한국소설 영어번역서에 부여된 주제명의 현황 분석과 자동분류에 관한 연구*

A Study on Analysis and Automatic Classification of Subject Headings in English Translations of Korean Fictions

성 유 경 (You Kyung Sung)**
남 영 준 (Young Joon Nam)***

목 차

초 록

이 연구는 492편의 한국소설 영어번역서에 부여된 주제명을 분석하고, 기계학습 기반 주제명 자동분류 모델의 성능 평가를 목표로 한다. 이를 위해 한국문학 디지털도서관과 WorldCat에서 서지데이터를 수집하였다. 주제명 빈도와 FAST 패싯별 주제명의 분포 등을 시각화하고, 다중 레이블 분류를 위한 주제명 라벨을 선정하였다. 분류 자질과 모델 아키텍처에 따라 모델의 성능을 검증한 결과, 요약문을 분류 자질로 사용한 딥러닝 모델이 가장 우수한 성능(F1=0.62, AUC=0.89)을 보였다. 모델의 성능을 평가한 결과, 10개의 라벨 중 9개에서 AUC 값이 0.8 이상으로 분류 성능이 우수함을 확인하였다. 또한 ROC 커브와 혼동 행렬을 근거로 성능이 낮은 일부 라벨과 라벨 간 연관성을 밝혔다. 이 연구는 한국문학 번역작품을 대상으로 주제별 정량 분석을 수행하고, 소설의 주제 분류에서 딥러닝 모델의 활용 가능성을 검토한 기초연구이다.

ABSTRACT

This study analyzes the subject headings of 492 English translations of Korean fictions and evaluates machine learning-based automatic classification models. Bibliographic data were collected from the Digital Library of Korean Literature and WorldCat. Subject heading frequencies and FAST facet distributions were visualized, and key labels were selected for multi-label classification. Among various models, deep learning models using summaries as features showed the highest performance (F1 = 0.62, AUC = 0.89), with AUC values above 0.8 for 9 out of 10 labels. Additionally, based on ROC curves and confusion matrices, the study identified labels with lower performance and explored the relationships between certain labels. This study demonstrates the potential of deep learning models for classifying subjects in translated Korean literature.

키워드: 번역문학, 한국소설, 주제명, 다중 레이블 분류, 자동 주제 분류
Translated Literlature, Korean Fictions, Subject Headings, Multi-label Classification, Automated Subject Classification

---

* 이 논문은 2022년도 중앙대학교 CAU GRS 지원에 의하여 작성되었음.
** 중앙대학교 문헌정보학과 대학원 석사과정(youkyung98@cau.ac.kr / ISNI 0000 0005 2422 5906) (제1저자)
*** 중앙대학교 문헌정보학과 교수(namyj@cau.ac.kr / ISNI 0000 0004 6471 6884) (교신저자)
논문접수일자: 2025년 1월 24일 최초심사일자: 2025년 2월 4일 게재확정일자: 2025년 2월 21일
한국문헌정보학회지, 59(1): 599-624, 2025. http://dx.doi.org/10.4275/KSLIS.2025.59.1.599

1. 서 론

한국문학의 우수성을 세계에 알리기 위해서는 문학작품의 번역출간이 필수적이다. 이러한 노력의 일환으로 우리나라 문학작품은 주요 국제 문학상을 수상하였으며, 이는 한국문학의 번역과 해외 출판을 촉진하는 계기가 되었다. 한국문학 번역작품에 대한 외국인의 관심과 수요가 증가하면서 전 세계를 대상으로 한 번역문학 전문 도서관 서비스 정책과 전략이 필요하게 되었다. 한국문학 번역작품에 대한 도서관 서비스는 단순한 한국문학 홍보를 넘어서 이용자가 한국문학 번역의 역사와 현황, 전망 등을 파악할 수 있도록 지원하는 새로운 역할이 요구되었다. 특히, 어떤 주제의 문학작품이 주로 번역되었는지 분석하는 것은 한국문학 번역의 흐름과 이용자의 요구를 구체적으로 파악하고 도서관 서비스를 개발하는 데 주요한 임무가 되었다.

