Journal of Next-generation Convergence Information Services Technology Vol.15, No.3, June (2026), pp.311-318
ISSN 2384-101X(p) 2672-1163(e) *http://dx.doi.org/10.29056/jncist.2026.06.01*

문맥적 임베딩과 거리 학습을 이용한 저자 식별 모델 개발

Development of an Authorship Identification Model Using Contextual Embeddings and Metric Learning

김광영[1], 설재욱[2], 이혜진[3]*

Gwang-Young Kim[1], Jae-Wook Seol[2], Hye-Jin Lee[3]*

요 약

최근 국내외 학술 데이터베이스의 급격한 증가에 따라 동명이인(Homonyms)과 동인이명(Synonyms)을 구별하는 저자 식별(Author Name Disambiguation) 문제는 학술 정보를 이용한 다양한 정보 서비스의 정확도를 높이기 위해 필수적인 과제로 대두되었다. 기존의 규칙 기반이나 단순 통계적 매칭 방식은 언어의 맥락적 의미를 파악하는 데 한계가 있으므로, 본 논문에서는 사전 학습된 한국어 언어 모델을 기반으로 대조 학습(Contrastive Learning)을 적용하여 저자 간의 미세한 의미론적 유사도를 정밀하게 측정하는 새로운 식별 모델을 제안한다. 제안 모델은 논문의 메타데이터(저자명, 소속, 제목, 초록 등)를 통합된 텍스트 특징으로 변환하고, 이를 고차원 벡터 공간에 매핑하여 동일 저자 여부를 판단한다. 또한, 신규 저자에 대해 실시간으로 신규 고유 ID를 발급하는 증분 학습(Incremental Learning) 메커니즘을 적용하여 시스템의 확장성을 확보하였다. 실험 결과, 제안된 모델은 정확도 0.962, 정밀도 0.958, 재현율 0.962, F1-score 0.957의 높은 성능을 나타냈다.

핵심어 : 저자 식별, 연구자 식별, 대조 학습, 증분 학습

Abstract

With the recent rapid growth of domestic and international academic databases, the problem of author name disambiguation to distinguish between homonyms and synonyms has emerged as an essential task to improve the accuracy of various information services using academic information. Since existing rule-based or simple statistical matching methods have limitations in capturing the contextual meaning of language, this paper proposes a new disambiguation model that precisely measures the fine-grained semantic similarity between authors by applying contrastive learning based on a pre-trained Korean language model. The proposed model converts the metadata of papers (author name, affiliation, title, abstract, etc.) into integrated text features and maps them into a high-dimensional vector space to determine whether they are the same

1 Data Curation Center, Korea Institute of Science and Technology Information, Seoul, Korea
e-mail: glorykim@kisti.re.kr
2 Data Curation Center, Korea Institute of Science and Technology Information, Seoul, Korea
e-mail: wodnr754@kisti.re.kr
3 Data Curation Center, Korea Institute of Science and Technology Information, Seoul, Korea
e-mail: hyejin@kisti.re.kr (Corresponding author)

* 이 논문은 「과학기술 AI 활용 강화를 위한 지능형 데이터 인프라 구축」(K26L2M1C1) 및 「AI 기술 활용 현장대응 의사결정 지원 플랫폼 실증 스케일업」(N25NM038)의 지원으로 작성되었습니다.

Received(March 23, 2026), Review Result(1st: April 10, 2026), Accepted(June 13, 2026), Published(June 30, 2026)

author. In addition, the scalability of the system was secured by applying an incremental learning mechanism that issues new unique IDs to new authors in real time. As a result of the experiment, the proposed model demonstrated high performance with an accuracy of 0.962, precision of 0.958, recall of 0.962, and an F1-score of 0.957.

