학습 443,488건, 검증 56,392건으로 산업분류 118개, 특허분류 597개의 라벨을 가지며 관측된 매핑쌍은 6,130개로 구성된다. 각 라벨은 모두 계층적인 구조를 가지지만, 계층이 깊어질수록 매핑쌍이 급격히 증가하고 라벨당 샘플이 감소하기 때문에 이를 고려하여 중분류, 서브클래스 단위로 라벨을 구성한다. 입력 텍스트는 제목, 초록, 청구항을 결합하여 하나의 입력문으로 구성한다.

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

표 1. 데이터셋 구성
Table 1. Dataset Construction

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

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[21] 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[29], Qwen3-Embedding-0.6B[30] 두 가지 백본에 대해 각각 평가하고, 사전학습 비교에는 KoB, 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. 지표로는 상위 k개 후보 안에 정답이 포함되는 비율인 Pair-R@1, R@3, R@5와, 정답 순위의 역수 평균인 평균 역순위(MRR)을 사용한다.

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

MRR과 매핑쌍 단위로 평균한 MRR macro를 함께 보고하며, macro는 희소한 매핑쌍에서도 동일한 가중치를 부여하므로 롱테일 구간의 성능을 반영한다.

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

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)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 다음과 같이 무작위로 뽑은 신규 문서의 점수가 기존 문서보다 높을 확률로 정의된다.

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

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 τ = 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 백본(Qwen3)은 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 2. 두 도메인(SCI-PAT, IND-PAT)에서의 고정 라벨 매핑 검색 성능. 모든 지표는 두 분류 축을 동시에 맞춘 쌍 단위 기준이며, 2단계 방식은 단일 후보만 산출하므로 R@{3,5}가 R@1과 동일하다. 학습이 필요한 2단계 및 제안 방법은 서로 다른 세 시드로 실행한 평균과 표준편차이며, 사전학습 방식은 미세조정을 거치지 않아 시드에 따른 차이가 없으므로 단일 값으로 보고한다.

Table 2. Closed-set mapping retrieval performance on the two domains (SCI-PAT, IND-PAT). All metrics are pair-level, requiring both classification axes to be correct simultaneously; since the two-stage method produces only a single candidate, R@{3,5} equal R@1. The two-stage and proposed methods, which require training, are reported as the mean and standard deviation over three random seeds, whereas the pre-trained method involves no fine-tuning and is therefore reported as a single seed-invariant value.

Domain	Method	Backbone	R@1	R@3	R@5	MRR	MRR_macro
SCI-PAT	two 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	two 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 3는 2단계 방식의 기반이 되는 KoBERT 분류기 성능이다. 과학분류는 정확도 0.684, 산업분류는 0.818, 특허분류는 0.718로 단일 분류체계 내에서는 비교적 안정적인 성능을 보인다. 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 2의 2단계 분류 결과가 이를 직접 보여준다. 과학-특허에서 두 분류 정확도의 곱(0.684 × 0.718 ≈ 0.491)은 표 3의 2단계 매핑 R@1(0.496)과 거의 일치하며, 산업 - 특허에서도 곱(0.818 × 0.718 ≈ 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.

표 3. 각 분류체계에 대한 KoBERT 분류기의 단일 분류 성능.

Table 3. 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

5.2 고정 라벨(Closed-set) 매핑 성능

표 2은 두 도메인의 고정 라벨 매핑 검색 성능이다. 학습을 포함하는 2단계 및 제안 방법은 서로 다른 세 시드로 실행한 평균과 표준편차로 보고하며,

미세조정을 거치지 않는 사전학습 방식은 시드에 따른 차이가 없어 단일 값으로 보고한다. 모든 방법에서 표준편차가 0.005 이하로 작아, 결과가 시드에 안정적임을 확인할 수 있다.

