BERT 모형을 이용한 주제명 자동 분류 연구*

A Study on Automatic Classification of Subject Headings Using BERT Model

이 용 구 (Yong-Gu Lee)**

목 차

초 록

이 연구는 딥러닝 기법의 전이학습 모형인 BERT를 이용하여 주제명의 자동 분류를 실험하고 그 성능을 평가하였으며, 더 나아가 주제명이 부여된 KDC 분류체계와 주제명의 범주 유형에 따른 성능을 분석하였다. 실험 데이터는 국가서지를 이용하여 주제명의 부여 횟수에 따라 6개의 데이터셋을 구축하고 분류 자질로 서명을 이용하였다. 그 결과, 분류 성능으로 3,506개의 주제명이 포함된 데이터셋(레코드 1,539,076건)에서 마이크로 F1과 매크로 F1 척도가 각각 0.6059와 0.5626 값을 보였다. 또한 KDC 분류체계에 따른 분류 성능은 총류, 자연과학, 기술과학, 그리고 언어 분야에서 좋은 성능을 보이며 종교와 예술 분야는 낮은 성능을 보였다. 주제명의 범주 유형에 따른 성능은 '식물', '법률명', '상품명'이 높은 성능을 보인 반면, '국보/보물' 유형의 주제명에서 낮은 성능을 보였다. 다수의 주제명을 포함하는 데이터셋으로 갈수록 분류기가 주제명을 제대로 부여하지 못하는 비율이 늘어나 최종 성능의 하락을 가져오기 때문에, 저빈도 주제명에 대한 분류 성능을 높이기 위한 개선방안이 필요하다.

ABSTRACT

This study experimented with automatic classification of subject headings using BERT-based transfer learning model, and analyzed its performance. This study analyzed the classification performance according to the main class of KDC classification and the category type of subject headings. Six datasets were constructed from Korean national bibliographies based on the frequency of the assignments of subject headings, and titles were used as classification features. As a result, classification performance showed values of 0.6059 and 0.5626 on the micro F1 and macro F1 score, respectively, in the dataset (1,539,076 records) containing 3,506 subject headings. In addition, classification performance by the main class of KDC classification showed good performance in the class General works, Natural science, Technology and Language, and low performance in Religion and Arts. As for the performance by the category type of the subject headings, the categories of plant, legal name and product name showed high performance, whereas national treasure/treasure category showed low performance. In a large dataset, the ratio of subject headings that cannot be assigned increases, resulting in a decrease in final performance, and improvement is needed to increase classification performance for low-frequency subject headings.

키워드: 자동 분류, 딥러닝, BERT 모형, 주제명 자동 부여, 자동 주제 분류
Automatic Classification, Deep Learning, BERT Model, Automated Subject Indexing, Automated Subject Classification

* 본 연구는 2022년도 국립중앙도서관이 지원한 연구과제 "국가서지를 활용한 주제명 자동 분류 적용방안 연구"의 일부 내용을 수정・보완하였음.

** 경북대학교 문헌정보학과 부교수(yglee@knu.ac.kr / ISNI 0000 0004 6437 6752)

논문접수일자: 2023년 4월 21일 최초심사일자: 2023년 4월 30일 게재확정일자: 2023년 5월 19일
한국문헌정보학회지, 57(2): 435-452, 2023. http://dx.doi.org/10.4275/KSLIS.2023.57.2.435

1. 서 론

최근 많은 분야에서 딥러닝 기법을 적용하려는 노력이 계속되고 있다. 이는 딥러닝이 최신 기법이기 때문이기도 하지만 다른 기계학습 기법에 비해 매우 탁월한 성능을 보이기 때문이다. 특히 기계가 인간처럼 이해하여 처리하기 어렵다고 간주했던 이미지와 동영상 형식의 데이터 처리 분야인 비전(vision) 영역에서 딥러닝 기법이 인간의 능력을 넘어서는 성능을 보이고 있다. 이러한 현상은 자연 언어 처리와 같은 텍스트 영역에서도 비슷한 실정이다.

도서관 분야에서도 인공지능과 딥러닝 기법을 적용하고자 하는 사례들이 해외의 국가대표 도서관을 중심으로 보고되고 있다. 최근 미국이나 일본, 네덜란드, 독일 등의 주요 국가대표도서관을 중심으로 문헌 분류 혹은 주제어 부여 업무를 자동화 하기 위해 인공지능과 기계학습 기술을 적용한 다양한 과업들이 추진 중이다. 이러한 시류의 선상에서 우리나라 국립중앙도서관도 인공지능 기술을 활용하여 사서의 업무를 지원하기 위한 도구 개발과 시범 적용을 위한 가능성을 검토하여 장기적으로 도서관에 인공지능 기술의 도입에 대한 방안을 연구하였다(오원석, 2021). 추가적으로 국립중앙도서관은 인공지능 기반 요약 및 검색 서비스를 시범적으로 개시하였다(국립중앙도서관, 2021).

한편, 국립중앙도서관은 2002년 주제명표목표를 개발하고 2013년 「국립중앙도서관 주제명표목표 개방을 위한 고품질화 연구」를 거쳐 현재 외부에 개방하고 있으며, 2003년부터 국립도서관 장서에 대해 주제명을 부여하여 국가서지 데이터를 구축하고 있다(최윤경, 정연경, 2014; 백지원, 정연경, 2014). 이는 도서관이나 검색 시스템이 이용자의 정보 요구에 대한 탐색을 원활하게 지원하기 위한 노력의 일환으로 볼 수 있다.

사서의 전문성을 기반으로 하는 주제명 부여 과정은 개인의 경험과 역량에 따라 그 정확성과 일관성의 편차가 존재하는 것을 알 수 있다. 서로 다른 기관의 사서 또는 색인자가 동일한 색인 도구(주제명표목표나 시소러스)를 이용하여 같은 문헌에 부여한 색인어가 얼마나 일치하는지를 조사한 연구(Tonta, 1991; Reich & Biever, 1991)에서 낮은 수준의 일관성을 보였으며, 이는 도서관의 색인 전문가와 일반 이용자 사이에서도 유사한 결과를 보였다(Saarti, 2002). 이러한 편차는 다양한 요인에서 기인하는데, 이를 해결하기 위한 기본적인 방법 중 하나로 교육을 통해 색인자의 숙련 수준을 높이는 것이다. 다른 방법은 기계적 처리 방법을 함께 활용하여 색인 업무의 효율성과 용어 선정 등의 정확성을 높이는 것이다. 인공지능이나 딥러닝의 성능 향상이 나날이 발전함에 따라 이러한 문제를 해결하기 위한 방안을 개발할 필요가 있다.

이에 본 연구는 국립중앙도서관이 수행하고 있는 주제명 부여 업무에 대해 딥러닝 기법을 활용하여 자동 분류 또는 자동 부여의 적용 가능성을 제시하고자 하였다. 이를 위해 국립중앙도서관의 서지데이터가 지닌 서명을 기본적인 분류 자질로 이용하여 최근 딥러닝 분야에서 개발된 전이학습(transfer learning) 모형인 BERT(Bidirectional Encoder Representations from Transformers)를 적용하여 주제명 자동 분류 알고리즘을 구현하였다. 더 나아가 기계

학습에서 이루어지는 분류 모형의 성능 평가뿐만 아니라 주제명이 부여된 KDC 분류체계에 따른 평가, 그리고 주제명의 범주 유형에 따른 평가 등 다각도로 자동 분류 결과를 평가하고 검증하여 개선 방안을 도출하고자 하였다. 특히 서명은 메타데이터 요소 중에 거의 모든 서지 레코드에 존재하여 도서관 환경에서는 중요하고 기본적인 분류 자질이기에 이를 활용하여 주제명 분류 성능을 알아보는 것은 큰 의미가 있다고 할 수 있다.