Keyword : Author Name Disambiguation, Researcher Name Disambiguation, Contrastive Learning, Incremental Learning

1. 서론

디지털 학술 정보의 폭발적인 증가는 연구자들에게 방대한 지식을 제공하지만, 동시에 데이터의 신뢰성과 검색 정확도를 저해하는 요인을 발생시켰다. 그중 가장 핵심적인 문제는 저자 이름의 모호성(Ambiguity)이다. 예를 들어, ‘김철수’라는 동일한 이름을 가진 서로 다른 연구자들이 존재하거나, 한 명의 연구자가 소속 변경이나 영문 표기법의 차이로 인해 다르게 표기되는 경우가 빈번하다 [1].

전통적인 저자 식별 방법은 이름, 소속, 이메일 등의 서지 정보를 단순 문자열 매칭하거나 규칙 기반 알고리즘을 사용하였다. 그러나 이러한 방법은 오타, 약어 사용, 정보 누락 등이 포함된 실제 데이터 환경에서 성능 저하를 보였다. 최근 기계학습 및 딥러닝 기술의 발전으로 랜덤 포레스트(Random Forest)나 SVM(Suppot Vector Machine) 등을 활용한 연구가 진행되었으나, 여전히 고차원의 의미적 맥락을 파악하는 데에는 한계가 있었다 [2].

본 논문에서는 문장 표현 추출을 위해 사용한 문장 임베딩 모델은 BM-K가 제안한 ‘Sentence-Embedding-is-all-you-need’ 모델을 활용하였다. 해당 모델은 PyTorch 기반으로 구현되었으며, 공개된 GitHub 저장소(https://github.com/BM-K/Sentence-Embedding-is-all-you-need)의 사전 학습된 가중치를 사용하였다. 이에 본 논문에서는 한국어 문장 임베딩에 BM-K모델을 활용하여 저자 식별 문제를 의미론적 유사도 측정하였다 [7].

본 논문에서는 첫째, 저자명 외에 소속, 논문 제목, 키워드, 초록 등 다양한 메타데이터를 결합한 통합된 요소들을 고려하여 식별 정확도를 높였다. 둘째, 대조 학습 손실 함수를 도입하여 동일 저자의 논문 간 벡터 유사도를 계산했다. 셋째, 사전 등록되지 않은 저자에 대해 자동으로 신규 ID를 부여하고 데이터베이스를 갱신하는 온라인 클러스터링 방법을 제안하여 자동으로 신규 저자에 대해서 저자 ID를 발급할 수 있도록 하였다.

2. 관련 연구

2.1 저자 식별

저자 식별 문제는 크게 두 가지 접근 방식으로 나뉜다. 첫 번째는 지도 학습 기반으로, 사전에

레이블링 된 데이터를 통해 분류기를 학습시키는 방식이다. Ferreira et al. [3]은 저자 간의 공저자 네트워크와 인용 정보를 활용한 식별 모델을 제안하였다. 두 번째는 비지도 학습 기반의 클러스터링 방식으로, 별도의 학습 데이터 없이 유사도 척도만을 이용하여 군집화를 수행한다.

본 논문에서는 지도 학습된 임베딩 모델을 활용하되, 추론 단계에서는 임계치 기반의 클러스터링을 수행하는 하이브리드 접근 방식을 취하였다.

2.2 의미론적 문장 임베딩

BERT(Bidirectional Encoder Representations from Transformers)의 등장은 자연어 처리 분야에 혁신을 가져왔으나, 문장 간의 유사도를 측정하는 데 있어서는 [CLS]토큰이나 평균 풀링(Mean Pooling) 만으로는 비효율적인 이방성(Anisotropy) 문제가 발생하였다 [4]. 이를 해결하기 위해 Reimers et al. [5]은 샴 네트워크(Siamese Network) 구조의 Sentence-BERT를 제안하였다.

본 논문에서는 이를 한국어 데이터에 맞게 개선하고 대조 학습을 적용한 BM-K모델을 기반으로 문장 임베딩을 구축하였다.