이러한 필요성에 따라 우리나라에서 번역작품에 대한 서지학적 연구는 1990년대에 본격적으로 이루어졌다. 박온자(1933)의 연구, 한국문학번역금고와 고려대학교 민족문화연구원이 공동으로 발간한『한국문학 번역서지 목록』(1998)이 이에 해당한다. 2001년에 한국문학번역원이 공식 출범하였고, 2007년에는 번역전문 도서관을 개관하였다. 이후 2015년에 개설된 한국문학 디지털도서관(library.ltikorea.or.kr)은 한국문학 번역과 관련된 공식 데이터베이스를 구축하고, 이를 기반으로 국내외 이용자를 대상으로 한 서비스를 제공하였다. 최근 해외에서는 워싱턴대학교 동아시아도서관의 이효경 사서가 주도하여 번역작품의 종합목록인 〈A comprehensive list of English translations of Korean literature〉가 출간되었다(Cho, 2022). 이처럼 학계에서는 한국문학 작품에 대한 서지와 동향 분석 연구가 지속적으로 이루어지고 있다.

한국문학의 세계화는 번역을 통해 주로 이루어지는데, 모든 작가와 작품을 번역하기는 현실적으로 한계가 있기 때문에 번역작품의 선정도 중요한 논의 대상이 되었다. 2006년에 개최된 〈한국문학의 세계화의 현실과 전망〉 토론회에서는 번역 과정의 편향성, 번역 지원 대상 선정의 공정성, 우선 번역 대상 작가와 작품 선정 등이 주요 쟁점으로 다루어졌다(윤지관, 2006). 이후 한국문학번역원의 다양한 지원 사업이 지속되면서 세계 문학상과 출판계에서 주목받는 한국문학 작품의 범위가 확대되었다. 이러한 맥락에서 주제명을 활용하여 현재까지 번역된 작품의 동향을 분석할 필요가 있다.

이 연구에서 '한국소설'의 '영어권' 번역을 연구범위로 설정한 것은 한국소설 영어번역서가 한국문학 번역의 가장 대표적인 영역이기 때문이다. 한국문학번역원의 2023년 사업연감에 따르면, 영어(525건)와 소설(1,581건)이 각각 언어권별, 문학 장르별 최다를 기록했다(한국문학번역원, 2024).

소설의 주제 분류는 문학 형식뿐만 아니라 등장인물, 소재, 배경, 장르 등 다양한 내용적 특성을 종합적으로 고려하는 방식으로 이루어져야 한다(노지현, 2010). 최근 한국문학이 국제적으로 주목받는 사례가 증가하면서, 한국소설의 영어번역서 또한 지속적으로 증가할 것으로 전망된다. 번역서의 양적 증가에 따라 작품의 주제를 정량적으로 분석하고 효율적으로 분

류하는 작업은 한국문학 번역의 동향을 이해하고 관련 연구 및 정보서비스의 질을 향상시키기 위한 필수적인 과제이다.

이와 관련하여 수작업으로 자료의 주제를 분류하는 기존의 방식을 보완하기 위해, 자동화된 주제 분류 시스템에 대한 논의가 활발히 진행되고 있다. 2024년 3월 CEAL(Council on East Asian Libraries) 컨퍼런스의 〈AI Tools Applied to Metadata Enhancement and Manipulation〉 프로그램에서는 파인튜닝(Fine-tuning), ChatGPT 등을 활용한 도서 분류 실험 결과가 발표되었다. 도시샤 대학교의 Harada Takashi는 일본 국회도서관의 44,900쌍의 NDC 코드와 서지데이터를 대상으로 자동분류를 수행한 결과, 'NDC Section Match'와 'NDC Main Class Match'의 정확도(accuracy)가 각각 11.1%, 29.3%로 나타났다고 보고했다. 발표자는 현재 기술만으로는 도서관의 목록 작업을 완전히 대체할 수 없지만, 양질의 데이터 확보를 통해 자료를 자동 분류하는 데 인공지능 기술을 활용할 수 있다고 전망했다(CEAL, 2024).

