OpenAlex 글로벌 저자 식별 모델의 한국 학술 데이터 적용성 평가 및 특성 최적화 연구*

Applicability Evaluation and Feature Optimization of the OpenAlex Global Author Disambiguation Model for Korean Scholarly Data

정 형 상 (Hyeong-Sang Jeong)**
곽 승 진 (Seung-Jin Kwak)***

초 록

저자명 식별은 학술 정보 시스템의 핵심 과제이나, 영문 중심인 OpenAlex 모델의 국내 학술 생태계 적용성에 대한 검증은 미비하다. 본 연구는 KISTI OCEAN 데이터베이스의 2023~2024년 논문 54,049건을 활용해 OpenAlex 모델의 한국 데이터 적용성을 평가하고, 한국어 특성에 맞춘 7개 특성 최적화를 수행하였다. 단계적 실험 결과, F1 점수는 0.852(v1-1)에서 0.860(v2-2)으로 향상되었으며, 정답셋 보정 후에는 정확도 0.930, F1 점수 0.931을 달성하였다. 또한 ORCID 기반 교차 검증에서 F1 점수 0.892를 기록하여 모델의 신뢰성을 확인하였다. 특히 대규모 데이터의 효율적 관리를 위해 증분적 처리 방식을 도입하고 수작업 검증을 결합한 최적화 공정을 제안하였으며, 최종적으로 국내 저자 183,105명을 109,205개 식별자로 그룹화하는 파이프라인을 구축하여 실무적 타당성을 검증하였다.

ABSTRACT

Author Name Disambiguation(AND) is a critical task in scholarly information systems; however, the applicability of the English-centric OpenAlex model to the Korean academic ecosystem has yet to be fully validated. This study evaluates OpenAlex's performance using 54,049 papers (2023-2024) from KISTI's OCEAN database and optimizes seven features tailored to Korean linguistic characteristics. Stepwise experiments demonstrate that the F1-score improved from 0.852 (v1-1) to 0.860 (v2-2), ultimately achieving an accuracy of 0.930 and an F1-score of 0.931 after ground-truth refinement. Cross-validation with ORCID yielded an F1-score of 0.892, confirming the model's reliability. Specifically, we propose an optimization process that combines incremental processing with manual verification to manage large-scale data efficiently. Finally, the study validates a pipeline that successfully clusters 183,105 author records into 109,205 unique identifiers, verifying its practical feasibility and scalability for Korean scholarly metadata.

키워드: 저자명 식별, OpenAlex, 한국어 저자명, 특성 최적화, OCEAN, 학술 데이터베이스
Author Name Disambiguation, OpenAlex, Korean Author Names, Feature Optimization, OCEAN, Scholarly Database

---

* 이 연구는 충남대학교에 의해 지원되었음.
** 충남대학교 문헌정보학과 석사과정(jhscjs@o.cnu.ac.kr) (제1저자)
*** 충남대학교 문헌정보학과 교수(sjkwak@cnu.ac.kr / ISNI 0000 0004 6812 0586) (교신저자)
논문접수일자 : 2026년 2월 20일 논문심사일자 : 2026년 2월 23일 게재확정일자 : 2026년 3월 2일
한국비블리아학회지, 37(1): 387-410, 2026. http://dx.doi.org/10.14699/kbiblia.2026.37.1.387

1. 서 론

디지털 학술 정보의 폭발적인 증가는 연구자들에게 방대한 지식에 대한 접근성을 제공하였으나, 동시에 '저자명 중의성(Author Name Ambiguity)'이라는 고질적인 문제를 심화시켰다. 동명이인을 구별하고 동일 저자의 다양한 성명 표기를 하나의 실체로 통합하는 저자 식별(Author Name Disambiguation, AND) 기술은 단순한 데이터 정제 작업을 넘어, 연구 성과의 정확한 평가와 기관별 성과 분석, 그리고 학술 네트워크 분석의 신뢰성을 좌우하는 핵심 요소로 그 중요성이 날로 증대되고 있다.

학술 정보 시스템에서 저자명 식별은 동일한 이름을 가진 여러 저자를 구분하거나, 다양한 이름 변형으로 발표하는 동일 저자를 통합하는 기본적인 과제이다(Ferreira & Laender, 2023). 저자명 중의성 문제는 학문적 성과의 왜곡된 귀속과 서지 통계의 오류를 야기하며, 나아가 디지털 정보 시스템 내에서의 검색 정확도와 전문가 식별 성능을 저해하는 핵심 요인으로 지목되고 있다(Sanyal et al., 2021).

이러한 저자명 중의성 해소를 위해 국제적으로는 ORCID, ISNI, VIAF 등 다양한 식별자 체계가 운용되고 있다. 특히 ISNI는 ISO 27729 표준에 기반한 다양한 분야의 인명 식별자로서, 학술・출판・음악・방송 등 다양한 창작 영역을 포괄하는 브리지 식별자 역할을 수행하고 있다. 국내에서는 국립중앙도서관과 국회도서관이 ISNI 등록기관으로 운영되고 있으나, ISNI의 실질적 활용에 대한 국내 기관들의 인식은 아직 초기 단계에 있는 것으로 파악된다.

OpenAlex는 저자명, 발표 기록, 인용 패턴, 그리고 이용 가능한 경우 ORCID를 활용하여 저자를 식별하는 알고리즘을 사용하며, 2023년 7월에 더 정확한 머신러닝 모델을 갖춘 새로운 저자 식별 시스템으로 전환하여 모든 저자에게 새로운 저자 ID를 할당하였다(OpenAlex, 2023). 그러나 이러한 개선된 모델도 국가 및 지역과 저자 집단에 따라 성능 편차가 나타나는 문제가 보고되고 있다(Zhao & Chen, 2025). 학술 데이터베이스 내 저자 식별에 있어 중국인 저자는 이름의 높은 중복성으로 인한 부정확한 할당 문제가 두드러지며(Treeratpituk & Giles, 2012), 한국인 저자 또한 로마자 표기법의 변이와 성・이름 순서의 혼용 등으로 인해 식별 효율성이 저해되는 유사한 도전 과제에 직면해 있다. 한국은 김(Kim), 이(Lee), 박(Park) 등 특정 성씨에 인구가 편중되어 있으며, 동일한 한글 성명이라도 영문 로마자 표기법에 따라 다양한 변형(예: 김/Gim/Kim)이 존재한다(Kim, 2018).

이러한 언어적 특성(Feature)은 영문 데이터 중심으로 설계된 글로벌 저자 식별 모델을 국내 데이터에 적용할 때 예측 성능을 저해하는 주요 요인이 된다. 현재까지 OpenAlex와 같은 글로벌 학술 그래프가 국내 학술 생태계 데이터에서 어느 정도의 효용성을 갖는지, 그리고 한국어 특화 특성을 통해 얼마나 성능을 개선할 수 있는지에 대한 체계적인 검증은 매우 미비한 실정이다. 이에 본 연구는 국내 최대 학술 인용 색인 시스템인 OCEAN 데이터를 활용하여 글로벌 저자 식별 모델의 국내 적용성을 평가하고, 한국적 맥락에 최적화된 식별 파이프라인을 제안하고자 한다. 이에 본 연구는 세 가지 연구 질문을 설정하고 이를 규명하고자 한다.

- RQ1. 영문 기반의 OpenAlex 모델이 한국인 저자명과 국내 학술 데이터 환경(OCEAN)에서도 유효한 성능을 유지하는가?
- RQ2. 한국인 저자 식별 시 가장 변별력이 높은 핵심 특성은 무엇이며, 한국어 특성을 반영한 최적화를 통해 어느 정도의 성능 향상을 기대할 수 있는가?
- RQ3. 대규모 국내 학술 데이터 처리를 위한 OpenAlex 기반의 저자 식별 파이프라인은 어떻게 설계되어야 하는가?

본 연구 수행을 위해 KISTI의 국내 학술지 인용색인 분석 시스템(OCEAN) 데이터베이스에서 2023~2024년 발행된 국내 논문 54,049건을 추출하였다. 또한, KISTI 인물 ID(KISTI_PSON_ID)와 ORCID를 정답셋으로 구축하여 OpenAlex 모델의 성능을 정밀하게 평가하였다. 나아가 한국어 성명의 특수성을 반영한 특성 재정의와 단계별 하이퍼파라미터 최적화를 수행하였으며, 최종적으로 국내 학술 논문 전반에 적용 가능한 저자 식별 파이프라인을 제안하고자 한다.

2. 관련 연구

2.1 저자 이름 식별의 일반적 도전 과제

저자 이름 식별은 학술 출판 시스템의 기본 문제로, 동일한 이름을 가진 여러 저자를 구분하거나 다양한 이름 변형으로 발표하는 동일 저자의 기록을 통합해야 한다(Ferreira & Laender, 2023). 기존 식별 방법론은 규칙 기반 방식과 머신러닝 기반 방식으로 분류되며(Zhai et al., 2019), 최근 딥러닝 기반 접근법의 등장에도 불구하고 특성 기반 지도 학습 방식이 널리 활용되고 있다. 이러한 방법론의 체계적 평가를 위해 S2AND(Subramanian et al., 2021) 등의 벤치마크가 제안된 바 있다.

2.2 특성 기반 저자 식별 방법

저자 식별에 사용되는 주요 특성으로는 공저자 관계, 논문 제목, 출판 제목 등의 인용 내부 특성과 이메일, 소속 기관 등의 인용 외부 특성이 있다(강인수 외, 2008). 조직/기관 특성은 동명 저자를 식별하는 데 있어 높은 인식도를 가지며, 대부분의 경우 동일한 기관의 동명 저자는 동일 인물로 판단할 수 있다(Zhai et al., 2019).