3. 제안 모델

3.1 데이터셋 구성 및 전처리 작업

본 논문에서는 저자 식별 모델의 정확도 확보하기 위해 총 90만 건의 대규모 국내 논문, 보고서 및 특허 데이터에서 연구자별로 군집화된 데이터셋을 사용하였다. 원본 데이터는 정형화된 메타데이터 형식을 따르며, [표 1]과 같이 문서유형, 문서번호, 저자명(국/영), 소속기관(국/영/ID), 키워드(국/영), 공동저자(국/영), 저널명(국/영), 출판사(국/영), 제목(국/영), 언어, 전자우편주소, 출판년도 등 총 25개의 세부 필드로 구성되었다.

데이터 전처리 단계에서는 모델이 저자의 연구 분야와 소속 정보를 문맥적으로 이해할 수 있도록 ‘특징 통합(Feature Unification)’ 과정을 수행하였다.

첫째, 국문과 영문이 별도 컬럼으로 존재하는 경우하고 영문저자명이 없을 경우에도 국문저자명 및 영문저자명(예: 국문저자명: 김병주, 영문저자명: (Null)) 이를 하나의 문자열로 결합하여 정보 손실을 방지를 위해 필드 병합 및 정제를 수행하였다.

또한 국문기관명: 주식회사 포스코와 국문 제목 ‘리튬 이차 전지용 음극...’ 등의 정보는 연구자의 정체성을 나타내는 핵심 단서이므로 같이 적용할 수 있도록 하였다.

둘째, BERT 기반 모델의 입력으로 활용하기 위해, 선별된 주요 필드(저자명, 소속기관, 이메일, 저널/출판사, 제목, 키워드, 초록)를 [SEP]토큰으로 연결하여 하나의 긴 텍스트 시퀀스(Text

Sequence)로 자동 입력 시퀀스를 생성하였다.

[표 1] 저자 학습 데이터셋 구성
[Table 1] Author Training Dataset

필드명	데이터 예시	비고
저자 정보	김병주 (Kim Byung-joo)	한/영 이름 병합 처리
소속 기관	주식회사 포스코 (POSCO)	소속명 식별의 핵심 피처
연구 키워드	리튬 이차 전지; 음극; 제조 방법	연구 분야(Context) 정보
공동 저자	이강호; 박세민; 우정규	공저자 네트워크 정보
제목	리튬 이차 전지용 음극, 이의 제조 방법...	텍스트 임베딩의 주 대상
...	...	...

이와 같이 [표 1]은 저자 학습 데이터셋으로 구성되어 한글/영어 저자명은 병합하고 소속 기관명, 연구 키워드, 공동 저자, 논문 제목까지 포함하여 벡터 공간을 구성하였으며 구축된 90만 건의 텍스트 데이터는 분리 과정을 거쳐 고차원 벡터 공간에 매핑하여, 동일한 저자의 문서들은 벡터 공간상에서 가깝게 위치하도록 학습시켰다.

3.2 시스템 아키텍처 및 온라인 클러스터링

제안된 저자 식별 시스템의 전체 프레임워크는 [그림 1]과 같이 학습 데이터 준비, 임베딩 모델 학습 및 최적화, 예측 및 증분 식별과 같이 3단계로 구성하였다.

첫째, 데이터 준비 단계에서는 수집된 서지 정보로부터 저자 식별에 필요한 핵심 특징을 추출하였다. 저자명, 소속 기관, 논문 제목, 초록 등의 텍스트 필드를 결합하여 통합 특징(Unified Feature)을 생성하였으며, 이를 저자 ID별로 그룹화하여 학습 데이터셋과 검증 데이터셋으로 분할하였다.