이 연구는 다음 조건을 만족하는 한국소설 영어번역서를 대상으로 한다. 첫째, 원작이 1900년 이후에 출판된 것으로 연구범위를 한정한다. 번역문학에서는 고전문학과 근현대문학을 구분하여 다루는 것이 일반적이다(이종호, 2022). 이 연구는 상대적으로 자료의 비중이 높고 다양한 역사·문화적 맥락을 반영한 근현대 작품을 대상으로 한다. 둘째, 번역 출발어가 한국어이고 목표어가 영어인 번역문학을 대상으로 한다. 한국어로 창작된 소설의 영어번역서로 한정하며, 한국계 미국인 작가가 영어로 직접 창작한 경우와 같은 디아스포라 문학은 제외한다. 이는 이 연구가 한국문학의 정체성에 관한 큰 담론보다는 작품에 부여된 주제명 현황을 실증적으로 파악하는 데 초점을 두기 때문이다.

이 연구의 목적은 주제명 데이터를 활용하여 한국소설의 번역 양상을 분석하고, 작품의 주제를 자동으로 분류하는 모델의 활용 가능성을 탐구하는 것이다. 이를 위해 한국문학 디지털도서관과 WorldCat에서 서지데이터를 수집하고, 작품에 부여된 주제명을 정량적으로 분석한다. 또한 분류 자질과 모델 아키텍처에 따라 주제명 라벨을 자동분류하는 모델의 성능을 검증하고, 라벨별 분류 성능을 평가한다.

2. 이론적 배경

2.1 한국문학 디지털도서관

한국문학번역원은 한국문학의 번역·출판을 지원하고 이를 국제적으로 알리는 역할을 수행하고 있다. 번역·출판 지원 건수는 해마다 꾸준히 증가하여, 2023년까지 총 42개 언어권에서 2,527건의 번역 지원과 44개 언어권에서 2,032건의 출판 지원이 이루어졌다(한국문학번역원, 2024). 이 기관이 운영하는 한국문학 디지털도서관에서는 다국어 서지 정보, 작가 정보, 보도자료 등을 검색할 수 있다. 소장 자료는 KDC와 기관 자체의 분류 기준에 따라 정리되며(한국문학번역원, 2023), 개별 작품은 수작업으로 분류하여 등록된다. 2021년에 한국문학번역원은 '한국문학 해외진출 통합 플랫폼 구축'을 주요 기관 과제로 선정하였다(곽효환, 2022). 이러한 맥락에서 디지털도서관은 전 세계 이용자가 한

국문학 번역 관련 정보를 수집, 검색, 활용할 수 있는 온라인 장으로 자리잡고 있다.

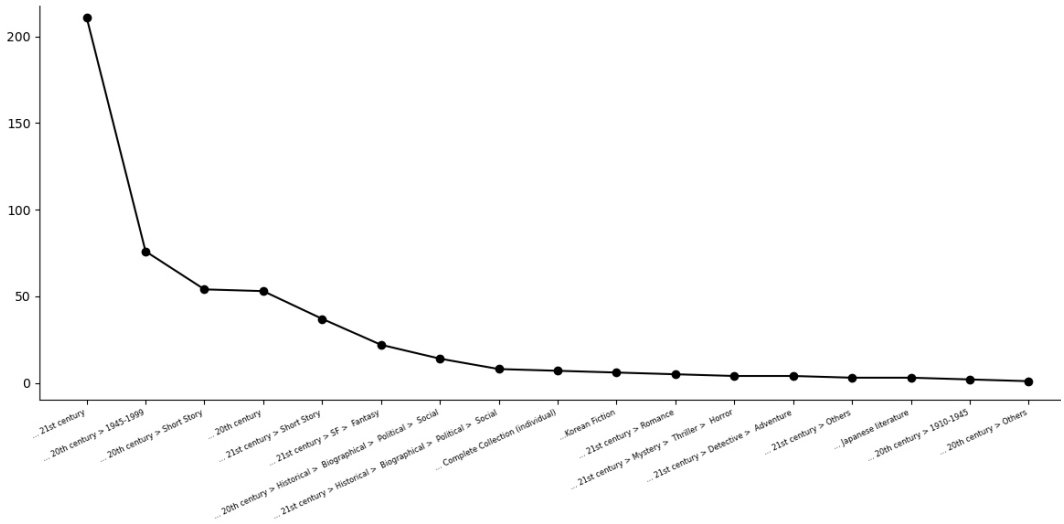
한국문학 디지털도서관의 자료 분류체계에서 'Korean Fiction'(한국 소설)은 다음의 기준에 따라 세분된다. 먼저, 원작의 출간연도에 따라 'Goryeo Dynasty', 'Joseon Dynasty', '20th century', '21st century'로 분류한다. 이중에서 근현대문학은 다시 'Short Story', 'Historical, Biographical, Political, Social', 'Romance', 'Detective, Adventure', 'SF, Fantasy', 'Mystery, Thriller, Horror', 'Others', '1910-1945', '1945-1999'로 분류된다. 하나의 작품이 복수의 분류항에 해당할 수 있다.