공저자 네트워크는 저자의 정체성을 드러내는 직관적인 특성 중 하나이며(Seol et al., 2016), 학술지 주제 분류 및 의미론적 범주 역시 저자가 특정 학문 영역 내에서 지속적으로 연구를 수행한다는 경향성을 반영하는 유용한 특성으로 활용된다. Vishnyakova et al.(2016)의 분석에 따르면, 저자의 전체 성명과 이메일, 소속 기관명, 그리고 공저자 정보가 식별 모델의 성능을 결정짓는 핵심 자질로 확인되었다.

최근에는 개별 특성에 가중치를 자동 도입하거나(Xu et al., 2018), 미분 가능한 특성 선택 알고리즘을 통해 최적화된 조합을 탐색하는 시도가 이루어지고 있다(Fang et al., 2023). 저자 식별 문제는 논문의 개별 속성 정보와 논문 간의 관계 정보를 통합적으로 고려할 때 가장 효과적으로 해결될 수 있으며(Tang et al., 2012), 본 연구에서 활용한 OpenAlex 모델 역시 텍스

트 기반의 특성과 그래프 기반의 관계 특성을 복합적으로 수용한다는 점에서 이러한 통합적 접근의 이론적 궤를 같이한다.

2.3 한국어 이름 식별의 특수성

한국 인명 식별은 로마자 표기법의 비일관성과 언어 구조적 특성으로 인해 고유한 기술적 난제를 수반한다. 선행 연구에 따르면, 아시아권 저자는 중복성이 높아 저자 식별 모델의 변별력을 저해하는 핵심 요인이 된다(Treeratpituk & Giles, 2012; Zhao & Chen, 2025). 특히 서구식 영문 참고문헌 표기 방식은 한국 인명의 고유한 변이형을 충분히 포착하지 못하며, 이는 결과적으로 저자 식별의 정밀도(Precision)와 재현율(Recall)을 동시에 저하시키는 원인이 된다(Kim, 2018). 또한, 동아시아 인명은 성과 이름의 표기 순서가 혼용되는 경우가 매우 빈번하여(Treeratpituk & Giles, 2012), 성 기반 식별 알고리즘 설계 시 이러한 가변성을 고려해야 한다.

국내외 학계에서는 이러한 난제를 극복하기 위해 다양한 알고리즘적 시도를 지속해 왔다. Seol et al.(2016)은 약 9만 건의 IT 분야를 대상으로 공저자 네트워크 확장 기법과 SVM 알고리즘을 결합하여 94.79%의 F1-score를 달성함으로써 정교한 특성 엔지니어링의 중요성을 입증하였다. 또한 서구와 비서구권 인명의 표기 방식 차이를 극복하기 위해 민족적 특성을 고려한 모델링의 필요성이 강조되고 있다. Louppe et al.(2016)은 준지도 학습 시 민족적 배경을 반영한 특성을 추가할 경우 비서구권 저자의 식별 성능이 유의미하게 향상됨을 입증하였는데, 이는 단순한 텍스트 매칭을 넘어 문화적 맥락을 고려한 중의성 해소가 필수적임을 시사한다.

한편, 저자 식별 문제의 근본적 해결을 위해서는 알고리즘적 접근뿐만 아니라 표준화된 식별자 체계의 구축과 확산이 병행되어야 한다. 국내에서는 ISNI 기반의 이종 플랫폼 간 데이터 융합(이승민 외, 2019), 분야별 기관 협력을 통한 식별자 보급(오상희 외, 2019), 기관 간 메타데이터 공동 활용을 위한 기술 요소 도출(박진호 외, 2020) 등의 연구가 수행되어, 기술적 방법론과 식별자 인프라의 상호 보완적 발전 필요성을 시사하였다.

따라서 본 연구가 한국어 특수성을 반영하여 OpenAlex 모델의 특성을 재정의한 접근 방식은, 기존 글로벌 모델의 한계를 보완하고 식별 정확도를 극대화하는 동시에 향후 표준 식별 체계와의 연계성을 높이기 위한 논리적 타당성을 확보한다.

2.4 OpenAlex 및 글로벌 데이터베이스의 지역별 성능 차이

OpenAlex는 기존 Microsoft Academic Graph (MAG)를 계승하여 구축된 개방형 학술 지식 그래프로(Priem et al., 2022), Crossref, PubMed, ORCID 등 다양한 소스로부터 수집된 정보를 통합하여 고도화된 저자 식별을 수행한다(OpenAlex, 2023).

최근 연구에 따르면 OpenAlex의 저자 식별 성능은 지역 및 분야별로 뚜렷한 편차를 보인다. Zhao와 Chen(2025)은 중국 학자 그룹을 분석하여 Clarivate에서는 과소 병합과 과대 병합이, OpenAlex에서는 식별자 중복 부여와 미

부여가 주요 문제로 나타남을 확인하였으며, 두 시스템이 각기 다른 데이터 오류 특성을 지니고 있음을 입증하였다. 그러나 선행 연구는 주로 중국 및 미국 학자 데이터에 편중되어 있으며, 한국 학술 데이터의 특성과 로마자 표기 변이 등을 고려한 OpenAlex 모델의 체계적인 성능 평가는 아직 미비한 실정이다.

3. 방법론

3.1 OpenAlex 저자 식별 모델의 구조 및 특성

OpenAlex의 저자 식별 모델은 Gradient Boosting 알고리즘의 구현체인 XGBoost 기반 지도 학습 모델이다. 본 모델은 두 저자 레코드 쌍의 동일 실체 여부를 판별하는 이진 분류 작업을 수행한다. 저자 간 유사성 측정을 위해 7개의 핵심 특성을 입력으로 활용하며, 방대한 서지 데이터 내 복합적 관계성을 정밀하게 학습한다. 상세 정의와 산출 방식은 〈표 1〉에 정리하였다.

실제 추론 과정에서 본 모델은 파이썬 객체인 Disambiguator.pkl 파일을 호출하여 동작한다. 입력된 7개의 특성을 바탕으로 predict_proba 함수를 호출하여 두 저자 레코드가 동일인일 확률값을 산출한다. 최종적으로 설정된 임계치를 적용하여 판정을 내린다. 즉, 산출된 확률값이 임계치를 초과할 경우에만 두 저자를 동일로 식별한다. 이러한 확률 기반 방식은 대규모 서지 데이터베이스에서도 연산 효율성을 보장하며, 연구 및 서비스의 목적에 따라 임계치를 미세 조정함으로써 정밀도와 재현율 사이의 최적의 균형점을 도출할 수 있다는 실무적 장점을 지닌다.

이러한 OpenAlex의 모델 설계는 S2AND (Subramanian et al., 2021) 등 현대적 저자 식별 시스템의 표준적 프레임워크를 충실히 따르고 있으며, 검증된 핵심 특성들을 변수로 포함하고 있어 모델의 구조적 타당성을 확보하고 있다.

〈표 1〉 OpenAlex 저자 식별 모델에 사용되는 7개 특성

특성명	정의 및 상세내용	값 범위	산출산식(Metric)
inst_per	소속기관 일치성	0 또는1	두 저자의 소속 기관 교집합 존재 여부(Binary)
concepts_shorter_per	연구주제 유사도	0.0 ~ 1.0	최상위분야(Level 0)를 제외한 연구개념(Concept)의 Jaccard 유사도
coauthors_shorter_per	공저자 네트워크 유사도	0.0 ~ 1.0	성명 길이 6자이상의 공저자 리스트에 대한 Jaccard 유사도
exact_match_len	성명 완전 일치 강도	0 이상의 정수	성명 완전 일치 여부(0 또는1) × 성명의 전체 길이
exact_match_spaces	성명 구성 복잡도	0 이상의 정수	성명 완전 일치 여부(0 또는1) × 성명내 공백(Space) 수
citation_per	인용 네트워크 유사도	0.0 ~ 1.0	두 저자가 참조한 문헌리스트의 Jaccard 유사도
citation_work_match	상호인용관계	0 또는1	두 논문 간의 직접적인 인용 또는 피인용관계 존재 여부

3.2 한국 데이터 적용을 위한 특성 재정의

공저자 관계는 저자 식별에 있어 가장 강력하고 직관적인 특성 중 하나이다. Fan et al.(2011)은 공저자 그래프 기반의 GHOST 프레임워크를 통해, 동명이인 문제 해결에 있어 네트워크 내 유효 경로 분석이 높은 정밀도와 재현율을 달성함을 입증하였다. 이러한 이론적 근거를 바탕으로, 본 연구는 OpenAlex의 범용 식별 모델을 국내 학술 데이터 환경(OCEAN)에 최적화하기 위해 한국어 성명의 언어적 특수성을 반영한 특성 재정의를 수행하였다. 주요 수정 사항 및 실험 단계별 특성 최적화 내용은 다음과 같다.

첫째, 성명 일치 강도(exact_match_len) 특성 최적화를 수행하였다. 원 모델은 로마자 표기의 길이를 기준으로 매칭 강도를 산출한다. 초기 실험(v1-1)에서는 직관적인 한글 성명의 음절 수(3자)를 적용하였으나, v2-1 단계부터는 원 모델의 설계 의도 유지 및 변별력 강화를 위해 한글 성명의 로마자 표기 길이를 기준으로 재전환하였다.

둘째, 성명 복잡도(exact_match_spaces) 특성을 한국어 환경에 맞게 재구조화하였다. 원 모델은 성명 내 공백 수를 통해 이름의 희소성을 파악한다. 공백이 거의 존재하지 않는 한글 성명의 특성을 고려하여, v1-1에서는 음절별 정보 가치를 반영하기 위해 상숫값(2)을 일괄 적용하였다. 이후 v2-1에서는 이름의 음절 수에 비례하여 정보량이 증가한다는 가설을 바탕으로 '음절 수 - 1'로 특성을 재정의함으로써 수치의 동적 범위를 확보하였다.

셋째, 공저자 네트워크 유사도(coauthors_shorter_per) 산출 방식을 개선하였다. v1-1에서는 KISTI 인물 ID를 매칭의 기준으로 삼았으나, 식별자 미부여 저자에 대한 사각지대를 해소하기 위해 v1-2부터는 성명 텍스트 직접 비교 방식을 병행 도입하여 네트워크 비교의 범위를 극대화하였다.