최상위 정확도(R@1)에서, 과학-특허 도메인은 2단계 분류(0.496)와 제안 방법(ko-sroberta 0.497, qwen3 0.503)이 유사한 수준이다. 두 분류기가 각각 충분히 학습되면 그 결합만으로도 상당수의 매핑쌍을 맞출 수 있기 때문이다. 반면 산업-특허 도메인에서는 제안 방법(ko-sroberta 0.614, qwen3 0.623)이 2단계(0.578)를 R@1에서부터 뚜렷이 앞선다.

상위 후보를 함께 고려하면 격차가 분명해진다. 제안 방법은 R@5에서 0.84~0.91에 이르는 반면, 2단계 분류는 단일 후보만 산출하므로 R@5가 R@1과 동일한 값(과학 0.496, 산업 0.578)에 머문다. 순위 품질을 반영하는 MRR에서도 제안 방법이 2단계를 크게 앞선다. 이는 제안 방법이 정답을 최상위에 두지 못하더라도 상위권에 일관되게 배치함을 의미하며, 복수의 관련 매핑을 후보로 제시해야 하는 실제 활용에서 유리하다.

한편 KoBERT 모델의 사전학습 베이스라인은 R@1이 0.11~0.171으로 가장 낮아, 미세조정 없이 사전학습 표현만으로는 매핑 정렬이 충분하지 않음을 보인다. 주목할 점은 백본의 영향이 학습 전후로 달라진다는 것이다. 사전학습 단계에서는 더 큰 백본일수록 높은 성능을 보여, qwen3-0.6b가 ko-sroberta를 앞서고(R@1 0.303 대 0.213) qwen3-8b가 가장 높다. 그러나 제안 학습을 거치면 ko-sroberta와 qwen3-0.6b의 성능이 거의 같은 수준으로 수렴한다(R@1 0.497 대 0.503). 이는 제안 프레임워크가 백본의 초기 표현력 차이에 크게 의존하지 않으며, 비교적 경량의 한국어 임베딩 모델만으로도 충분한 매핑 성능을 얻을 수 있음을 시사한다.

이러한 수렴 경향은 더 큰 백본을 학습할 필요성을 줄인다. qwen3-8b는 사전학습 단계에서 가장 높은 성능을 보이나 이는 미세조정 이전의 표현력 차이일 뿐이며, 제안 학습 후 백본 격차가 수렴하는 패턴을 고려하면 추가 학습으로 얻을 이득은 제한적일 것으로 예상된다. 이에 본 논문은 학습 비용 대비 효용을 고려하여 qwen3-8b는 사전학습 성능만 비교에 포함하고, 제안 방법은 경량 백본인 ko-sroberta와 qwen3-0.6b를 중심으로 평가하였다.

특히 희소 매핑을 동일 가중으로 평가하는 MRR_macro에서 격차가 가장 크다. 제안 방법은 과학 - 특허에서 0.437~0.480, 산업 - 특허에서 0.515~0.523인 반면, 2단계는 각각 0.285와 0.321에 그친다. 이는 앞서 5.1에서 확인한 특허·산업 분류기의 낮은 Macro-F1, 즉 롱테일 한계가 매핑을 단위로 정렬하는 제안 방식에서 완화됨을 보여준다.

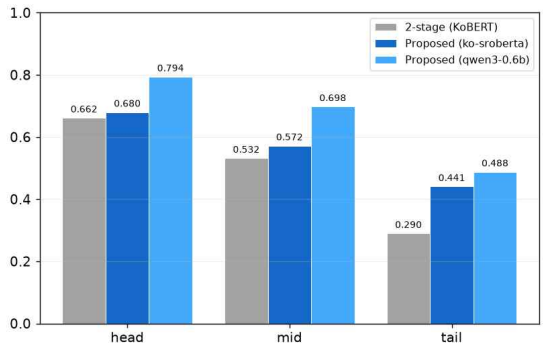


그림 2. 과학-특허 도메인의 빈도(q) 구간별 매핑 성능(MRR). 매핑쌍을 학습 빈도에 따라 head(0.33<q)·mid(0.33<q<0.66)·tail(q>0.66) 세 구간으로 나누어 측정하였다.