2. 이론적 배경

2.1 딥러닝과 전이학습 BERT

인공지능의 딥러닝 기법은 비전 분야에서 획기적인 성과를 가져오며 연구뿐만 아니라 실무 영역에서까지 화두가 되고 있다. 대표적으로 이미지 처리에 탁월한 성능을 보이는 합성곱 신경망(CNN) 방법(LeCun et al., 1998)을 필두로 하여 새롭고 혁신적인 다른 모형들이 등장하였다. 이 방법은 또한 분야를 달리하여 텍스트 처리에서도 다양한 시도를 통해 상당한 성과를 가져왔다(Kim, 2014).

텍스트나 자연언어 처리에서는 비전 영역의 이미지와 달리 출현 단어의 순서 또는 순차(sequential)적인 특성을 반영하는 딥러닝 모형이 필요하다. 이러한 배경에서 출현한 모형으로 순환 신경망(RNN)이 있으며(Mikolov et al., 2011), 이를 개선한 다양한 모형으로 LSTM (Hochreiter & Schmidhuber, 1997)과 GRU (Cho et al., 2014) 등이 등장하였다. 또한 텍스트 처리 분야에서 딥러닝의 지속적인 변화 발전에 따라 더 좋은 성능의 트랜스포머(transformer) 모형이 등장하였다.

트랜스포머 모형(Vaswani et al., 2017)은 2017년 구글의 "Attention is all you need" 논문에서 발표한 모형으로 시퀀스-투-시퀀스(sequence-to-sequence) 모형에 해당하며 입력 시퀀스 데이터를 압축 변환하여 출력 시퀀스로 보내 이를 통해 최종 결과물을 생성하는 구조를 갖는다. 이때 전자의 압축 변환하는 과정을 인코딩(encoding)이라고 하며, 후자의 과정을 디코딩(decoding)이라고 한다. 따라서 트랜스포머 모형은 크게 앞부분의 인코더(encoder)와 뒷부분의 디코더(decoder)로 구성된다.

이후 2018년에 구글은 트랜스포머 모형의 인코더를 활용한 언어 모형(language model)인 BERT를 발표하였다. BERT 모형(Devlin et al., 2018)은 위키피디아와 도서의 많은 텍스트를 레이블 되지 않은 상태로 이용하여 사전학습(pretraining)하였으며, 미세조정(fine-tuned)을 통해 다양한 자연언어 처리 영역에서 당시 최고 수준의 결과에 대비 실제 값으로 최대 7.7%의 성능 향상을 가져오는 것으로 알려졌다.

구글은 영어판으로 인코더 레이어와 self-attention 헤드에 따라 두 종류의 모형을 공개하였다. 이후 70여 개의 언어의 텍스트를 이용하여 학습한 다국어 버전의 BERT를 추가하였는데, 이 안에는 한국어도 포함한다. 최근에는 딥러닝 기술의 발전으로 T5(Text-to-Text Transfer Transformer)나 BIGBIRD, Longformer와 같이 보다 개선된 모형들이 등장하고 있다.

BERT는 입력된 문장이나 텍스트에 대해 단어 임베딩(word embedding) 방법을 제공하는

마스크 언어 모형(masked language model)이며, 여기서 임베딩은 자연언어 텍스트에 나타난 특정 단어를 밀집된 벡터화로 표현(dense representation)하여 그 단어가 가지는 특징이나 문맥적 의미를 추출하거나 표현하는 방법을 말한다. 가장 대표적인 예로 Word2Vec나 Doc2Vec가 해당한다. 이들 방법이 각 단어에 대해 하나의 벡터 표현을 생성하는 문맥 독립적이라면, BERT에서 사용한 단어 임베딩은 같은 단어에 대해 그 단어가 사용된 문맥에 따라 서로 다른 벡터 표현을 가능하게 하는 문맥 의존적인 특징을 갖는다. 또한 임베딩의 수준을 단어 차원이나 문장(문헌) 차원으로 나누어 볼 수 있는데, BERT는 문장 수준의 임베딩까지 가능하기에 다양한 자연 언어 처리 과업에서 뛰어난 성능을 보여 널리 쓰이고 있다(Devlin et al., 2018). BERT는 자동 분류, 질의 응답, 자동 요약, 자연어 추론, 개체명 인식, 감성 분석 등의 자연 언어 처리 분야에서 탁월한 성능을 가져오기 때문에 이들 분야의 많은 연구에서 활용되고 있다. BERT를 이용한 국내·외의 연구는 이학 및 공학 분야뿐만 아니라 다수의 다른 학문 분야에서도 상당히 빠르게 증가하고 있다.

문헌 자동 분류 관련 국내의 주요 논문을 중심으로 선행 연구를 살펴보면, 다음과 같다. 황상흠, 김도현(2020)은 국가과학기술지식정보서비스(NTIS)에 등록된 인공지능 및 지능형 로봇 분야의 기술 문서 7천여 건을 실험 데이터로 이용하였으며, 분류 범주로 33개의 중분류기술명을 사용하였다. BERT 모형으로 SK T-Brain의 KoBERT를 적용하였으며 50%대의 F 점수를 보여 다른 연구들에 비해 상대적으로 낮은 성능의 결과를 보였다. 김인후, 김성희(2022)는 문헌정보학 분야의 7개 학술지의 논문 5천여 건의 초록을 대상으로 연구재단의 학술연구분야 분류표의 소분류명 13개의 범주로 전문가 검증 없이 저자가 직접 수작업 분류하여 실험 데이터를 구축하고, KoBERT를 이용하였다.

문헌 분류 이외에 BERT를 이용한 연구로 박진우 외(2022)는 특허 문헌의 자동 분류를 위해 구글 BERT 기본 모형을 기반으로 특허 분야의 120GB 규모 자연어 데이터를 사전학습한 KorPatBERT 언어 모형을 자체 개발하고, 분류 범주로는 각 특허 문헌에 부여된 선진특허분류(Cooperative Patent Classification) 기호를 사용하였다. 엄기홍, 김대식(2021)은 온라인 상의 정치 여론을 분석하고자 뉴스 기사의 댓글에 대해 긍부정의 감성 분석을 수행할 수 있는 '온라인 댓글 분류기'를 개발하였다. 3인의 연구자가 수작업으로 분류한 긍정과 부정 댓글로 구성된 만 3천여 개의 데이터셋을 구축하고, 댓글의 감성을 자동 분류하기 위해 KoBERT를 사용하여 93.04%의 정확도를 얻었다.

2.2 딥러닝 기반 주제명 자동 분류

주제명이나 주제어 또는 색인어를 자동 분류하거나 부여하기 위한 연구는 미국 의학도서관의 주제명표목표인 MeSH(Medical Subject Headings)와 관련하여 다수 수행되었다. 실제로 미국 의학도서관에서는 1990년대 중반부터 주제 색인 도구인 MTI(Medical Text Indexer)를 사용해 오고 있다(Mork, Jimeno-Yepes, & Aronson, 2013). MTI는 UMLS(Unified Medical Language System) 메타시소러스, 관련 인용문헌에 부여된 MeSH 용어, 포괄적인 색인 규

칙과 클러스터링 방법 등을 적용하여 적절한 MeSH 용어를 추천한다.

딥러닝과 관련하여 MeSH 주제명을 자동으로 분류하는 주요 연구들은 유럽 연합이 지원한 BioASQ(http://bioasq.org) 프로젝트에서 다루어졌다. 현재 BioASQ에서는 생의학 문헌에 대해 의미적 색인과 질의 응답과 관련된 과업을 조직하고 이를 위한 데이터셋 구축과 워크숍을 주관하고 있다. 딥러닝과 같은 기계 학습 모형을 사용하여 자동으로 MeSH의 주제명을 부여하는 과업(BioASQ Task a on Large-Scale Online Biomedical Semantic Indexing)은 지난 10년 동안 진행하여 2022년에 종료되었다. 이 과업을 위해 매년 작게는 4백만 건에서 많게는 1천 3백만 건의 PubMed 레코드로 구성된 데이터셋을 제공하였으며 참여 연구자들은 매년 해당 데이터셋에 대해 자신들의 딥러닝 모형을 구현하여 분류 성능을 높이고자 노력해왔다. 그 결과 이 과업에서 다음과 같은 MeSH 관련 주요 연구 성과들을 도출하였다.