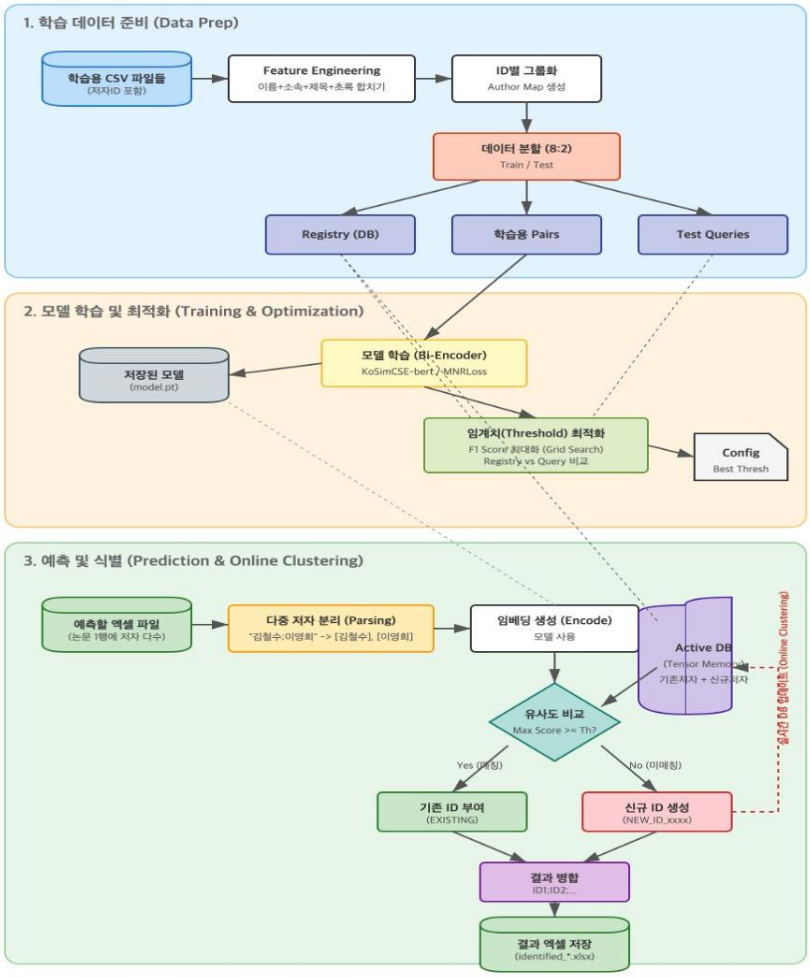
둘째, 학습 및 최적화 단계에서는 이중 인코드(Bi-Encoder) 구조의 BM-K모델을 사용하여 저자 간의 의미론적 유사도를 학습하였다. 동일 저자의 논문 쌍은 벡터 공간에서 가깝게, 다른 저자의 논문은 멀어지도록 대조 학습(Contrastive Learning)을 수행하였다. 학습 완료 후, 검증 데이터셋을 활용하여 F1 Score를 최대화하는 코사인 유사도 임계치(Threshold)를 실험적으로 결정하였다.

셋째, 예측 및 증분 식별 단계에서는 실제 운영 환경을 고려한 온라인 클러스터링 메커니즘을 적용하였다. 시스템은 입력된 논문의 저자들을 개별 단위로 분리하여 벡터화한 뒤, 메모리 상에 상주하는 실제 저자 데이터베이스와 실시간으로 유사도를 비교하였다.

이 과정에서 산출된 최대 유사도 점수가 임계치 이상일 경우 기존 저자 ID를 부여하였다. 반면, 임계치 미만일 경우 데이터베이스에 없는 신규 연구자로 판단하여 새로운 식별자를 발급하였다.

본 논문에서의 핵심적인 특징은 피드백 루프(Feedback Loop)구조에 있다. [그림 1]의 하단 점선

화살표와 같이, 신규 발급된 저자의 임베딩 벡터는 즉시 저자 데이터베이스에 등록되어 업데이트 되었다. 이를 통해 이후 입력되는 동일한 신규 저자의 논문을 즉각적으로 식별할 수 있는 증분 학습(Incremental Learning) 효과를 구현하였다.