Fig. 2. Mapping performance (MRR) by frequency bin on the science - patent domain. Mapping pairs are divided into head, mid, and tail bins according to training frequency.

그림 2와 같이 롱테일 격차는 빈도 구간별 성능에서 더 분명히 드러난다. 매핑쌍의 분포는 소수의 빈번한 결합과 다수의 희소한 결합으로 이루어진 전형적인 롱테일 구조를 가지며, 이로 인해 세 방법 모두 빈도가 높은 head 구간에서는 비교적 높은 성능을 보이지만 구간이 희소해질수록 성능이 하락한다. 다만 그 하락의 폭은 방법에 따라 크게 다르다. 2단계 분류는 tail 구간에서 성능이 급격히 떨어지는 반면, 제안 방법은 같은 구간에서도 성능을 상당 부분 유지하여 두 방식의 격차가 tail에서 가장 크게 벌어진다.

이러한 차이는 두 방식의 구조적 특성에서 비롯된다. 2단계 분류는 각 축의 분류기가 희소 라벨을 충분히 학습하지 못하며, 두 축이 모두 희소한 매핑쌍에서는 이러한 한계가 누적되어 성능이 급락한다.

반면 제안 방법은 매핑을 두 코드 임베딩의 합성으로 표현하므로 개별 축의 정보를 공유할 수 있다. 따라서 어떤 매핑쌍이 드물게 관측되더라도, 그 축을 구성하는 라벨이 다른 결합에서 충분히 학습되었다면 합성을 통해 해당 매핑의 위치를 보다 안정적으로 추정한다. 이는 롱테일 환경에서 제안 방법이 갖는 핵심 이점이다.

5.3 확장 라벨(Open-set) 매핑 성능

표 4. 학습데이터에서 관측되지 않은 매핑쌍에 대한 확장 라벨 탐지 성능.

Table 4. Open-set detection performance on mapping pairs not observed in the training data.

Domain	Backbone	AUROC	FPR@95TPR
SCI-PAT	two stage	0.735	0.778
	ko-sroberta	**0.823**	**0.592**
	qwen3-0.6b	0.822	0.605
IND-PAT	two stage	0.712	0.818
	ko-sroberta	**0.783**	**0.689**
	qwen3-0.6b	0.781	0.792

표 4는 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).

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.

표 5. 제안 방법이 산출한 신규 매핑 후보 예시(과학 - 특허, qwen3-0.6b). 신규 매핑, 최근접 관측 매핑, 매핑 간 거리 δ

Table 5. Examples of novel mapping candidates produced by the proposed method (science – patent, qwen3-0.6b): the novel mapping, its nearest observed mapping, and the distance between them.

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

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 5는 그 거리에 따라 후보를 가까운 경우(A), 중간(B), 먼 경우(C)로 나누어 예시한 것이다.

가까운 경우(A)는 동일한 과학분류를 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 합성 표현이 특허 축의 코드 임베딩에서 분류 간 인접성을 보존하며, 그 결과 한 축을 고정한 채 의미적으로 가까운 다른 축으로의 신규 결합을 합리적으로 제안함을 의미한다.

반면 먼 경우(C)는 기존의 어떤 관측 매핑과도 충분히 가깝지 않은 결합으로, 거리 δ가 크게 나타

난다. 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. 결론

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 손실로 문서와 매핑을 공동 임베딩 공간에 정렬한다. 이를 통해 앞단 분류의 오차가 매핑으로 전이되는 2단계 방식의 구조적 한계를 제거하고, 두 분류 축의 의미를 분리해 담은 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 큰 향상을 보였다. 이는 각 축의 분류 오차가 누적되는 2단계 방식과 달리, 합성 표현이 개별 축의 정보를 공유하여 드물게 관측된 결합에 대해서도 안정적인 표현을 형성하기 때문이다. 또한 학습에서 관측되지 않은 신규 결합을 탐지하는 개방형 평가에서도, 매핑 좌표로부터의 거리를 직접 측정하는 제안 방법이 분류 확신도에 의존하는 2단계 방식 대비 우수하였다.