DeepMeSH(Peng et al., 2016)는 먼저 MeSH 용어 후보군을 생성하고 예측하기 위해 단어빈도와 역문헌빈도를 이용하여 벡터를 생성하고 문헌 임베딩 벡터(doc2vec)를 결합하여 D2V-TFIDF 벡터를 만들고 이를 이진 분류기와 kNN 분류기에서 사용함으로써 MeSH 용어를 순위화하여 추천하였다. 그 결과 MTI보다 12% 높은 마이크로 F척도 값을 얻었다.

AttentionMeSH(Jin et al., 2018)는 생의학 논문에 나타난 표제와 초록 및 학술지명을 입력하여 문헌에 나타난 단어에 대해 임베딩하고 이를 단어의 문맥을 인식하기 위한 BiGRU(Bi-directional Recurrent Gated Unit)에 입력하고 어텐션 메카니즘을 사용하였다. 2020년에 국립중앙도서관에서 수행한 오원석(2021)의 연구도 AttentionMeSH 패키지를 그대로 사용하였는데, 다수의 문학류 학습 데이터의 사용과 온라인 서점 주제어를 대상으로 한 점에서 한계를 가진다.

BERTMeSH(You et al., 2021)는 논문의 표제 및 초록 그리고 본문의 일부 텍스트에 대해 미세 조정된 BioBERT 모형을 적용하여 임베딩 표현을 추출하고 다중 레이블 어텐션(multi-label attention)을 사용하는 전이 학습 전략을 사용하였다. 그 결과 앞의 연구들에 비해 더 좋은 성능을 가져왔다.

3. 실험 설계 및 성능 평가

이 연구는 주제명 자동 분류용 실험 데이터로 국립중앙도서관이 소장한 장서에 대해 분류와 편목 업무를 통해 생산한 국가서지의 일부를 사용하였다. 국가서지 데이터는 서지 레코드를 기반으로 하기에 도서의 서명과 저자명이 기본적으로 포함되어 있으며, 도서관 이용자에게 더욱 풍부한 정보를 제공하기 위해 일차적으로 목차를 제공하고 있으며, 필요에 따라 원문을 추가적으로 구축하고 있다.

이 연구는 주제명표목표에 수록된 주제명의 자동 분류를 위하여 구글에서 개발한 딥러닝 사전학습 모형인 다국어 BERT를 적용하였다. KoBERT와 같이 국내에서 주로 한국어 텍스트를 대상으로 사전 학습한 BERT 모형이 다수 개발되어 있는데, 이들은 대체로 구글의 다국어 BERT보다 더 좋은 성능을 보인다. 다만 구글의 다국어 BERT는 다양한 딥러닝 패키지

에서 사용할 수 있을 뿐 아니라 새로운 한국어 BERT 모형을 만들어 비교하는 베이스라인 모형이 되기에 이 연구에서는 기준 모형인 구글의 다국어 BERT를 이용하여 실험하였다. 실험에 사용한 소프트웨어 라이브러리는 허깅 페이스(Hugging Face) 사의 Transformers 패키지를 사용하였다.

전체적인 실험과정은 먼저 구글의 다국어 BERT를 이용하여 국가서지 데이터에서 추출한 분류 자질에 대해 문장 수준 임베딩(sentence embedding)을 처리하고, 이를 딥러닝 분류기에 입력하여 최종적으로 주제명을 부여한다. 여기서 딥러닝 분류기는 BERT의 결과물인 임베딩 벡터 pooler_output에 대해 선형 변환을 적용하는 레이어에 해당한다. 실험에 사용된 분류 자질은 대부분의 서지레코드에 포함되어 있는 짧은 원문 성격의 서명을 선정하였다.

주제명표목표에 수록되어 있는 각각의 주제명은 서지데이터에 균등하게 부여되지 않는다. 어떤 주제명은 흔하게 사용되며 다른 주제명은 소수의 서지에 부여되거나 심지어 전혀 부여되지 않는 사례도 있다. 이는 지식 세계에서는 자연스러운 모습이지만, 자동 분류 데이터셋 측면에서는 분류하고자 하는 범주의 분포가 균등한 것이 유리하다. 그러나 주제명 간의 분포의 차이를 인위적으로 가공할 수는 없고 실제 데이터의 모습을 그대로 적용하는 것이 바람직하다.

BERT 모형은 토크나이저로 워드피스(wordpiece) 방법을 사용한다. 워드피스는 바이트 쌍 인코딩(byte pair encoding: BPE) 알고리즘에 기초하는데, 이 BPE는 데이터에서 자주 등장하는 문자열 쌍을 데이터 내에 등장하지 않는 바이트로 대체하여 데이터를 압축한다. BPE 기법을 이용한 토큰화는 텍스트나 말뭉치에서 자주 나타나는 문자열을 서브워드(subword) 또는 부분 문자열로 대체하고 분리하여 어휘집합으로 구축한다. 따라서 워드피스의 토큰화 기법은 분석 대상 텍스트에 쓰인 언어에 대한 지식을 요구하지 않는다. 즉 한국어나 영어에 대해 그 언어가 필요로 하는 사전 처리 등을 하지 않아도 된다.

기계 학습에서 실험 데이터를 구축할 때 학습 데이터의 규모는 최종 성능에 많은 영향을 미치므로 적정한 크기를 가져야 한다. 크기의 기준은 분야나 데이터의 품질 등에 따라 다르므로 획일해서 한 가지로 설정할 수 없다. 국가서지 측면에서 보면 주제명의 부여는 장서의 주제에 따라 다르므로 균형적이지 않아 이 연구에서는 주제명의 부여 횟수에 따라 그 기준을 다양하게 설정하고 실험 데이터셋을 추출하였다. 국가서지 데이터와 주제명 부여 데이터의 통계를 참조하여 서명 중심의 분류를 위한 데이터셋을 구축한 현황은 〈표 1〉과 같다.

주제명 부여 횟수에 해당하는 출현 횟수를 기준으로 최저 5,000회 이상 부여된 주제명은 25개이며, 최저 100회 이상 출현한 횟수를 기준으로 3,506개의 주제명을 분류 범주로 사용하였다. 이들이 부여된 서지데이터의 비율을 보면 5,000회 이상의 25개 주제명이 전체 실험 대상 국가서지 데이터 122만여 건 중 13.7%에 해당하는 287,253건에 부여되었으며, 최저 100회 이상의 부여된 3,506개의 주제명은 대상 서지데이터의 73.4%에 해당하는 1,539,076건에 부여되었다. 이러한 결과를 주제명 자동 부여의 실무 측면에서 통계적으로 추정하면 보면, 부여 횟수 상위 100회 이상의 3,506개 주제명과

〈표 1〉 서명 중심 데이터 현황

세트명	부여 횟수(최저)	주제명 수	서지 레코드(건)	전체 비율
Title25	5,000	25	287,253	13.70%
Title46	3,000	46	372,971	17.80%
Title254	1,000	254	709,406	33.80%
Title603	500	603	947,581	45.20%
Title1106	300	1,106	1,139,906	54.30%
Title3506	100	3,506	1,539,076	73.40%

이들이 부여된 국가서지 전체에 73.4%를 학습하면 앞으로 추가될 서지데이터의 대략 73.4%에 대해 3,506개 주제명을 추천할 수 있다는 의미로 볼 수 있다.