3.3 정답셋 구축 및 ORCID의 역할

ORCID 식별자를 보유한 저자 집단은 식별 오류율이 유의미하게 낮은 것으로 보고되어(Zhou & Sun, 2024), 고도화된 저자 식별 모델의 정답셋 구축 과정에서 ORCID의 결정적 중요성이 확인되었다.

본 연구의 성능 평가를 위한 정답셋 구축 과정에서는 KISTI 인물 ID를 주 평가지표로 설정하였으며, 모델의 독립적 검증 및 객관성 확보를 위해 ORCID를 보조 평가지표로 병행 활용하였다. 이러한 다중 검증 체계는 Zhang et al.(2023)이 LAGOS-AND 데이터셋을 개발하며 ORCID와 DOI의 결합을 통해 대규모의 고신뢰도 정답 데이터를 자동으로 추출한 방법론에 근거를 두고 있다.

또한 ORCID 기반의 정답셋은 국제적으로 통용되는 표준 데이터를 포함하므로(Fernández-Marcial et al., 2023; Zhang et al., 2023), 본 연구에서 도출된 평가 결과의 대외적 신뢰성과 일반화 가능성을 동시에 제고할 수 있다.

한편, 국제 표준 인명 식별자인 ISNI는 학술뿐 아니라 출판, 음악, 방송 등 광범위한 창작 영역을 포괄하는 범용 식별자이나(이승민 외, 2019), 국내 기관들의 ISNI 활용이 아직 초기 단계에 머물러 있고(오상희 외, 2019), ORCID가 연구 및

학술활동에 특화된 식별체계라는 점에서 본 연구의 학술 데이터 검증 목적에 보다 직접적으로 부합한다고 판단하였다.

3.4 OCEAN 데이터베이스 매핑 및 전처리

본 연구는 단순한 텍스트 매칭의 한계를 극복하고 저자 식별의 정밀도를 확보하기 위해, 의미론적 내용 유사성과 네트워크 구조를 복합적으로 고려한다(Müller, 2018). OCEAN 데이터베이스로부터 OpenAlex 모델의 입력 특성을 생성하기 위해 설계한 데이터 소스별 매핑 현황은 다음과 같다.

첫째, 소속 기관 일치성(inst_per) 산출을 위해 저자 메타데이터 내의 'KISTI 기관 ID'를 식별자로 활용한다. 기관 식별자가 고유번호 형태로 존재하지 않는 레코드의 경우, 데이터 누락을 방지하고 식별 범위를 극대화하기 위해 '소속 기관 국문명' 및 '영문명' 텍스트를 상호 보완적으로 참조하여 기관의 일치 여부를 판별한다.

둘째, 연구 주제 유사도(concepts_shorter_per) 측정을 위해 저널 메타데이터의 다중 분류 체계를 활용한다. 이는 저자가 특정 학문 분야 내에서 지속적으로 연구를 수행한다는 의미론적 맥락을 반영하기 위함이다. 구체적으로 듀이 십진분류법, KISTI 주제 코드, Web of Science 및 Scopus의 주제 분류 데이터를 통합 비교함으로써 저자의 학문적 도메인 일치성을 정밀하게 측정한다.

셋째, 인용 네트워크 특성(citation_per, citation_work_match) 생성은 OCEAN의 참고문헌 데이터셋을 기반으로 수행된다. 개별 논문의 인용 리스트를 대조하여 저자 간의 네트워크 및 상호 인용 관계를 식별하며, 이는 소속 기관이나 이메일 정보가 불분명한 상황에서도 저자의 정체성을 규명하는 강력한 지표로 활용된다.

3.5 테스트 데이터셋 구축 및 레이블링

본 연구의 성능 검증에 활용된 OCEAN 데이터베이스의 인물 식별 체계는 강인수 외(2009)가 KISTI에서 구축한 대용량 저자 식별 평가셋(KISTI-AD-E-01-TestSet)의 방법론적 계보를 잇고 있다. 해당 선행 연구는 웹 검색 및 수작업 식별 과정을 거쳐 총 6,921명의 실세계 저자 식별자를 확보하였으며, 이는 국내 학술 정보 서비스 내 고신뢰 전거 데이터를 구축하기 위한 선구적인 시도로 평가받는다.

이러한 방법론적 정통성을 바탕으로, 본 연구는 최신 학술 생태계에서의 모델 유효성을 검증하고자 데이터의 시계열적 범위를 확장하였다. 구체적으로 OCEAN 데이터베이스에 수록된 2023~2024년 사이의 최신 국내 발행 논문을 표집 대상으로 삼았으며, 다음과 같은 체계적 절차를 거쳐 실험용 테스트 데이터셋을 구축하였다.

1) 데이터 추출 및 필터링: 학술지 관리 번호의 식별 체계를 기반으로 2023년과 2024년에 해당하는 논문 레코드를 총 54,049건을 추출, 데이터 정제 과정을 거쳐 최종 54,036건을 분석 데이터로 확보하였다.

2) 분석 대상 저자 한정: 식별의 정확도를 확보하기 위해 각 논문에서 제1저자를 추출하였으며, 이를 통해 총 49,433건의 저자 레코드를 확보하였다. 이 과정에서 KISTI 인물 ID가

없는 4,603건(약 8.5%)의 저자 레코드는 실험의 통제성을 위해 전처리 과정에서 제외하였다.

3) 동명이인 후보군 도출: 추출된 제1저자 중 한글 성명이 동일한 저자들의 논문 쌍을 구성하여, 저자 식별 모델의 주요 과제인 동명이인 판별 후보군 96,003건을 생성하였다.

4) 중복 제거 및 최종 데이터셋 확정: 쌍의 중복을 제거하고 데이터의 정합성을 검토하여 49,401건의 테스트 쌍을 확정하였다.

5) 정답셋 레이블링: 확정된 데이터 쌍에 대해 KISTI 인물 ID를 기준으로 레이블링을 수행하였다. 두 레코드의 ID가 일치할 경우 동일인, 불일치할 경우 비동일인으로 정의하였다. 최종 데이터셋은 동일인 10,910건, 비동일인 38,491건으로 구성되었으며, 약 1:3.5 비율의 불균형 특성을 나타냈다.

3.6 특성 추출 결과 및 데이터 가용성 분석

저자 식별 과정에서 소속 기관이나 이메일 정보가 부재하거나 불확실한 경우, 인용 네트워크 기반 정보가 핵심적인 보완 자료로 기능한다(Levin et al., 2012). 이에 본 연구는 citation_per 및 citation_work_match를 주요 변수로 채택하여 식별 모델의 견고성을 확보하고자 하였다. 최종 선정된 49,401건의 데이터 대상으로 7개 핵심 특성을 추출한 결과는 〈표 2〉와 같다.

각 특성별로 유효값을 보유한 데이터의 비중이 상이하게 나타났으며, 이는 국내 학술 메타데이터의 구축 수준과 정보의 완전성에 따른 차이로 해석된다.

1) 성명 기반 특성의 완전성(100.0%): 성명 매칭 강도(exact_match_len)와 복잡도(exact_match_spaces)는 모든 데이터 쌍에서 100%의 유효값을 확보하였으며, 본 연구는 이를 기반으로 한국어 성명의 정보 밀도 한계를 극복하기 위해 영문 표기 체계를 활용한 특성 공학을 적용하였다.

2) 주제 및 개념 유사성의 활용성(34.6%): 연구 분야 유사도(concepts_shorter_per)는 17,085건(34.6%)의 유효 건수를 기록하며 네트워크 기반 특성 중 가장 높은 가용성을 보였다.

3) 서지 네트워크 특성의 희소성(2.8%~28.5%): 반면, 소속 기관 일치성(inst_per) 28.5%, 공저자 네트워크(coauthors_shorter_per) 11.2%, 상호 인용 관계(citation_work_match) 2.8%로, 직접적인 서지 연결을 활용한 특성들은 상

〈표 2〉 특성별 추출 결과 및 가용성 통계(N = 49,401)

특성	비고	건수(0 이상)
exact_match_len	한글 이름 글자 수로 계산(대부분3)	49,401
exact_match_spaces	한글 한글자마다 의미있다고 판단하여 2로 일괄적용	49,401
inst_per	KISTI 기관 ID 사용(없으면 기관명 한글/영문 사용)	14,061
coauthors_shorter_per	두 논문의 KISTI 인물ID 비교(1저자 제외)	5,540
citation_work_match	두 논문의 참고문헌(논문번호) 비교	1,381
citation_per	두 논문의 참고문헌(논문번호, DOI) 비교	6,296
concepts_shorter_per	저널의 DDC, KISTI 주제, WOS, SCOPUS 주제 비교	17,085

대적으로 높은 희소성을 보였다. 이는 국내 학술 메타데이터의 표준화 및 연계율이 여전히 개선 중인 단계임을 보여준다.

이처럼 한국 학술 데이터에서 관찰되는 소속 기관 및 공저자 정보의 낮은 가용성은 Sanyal et al.(2021)이 저자 식별의 핵심 난제로 지적한 '메타데이터의 희소성' 문제와 궤를 같이하며, 고도화된 식별 모델의 실용화를 위해서는 알고리즘 개선뿐만 아니라 국가적 차원의 표준 메타데이터 인프라 확충이 필수적임을 시사한다.

3.7 저자 식별자 부여 파이프라인의 설계 및 운용

학술 데이터의 폭발적 증가로 인해, 전체 데이터셋을 매번 재학습하는 방식은 자원 효율성 측면에서 한계에 직면하고 있다(Chen et al., 2023). 본 연구는 이러한 요구를 반영하여 OpenAlex 모델을 기반으로 국내 학술 논문의 특수성을 고려한 저자 식별자 부여 파이프라인을 구축하였다. 본 시스템은 신규 유입 데이터만을 선택적으로 처리하는 증분적 처리 방식을 채택하여 시스템의 효율성과 유지보수성을 극대화하였다.