[그림 1] 시스템 프레임워크

[Fig. 1] system framework

3.3 모델 구조 및 학습

본 논문에서 학습 데이터는 동일 저자가 작성한 서로 다른 논문 쌍(Positive Pair)으로 구성하였다. 학습 목적 함수로는 다중 부정 랭킹 손실(Multiple Negatives Ranking Loss)를 채택하였다. 이 손

실 함수는 배치(Batch) 내에서 정답 쌍(Anchor-Positive)의 거리는 좁히고, 나머지 모든 샘플(Negatives)과의 거리는 멀어지도록 학습하여 벡터 공간 내에서 저자별 군집화가 자연스럽게 이루어지도록 하였다.

$$L(x_i, y_i) = -\log \frac{e^{Sim(x_i, y_i)/\tau}}{\sum_{j=1}^{N} e^{Sim(x_i, y_j)/\tau}} \quad (1)$$

여기서 [수식 1]에서 Sim은 코사인 유사도이며, taur는 온도(Temperature) 매개변수이다.

3.4 증분 식별 및 신규 ID 발급

본 논문에서의 모델은 기존 데이터베이스(Registry)에 존재하는 저자 벡터와의 코사인 유사도를 계산하였다. 최적화된 임계치 이상의 유사도를 보일 경우 기존 ID를 부여하였으며, 매칭되는 저자가 없을 경우 새로운 식별자를 발급하였다. 신규 발급된 저자 정보는 즉시 임베딩 텐서에 업데이트되어, 이후 입력되는 동일 저자의 논문을 즉시 식별할 수 있는 온라인 학습(Online Learning) 효과를 반영하도록 구현하였다.

4. 실험 결과

4.1 실험 환경

본 실험은 PyTorch 프레임워크 기반으로 수행되었으며, A6000 리눅스 서버에 CUDA(NVIDIA GPU) 가속을 지원하는 환경에서 진행되었다. 데이터셋은 국내 학술 논문 데이터베이스에서 추출한 대규모 저자 군집화된 데이터를 사용하였으며, 학습셋과 검증셋을 8:2 비율로 분할하여 사용하였다. 대용량 데이터 처리를 위해 데이터 개수에 따라 멀티 프로세싱 적용하였다. 전체 학습 시간은 약 14시간이며 배치(batch) 크기는 16 및 2에폭(epoch) 정도로만 수행하였다.

4.2 임계치 최적화 및 실험 결과

저자 식별의 성능을 결정짓는 중요한 하이퍼파라미터인 유사도 임계값을 결정하기 위해 0.6에서 0.95 사이의 범위에서 0.05 단위로 그리드 서치(Grid Search)를 수행하였다. 검증 데이터셋에 대한 F1 Score를 기준으로 최적의 임계값을 탐색하였으며, 그 결과 0.75 부근에서 가장 균형 잡힌 성능을 보임을 확인하였다.

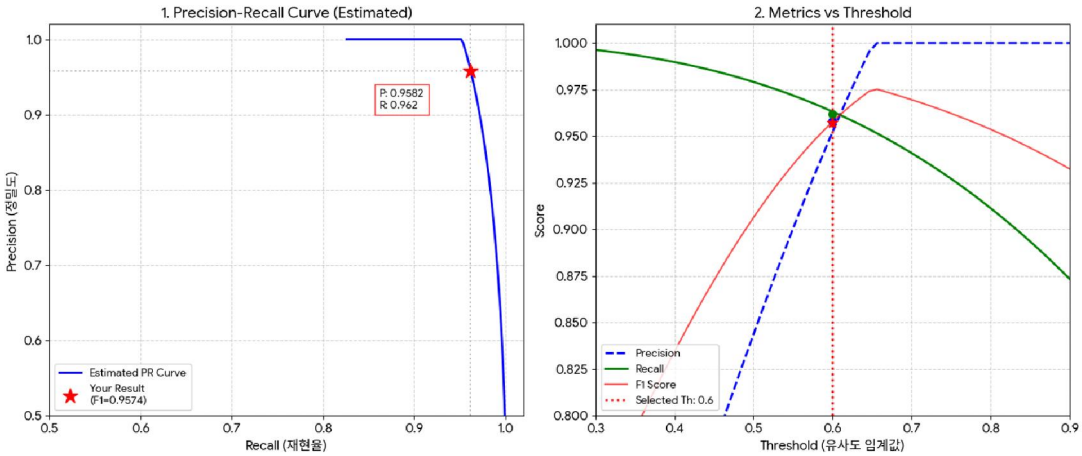
최적화된 모델을 사용하여 테스트 데이터셋에 대한 성능 평가를 수행하였다. 평가는 식별된 저

자 ID와 실제 정답 ID를 비교하여 이루어졌으며, 주요 성능 지표는 다음과 같다.