기계학습 방법 중의 지도학습에 의한 자동 분류 실험의 데이터셋은 학습 세트(train set), 검증 세트(validation set), 테스트 세트(test set)로 구성된다. 일반적으로 이들의 비율은 학습 세트와 테스트 세트를 80:20 비율로 나누며, 학습하는 과정에서 분류기의 성능을 검증을 위해서 전체의 80%에 해당하는 학습 세트를 다시 80:20 비율로 나누어 20%에 해당하는 데이터를 검증 세트를 구성한다. 이에 본 연구에서도 같은 비율로 각각의 데이터셋을 구성하였다. 예를 들어 최저 300회 이상 부여된 주제명에 대한 데이터셋은 전체 1,139,906건의 서지데이터에 대해 무작위로 729,540건을 추출하여 학습 세트로, 182,385건은 검증 세트 그리고 전체의 20%에 해당하는 227,981건은 테스트 세트로 구성하였다.

자동 분류 모형의 성능은 다양한 방법으로 평가하는데 해당 모형이 수행하는 과업이 분류인지 회귀인지에 따라 성능 평가 척도는 여러 종류로 나누어진다. 회귀 모형 과업의 경우 평균 제곱근 오차(Root Mean Square Error), 평균 절대 오차(Mean Absolute Error) 등에 기반하여 평가하며, 분류 모형 과업은 정확도(Accuracy), 재현율(Precision), 정확률(Recall), F1 척도, ROC(Receiver Operating Characteristic), AUC(Area Under the ROC Curve)를 활용하여 분류 모형의 성능을 평가한다. 재현율과 정확률은 상호 보완적인 척도여서 이 중 하나만으로 분류 성능을 완전하게 평가하기 어려워 정확률과 재현율의 조화 평균인 F1 척도를 많이 사용하며 이 연구에서도 이 척도를 중심으로 분류 성능을 평가하였다. F1 척도의 경우 최종 성능을 판단하기 위해 그 관점을 분류 범주 중심으로 보느냐 분류 문헌 중심으로 보느냐에 따라 마이크로 평균 F1(micro-average F1: microF1) 척도나 매크로 평균 F1(macro-average F1: macroF1) 척도의 적용이 가능하다.

이 연구에서 수행하는 분류 대상이 주제명으로 문헌정보학 분야나 도서관 영역에 해당하므로 주제명의 자동 분류 결과를 KDC 분류체계의 주제 분야와 주제명의 유형 및 범주에 따라 평가에 대한 분석을 수행하였다. 이를 위해 기계학습의 기본적인 성능 평가 방법과 함께, KDC 분류체계에 따른 분석으로 분석 단위를 KDC의 주류 및 강목 수준의 일치 정도를 분석함으로써 이를 통해 주제 분야별 자동 분류의 용이

함과 성능 차이를 파악하고자 하였다. 또한 주제명의 범주 유형에 따른 분석은 주제명의 범주별 종류에 따른 분류 일치도 분석을 하여 해당 주제명이 자동 분류의 용이함과 난해함 정도를 파악하고자 하였다.

4. 자동 분류 결과 및 분석

4.1 딥러닝 하이퍼파라미터의 최적화

딥러닝 같은 인공 신경망은 태생부터 사용자가 모형을 최적으로 학습하도록 구현하기 위해 입력하거나 정해주어야 하는 변수가 있는데 이를 하이퍼파라미터라 하며, 이들 값에 따라 모형의 최종 성능이 크게 좌우될 수 있다. 이러한 하이퍼파라미터로는 대표적으로 은닉층의 수, 에포크(epoch), 배치 크기(batch size), 학습률(learning rate) 등이 있다. 예를 들어 학습 데이터에 대해 에포크나 배치 크기를 크게 하거나 작게 하느냐에 따라, 신경망의 학습이 최적화 되어 성능이 수렴하거나 모형의 파라미터가 발산하여 최종 결과를 얻기 어려울 수 있다. 이들 하이퍼파라미터를 최적화하기 위한 사전 실험을 수행하였는데, 은닉층의 수는 BERT-base 모형에서 이미 고유값 768로 최적화를 되었으므로 그대로 적용하였다. BERT 모형을 제안한 논문(Devlin et al., 2018)에서 다수의 데이터셋을 대상으로 실험하여 이들 하이퍼파라미터에 대한 수치를 권고하였는데 에포크는 2에서 4회, 배치 크기는 16에서 32, 그리고 학습률은 2e-5에서 5e-5 정도의 범위를 제시하였다. 이 연구는 서명 중심으로 자동 분류를 하기에 논문의 권고 수치를 포함하여 배치 크기는 8부터 32까지, 학습률은 6e-6까지 확장하여 실험하였다. 에포크를 100회라는 큰 값을 적용하여 사전 실험한 결과 〈표 2〉와 같다. 전이학습 모형이며 미세조정을 적용하여 하이퍼파라미터의 변화에 따라 분류 성능이 크게 달라지지는 않지만, BERT 논문(Devlin et al., 2018)에서 제시한 배치 크기, 학습률 등에서 차이가 나타나는 것을 알 수 있다. 학습률의 경우 대체로 권고된 수치보다 더 작은 값에서 좋은 성능을 보였으며 이는 실험 대상의 차이에서 오는 것으로 생각해볼 수 있다. 향후 전체 실험에서는 배치 크기 8과 학습률 1e-5를 적용하였다. 에포크의 경우 소수의 범주를 갖는 데이터셋과 수백 수천 개 이상의 범주를 갖는 데이터셋에서 최적화 값이 다르므로 모든 데이터셋에 100회로 설정하여 최고의 성능을 내는 순간의 모형을 저장하고 테스트 세트를 사용하여 최종 성능으로 평가하였다.

〈표 2〉 하이퍼파라미터 최적화를 위한 실험 성능(에포크 100회, microF1 기준)

배치 크기	학습률	Title25
32	5e-5	0.8132
	3e-5	0.8176
	2e-5	0.8158
	1e-5	0.8172
	8e-6	0.8166
	6e-6	0.8167
8	5e-5	0.8093
	3e-5	0.8138
	2e-5	0.8151
	1e-5	**0.8184**
	8e-6	0.8182
	6e-6	0.8179

4.2 서명 자질의 분류 성능 및 결과 분석

서지 레코드는 많은 정보를 담고 있지만, 텍스트 내지 자연어로 된 문장이 포함된 구성요소는 소수에 불과하다. 또한 이들 요소 중에 주제를 나타내는 대표적인 요소가 표제 또는 서명에 해당한다. 출현 횟수를 기준으로 주제명 수에 따른 각각의 데이터셋에서 서명을 이용한 단일 분류 자질에 대한 성능을 보면, 〈표 3〉과 같다.

주제명 자동 분류에 따른 서명의 성능을 마이크로 평균 F1 기준으로 보면, 고빈도에 해당하는 소수의 주제명에서 자동 분류의 성능이 매우 높고, 저빈도 다수의 주제명으로 갈수록 성능이 차츰 떨어지는 것을 알 수 있다. 또한 그 차이가 20%를 보일 정도로 크다. 이는 소수의 범주로 분류할 때 성능이 높은 자동 분류의 일반적인 경향과 일치하는 것으로 판단되며, 고빈도로 갈수록 학습에 사용되는 데이터가 증가하는 출현 횟수를 기준으로 설정한 이 연구 데이터셋의 특성에서 기인한다고도 볼 수 있다. 즉 세트명 Title25는 5,000번 이상 출현하여 학습할 수 있는 데이터가 그만큼 큰 데 반해, 세트명 Title3506은 100회 출현하여 학습할 수 있는 데이터가 그만큼 작아 전체적인 분류 성능에서 차이를 보이고 있음을 알 수 있다.