증분적 처리 과정에서는 동일 저자의 기록이 여러 클러스터로 쪼개지는 파편화 문제가 발생할 수 있다(Esperidião et al., 2014). 본 연구의 파이프라인은 이를 최소화하면서도 오분류를 방지하기 위해 0.9 이상의 고신뢰 임계치를 적용하고, 기준 미달 시 신규 식별자를 발급하여 클러스터의 순도(Purity)를 엄격하게 유지하도록 설계하였다.

최종적으로 본 시스템은 Python 3 환경에서 구축되었으며, OCEAN 데이터베이스의 기초 서지 데이터(논문, 저자, 저널, 참고문헌) 4종을 유기적으로 결합하여 분석을 수행한다. 이러한 통합적 접근은 개별 데이터 소스의 한계를 보완하고, 한국 저자 식별의 정밀도를 실무적인 수준까지 끌어올리는 데 기여한다. 파이프라인의 핵심 처리 공정은 다음과 같은 7단계 알고리즘으로 구성된다.

1) 증분 처리 시점 결정: 마지막으로 처리된 논문의 고유번호를 식별하여 신규 반입 데이터의 시작 지점을 자동으로 설정한다.

2) 저자 메타데이터 추출: 대상 논문으로부터 저자 목록 및 관련 서지 정보를 개별 레코드 단위로 추출한다.

3) 성명 정규화: 한글 성명은 원형을 유지하고, 영문 성명은 소문자 변환, 공백 제거를 적용하며 영문 성명 부재 시 로마자 표기법 변환을 통해 매칭 풀을 확장한다.

4) 후보군 검색: 통합 저자 전거 저장소로부터 정규화된 성명이 일치하는 비교 대상 저자군을 조회한다.

5) 특성 생성 및 확률 추론: 추출된 후보군을 대상으로 7개 핵심 특성값을 산출하고, 학습된 OpenAlex 모델을 통해 동일인 여부에 대한 통계적 확률을 도출한다.

6) 식별자 할당 알고리즘: 예측 확률이 0.9(고신뢰도 임계값) 이상인 후보 중 최댓값을 보유한 저자의 기존 식별자를 상속한다. 적절한 매칭 대상이 존재하지 않을 경우 새로운 독립 개체로 판단하여 신규 식별자를 발급한다.

7) 데이터 적재 및 이력 관리: 최종 판별된 저자 식별 정보는 저자 전거 데이터셋에 적재하며, 분석에 사용된 특성 데이터는 별도로 보관하여 모델의 재학습 및 성능 모니터링에 활용한다.

신규 식별자 체계는 'OPALU'와 10자리의 일련번호를 조합한 형태(예: OPALU0000000001)로 정의하였으며, 배치 스케줄링 기반의 주기적 실행을 통해 지속적인 저자 식별자 업데이트가 가능한 구조를 확보하였다.

3.8 평가지표

모델의 판별 성능을 다각도에서 분석하기 위해 정확도, 재현율, 정밀도, F1 점수를 평가지표로 채택하였다. 각 지표는 혼동 행렬의 구성 요소인 TP(True Positive), FP(False Positive), FN(False Negative), TN(True Negative)을 기반으로 산출된다.

4. 실험 및 결과

4.1 실험 환경 및 데이터셋

본 연구의 실험에 활용된 OCEAN 데이터베이스는 논문, 저자, 참고문헌, 저널의 네 가지 논리적 범주로 구성되며, 각 범주별 주요 필드 구성은 〈표 3〉과 같다.

수집된 데이터 범주들은 저자 식별을 위한 특성 추출의 원천 데이터로 활용되었다. 실험 대상 기간은 2023년부터 2024년까지로 설정하였으며, 해당 기간 내에서 총 54,049건의 논문 레코드를 추출하였다. 이후 제1저자 추출 및 동명이인 후보군 생성 절차를 거쳐 최종적으로 49,401건의 저자 쌍으로 구성된 데이터셋을 구축하였다. 전체 데이터셋 중 동일인은 10,910건, 비동일인은 38,491건으로 구성되었다.

4.2 Baseline 성능 평가 및 분석 (v1-1)

OpenAlex 공개 모델을 국내 데이터 환경에 이식한 초기 실험(v1-1)의 결과는 〈표 4〉 및 〈표 5〉와 같다. 해당 실험에서는 한글 성명의 음절 수를 exact_match_len 특성으로 활용하고, exact_match_spaces에는 상숫값 2를 일괄 적용하였다. 또한, 공저자 네트워크 유사도

〈표 3〉 OCEAN 원천 데이터 구성 및 주요 항목

데이터 범주	주요 항목 및 식별자
논문(Article) 데이터	논문번호, 학술지코드, DOI
저자(Author) 데이터	저자순번, 기관ID, KISTI 인물ID, 국/영문성명
참고문헌(Reference) 데이터	참고문헌번호, 인용및피인용논문번호, DOI
저널(Journal) 데이터	학술지코드, DDC, KISTI 주제코드, WoS/Scopus 분류정보

〈표 4〉 v1-1 모델의 혼동 행렬 분석(KISTI 인물 ID 기준)

실제값	총빈도	TP/FP (〉0.5)	FN/TN (〈0.5)	TP/FP (〉0.9)	FN/TN(〈0.9)
Y(동일인)	10,910	10,576 (96.9%)	334 (3.0%)	9,702 (88.9%)	1,208 (11.0%)
N(비동일인)	38,491	11,890 (30.8%)	26,601 (69.1%)	7,607 (19.7%)	30,884 (80.2%)

〈표 5〉 v1-1 모델 성능 지표 요약

평가지표	임계값(Threshold)〉0.5	임계값(Threshold)〉0.9
정확도(Accuracy)	0.830	0.846
재현율(Recall)	0.969	0.889
정밀도(Precision)	0.758	0.818
F1 점수(F1 Score)	0.851	0.852

(coauthors_shorter_per) 산출을 위해 KISTI 인물 ID를 비교 기준으로 채택하였다.

베이스라인 모델(v1-1)의 성능을 분석한 결과, 결정 임계치 0.5 기준으로 0.969의 높은 재현율을 기록하였으나, 정밀도는 0.758에 머물러 비동일인 그룹 중 약 30.8%(11,890건)가 동일인으로 오인되는 '혼합 인용(Mixed Citation)' 문제(Cota et al., 2010)가 확인되었다.

임계치를 0.9로 상향 조정할 경우, 정밀도는 0.818로 개선되는 양상을 보였으나 재현율이 0.889로 하락함에 따라 F1 점수는 0.852를 기록하였다. 이는 임계치 0.5 기반의 F1 점수(0.850)와 비교할 때 성능 향상이 미미한 수준(0.002)임을 시사한다. 이러한 결과는 단순히 결정 임계치를 조정하는 것만으로는 정밀도 향상에 따른 재현율 손실을 극복하기 어려우며, 글로벌 모델이 한국 인명의 특수성을 정교하게 포착하는 데 구조적 한계가 있음을 보여준다.

또한, Seol et al.(2016)의 연구에서 지적된 국내 학술 데이터의 고질적인 메타데이터 희소성(Sparsity) 문제는 모델이 성명 이외의 변별력 있는 단서를 찾는 데 걸림돌이 된다. 베이스라인 모델이 한글 성명의 낮은 정보 밀도를 극복하지 못한 상태에서 소속 기관이나 이메일 등 보완적 특성마저 불충분하게 활용됨에 따라, 임계치 조정만으로는 실무적 요구 수준에 부합하는 정밀도 확보에 한계가 있음이 드러났다. 이는 한국어 성명의 언어적 특성을 역이용한 단계적 특성 최적화와 로마자 표기 체계 도입의 필연성을 강력히 뒷받침한다.

4.3 공저자 네트워크 특성 산출 방식 개선 및 성능 분석 (v1-2)

v1-2 실험에서는 공저자 네트워크 유사도(coauthors_shorter_per) 특성의 산출 방식을 KISTI 인물 ID 기반에서 성명 텍스트 직접 대조 방식으로 전환하여, 미식별 저자에 대한 비교 범위를 확장하였다. 그 결과 〈표 6〉에서와 같이 결정 임계치 0.9 기준으로 F1 점수는

〈표 6〉 v1-2 모델의 성능 평가 지표(공저자 네트워크 특성 개선)

평가지표	임계값(Threshold)〉0.5	임계값(Threshold)〉0.9
정확도(Accuracy)	0.830	0.847
재현율(Recall)	0.970	0.891
정밀도(Precision)	0.758	0.819
F1 점수(F1 Score)	0.851	0.854

0.852에서 0.854로 소폭 향상되었으나, 한국 인명의 중복성으로 인해 텍스트 기반 비교의 성능 향상 폭은 제한적이었다.

이러한 결과는 KISTI 인물 ID 기반 매칭이 갖는 데이터 희소성 문제를 성명 텍스트 비교 방식으로 보완함으로써, 네트워크 기반의 변별력이 한국 데이터 환경에서 제한적이나마 긍정적으로 작용했음을 시사한다. 다만, 전체적인 성능 향상 폭이 크지 않은 점은 한국 인명의 중복성으로 인해 텍스트 기반 비교 역시 동명이인 오판의 가능성을 내포하고 있기 때문으로 분석된다. 아울러, 본 실험의 데이터셋이 2023~2024년의 2개년 논문으로 한정되어 있어, 공저자 간 협력 관계가 장기간에 걸쳐 충분히 축적되지 못한 점도 부분적인 원인으로 작용하였을 가능성이 있다. 공저자 네트워크 특성은 본질적으로 저자 간 반복적 협력이 누적될수록 변별력이 강화되는 특성이므로, 보다 넓은 시계열 범위의 데이터에 적용할 경우 해당 특성의 기여도가 증가할 수 있을 것으로 예상된다.