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 연구-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, 2000.

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

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

[4] WIPO, “Guide to the International Patent Classification”, World Intellectual Property Organization, 2024.

[5] 통계청 “한국표준산업분류”, https://kssc.mods.go.kr:8443/ksscNew_web/index.jsp#

[accessed: Jun. 24, 2026]

[6] Hammarfelt, B. “Linking science to technology: The patent paper citation and the rise of patentometrics in the 1980s”. Journal of Documentation, 77(6), 1413–1429, 2021. https://doi.org/10.1108/JD-12-2020-0218

[7] 곽지호, 정유철. “저자원 텍스트분류의 성능 향상을 위한 KoBERT기반 카테고리 매핑과 LLM 결합 연구”. 한국정보기술학회논문지, 23(1), 2025, 1-11, 2025. 10.14801/jkiit.2025.23.1.1

[8] Li, N., Kang, B. & De Bie, T. “Your Next State-of-the-Art Could Come from Another Domain: A Cross-Domain Analysis of Hierarchical Text Classification.” Mach Learn 115, 80 (2026). https://doi.org/10.1007/s10994-025-06993-w

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

[10] 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://arxiv.org/abs/2401.11705

[11] 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://arxiv.org/abs/2502.14570

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

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

[14] Fall, C. J., Törcsvári, A., Benzineb, K., and Karetka, G. “Automated Categorization in the International Patent Classification”. ACM SIGIR Forum, 37(1), 10–25, 2003. https://doi.org/10.1145/945546.945547

[15] Im, C. J., Kim, D. W., and Mandl, T. “Text Classification for Patents: Experiments with Unigrams, Bigrams and Different Weighting Methods”. International Journal of Contents, 13(2), 66–74, 2017. https://doi.org/10.5392/IJoC.2017.13.2.066

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

[17] 박진우, 심우철, 이상헌, 고봉수, 노한성. “한국어 특허 문장 기반 CPC 자동분류 연구 —인공지능 언어모델 KorPatBERT를 활용한 딥러닝 기법 접근—”. 지식재산연구, 17(3), 209–256, 2022. https://doi.org/10.34122/jip.2022.17.3.209

[18] 특허청 “산업(KSIC)-특허(IPC) 연계표” https://www.kipo.go.kr/ko/kpoContentView.do?menuCd=SCD0200272. [accessed: Jun. 24, 2026]

[19] Lybbert, T. J. Getting Patents and Economic Data to Speak to Each Other : [World Intellectual Property Organization], 2012. https://doi.org/10.34667/TIND.28868

[20] 지식재산처. “기술-품목-특허 연계표”. https://www.kipo.go.kr/ko/kpoContentView.do?menuCd=SCD0200273 [accessed: Jun. 24, 2026]

[21] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. “BERT: Pre-training of Deep

Bidirectional Transformers for Language Understanding". Proceedings of NAACL-HLT 2019, 4171–4186, 2019. https://doi.org/10.18653/v1/N19-1423

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

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

[24] Robinson, J. D., Chuang, C.-Y., Sra, S., and Jegelka, S. "Contrastive Learning with Hard Negative Samples". Proceedings of the 9th International Conference on Learning Representations, 2021. https://arxiv.org/abs/2010.04592

[25] Xiong, L., Xiong, C., Li, Y., Tang, K.-F., Liu, J., Bennett, P., Ahmed, J., and Overwijk, A. "Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval". Proceedings of the 9th International Conference on Learning Representations, 2021. https://arxiv.org/abs/2007.00808

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

[27] AI-Hub. "과학기술표준분류 대응 특허 데이터". AI-Hub, https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=71531, Accessed: Jun. 24, 2026.

[28] AI-Hub. "특허 분야 자동분류 데이터". AI-Hub, https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=547, Accessed: Jun. 24, 2026.

[29] jhgan. "ko-sroberta-multitask". Hugging Face Model Repository, Hugging Face, https://huggingface.co/jhgan/ko-srober a-multitask Accessed: Jun. 24, 2026.

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

저자소개