구체적으로 분류 성능은 서명 자질에서는 주제명 상위 25개의 데이터셋(Title25)에서 0.8184(마이크로 F1)로 높게 나타났으며, 상위 3,506개의 세트(Title3506)에서는 0.6100(마이크로 F1)로 나타났다. 마이크로 평균 정확률과 재현율 측면에서 보면 대부분의 데이터셋에서 재현율에 비해 상대적으로 정확률이 높은 것을 볼 수 있다. 해외에서 MeSH 주제명의 자동 분류에서 가장 좋은 성능을 보인 BERTMeSH(You et al., 2021)가 마이크로 평균 F1 기준으로 69.2%의 성능을 보였다. 이를 실험에 사용한 자질(표제, 초록, 본문)과 데이터셋의 규모, 분류할 주제명의 수 등으로 인해 이 연구와 직접적인 비교는 어렵지만 BERTMeSH가 부여 빈도에 상관없이 비교적 일관된 성능을 보인 반면, 이 연구는 저빈도의 주제명으로 갈수록 성능이 하락하는 것을 볼 수 있다. 다만 이 연구의 Title3506 세트의 경우 최소 100회 부여된 주제명을 대상으로 하는데, MeSH 관련 해외 연구들이 50~60%대의 F1 성능을 보이기에 많은 차이가 없는 것으로 보인다. 또한 저빈도 부여 주제명 보다 고빈도 부여 주제명의 분류 성능이 크게 높은 측

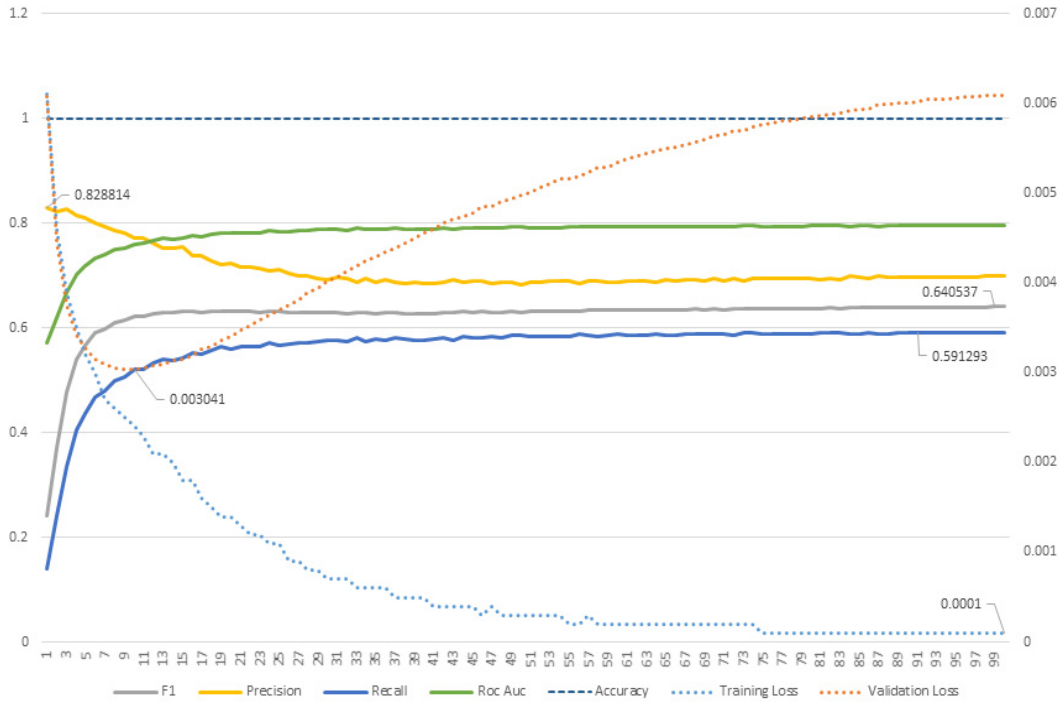
〈표 3〉 서명 자질의 분류 성능

세트명	마이크로 평균				매크로 평균				Accuracy
	F1	Precision	Recall	ROC_AUC	F1	Precision	Recall	ROC_AUC	
Title25	0.8184	0.8560	0.7840	0.8886	0.7008	0.7719	0.6728	0.8328	0.9827
Title46	0.7697	0.8370	0.7125	0.8543	0.6533	0.7387	0.6242	0.8101	0.9884
Title254	0.6952	0.7829	0.6251	0.8121	0.6113	0.6930	0.5752	0.7872	0.9972
Title603	0.6568	0.7529	0.5851	0.7923	0.5875	0.6933	0.5422	0.7709	0.9986
Title1106	0.6462	0.7074	0.5947	0.7972	0.6073	0.6730	0.5666	0.7831	0.9992
Title3506	0.6100	0.6861	0.5491	0.7745	0.5626	0.6662	0.5091	0.7545	0.9997

면은 이들 고빈도 주제명이 실무적으로 자주 부여되기에 현장 적용 가능성에 대하여 시사하는 바가 크다고 할 수 있다.

한편 서지데이터에 대해 주제명의 부여가 불균형적인 모습을 보이므로 이를 반영하여 각 주제명의 성능에 대해 동등하게 중요도를 갖도록 매크로 평균을 적용하였다. 각 데이터셋에 대해 매크로 F1 값은 주제명 수가 가장 적은 Title25 세트에서 0.7008, 가장 큰 Title3506 세트에서 0.5626의 성능을 보인다. 다만 전체적으로 마이크로 F1과 비교하면 다소 많이 낮은 값을 가진다. 정확도의 경우 데이터셋 사이에서 상대적으로 다른 평가 척도에 비해 큰 변화가 없는 것으로 보이는데, 이는 주제명의 분포가 불균형적이고 부여해야 할 범주가 많아 분자와 분모에 들어가는 true negative 값이 큰 값을 가지기 때문이다. 이는 실제 분류할 주제명이 많은, 즉 범주 수가 많은 큰 데이터셋으로 갈수록 F1 값이 낮아지는데 반해 정확도는 커지고 있는 것을 보면 알 수 있다.

최소 부여횟수가 300회 이상인 주제명 1106개의 데이터셋(세트명 Title1106)에 대해 학습 횟수인 에포크(epoch)에 따른 학습률 8e-6에서 검증 세트의 성능을 다양한 평가 척도로 나타내면, 〈그림 1〉과 같다. 초기 에포크에서는 정확률이 높지만 더 많은 학습이 이루어질수록 낮아지며, 반대로 재현율은 0.2 미만의 아주 낮은 값에서 점차 높아지는 것을 볼 수 있다. BERT 논문(Devlin et al., 2018)에서 권고한 소수의 에포크로는 정확한 학습이 되지 않아 최적의



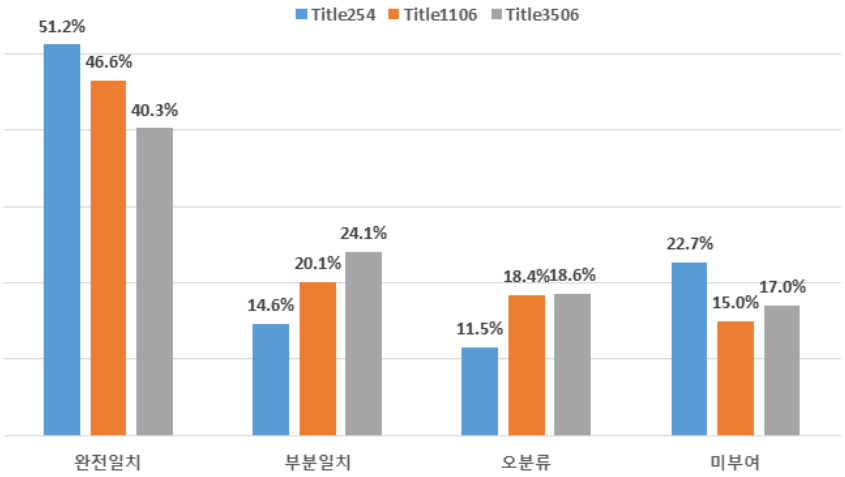
〈그림 1〉 에포크에 따른 Title1106 데이터셋의 성능

F1 값을 가져오지 않는 것을 알 수 있다. 두 값의 조화평균이 F1 척도 값은 15회부터 25회 정도에서 0.63 이상의 값을 유지하다 살짝 주춤하며 60회 근처에 다시 회복하며 90회대 후반부에서 최고점을 찍었다. 이는 일반적인 딥러닝 분류 결과와 다른 모습이다. 그림을 보면 학습 데이터에 대한 손실(training loss)은 지속적으로 감소하며 검증 데이터에 대한 손실(validation loss)은 감소하다 과적합으로 인해 다시 상승하는 모습을 보인다. 일반적으로 딥러닝에서는 검증 손실 값이 최소가 되는 순간에 학습을 멈추게 되는데, 여기서는 10회 에포크(validation loss가 0.003으로 최저)에 해당한다. 하지만 이 때의 F1 값은 0.6215로 최고점 대비 다소 낮은 것을 알 수 있다. 이는 이 연구의 데이터셋의 특성에서 기인하는 것으로 보인다. 즉 짧은 문장으로 이루어진 서명과 그에 따른 임베딩 표현에 대해 다양한 주제명이 부여되기에 검증 손실의 최소 단계를 지난 후에 높은 빈도의 주제명에 과대적합이 되면서 성능이 다소 향상되는 것으로 파악해 볼 수 있다.