[표 2] 성능 지표

[Table 2] Evaluation Metrics

평가 지표 (Metric)	값 (Value)	비고
Accuracy (정확도)	0.9620	전체 예측 중 정답 비율
Precision (정밀도)	0.9582	Weighted Average
Recall (재현율)	0.9620	Weighted Average
F1 Score	0.9574	정밀도와 재현율의 조화 평균



[그림 2] 성능 지표 그래프

[Fig. 2] Evaluation Metrics Graph

실험 결과, [표 2] 및 [그림 2]와 같이 제안한 모델은 0.9620의 높은 정확도와 0.9574의 F1 Score를 기록하였다. 이는 동명이인 구별뿐만 아니라, 소속이나 이메일 정보가 일부 누락된 경우에도 텍스트의 맥락 정보(제목, 초록 등)를 통해 성공적으로 저자를 식별하였다. 특히 Precision(0.9582)과 Recall(0.9620)의 차이가 크지 않아, 모델이 과적합 되거나 특정 클래스에 편향되지 않고 안정적인 성능을 보였음을 확인하였다.

5. 결론

본 논문에서는 문장 임베딩과 대조 학습 기법을 활용하여 한국어 저자 식별 모델을 제안하고 구현하였다. 제안된 모델은 서지 정보를 통합하여 의미론적 벡터를 생성함으로써, 단순 키워드 매칭의 한계를 극복하고 95.74%의 F1 Score라는 높은 식별 성능을 달성하였다. 특히, 연구 과정에서 대용량 학술 데이터를 효율적으로 처리하고, 신규 연구자를 실시간으로 데이터베이스에 추가하여

바로 사용할 수 있도록 자동 시스템화하였다.

향후 연구에서는 동명이인이 매우 많은 특정 이름(e.g., Kim, Lee)에 대한 세밀한 구분을 위해 그래프 신경망을 도입하여 공저자 네트워크 정보를 결합하는 방향으로 모델을 고도화할 계획이며 실제 최근 10년 간의 학술 논문데이터를 이용하여 자동 저자 식별 프로세스를 통해 발급 서비스를 제안할 예정이다.

References

[1] V. I. Torvik and N. R. Smalheiser, "Author name disambiguation in MEDLINE," *ACM Transactions on Knowledge Discovery from Data*, vol. 3, no. 3, pp. 1-29, Jul. 2009, doi: 10.1145/1552303.1552304.

[2] A. A. Ferreira, M. A. Gonçalves, and A. H. F. Laender, "A brief survey of automatic methods for author name disambiguation," *SIGMOD Record*, vol. 41, no. 2, pp. 15-26, Jun. 2012, doi: 10.1145/2334510.2334514.

[3] H. Han, L. Giles, H. Zha, C. Li, and K. Tsioutsiouliklis, "Two supervised learning approaches for name disambiguation in author citations," in *Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*, Tucson, AZ, USA, Jun. 2004, pp. 296-305, doi: 10.1145/996350.996426.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Minneapolis, MN, USA, Jun. 2019, pp. 4171-4186, doi: 10.18653/v1/N19-1423.

[5] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Hong Kong, China, Nov. 2019, pp. 3982-3992, doi: 10.18653/v1/D19-1410.

[6] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple contrastive learning of sentence embeddings," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online and Punta Cana, Dominican Republic, Nov. 2021, pp. 6894-6910, doi: 10.18653/v1/2021.emnlp-main.552.

[7] Y. J. Ham, "KoSimCSE: Korean sentence embeddings using contrastive learning," *GitHub.com*, https://github.com/BM-K/Sentence-Embedding-is-all-you-need (accessed Feb. 3, 2026).