4.4 성명 매칭 특성 최적화 및 성능 분석 (v2-1)

v2-1 실험에서는 한국 인명의 높은 중복성으로 인한 식별 오류를 최소화하고 한국어 환경의 특수성을 반영하기 위해 성명 관련 두 가지 핵심 특성을 재정의하였다. 우선, 성명 일치 강도(exact_match_len)의 산출 기준을 기존의 한글 음절 수에서 영문 로마자 표기 길이로 변경하였다. 또한, 성명 내 공백 수(exact_match_spaces) 특성 역시 고정된 상수 값(2) 대신 '한글 음절 수 - 1'의 수식을 적용하여 성명의 정보량을 동적으로 반영하도록 수정하였다.

〈표 7〉과 같이 v2-1 실험에서 가장 주목할 만한 성과는 재현율의 현저한 향상이다. 결정 임계치 0.5 기준 재현율은 0.969(v1-1)에서 0.982로 상승하였으며, 0.9 기준으로는 0.889에서 0.919로 대폭 개선되었다.

이러한 성능 향상은 영문 로마자 표기가 한글 성명 대비 다양한 문자열 길이를 지니므로 모델에 더욱 풍부한 식별 정보를 제공하기 때문으로 해석된다. 다만, 정밀도는 0.801로 v1-1(0.818) 대비 소폭 하락하였는데, 이는 재현율 향상 과정에서 발생하는 전형적인 상충 관계 현상으로 판단된다.

4.5 복합 특성 최적화 및 최종 성능 분석 (v2-2)

v2-2 실험에서는 앞선 단계에서 유효성이 입증된 성명 매칭 특성 최적화(v2-1)와 공저자 네트워크 산출 방식 개선(v1-2) 전략을 통합 적용

〈표 7〉 v2-1 모델의 성능 평가 지표(성명 매칭 특성 최적화)

평가지표	임계값(Threshold)〉0.5	임계값(Threshold)〉0.9
정확도(Accuracy)	0.828	0.845
재현율(Recall)	0.982	0.919
정밀도(Precision)	0.751	0.801
F1 점수(F1 Score)	0.851	0.856

하였다. 구체적으로 exact_match_len(영문 성명 길이), exact_match_spaces(한글 음절 수 기반 가중치), 그리고 coauthors_shorter_per(성명 텍스트 직접 대조)의 세 가지 핵심 특성을 동시 개정하여 모델의 식별력을 극대화하였다. 해당 복합 특성 최적화에 따른 결정 임계치별 성능 평가는 〈표 8〉과 같다.

〈표 8〉에서 확인되는 바와 같이, v2-2 모델은 모든 임계치 구간에서 실험군 중 가장 우수한 F1 점수를 기록하며 최적 모델로서의 유효성을 입증하였다. 결정 임계치 0.5 기준으로는 0.983의 압도적인 재현율을 확보하였으며, 임계치를 0.9로 상향 조정할 경우 정밀도가 0.805까지 상승하며 F1 점수 0.860을 달성하였다. 이는 초기 베이스라인 모델(v1-1)의 F1 점수(0.852) 대비 유의미한 향상을 보인 결과이며, Sanyal et al.(2021)이 정의한 쌍 단위 평가 체계 내에서 정밀도와 재현율 사이의 최적의 균형점을 도출한 것으로 평가된다.

이러한 성과는 개별 특성의 정교화가 상호 보완적으로 작용하여 한국 인명의 중의성 문제를 효과적으로 해소했음을 시사한다. 특히 성명 특성 기반의 높은 재현율 위에 공저자 네트워크 정보가 결합됨으로써 v2-1에서 발생했던 정밀도 하락 문제를 성공적으로 상쇄하였다.

4.6 버전별 성능 비교 및 종합 고찰

본 연구에서 제안한 네 가지 모델 버전의 성능을 종합적으로 비교 분석한 결과는 〈표 9〉와 같다. 특히 글로벌 모델이 간과하기 쉬운 한국어 데이터의 구조적 한계를 극복하기 위해 수행된 단계별 최적화의 효용성을 검증하고자 하였으며, 모든 수치는 모델의 엄밀성 확보를 위해 고신뢰 결정 임계치인 0.9를 기준으로 산출하였다.

단계적 실험 결과, 한국 인명 특성 반영(v2-1)으로 재현율이 0.889에서 0.919로 향상되었고, 공저자 네트워크 개선(v1-2)이 정밀도 하락을 보완하여, 최종 복합 모델(v2-2)이 F1 0.860으로 가장 균형 잡힌 성능을 달성하였다.

〈표 8〉 v2-2 모델의 성능 평가 지표(복합 특성 최적화)

평가지표	임계값(Threshold)〉0.5	임계값(Threshold)〉0.9
정확도(Accuracy)	0.837	0.850
재현율(Recall)	0.983	0.922
정밀도(Precision)	0.761	0.805
F1 점수(F1 Score)	0.858	0.860

〈표 9〉 저자 식별 모델 버전별 성능 비교 종합(임계값 0.9 기준)

버전	주요 변경 특성 및 전략	정확도	재현율	정밀도	F1-score
v1-1	Baseline	0.846	0.889	0.818	0.852
v1-2	공저자 네트워크특성 고도화	0.847	0.891	0.819	0.854
v2-1	성명 매칭 특성 최적화	0.845	0.919	0.801	0.856
v2-2	복합최적화(v1-2, v2-1 통합적용)	0.850	0.922	0.805	0.860

각 특성의 기여도를 보다 구체적으로 분석하면, 성명 매칭 특성(exact_match_len, exact_match_spaces)의 재정의가 성능 향상에 가장 큰 영향을 미친 것으로 판단된다. v2-1에서 성명 일치 강도의 산출 기준을 한글 음절 수에서 영문 로마자 표기 길이로 전환하고, 성명 복잡도를 '한글 음절 수 - 1'의 동적 수식으로 변경한 결과, 재현율이 0.889에서 0.919로 3.0%p 향상되었다. 이는 한글 성명의 낮은 정보 밀도를 영문 표기의 다양한 문자열 길이로 보완함으로써 모델의 변별력이 크게 개선되었음을 의미한다. 반면, 공저자 네트워크 유사도(coauthors_shorter_per)의 산출 방식 개선(v1-2)은 F1 기준 0.002의 소폭 향상(0.852 → 0.854)에 그쳤으나, v2-2에서 성명 특성과 결합되었을 때 정밀도가 0.801에서 0.805로 회복되는 보완적 역할을 수행하였다. 인용 네트워크 특성(citation_per, citation_work_match)은 데이터 가용성이 각각 12.7%, 2.8%로 낮아 직접적인 성능 기여도 측정에는 한계가 있었으나, 소속 기관(inst_per) 특성과 함께 모델의 기본 판별 구조를 지탱하는 기저 특성으로 기능하였다. 이러한 분석 결과는 〈표 9〉의 버전별 성능 변화 추이와 일관되며, 한국어 데이터 환경에서는 성명 관련 특성의 최적화가 가장 효과적인 성능 개선 전략임을 정량적으로 뒷받침한다. 이는 글로벌 식별 모델을 특정 국가의 학술 생태계에 이식할 때, 해당 언어 및 데이터 구조의 맥락을 반영한 특성 공학이 필수적임을 시사한다.

4.7 오탐지(False Positive) 분석 및 정답셋 정밀화

최적 모델(v2-2)에서 발생한 오탐지 8,595건을 대상으로, 모델의 오판 원인을 규명하고 정답셋의 완전성을 검증하기 위한 수작업 전수 조사를 실시하였다. 이는 정량적 평가지표가 포착하지 못하는 실제 데이터의 노이즈와 모델의 판별 논리를 심층적으로 파악하여 분석 결과의 신뢰도를 담보하기 위함이다.

이러한 정답셋 보정 과정의 방법론적 엄밀성을 확보하고 연구자의 주관적 개입을 통제하기 위해, 본 연구는 단순 텍스트 매칭을 넘어선 다각적인 교차 검증 프로토콜을 적용하였다. 구체적으로 1) 소속 기관명의 하부 기관(학부/학과 단위) 일치 여부, 2) 세부 연구 주제(Concept)의 의미론적 부합성, 3) 논문 발행 연도의 연속성, 4) 기업·연구소의 하위 부서 등 4단계의 검증 기준(〈표 10〉 참조)을 수립하여 판별의 객관성을 담보하였다.

〈표 10〉 v2-2 오탐지 사례에 대한 수작업 검증 결과

구분	검증 세부 기준	건수	비율
X	실제오탐지: 소속대학명만 일치, 하부기관이 상이함	2,950	34.3%
1	잠재적동일인: 소속기관명(하부기관단위)-연구주제(Concept) 일치	665	7.7%
2	잠재적동일인: 소속기관명(하부기관단위)-논문발행연도일치	1,935	22.5%
3	잠재적동일인: 소속기관명(학부/학과단위) 완전일치	219	2.5%
4	잠재적동일인: 동일기업·연구소 소속 수기 확인	2,826	32.9%
합계		8,595	100%

모델은 동일인으로 예측하였으나 KISTI 인물 ID 기준으로는 비동일인으로 분류된 사례들을 분석한 결과는 〈표 10〉과 같다. 분석 결과, 전체 오탐지 사례 중 34.3% (2,950건)는 실제 비동일인으로 확인되었으나, 나머지 65.7% (5,645건)는 서지 정보상 동일인으로 판단되는 사례였다.

이는 기존 KISTI 인물 ID 기반 정답셋에 미반영된 동일인 레코드가 존재함을 시사하며, 모델이 '데이터 정화(Data Cleaning)' 기능을 수행할 수 있음을 입증한다. 특히, 보정 대상으로 편입된 5,645건의 잠재적 동일인 사례는 모두 위에서 수립한 4단계 검증 기준 중 하나 이상을 충족하는 경우에 한정되며, 판단 근거가 명확하지 않은 사례는 원래의 '비동일인' 레이블을 유지하였다. 이러한 보수적 접근은 정답셋 보정이 모델 성능을 과대 추정하는 방향으로 편향되지 않도록 하기 위한 것이다.