추가적으로 정확도는 앞서 설명하였듯이 매 에포크마다 거의 차이가 없는 것을 알 수 있다. 대체로 실험 데이터가 규모(주제명 수)가 클수록 많은 수의 에포크를 수행해야 최고 성능을 가져오며 작은 규모의 서명 데이터셋은 BERT 논문에서 권고한 대로 약 5회 정도의 에포크에서 가장 높은 성능을 보여주었다.

자동 분류 모형의 평가는 분류 성능을 통해 서로 다른 모형이나 사용된 자질을 비교하는 데 초점을 두고 있어 실제 올바르게 분류한 범주나 주제명을 보여주거나 제시하지 않는 경향이 있다. 이 연구는 실제 주제명과 자동 분류 모형이 올바르게 분류한 주제명을 비교하여 추가적인 분석을 하였다. 「국립중앙도서관 주제명표목 업무지침(2021)」에 따르면, 하나의 서지 레코드에 최소 1개에서 3개까지 주제명을 부여하도록 원칙으로 규정하고 있다. 이 연구에 사용된 데이터의 실제 현황을 보면 주제명이 서지데이터에서 최소 1개부터 최대 12개까지 부여된 것을 알 수 있다. 하나의 서지 레코드에 다양한 수의 주제명이 부여되기에 기계에 의해 자동으로 부여된 주제명과 어떤 차이가 나는지 파악하기 위해 동일 서지 레코드에 기계와 인간이 부여한 주제명의 일치 정도를 파악해 볼 필요가 있다. 예를 들어 주제명 2개가 부여된 서지 레코드에 기계가 2개 모두를 정확하게 부여하거나 하나 또는 2개 이상, 더 나아가 하나도 부여하지 못할 수도 있다. 서명과 저자명을 이용한 분류에서 서지 레코드를 기준으로 자동 분류 결과의 일치 정도에 따른 비율을 그림으로 나타내면, 〈그림 2〉와 같다. 여기서 완전일치는 기계가 정답과 동일하게 주제명을 부여한 경우를 뜻한다. 부분일치는 정답과 기계가 부여한 주제명이 서로 부분집합인 경우를 말하며 이때는 기계가 더 적게 또는 더 많이 주제명을 부여한 경우이다. 오분류는 기계가 주제명을 자동 부여했지만, 정답과 하나도 일치하지 않는 경우를 뜻하며, 미부여는 기계가 주제명을 전혀 부여하지 않은 경우를 뜻한다. 다만 여기서 일치정도는 단순 일치 비율로 마이크로 평균 척도와 같은 평가지표와는 그 의미가 다르다.

하나의 서지 레코드에 실무자가 부여한 주제명과 기계가 완전히 동일하게 부여한 경우인 완전 일치가 주제명 상위 254개 기준의 데이터셋



〈그림 2〉 데이터셋에 따른 자동 분류 일치정도

(Title254)에서 51.2%로 가장 높았으며, 1,106개의 데이터셋(Title1106)에서는 46.6%, 3,506개 세트(Title3506)에서는 40.3%로 나타나서 자동 분류 데이터의 주제명 수가 커질수록 낮아지는 것을 알 수 있다. 서지 레코드에 부여된 주제명 중 일부만 부여한 경우인 부분 일치는 각각의 데이터셋에 대해 14.6%, 20.1%, 24.1%로 나타나서 주제명 수가 커질수록 반대로 높아지는 것을 알 수 있다. 기계가 주제명을 부여하긴 했지만 서지 레코드의 정답인 주제명과 하나도 맞지 않는 경우인 불일치는 각각의 데이터셋에 대해 11.5%, 18.4%, 18.6%로 나타나서 주제명 수가 커질수록 낮아지는 것을 알 수 있다. 더 나아가 불일치긴 하지만, 기계가 아예 주제명을 부여하지 않은 미부여 불일치가 254개 세트(Title254)에서 22.7%로 나타났고, 나머지 두 세트에서는 각각 15.0%와 17.0%로 나타났다.

주제명의 부여 분포가 장서에 따라 다르고 되도록 많은 주제명을 자동 분류해야 하는 관점에서 보면, 3,506개의 데이터셋(Title3506)에 대해 분류 정도를 좀 더 자세히 분석해 볼 필요가 있다. 특히 100회에서 200회 사이의 저빈도 주제명을 제대로 학습할 수 없는 상황이 될 수 있어 이들 주제명의 오분류를 파악해야 한다. 그 이유는 되도록 많은 주제명을 학습하여 자동 분류시스템 또는 추천 시스템이 분류할 수 있는 주제명의 수를 늘리는, 또 다른 의미로 자동 분류의 대상이 되는 주제명의 범위를 넓히는 것이 또 다른 중요한 목표가 될 수 있기 때문이다. 3,506개 데이터셋(Title3506)에서 100회에서 200회 사이의 저빈도 주제명의 개수는 1,771개로 대략 51%에 해당하므로 이들 주제명이 이 세트의 전체 성능을 크게 좌우하여 영향을 미치게 된다. 이들 저빈도 주제명의 분류 성능을 분석하면, 기계의 자동 분류에서 정확히 일치하여 부여한 비율이 48%에 해당하며, 부여는 하였지만 잘못 분류한 불일치가 34%, 미부여는 18%로 나타났다. 이는 저빈도 부여 주제명이 학습데이터의 부족으로 적절한 학습이 이루어지지 않음을 의미하며, 특히 10회와 같은 낮

은 에포크에서는 이들 주제명의 학습이 잘 이루어지지 않아 이 비율이 47%까지 높아지는 경향을 보인다. 따라서 저빈도 부여 주제명의 분류 성능을 높이기 위한 적절한 조치가 필요함을 알 수 있다.

주제명 출현 횟수에 의한 생성된 데이터셋에 대해 KDC 주류에 따른 분류 일치 정도 성능을 계산하면 〈표 4〉와 같다. 주제명 254개의 데이터셋(Title254)에서는 총류(0XX), 사회과학(3XX), 기술과학(5XX), 역사(9XX) 등이 좋은 성능을 보이며 종교(2XX), 예술(6XX), 문학(8XX)이 낮은 성능을 보였다. 특히 문학과 사회과학의 경우 테스트 데이터셋에서 각각 24.4%와 31.2%의 비율로 두 분야가 절반 이상에 해당함에도 불구하고 문학의 경우는 일치도가 상대적으로 낮고 사회과학은 높다. 문학 분야의 도서의 경우 장르 관련 주제명이 다수 부여되는데 서명만으로는 이러한 주제명에 대한 정확히 부여가 어렵다. 예를 들어 『김영랑: 최고의 순수 서정 시인』 도서에 대해 '한국 근대문학[韓國近代文學]'이 아닌 '한국 현대시[韓國現代詩]'라는 주제명을 부여하였다. 문학의 경우 정답과 동일하게 부여되는 주제명 분류가 잘 이루어지지 않음을 알 수 있다. 반면 사회과학 분야의 경우 서명이 비교적 그 분야를 잘 나타내기에 보다 더 좋은 성능을 보인다. 주제명 3,506개의 데이터셋(Title3506)에서는 총류(0XX), 자연과학(4XX), 기술과학(5XX), 언어(7XX) 등이 좋은 성능을 보이며 종교(2XX), 예술(6XX)은 낮은 성능을 보였다. 낮은 성능의 주제 분야는 이들 주제에 소수의 주제명이 포함되는 것도 성능에 영향을 미치는 것으로 보인다.