한국문학 디지털도서관에 수록된 한국소설 영어번역서는 496편(2024년 9월 기준)이며, 주제 분류 현황은 〈그림 1〉과 같다. 작품당 평균 1.02개의 분류항에 해당하였는데, 국립중앙도서관은 소설에 대해 평균 1.1개의 주제명을 부여하고 있는데(노지현, 2015), 이와 유사한 수치를 보였다. 'Korean Fiction 〉 21st century'가 가장 큰 비중을 차지하였고, 그 뒤를 이어 'Korean Fiction 〉 20th century 〉 1945-1999', 'Korean Fiction 〉 20th century 〉 Short Story'에 속하는 작품의 비율이 높게 나타났다.

이외에도 디지털도서관은 2024년에 'Themes'라는 메뉴를 추가하여 유사한 주제의 번역작품을 소개하는 서비스를 제공하고 있다. 이와 같은 소설의 주제 분류체계와 큐레이션 서비스는 이용자가 해당 번역작품의 주제를 파악하는 유용한 정보원이다.

2.2 기계학습 기반 자동분류

기계학습(machine learning)은 컴퓨터 프로그램이 특정 과제(T)를 수행함에 있어 특정 성능 지표(P)에서 향상되는 경험(E)을 보이면, 해당 프로그램은 T와 P에 관하여 E를 학습했다고 보는 개념이다(Mitchell, 1997). 기계학습의



〈그림 1〉 한국문학 디지털도서관의 한국소설 영어번역서 분류 현황

주요 방법론인 지도학습(supervised learning)은 입력값과 정답 라벨을 활용하여 모델을 학습시키는 방식으로, 모델의 예측값과 실제 정답 간의 차이를 최소화하는 방향으로 반복 학습이 이루어진다. 따라서 지도학습 기반 자동분류에서는 모델 아키텍처의 유형, 분류 자질, 분류 라벨, 성능 지표 등을 명확히 정의해야 한다.

분류는 지식을 개념화하여 표현하고 조직하기 위한 도구다. 이 개념에 대응하는 영문 표현으로 'classification'과 'categorization'이 혼용되는데, 특정한 기준에 따라 유사한 속성의 개체들을 그룹화한다는 공통점이 있다. 그러나 전자의 개념에서는 그룹 간 경계가 상호배타적인 반면, 후자의 개념에서는 그룹 간 경계가 유동적이고 맥락에 따라 개체 분류의 유연성이 높다(Jacob, 2004). 자동분류 방식은 크게 다중 클래스(multi-class)와 다중 레이블(multi-label)로 나뉜다. 다중 클래스 분류(classification)에서는 각 개체가 여러 클래스 중 하나에만 속하는 것과 달리, 다중 레이블 분류(categorization)는 하나의 개체가 여러 라벨에 동시에 속할 수 있다.

자연어 처리에 특화된 파이썬 라이브러리인 spaCy는 텍스트 분류를 위한 모델 파이프라인인 Text Categorizer를 제공한다. 전통적인 기계학습 모델의 아키텍처인 TextCatBOW는 BoW(Bag of Words) 방식을 기반으로 텍스트를 벡터화하며, 단어의 출현 빈도와 존재 여부를 기준으로 문서를 수치화한다. 이 모델은 구현이 단순하고 계산 속도가 빠르다는 장점이 있다. 딥러닝 모델의 TextCatCNN은 트랜스포머 기반의 RoBERTa와 CNN(Convolutional Neural Network)을 결합한 하이브리드 모델 아키텍처다. 사전 학습 모델인 RoBERTa는 텍스트의 문맥을 반영하여 각 단어와 문장의 임베딩 벡터를 출력한다. 이 출력값을 받은 CNN은 단어 임베딩 행렬에서 연속된 벡터를 탐색하여 의미적으로 중요한 패턴을 감지한다. 또한 Text Categorizer는 다중 클래스와 다중 레이블 분류를 모두 지원하는데, 문학작품은 하나의 자료가 여러 주제에 속할 수 있으므로 이 연구에서는 textcat_label을 사용하여 다중 레이블 분류를 실험한다.