나아가, 보정 결과의 타당성은 독립적인 ORCID 기반 교차 검증(4.8절)을 통해 간접적으로 확인된다. ORCID 정답셋에서 F1 점수 0.892를 기록한 것은, KISTI 인물 ID 기반 보정 전 성능(F1 0.860)보다 높고 보정 후 성능(F1 0.931)보다는 낮은 수준으로, 보정된 정답셋이 모델의 실질적 식별 능력을 보다 정확하게 반영하고 있음을 시사한다. 서지 메타데이터의 높은 정합성을 근거로 식별된 잠재적 동일인 사례를 정답셋에 반영하여, 보정된 정답셋은 동일인 16,555건, 비동일인 32,846건으로 재구성되었으며, 이에 따른 최종 평가 결과는 〈표 11〉과 같다.

이러한 보정 과정을 거친 최종 성능 수치는 모델이 실제 운영 환경에서 발휘할 수 있는 실질적인 식별 정확도를 대변하며, 향후 고도화된 전거 데이터 구축을 위한 피드백 루프로 활용될 수 있다.

〈표 11〉의 결과는 두 가지 시사점을 제공한다. 첫째, 모든 버전에서 정답셋 보정 후 성능이 향상되었으며, 이는 기존 KISTI 인물 ID 기반 정답셋에 내재된 레이블 오류가 전반적으로 모델 성능을 과소 평가하고 있었음을 확인시켜 준다. 둘째, 보정 후에도 v2-2가 가장 높은 F1 점수를 기록함으로써, 특성 최적화의 효과가 정답셋 보정과 독립적으로 유효함을 입증한다. 즉, 성능 향상은 정답셋 보정만의 산물이 아니라, 한국어 특성에 맞춘 특성 재정의가 실질적으로 기여한 결과임을 알 수 있다.

강인수 외(2009)가 제안한 평가 프레임워크를 바탕으로, 정답셋의 품질을 F1 점수 0.931 수준으로 끌어올림으로써, 대규모 학술 정보 인프라에서 데이터 정화 도구로서의 모델 활용 가능성을 입증하였다.

최종 분석 결과, 결정 임계치 0.9 기준 F1 점수는 0.931을 기록하며 보정 전(0.860) 대비 0.071의 비약적인 향상을 나타냈다. 이는 본 연

〈표 11〉 정답셋 보정 후 모델 최종 성능 지표

버전	주요 변경 특성 및 전략	정확도	재현율	정밀도	F1-score
v1-1	Baseline	0.924	0.915	0.933	0.924
v1-2	공저자 네트워크특성 고도화	0.926	0.916	0.934	0.925
v2-1	성명 매칭 특성 최적화	0.923	0.944	0.906	0.925
v2-2	복합최적화(v1-2, v2-1 통합적용)	0.930	0.949	0.914	0.931

구에서 최적화한 OpenAlex 모델이 기존 인물 식별 체계가 포착하지 못한 저자 간의 동일성을 정교하게 식별해낼 수 있음을 입증하는 결과이다. 특히, 정답셋 보정 후의 성능 지표는 모델이 실제 데이터의 노이즈를 극복하고 고도로 일관된 판별 기준을 유지하고 있음을 보여준다. 결과적으로 본 모델은 한국 학술 데이터의 특수성을 효과적으로 반영하고 있을 뿐만 아니라, 방대한 서지 데이터로부터 잠재적인 동일 저자 레코드를 자동 발견하여 데이터의 무결성을 개선하는 데 기여한다. 이는 실무적인 저자 전거 구축 및 관리 공정에서 인간의 개입을 최소화하고 식별 정확도를 극대화할 수 있는 보조적 도구로서 매우 높은 활용 가치를 지님을 시사한다.

4.8 ORCID 기반 교차 검증 및 모델 신뢰성 분석

KISTI 인물 ID 기반 정답셋의 국한성을 탈피하고 모델의 독립적 타당성을 검증하기 위해, ORCID를 정답셋으로 활용하여 OpenAlex 모델의 성능을 재평가하였다. ORCID 기반 데이터셋은 동일인 15,497건, 비동일인 13,950건으로 구축되었으며, 실험에는 가장 우수한 성능을 보였던 v2-2의 특성 설정을 동일하게 적용하였다.

〈표 12〉에 제시된 바와 같이, ORCID 기반 검증 결과 OpenAlex 모델은 결정 임계치 0.9 기준 정확도 0.891, 재현율 0.894, 정밀도 0.889, F1 점수 0.892를 달성하였다. 이는 보정 전 KISTI 인물 ID 기반 평가 결과(F1 점수 0.860)를 상회하는 수치로, 국제적으로 신뢰할 수 있는 외부 표준 데이터셋에서도 본 모델이 높은 안정성과 식별력을 유지함을 입증한다.

특히 주목할 점은 재현율(0.894)과 정밀도(0.889)의 균형 잡힌 분포이다. 이는 모델이 특정 클래스에 편향되지 않고 동일인 식별과 비동일인 판별 모두에서 고른 판별 능력을 갖추었음을 시사한다. 결과적으로 본 연구를 통해 최적화된 OpenAlex 모델은 국내외 다양한 식별 체계와 결합하여 고신뢰도의 저자 전거 데이터를 구축할 수 있는 실무적 타당성을 확보하였다고 판단된다.

4.9 국내 학술 논문 대상 식별자 부여 결과 및 시스템 성능 분석

설계된 저자 식별 파이프라인의 실무 적용 가능성을 검토하기 위해, OCEAN 국내 논문 53,856건을 대상으로 식별자 부여 실험을 수행하였다. 본 실험은 구축된 모델이 실제 운영 환경의 비정형 메타데이터를 얼마나 유연하게 처리하는지 검증하는 데 중점을 두었으며, 최종 식별 결과 및 시스템 자원 활용 현황은 〈표 13〉과 같다.

〈표 12〉 ORCID 정답셋 기준 OpenAlex 모델(v2-2) 성능 지표

평가지표	임계값(Threshold)〉0.5	임계값(Threshold)〉0.9
정확도(Accuracy)	0.870	0.891
재현율(Recall)	0.974	0.894
정밀도(Precision)	0.806	0.889
F1 점수(F1 Score)	0.882	0.892

〈표 13〉 국내 논문 데이터 대상 식별자 부여 실험 결과 및 시스템 사양

항목	내용
테스트 환경	노트북(I7-1260, 16 CPUs, 32GB RAM)
대상 데이터	논문 53,856건, 저자 183,105건, 저널 2,416건, 참고문헌 860,405건
처리 시간	약 30시간
대상 저자수	183,105명
부여된 고유 식별자수	109,205개(OPENALEX_PSON_ID)

실험 결과, 총 183,105건의 저자 레코드가 109,205개의 고유 식별자로 그룹화되었음을 확인하였다. 이는 단일 고유 식별자당 평균 약 1.68개의 저자 레코드가 매핑된 결과로, 분산된 서지 데이터가 개체 식별 과정을 통해 체계적인 전거 데이터로 통합되었음을 시사한다.

시스템 자원 활용 측면에서, 약 18만 건의 저자 데이터를 처리하는 데 총 30시간이 소요되었으며 평균 CPU 점유율은 50~60% 수준을 유지하였다. 메모리 가용성 역시 안정적인 수치를 기록하여 중급 사양의 컴퓨팅 환경에서도 대규모 데이터 처리가 가능함을 입증하였다.

다만, 국내 논문 전체 데이터(약 200만 건)로 범위를 확장할 경우 예상 소요 시간은 약 50일로 산출되어, 향후 블로킹 기법 도입 및 알고리즘 최적화를 통한 처리 속도 향상이 요구된다.

5. 논 의

5.1 연구 질문에 대한 답변 및 고찰

- RQ 1. OpenAlex 모델의 한국 학술 데이터 적용 유효성

본 연구를 통해 OpenAlex 저자 식별 모델이 한국어 저자명과 국내 학술 데이터 환경에서도 우수한 판별 성능을 발휘함을 입증하였다. 초기 베이스라인 모델(v1-1)은 별도의 튜닝 없이도 결정 임계치 0.9 기준 F1 점수 0.852를 기록하였으며, 한국어 특성을 반영한 단계적 최적화를 거친 최종 모델(v2-2)은 F1 점수 0.860을 달성하였다. 나아가 정답셋 보정 후 v2-2 모델은 F1 점수 0.931을 기록하여, 특성 최적화와 정답셋 정밀화의 결합 효과를 확인하였다. 이는 Zhao와 Chen(2025)이 보고한 중국 학자 데이터 대비 건고한 수치이자, Seol et al.(2016)이 국내 IT 분야에서 SVM으로 달성한 94.79%의 F1 지표에 근접한 결과이다. 이는 적절한 특성 공학이 뒷받침될 경우, OpenAlex와 같은 글로벌 범용 모델이 특정 국가의 언어적 장벽을 넘어 도메인 특화 모델에 대등한 식별력을 발휘할 수 있음을 시사한다.

- RQ 2. 저자 식별 특성의 기여도 및 영향력 분석

한국 학술 데이터의 특성을 반영한 특성 재정의가 모델 성능 향상에 결정적인 역할을 수행함을 확인하였다. 특히 성명 일치 강도(exact_match_len)와 복잡도(exact_match_spaces)를 로마자 표기 체계로 전환한 전략이 재현율의 비약적 향상(0.889 → 0.919)을 견인하였다. 이는 고정된 음절 수로 인해 정보 밀도가 낮은

한글 성명과 달리, 다양한 문자열 길이를 지닌 영문 표기가 모델에 더욱 풍부한 식별 정보를 제공하기 때문이다.