여기서 데이터셋 중심으로 살펴보면 3,506개의 데이터셋(Title3506)은 254개나 1,106개 세트에 비해 일치도와 부분 일치도가 낮아지는데 미부여를 포함하는 불일치는 그 차이가 크지 않다. 일치 여부를 중심으로 살펴보면 완전일치의 경우 주제명 출현 횟수가 높은, 즉 소수의 주제명이 포함되고 학습 데이터가 풍부한 데이터셋에서 좋은 성능을 보인 총류(0XX), 사회과학(3XX), 기술과학(5XX) 분야가 다수의 주제명을 포함하는 데이터셋에서는 부분 일치나 불일치의 성능이 급격히 저하되는 것을 알 수 있다. 반면 문학(8XX)의 경우 상대적으로 그

〈표 4〉 KDC 주류에 의한 분류 일치 현황(%)

세트명	유형	0XX	1XX	2XX	3XX	4XX	5XX	6XX	7XX	8XX	9XX
Title254	완전일치	71.2	39.6	35.4	61.3	55.9	68.0	34.1	56.8	45.2	61.3
	부분일치	4.8	3.3	10.7	13.5	5.8	9.9	43.5	17.5	8.7	9.7
	불일치	24.0	57.1	53.9	25.1	38.3	22.1	22.4	25.7	46.0	28.9
Title1106	완전일치	54.4	36.4	31.7	52.1	52.8	59.6	31.3	55.0	42.9	42.1
	부분일치	19.5	11.1	18.6	20.1	13.6	15.5	40.1	20.8	14.7	19.8
	불일치	26.1	52.5	49.8	27.8	33.6	24.9	28.6	24.2	42.4	38.1
Title3506	완전일치	46.4	30.3	26.6	42.8	46.1	48.0	28.3	48.6	40.4	34.1
	부분일치	24.1	15.4	22.2	25.7	18.4	21.9	39.9	25.1	15.8	27.9
	불일치	29.5	54.2	51.2	31.5	35.5	30.1	31.7	26.3	43.7	38.1

성능이 45.2%에서 40.4%로 상대적으로 덜 저하되는 것으로 나타났다.

출현 횟수 기준으로 생성된 3개의 데이터셋에 대해 주제명의 범주 유형에 따른 일치도로 성능을 분석하면 〈표 5〉와 같다. 다수의 주제명을 포함하는 데이터셋(Title3506) 기준으로 보면 주제명의 범주 유형에서 고유 명사처럼 특정성이 높은 주제어 유형인 '식물', '법률명', '상품명'에서 높은 성능을 보였다. 다만, '법률명'과 '상품명'의 경우 254개 데이터셋(Title254)에서는 각각 97.66%와 89.84%의 일치도를 나타났으나 3,506개 데이터셋에서는 각각 80.98%와 73.58%의 일치도를 보여 다소 많이 감소하였다. 이를 통해 타 주제명 유형에 비해 상대적으로 데이터셋 변화에 따라 일치도의 변동이 큰 것을 알 수 있다. 또한 '국보,보물' 유형의 주제명은 다른 유형에 비해 상대적으로 낮은 성능을 보였으며, 특수한 유형이 아닌 일반적인 주제어의 일치도도 52.52%로 비교적 낮다. 즉 특수한 유형의 주제명이 전체적으로 성능이 나은 것을 알 수 있다.

5. 결 론

이 연구는 최신 딥러닝 기법인 전이학습 모형인 BERT를 이용하여 주제명 자동 분류를 실험하고 성능을 평가하였으며, 더 나아가 주제명이 부여된 KDC 분류체계와 주제명의 유형 및 범주에 따른 성능을 분석하였다. 이를 위해 국립중앙도서관의 서지데이터를 이용하여 주제명의 부여 횟수에 따라 6개의 실험 데이터셋을 구축하고 서명 정보를 자질로 하여 분류 실험을 수행하였다. 실험 데이터셋은 딥러닝 분류기가 분류할 범주 수에 해당하는 주제명 개수를 최소 25개(레코드 287,253건)부터 최대 3,506개(레코드 1,539,076건)로 구성하였다. 주제명이 3,506개로 구성된 데이터셋은 국가서지 데이터의 73.4%를 해당한다. 분류 성능은 마이크로 평균 F1을 중심으로 평가하였으며, 최종 성능 평가를 위한 딥러닝 모형의 하이퍼파라미터를 최적화하였다.

실험 결과를 정리하면, 우선 딥러닝 분류기의 분류 성능을 보면, 범주 수가 최소로 25개의

〈표 5〉 주제명의 범주 유형에 따른 일치도(%)

주제명 수 / 주제명의 범주 유형	254개		1,106개		3,506개	
	일치	불일치	일치	불일치	일치	불일치
국명 〈 지명	66.82	33.14	60.95	39.05	55.07	44.93
국보,보물 〈 기념물 〈 주제어	-	-	-	-	41.38	58.62
기념물 〈 주제어	-	-	61.70	38.30	57.60	42.40
동물 〈 생물 〈 주제어	-	-	67.97	32.03	57.06	42.94
법률명 〈 주제어	97.66	2.34	86.81	13.19	80.98	19.02
상품명 〈 주제어	89.84	10.16	80.87	19.13	73.58	26.42
식물 〈 생물 〈 주제어	-	-	-	-	88.81	26.42
주제어	58.05	41.95	57.51	42.49	52.52	47.48
지명	75.92	24.08	74.41	25.59	65.62	34.38
통일서명	68.22	31.78	69.59	30.41	66.18	33.82
행정구역 〈 지명	77.52	22.48	72.54	27.46	66.71	33.29

주제명으로 이루어진 데이터셋에서는 마이크로 평균 F1 척도 값이 0.8184를 보였으며, 범주 수가 많아질수록 분류 성능은 낮아져 3506개의 주제명(범주)을 갖는 데이터셋에서는 0.6059 값을 보였다. 이러한 결과는 데이터셋에서 사용한 분류 자질의 종류나 적용한 딥러닝 기법의 차이로 인해 직접적인 비교는 어렵지만 MeSH를 이용한 해외 연구의 분류 성능과 거의 유사한 것으로 보인다. 마이크로 평균 정확률과 재현율 측면에서 보면 대부분의 데이터셋에서 재현율에 비해 상대적으로 정확률이 더 좋았다.

둘째, 이 연구에서는 짧은 문장으로 이루어진 서명을 BERT를 이용하여 임베딩 표현하고 이를 분류 자질로 사용하였는데, 하이퍼파라미터 측면에서 긴 텍스트로 구성된 데이터셋을 사용한 BERT 논문(Devlin et al., 2018)에서 권고한 값과 다른 값으로 최적화 되었다. 또한 F1 척도 값을 기준으로 최고 능을 보인 에포크 횟수와 검증 손실 값이 최소인 지점이 차이가 나는 특징을 보였는데, 특히 6개 중에 규모가 큰 데이터셋에서 이러한 측면이 더 두드러졌다. 이는 추가 연구를 통해 검증할 필요가 있지만 서명만을 분류 자질로 적용할 때 고려해야 할 부분으로 보인다.