2.3 국내외 선행연구

한국문학 번역작품을 대상으로 한 국내외 연구는 2010년대 후반부터 활발히 진행되어왔다. 유건수 외(2023)는 한국문학번역원의 데이터베이스를 기반으로 886편의 현대문학 번역작품을 분석하였다. 시기별 출판량, 국가・장르별 번역 추이 등을 시각화하고, 국문학의 관점에서 한국문학 번역의 현황을 분석하였다. 신지선(2023)은 통번역학의 관점에서 한국소설 영어번역서의 주제를 분류하여 해외에 표출되는 한국의 이미지 변화를 포착하고자 했다. 소설의 주제는 '정치 사상, 사회 비판, 인간의 내면 성찰, 여성주의 관점, 실험적 글쓰기'로 자체 구분하였다. 이에 앞서 문헌정보학 분야에서는 한국문학번역원에 등록된 2,172건의 서지데이터를 바탕으로 시대별 출판량 추이, 작품・작가별 번역 빈도 등 양적인 특징을 태블로(Tableau)로 시각화한 바 있다(Hur & Yi, 2017). 이유식(2000)은 1990년대 후반까지의 한국문학 번역작품 서지목록을 종합하여 영어권으로 번역된 단행본 169종의 연대별, 장르별 출판 현황을 분석하였다.

이상의 논의를 종합하면, 선행연구는 국문학

과 통번역학, 문헌정보학 등 다학제적으로 진행되어왔다. 최근 들어서는 한국문학번역원의 데이터베이스를 활용해 한국문학 번역작품의 특징을 정량적, 정성적으로 분석하려는 연구들이 이루어졌다. 이에 비해 문헌정보학의 관점에서 번역작품의 주제별 정량 분석을 시도한 연구는 상대적으로 활발하게 이루어지지 않았다.

한편 문헌의 주제를 분류하기 위해서는 인간 작업자의 많은 시간과 노력이 수반된다.이에 도서관과 정보학 분야에서는 주제명이 부여된 기존 데이터를 활용하여 문헌의 주제명을 자동분류하는 인공지능 기술에 대한 관심이 높다(Kazi et al., 2021; Yulianti & Rahadianti, 2021). 이와 관련된 최신의 국내 연구 중 이용구(2023)는 국립중앙도서관의 국가서지를 활용하여 주제명 부여 횟수를 기준으로 6개의 데이터셋을 구축하고, BERT 기반 자동분류를 실험하였다. 모든 데이터셋에서 정확도는 0.98 이상으로 높은 성능을 보였으나, 문학 분야에서는 서명만으로 장르 관련 주제명을 정확히 예측하기 어려웠다.

텍스트 자동분류 실험에서 특정 도메인에 적합하면서 정답 라벨이 부여된 대규모 데이터셋을 확보하기가 어려운 경우, 소규모이면서 불균형한 데이터셋을 활용하는 연구가 적지 않다. 이용구(2013)는 kNN 분류기를 이용해 3,988개의 신문 기사를 8개 범주로 자동분류하는 실험을 수행했는데, 학습 세트에 범주별 최소 68개에서 최대 1,100개, 테스트 세트에서는 최소 6개에서 최대 106개의 문헌을 사용했다. Zhu et al.(2023)은 4개 범주로 이메일을 분류하는 실험에서 학습 세트를 범주당 50, 100, 150개로 나누고, 테스트 세트는 범주당 5개의 문헌으로 고정해 정확도를 비교했다. 대규모 데이터셋을 활용한 경우에 비해 모델의 학습 품질과 일반화 성능에서 한계가 있지만, 이상의 선행연구들에서는 비교적 적은 양의 학습 세트를 활용하고도 정확도 0.6 이상의 성능을 달성하였다. Rajput et al.(2023)은 고품질의 데이터와 경계가 뚜렷한 분류 라벨을 사용할 경우 276개 정도의 적은 데이터로도 일정 수준 이상의 모델 구축이 가능함을 검증하였다. 이를 통해 데이터셋의 규모가 자동분류의 성능을 결정하는 절대적 요인이 아님을 알 수 있다.