이러한 접근은 Cota et al.(2010)이 지적한 바와 같이 아시아계 저자명에서 빈번하게 발생하는 혼합 인용 문제를 해소하기 위한 실질적인 해법이 될 수 있다. 로마자 표기 변이를 단순한 데이터 오류가 아닌 정보 가용성을 높이는 매개체로 활용한 본 연구의 방향성이 유효했음을 시사한다. 한편, 공저자 네트워크(coauthors_shorter_per) 개선이 정밀도 향상에 기여한 점은, 공저자 정보가 저자 식별에서 가장 직관적이고 효과적인 특징이라고 정의한 Seol et al.(2016)의 견해와 일치한다. 선행 연구에 따르면 공저자 네트워크의 확장은 부족한 서지 정보를 극복하고 식별 성능을 개선하는 핵심 기제로 작용한다. 소속 기관(inst_per) 특성은 데이터 가용성(28.5%)의 한계로 인해 향후 데이터 정제 노력을 통한 보완의 여지가 남아 있으나, 결론적으로 본 연구는 선행 연구에서 강조한 기초 서지 속성(저자명, 공저자, 제목 등)의 체계적 결합이 별도의 학습 데이터 없이도 높은 식별력을 확보할 수 있음을 뒷받침한다.

- RQ 3. 증분형 저자 식별 파이프라인의 실효성

본 연구에서 제안한 증분적 처리 방식은 183,105건의 저자 레코드를 109,205개의 고유 식별자로 성공적으로 통합하며 시스템의 실무적 효용성을 증명하였다. 특히 0.9 이상의 고신뢰 임계치를 적용하여 클러스터의 순도를 엄격하게 유지하면서도 대규모 국내 학술 데이터를 효율적으로 처리할 수 있는 구조를 확보하였다. 이러한 증분적 접근은 자원 소모가 큰 전체 데이터 재학습의 대안으로서, 신규 논문이 수시로 유입되는 실제 학술 데이터 서비스 환경에 즉시 적용 가능한 실용적인 아키텍처임을 시사한다. 이는 향후 국가적 차원의 저자 전거 데이터 자동화 구축 및 지속 가능한 데이터 관리에 있어 핵심적인 기술적 토대가 될 것으로 기대된다.

5.2 연구의 한계점 및 학술적 시사점

학술 데이터의 방대함과 복잡성을 고려할 때, 자동화된 알고리즘만으로 완벽한 정답셋을 구축하는 데는 근본적인 한계가 존재한다. 대규모 학술 데이터베이스인 Aminer를 구축한 Zhang et al.(2018)은 전역적 및 지역적 임베딩을 결합한 학습 모델을 제시함과 동시에, 인간의 피드백을 식별 과정에 실시간으로 통합하는 '인간 참여형(Human in the loop)' 접근 방식이 식별 정확도 향상의 핵심임을 강조한 바 있다. 본 연구 역시 이러한 방법론적 흐름에 궤를 같이하여, OpenAlex 모델의 예측 결과에 대한 정밀한 수작업 검증을 병행하였다. 이를 통해 정답셋을 보정함으로써 식별 성능을 최적화(F1 점수 0.860 → 0.931)하였으며, 이는 자동화 모델의 구조적 한계를 보완하고 데이터의 신뢰성을 확보하기 위한 필수적인 공정임을 확인하였다. 이러한 성과에도 불구하고, 향후 시스템 고도화를 위해 고려해야 할 한계점과 학술적 시사점은 다음과 같다.

첫째, 정답셋의 완전성 및 신뢰도 확보 문제이다. 수작업 검증 과정에서 KISTI 인물 ID 기반 정답셋 중 5,645건의 보완 사례가 발견된 것

은, 대규모 서지 데이터 환경에서 무결한 정답셋 구축의 본질적 어려움과 국내 학술 정보 통합 전거 체계 마련의 시급성을 시사한다. 따라서 국내 학술 생태계에서도 연구자 식별자 등록을 독려하고 확충하는 노력이 병행되어야 한다(Zhang et al., 2023).

둘째, 시스템의 확장성 및 연산 효율성 문제이다. 현재 파이프라인은 5만 건 처리 시 약 30시간, 200만 건 규모 처리 시 약 50일이 소요되는 구조적 한계를 지닌다. 특히 전거 테이블 확장에 따른 비교 후보군의 기하급수적 증가가 병목 현상을 유발하므로, Kim et al.(2020)이 제안한 CNF 블로킹 기법 등을 파이프라인에 통합하여 연산 비용을 절감할 필요가 있다.

셋째, 원천 데이터의 희소성 및 메타데이터의 불완전성이다. 소속 기관(28.5%), 공저자(11.2%), 상호 인용(2.8%) 정보의 낮은 가용성은 모델의 잠재적 식별 능력을 제한하는 '정보 병목' 현상을 초래한다. 따라서 고도화된 알고리즘 도입과 함께, 기관 식별자 부여와 참고문헌 서지 데이터 정교화 등 서지 정보 인프라의 양적·질적 확충이 병행되어야 한다.

6. 결론 및 향후 과제

본 연구는 글로벌 저자 식별 모델인 OpenAlex를 국내 학술 데이터 환경에 체계적으로 이식하고, 한국어 및 국내 서지 구조의 특수성을 반영한 특성 공학(Feature Engineering) 최적화 전략을 제안하였다. OCEAN 데이터베이스의 2023~2024년 발행 국내 논문 54,049건을 기반으로 구축된 49,401건의 테스트 데이터 쌍을 통해 단계적 실험을 수행한 결과, 다음과 같은 핵심적 학술 성과를 도출하였다.

첫째, OpenAlex 모델의 국내 학술 데이터 적용 타당성을 입증하였다. 베이스라인 모델(v1-1)은 별도의 튜닝 없이도 결정 임계치 0.9 기준 F1 점수 0.852를 기록하여, 국내 학술 생태계에서도 실무적 활용이 가능한 수준의 기초 성능을 확보하고 있음을 확인하였다. 이는 글로벌 범용 모델이 고유의 전처리 과정만으로도 한국어 서지 환경에 유연하게 대응할 수 있음을 보여주는 지표이다.

둘째, 한국 인명의 언어적 특성을 고려한 특성 최적화의 유효성을 검증하였다. 성명 매칭 강도 및 복잡도 산출 방식을 로마자 표기 체계로 재정의하고 공저자 네트워크 비교 로직을 고도화함으로써, F1 점수를 0.860으로 향상시켰다. 특히 한글 음절에 비해 정보 밀도가 높은 영문 성명의 정보량을 활용한 전략은 재현율의 비약적인 상승을 견인하는 핵심 요인임을 확인하였다. 이는 아시아계 인명의 중복성 문제를 해결하기 위해 성명 표기 변이를 정보원으로 역이용한 본 연구의 접근법이 타당했음을 시사한다.

셋째, 오탐지 분석을 통한 정답셋 품질 개선 및 모델의 잠재력을 식별하였다. 수작업 검증 결과 기존 KISTI 인물 ID 기반 정답셋에서 5,645건의 미반영 사례를 발견하였으며, 이를 보정한 정답셋으로 재평가한 결과 F1 점수 0.931이라는 우수한 성능을 달성하였다. 이는 제안 모델이 단순한 판별기를 넘어, 기존 식별 체계의 누락과 오류를 능동적으로 탐지하여 데이터 무결성을 제고하는 '데이터 정화(Data Cleaning)' 도구로서의 가치를 지님을 증명한다.

넷째, 독립적 교차 검증을 통한 모델의 범용적 신뢰성을 확보하였다. 국제 표준 식별자인 ORCID를 독립적 정답셋으로 활용한 실험에서 F1 점수 0.892를 기록함으로써, 다양한 식별 기준 하에서도 모델의 판별 성능이 안정적으로 유지됨을 입증하였다. 이는 특정 기관의 데이터 편향성에서 벗어나 국제적 기준에서도 본 연구의 방법론이 유효하게 작동함을 의미한다.

다섯째, 대규모 데이터 처리를 위한 증분형 파이프라인의 실효성을 확인하였다. 제안된 시스템은 약 18만 명의 저자 레코드를 10만여 개의 고유 식별자로 성공적으로 통합하며 대규모 서지 데이터에 대한 확장성을 증명하였다. 특히 신규 데이터만을 처리하는 증분적 방식은 시스템 유지보수 비용을 획기적으로 절감할 수 있는 실질적인 아키텍처임을 확인하였다.

향후 연구에서는 다음과 같은 고도화 과제를 추진할 계획이다. 먼저, 시스템 확장성 극대화를 위해 수백만 건 이상의 대규모 연산 시 발생하는 병목 현상을 해결하고자 CNF 블로킹 기법을 파이프라인에 통합할 예정이다. 이를 통해 불필요한 비교 쌍을 사전에 제거함으로써 연산 효율성을 획기적으로 개선할 것이다. 또한, 소속 기관 및 참고문헌 데이터의 낮은 가용성을 극복하기 위해 기관 식별자 자동 보정 알고리즘과 텍스트 마이닝 기반의 메타데이터 확충 연구를 병행할 필요가 있다.

궁극적으로 국내 학술 커뮤니티의 ORCID 확대와 표준 메타데이터의 정교화는 본 연구의 성과를 극대화하는 동시에 국가적 차원의 저자 식별 인프라를 강화하는 근본적인 해결책이 될 것이다. 특히 Fernández-Marcial et al.(2023)이 강조한 바와 같이, 고유 식별자의 확산은 외부 시스템과의 상호운용성을 확보하여 데이터 통합의 객관성을 담보하는 핵심 기제가 되므로 국내 학술 생태계 내 ORCID 도입을 가속화해야 한다.

나아가 개별 연구자 식별을 넘어선 데이터 융합 관점의 접근도 요구된다. ISNI 기반의 저자 식별 체계 운용이 이종 데이터 간의 융합 효율성을 높이는 핵심 기제이며, 식별 데이터의 확산을 위해서는 국립중앙도서관을 중심으로 한 분야별 기관 간의 긴밀한 컨소시엄 협력이 필수적이다. 각 기관의 메타데이터 및 링크드 데이터 공유를 위한 기술적 요소들은, 본 연구의 식별 파이프라인이 향후 타 시스템과 상호운용성을 확보하고 고품질의 전거 데이터를 구축하는 데 있어 구체적인 기술적 가이드라인이 될 수 있다.