셋째, 데이터셋이 주제명이나 범주 수가 소수에서 다수로 늘어날수록 하나의 주제명에 대한 학습 데이터가 줄어들어 저빈도 부여 주제명도 동시에 늘어나는데, 이렇게 소수의 주제명 보다 다수의 주제명을 포함하는 데이터셋으로 갈수록 분류기가 주제명을 부여하지 못하는 비율이 늘어나 최종 성능의 하락을 가져왔다. 특히 부여 횟수 100회에서 200회 사이의 주제명에서 이 정도가 심화되어 분류 성능을 높이기 위한 적절한 방안이 요구된다.

넷째, KDC 분류체계에 따른 분류 성능을 보면 다수의 주제명을 포함하는 데이터셋을 기준으로 총류(0XX), 자연과학(4XX), 기술과학(5XX), 언어(7XX) 등이 좋은 성능을 보이며 종교(2XX), 예술(6XX)은 낮은 성능을 보였다. 낮은 성능의 주제 분야는 이들 주제에 소수의 주제명이 포함되는 것도 성능에 영향을 미치는 것으로 보인다. 또한 주제명의 범주 유형에 따른 성능은 '식물', '법률명'이나 '상품명'이 높은 성능을 보인 반면, '국보/보물' 유형의 주제명은 낮은 성능을 보였다.

이미 다른 연구에서 보이듯이 다른 기계학습 영역뿐만 아니라 딥러닝에서도 전이 학습을 적용한 BERT 모형이 매우 좋은 성능을 가져오고 있어, 이 연구에서도 이러한 딥러닝 방법을 적용하여 국가서지를 대상으로 주제명의 자동 분류를 위한 실험을 수행하였다. 후속 연구로 서명 이외에 저자, 목차, 원문 등 다수 자질을 활용하여 자동 분류 알고리즘의 성능을 높이거나 학습 데이터가 부족한 주제명이나 분류 성능이 낮은 KDC 주제영역과 주제명의 범주 유형에 대해 성능 향상 방안을 분석하고자 한다. 그렇게 함으로써 도서관에서 수행하고 있는 주제명 관련 업무에서 자동화 내지 보조 도구로써 적용 가능성을 끌어 올려 색인 업무의 효율성을 높이고 용어 선정의 정확성을 높일 수 있을 것으로 기대한다.

참 고 문 헌

국립중앙도서관 (2021. 4. 10.). 인공지능이 서비스하는 검색과 요약, 체험해 보세요. 출처: https://www.nl.go.kr/NL/contents/N50603000000.do?schM=view&id=38537&schBcid=normal0302

김인후, 김성희 (2022). 딥러닝 기반의 BERT 모델을 활용한 학술 문헌 자동분류. 정보관리학회지, 39(3), 293-310. https://doi.org/10.3743/KOSIM.2022.39.3.293

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

백지원, 정연경 (2014). 국립중앙도서관 주제명표목표 검색 시스템 개선 방안에 관한 연구. 정보관리학회지, 31(1), 31-51. https://doi.org/10.3743/KOSIM.2014.31.1.031

엄기홍, 김대식. (2021). 온라인 정치 여론 분석을 위한 댓글 분류기의 개발과 적용: KoBERT를 활용한 여론 분석. 한국정당학회보, 20(3), 167-191. https://doi.org/10.30992/KPSR.2021.09.20.3.167

오원석 (2021). 인공지능 기술을 활용한 사서 업무 지원 도구 개발에 관한 연구(979-11-6513-187-6). 국립중앙도서관.

최윤경, 정연경 (2014). 국립중앙도서관 주제명표목표의 고품질화 방안에 관한 연구. 한국문헌정보학회지, 48(1), 75-95. https://doi.org/10.4275/KSLIS.2014.48.1.075

황상흠, 김도현 (2020). 한국어 기술문서 분석을 위한 BERT 기반의 분류모델. 한국전자거래학회지, 25(1), 203-214. https://doi.org/10.7838/jsebs.2020.25.1.203

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078. https://doi.org/10.48550/arXiv.1406.1078

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805. https://doi.org/10.48550/arXiv.1810.04805

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Jin, Q., Dhingra, B., Cohen, W., & Lu, X. (2018). AttentionMeSH: simple, effective and interpretable automatic MeSH indexer. Proceedings of the 6th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering, 47-56. https://doi.org/10.18653/v1/W18-5306

Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 1746-1751. https://doi.org/10.3115/v1/D14-1181

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324. https://doi.org/10.1109/5.726791

Mikolov, T., Kombrink, S., Burget, L., Černocký, J., & Khudanpur, S. (2011, May). Extensions of recurrent neural network language model. In 2011 IEEE international conference on acoustics, speech and signal processing, 5528-5531. IEEE. https://doi.org/10.1109/ICASSP.2011.5947611

Mork, J. G., Jimeno-Yepes, A., & Aronson, A. R. (2013). The NLM Medical Text Indexer System for Indexing Biomedical Literature. BioASQ@ CLEF, 1.

Peng, S., You, R., Wang, H., Zhai, C., Mamitsuka, H., & Zhu, S. (2016). Deepmesh: deep semantic representation for improving large-scale MeSH indexing. Bioinformatics, 32(12), i70-i79. https://doi.org/10.1093/bioinformatics/btw294

Reich, P. & Biever, E. J. (1991). Indexing Consistency: the Input/Output Function of Thesauri. College & Research Libraries, 52(4), 336-342. https://doi.org/10.5860/crl_52_04_336

Saarti, J. (2002). Consistency of subject indexing of novels by public library professionals and patrons. Journal of Documentation, 58(1), 49-65. https://doi.org/10.1108/00220410210425403

Tonta, Y. (1991). A study of indexing consistency between library of congress and british library catalogers. Library Resources & Technical Services, 35(2), 177-185.

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. Advances in neural information processing systems, 30, 6000-6010. https://doi.org/10.48550/arXiv.1706.03762

You, R., Liu, Y., Mamitsuka, H., & Zhu, S. (2021). BERTMeSH: deep contextual representation learning for large-scale high-performance MeSH indexing with full text. Bioinformatics, 37(5), 684-692. https://doi.org/10.1093/bioinformatics/btaa837

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

Baek, Ji-Won & Chung, Yeon Kyoung (2014). A study on improving access & retrieval system

of the National Library of Korea subject headings. Journal of the Korean Society for Information Management, 31(1), 31-51. https://doi.org/10.3743/KOSIM.2014.31.1.031

Choi, Yoon Kyung & Chung, Yeon Kyoung (2014). A study on improvements for high quality in National Library of Korea subject headings list. Journal of the Korean Society for Library and Information Science, 48(1), 75-95. https://doi.org/10.4275/KSLIS.2014.48.1.075

Eom, Kihong & Kim, Dae-Sik (2021). Automated classification model for online public opinions in a political arena: KoBERT based sentiment analysis. Korean Party Studies Review, 20(3), 167-191. https://doi.org/10.30992/KPSR.2021.09.20.3.167

Hwang, Sangheum & Kim, Dohyun (2020). BERT-based classification model for Korean documents. The Journal of Society for e-Business Studies, 25(1), 203-214. https://doi.org/10.7838/jsebs.2020.25.1.203

Kim, In hu & Kim, Seonghee (2022). Automatic classification of academic articles using BERT model based on deep learning. Journal of the Korean Society for Information Management, 39(3), 293-310. https://doi.org/10.3743/KOSIM.2022.39.3.293

National Library of Korea (2021. 4. 10.). Try it: Search and summary served by artificial intelligence. Available: https://www.nl.go.kr/NL/contents/N50603000000.do?schM=view&id=38537&schBcid=normal0302

Oh, Wonseok (2021). A Study on the Development of Work Support Tools for Librarians Using Artificial Intelligence Technology(979-11-6513-187-6). National Library of Korea.

Park, Jinwoo, Sim, Woochul, Lee, Sanghun, Ko, Bongsoo, & Noh, Hansung (2022). A study on automatic CPC classification based on Korean patent sentence: a deep learning approach using artificial intelligence language model KorPatBERT. The Journal of Intellectual Property, 17(3), 209-256. https://doi.org/10.34122/jip.2022.17.3.209