이 연구의 실험대상인 한국소설 영어번역서는 500편 미만으로 소규모 데이터셋에 해당한다. 이에 따라 데이터 규모의 한계를 고려하면서도 이를 보완하기 위해 효과적인 모델 학습과 성능 평가 전략을 적용하여 자동분류 실험을 수행한다.

3. 연구방법

이 연구의 과제는 한국소설 영어번역서에 부여된 주제명을 정량적으로 분석하고, 주제명 라벨을 자동분류하는 것이다. 이를 위한 연구 절차는 (1) 데이터 수집 및 전처리, (2) 주제명의 현황 분석, (3) 모델의 학습 및 성능 검증, (4) 모델의 성능 평가 및 분석으로 구성된다. 실험 데이터셋의 규모를 고려하면 모든 주제명을 분류 라벨로 포함하는 것은 모델 학습 및 성능 평가에 제약이 많기 때문에, 주제명 라벨을 선별해야 한다. 이를 위해 주제명 데이터를 분석해 정답 라벨을 부여하는 작업을 먼저 수행한다.

3.1 데이터 수집 및 전처리

한국소설 영어번역서에 부여된 주제명의 현황을 분석하기 위해 한국문학 디지털도서관에서 496건의 서지데이터를 수집하였다. 개별 작품을 검토하여 61건을 제외하고, 57건을 추가하였다. 데이터를 정제한 결과 연구대상 작품 목록은 총 492편이다.

삭제된 데이터의 유형은 세 가지로 구분된다. 첫째, 명확하지 않거나 중복된 데이터를 정제하였다. OCLC 번호나 ISBN이 없어서 WorldCat에서 검색되지 않는 작품이나 서명, 출판사가 중복되는 작품 중 가장 먼저 출간된 것을 제외한 나머지를 삭제하였다. 또한, 〈Land〉(『토지』, 박경리 저)와 같이 시리즈로 출간된 데이터는 하나로 통합하였다. 이 과정에서 총 50건의 데이터를 제외하였다. 둘째, 번역의 출발어가 한국어가 아닌 작품은 연구대상에서 제외하였다. 예를 들어, 〈East Goes West〉(『동양선비 서양에 가시다』, 강용흘 저)는 원작 소설이 영어로 작성되었고, 이후 한국어로 번역된 사례이다. 이러한 기준에 따라 총 6편의 작품을 제외하였다. 셋째, 소설로 잘못 분류된 다른 장르의 작품을 제거하였다. 만화 〈Bad Friends〉(『나쁜 친구』, 앙꼬 저), 시집 〈I Heard Life Calling Me〉(『뒹구는 돌은 언제 잠깨는가』, 이성복 저), 사회과학 분야의 도서 〈The Gwangju Uprising〉(『5월의 사회과학』, 최정운 저) 등 총 5편의 작품을 제외하였다.

추가한 데이터의 유형은 세 가지이다. 첫째, 디지털도서관의 자료 분류체계에서 'Korean Literature(한국문학)'의 하위 분야 중 'Complete Collection 〉 Library 〉 Complete Collection & Library'에 속한 데이터를 검토하였다. 그 결과 단편 소설집, 복수 작가의 소설집 총 49편을 추가하였다. 둘째, 디지털도서관의 'Library Catalogue'에 속한 데이터를 검토하였다. 그 결과 〈People I Left in Shanghai〉(「상하이에 두고 온 사람들」, 공선옥 저)를 포함한 6건을 추가하였다. 셋째, 'Literature 〉 Chinese Literature'와 'History'로 각각 분류된 〈Chunja's Nanjing〉(『춘자의 남경』, 김혁 저), 〈King Sejong〉(『세종대왕』, 박종화 저), 이 2편을 연구대상 작품 목록에 포함하였다.

WorldCat에서 수집한 주제명 데이터는 전 세계에서 가장 보편적으로 사용되는 주제명 표목표인 LCSH(Library of Congress Subject Headings)를 중심으로 분석한다.