결과적으로 본 연구에서 제안한 모델과 파이프라인이 이러한 고신뢰 정답셋의 지속적인 확충 및 주기적 업데이트 체계와 유기적으로 결합된다면, 한국 학술 데이터의 국제적 가시성과 정보 연계성은 비약적으로 향상될 것으로 기대된다.

참 고 문 헌

강인수, 김평, 이승우, 정한민, 류범종 (2009). 저자 식별을 위한 대용량 평가셋 구축. 한국콘텐츠학회 논문지, 9(11), 455-464. https://doi.org/10.5392/JKCA.2009.9.11.455

강인수, 이승우, 정한민, 김평, 구희관, 이미경, 성원경, 박동인 (2008). 저자 식별을 위한 자질 비교. 한국콘텐츠학회 논문지, 8(2), 41-47. https://doi.org/10.5392/JKCA.2008.8.2.041

박진호, 곽승진, 이승민, 오상희 (2020). ISNI Korea 컨소시엄의 저작권 권리 단체 데이터 공동 활용을 위한 기술요소 도출 연구. 한국비블리아학회지, 31(1), 379-392. https://doi.org/10.14699/kbiblia.2020.31.1.379

오상희, 곽승진, 이승민, 박진호 (2019). 국내 분야별 인명정보 관리를 위한 저자식별체계인 ISNI 활용에 관한 연구 국립중앙도서관의 ISNI-Korea 컨소시엄 참여기관과 비참여기관을 대상으로 한 집단면담 연구방법 이용. 한국도서관·정보학회지, 50(2), 121-147. https://doi.org/10.16981/kliss.50.2.201906.121

이승민, 곽승진, 오상희, 박진호 (2019). ISNI 기반 데이터 융합을 위한 저자식별체계 운용에 관한 연구. 한국비블리아학회지, 30(1), 29-51. https://doi.org/10.14699/kbiblia.2019.30.1.029

Chen, B., Zhang, J., Zhang, F., Han, T., Cheng, Y., Li, X., Dong, Y., & Tang, J. (2023). Web-scale academic name disambiguation: The WhoIsWho benchmark, leaderboard, and toolkit. Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 3817-3828. https://doi.org/10.1145/3580305.3599930

Cota, R. G., Ferreira, A. A., Nascimento, C., Gonçalves, M. A., & Laender, A. H. F. (2010). An unsupervised heuristic-based hierarchical method for name disambiguation in bibliographic citations. Journal of the American Society for Information Science and Technology, 61(9), 1853-1870. https://doi.org/10.1002/asi.21363

Esperidião, L. V. B., Ferreira, A. A., Laender, A. H. F., Gonçalves, M. A., Gomes, D. M., Tavares, A. I., & de Assis, G. T. (2014). Reducing fragmentation in incremental author name disambiguation. Journal of Information and Data Management, 5(3), 293-307.

Fan, X., Wang, J., Pu, X., Zhou, L., & Lv, B. (2011). On graph-based name disambiguation. Journal of Data and Information Quality, 2(2), Article 10. https://doi.org/10.1145/1891879.1891883

Fang, Z., Zhuo, Y., Xu, J., Tang, Z., Jia, Z., & Zhang, H. (2023). Automatic author name disambiguation by differentiable feature selection. Journal of Information Science, 0(0). https://doi.org/10.1177/01655515231193859

Fernández-Marcial, V., González-Solar, L., & Vale, A. (2023). Is ORCID your ID? a case study at the faculty of arts and humanities of the university of Porto. Learned Publishing, 36(4), 564-576. https://doi.org/10.1002/leap.1562

Ferreira, A. A. & Laender, A. H. F. (2023). Automatic disambiguation of author names: Foundations, methods and open issues. Companion Proceedings of the 38th Brazilian

Symposium on Databases (SBBD 2023), 179-182. https://doi.org/10.5753/sbbd_estendido.2023.25633

Kim, K., Sefid, A., & Giles, C. L. (2020). Learning CNF blocking for large-scale author name disambiguation. Proceedings of the First Workshop on Scholarly Document Processing, 72-80. https://doi.org/10.18653/v1/2020.sdp-1.8

Kim, S. (2018). Disambiguation of Korean names in references. Journal of Information Science Theory and Practice, 6(2), 62-70. https://doi.org/10.1633/JISTaP.2018.6.2.5

Levin, M., Krawczyk, S., Bethard, S., & Jurafsky, D. (2012). Citation-based bootstrapping for large-scale author disambiguation. Journal of the American Society for Information Science and Technology, 63(5), 1030-1047. https://doi.org/10.1002/asi.22621

Louppe, G., Al-Natsheh, H. T., Susik, M., & Maguire, E. J. (2016). Ethnicity sensitive author disambiguation using semi-supervised learning. Knowledge Engineering and Semantic Web, 649, 272-287. https://doi.org/10.1007/978-3-319-45880-9_21

Müller, M. C. (2018). On the contribution of word-level semantics to practical author name disambiguation. Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, 367-368. https://doi.org/10.1145/3197026.3203912

OpenAlex (2023). Author disambiguation. OpenAlex Help Center. Available: https://help.openalex.org/hc/en-us/articles/24347048891543-Author-disambiguation

Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. 26th International Conference on Science and Technology Indicators (STI 2022), Granada, Spain.

Sanyal, D. K., Bhowmick, P. K., & Das, P. P. (2021). A review of author name disambiguation techniques for the PubMed bibliographic database. Journal of Information Science, 47(2), 227-254. https://doi.org/10.1177/0165551519888605

Seol, J. W., Lee, S. H., & Kim, K. Y. (2016). Author disambiguation using co-author network and supervised learning approach in scholarly data. International Journal of Software Engineering and Its Applications, 10(4), 73-82. https://doi.org/10.14257/ijseia.2016.10.4.08

Subramanian, S., King, D., Downey, D., & Feldman, S. (2021). S2AND: A benchmark and evaluation system for author name disambiguation. 2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL), 170-179. https://doi.org/10.1109/JCDL52503.2021.00029

Tang, J., Fong, A. C. M., Wang, B., & Zhang, J. (2012). A unified probabilistic framework for name disambiguation in digital library. IEEE Transactions on Knowledge and Data

Engineering, 24(6), 975-987. https://doi.org/10.1109/TKDE.2011.13

Treeratpituk, P. & Giles, C. L. (2012). Name-ethnicity classification and ethnicity-sensitive name matching. Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, 1141-1147.

Vishnyakova, D., Rodriguez-Esteban, R., Ozol, K., & Rinaldi, F. (2016). Author name disambiguation in MEDLINE based on journal descriptors and semantic types. Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016), 134-142.

Xu, X., Li, Y., Liptrott, M., & Bessis, N. (2018). NDFMF: An author name disambiguation algorithm based on the fusion of multiple features. 2018 IEEE 42nd International Computer Software and Applications Conference (COMPSAC), 2, 187-190. https://doi.org/10.1109/COMPSAC.2018.10226

Zhai, X., Han, H., Li, Z., & Ran, Y. (2019). Research on author name disambiguation based on fusion features and semantic fingerprints. Journal of Physics: Conference Series, 1302, 022013. https://doi.org/10.1088/1742-6596/1302/2/022013

Zhang, L., Lu, W., & Yang, J. (2023). LAGOS-AND: A large gold standard dataset for scholarly author name disambiguation. Journal of the Association for Information Science and Technology, 74(2), 168-185. https://doi.org/10.1002/asi.24720

Zhang, Y., Zhang, F., Yao, P., & Tang, J. (2018). Name disambiguation in AMiner: Clustering, maintenance, and human in the loop. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 1002-1011. https://doi.org/10.1145/3219819.3219859

Zhao, R. & Chen, Y. (2025). Accuracy assessment of OpenAlex and Clarivate Scholar ID with an LLM-assisted benchmark. arXiv. https://doi.org/10.48550/arXiv.2502.11610

Zhou, H. & Sun, M. (2024). Evaluating authorship disambiguation quality through anomaly analysis on researchers' career transition. arXiv. https://doi.org/10.48550/arXiv.2412.18757

• 국문 참고자료의 영어 표기
(English translation / romanization of references originally written in Korean)

Kang, In-Su, Kim, Pyung, Lee, Seungwoo, Jung, Hanmin, & You, Beom-Jong (2009). A large-scale test set for author disambiguation. Journal of the Korea Contents Association,

9(11), 455-464. https://doi.org/10.5392/JKCA.2009.9.11.455

Kang, In-Su, Lee, Seungwoo, Jung, Hanmin, Kim, Pyung, Koo, Heekwan, Lee, Mi-Kyung, Sung, Won-Kyung, & Park, Dong-In (2008). Features for author disambiguation. Journal of the Korea Contents Association. 8(2), 41-47. https://doi.org/10.5392/JKCA.2008.8.2.041

Lee, Seungmin, Kwak, Seung-Jin, Oh, Sanghee, & Park, Jin Ho (2019). A study on the management of name identifier system for ISNI-based data integration. Journal of the Korean Biblia Society for Library and Information Science, 30(1), 29-51. https://doi.org/10.14699/kbiblia.2019.30.1.029

Oh, Sanghee, Kwak, Seung-Jin, Lee, Seungmin, & Park, Jinho (2019). A study on the application of ISNI for the personnel information management: Having focused group interviews with participants and non-participants in the ISNI-Korea Consortium managed by National Library of Korea. Journal of Korean Library and Information Science Society, 50(2), 121-147. https://doi.org/10.16981/kliss.50.2.201906.121

Park, Jin Ho, Kwak, Seung Jin, Lee, Seungmin, & Oh, Sang Hee (2020). A study on derivation of technical elements for joint use of copyright rights group data by ISNI Korea Consortium. Journal of the Korean Biblia Society for Library and Information Science, 31(1), 379-392. https://doi.org/10.14699/kbiblia.2020.31.1.379