LCSH는 주표목과 세목으로 구성된 체계로, 주표목은 단일 단어 또는 형용사와 명사, 전치사구 등 두 개 이상의 단어로 구성되며 긴 줄표(-)를 통해 세목과 결합하여 주제의 구체성을 높일 수 있다(최윤경, 정연경, 2013). 이러한 LCSH의 논리적 결합 구조를 고려하여 노지현(2015)은 주표목에 가장 주된 주제가 반영되어 있다고 보고, 주표목을 근거로 주제명 데이터의 품질을 분석하였다. 이 연구에서도 이 선행연구와 같은 기준을 적용한다. 원칙적으로 세목이 포함된 표목은 분절하지 않고 전체를 하나의 주제명으로 간주하되, 분석 시에는 주표목을 중심으로 한다(노지현, 2015).

번역작품에 부여된 주제명의 현황 분석에서 주제명의 속성을 분석하는 도구로 FAST(Faceted Application of Subject Terminology)를 사용한다. FAST는 LCSH의 복합적인 구조를 단순화하여 주제 속성을 직관적으로 분석할 수 있도록

설계되었으며, 주제명을 일반주제명(Topical), 개인명(Personal Names), 단체명(Corporate Names), 회의명(Meetings), 사건(Named Events), 표제(Uniform Titles), 시대(Chronological), 장소(Geographical), 형식・장르(Form/Genre)의 9개 패싯으로 구분한다(노지현, 2015; OCLC, 2024). FAST 기반의 검색 도구인 searchFAST (fast.oclc.org/searchfast)를 활용하여 패싯별로 LCSH 주제명의 분포를 분석한다.

3.2 모델 학습 및 성능 평가 지표

자동분류 실험은 다음의 순서로 진행한다. 먼저, 분류 자질(서명, 요약문)과 모델 아키텍처(TextCatBOW, TectCatCNN)를 조합해 네 가지 유형의 모델을 학습시킨다. 학습・검증 세트의 80%를 학습 세트(train set)로, 20%는 검증 세트(validation set)로 분할한다. 각 모델의 성능을 검증 세트에서 비교해 최적 모델을 선정한 후, 해당 모델의 성능을 테스트 세트(test set)에서 평가한다.

데이터셋의 규모가 작은 한계를 보완하기 위해 모델 학습 및 성능 평가에서 세 가지 방법을 적용한다. 첫째, 학습・검증 세트와 테스트 세트를 분할하는 비율을 5:5로 적용한다. 8:2, 7:3의 비율을 설정하는 것이 일반적이나, 데이터셋의 규모 등을 고려하여 이 비율을 5:5, 6:4 등으로 유동적으로 조정할 수 있다(Muraina, 2022). 일례로, Yang et al.(2023)은 정답 라벨이 부여된 데이터셋이 한정된 환경에서 학습 세트와 테스트 세트를 5:5로 분할하여 실험을 수행하였다. 이처럼 소규모 데이터셋을 사용하는 자동분류 실험에서는 모델의 일반화 성능을 정확히 평가하기 위해 테스트 세트의 비중을 상대적으로 높게 적용할 수 있다.

둘째, 학습・검증 세트에 WorldCat의 영문 소설 서지데이터를 추가한다. 이 데이터에는 영문 원작뿐만 아니라 다양한 언어권에서 영문으로 번역된 소설이 포함되며, 한국소설 영어번역서도 이 범주에 속한다.

셋째, 테스트 세트의 규모가 작아서 성능 평가의 안정성이 저하될 것을 보완하기 위해, 두 개의 테스트 세트에서 모델의 분류 성능을 비교한다. 소규모 데이터셋에 대한 모델 성능 평가의 안정성을 보장하기 위해서 교차 검증을 사용할 수 있다(Matykiewicz & Pestian, 2012). 이 연구에서는 2-겹 교차 검증(2-fold cross-validation) 방식을 응용한다. 앞서 한국소설 영어번역서로 구성된 데이터셋을 5:5로 분할하여 두 개의 부분 집합으로 나눈 바 있다. 검증 단계에서 선정한 최적 모델이 학습・검증 세트와 테스트 세트로 각각 절반씩을 사용한다면, 또 다른 모델에서 두 부분 집합의 역할만 교차하여 동일한 조합으로 실험을 반복하는 것이다. 이때 서로 다른 테스트 세트에 대한 모델 간 성능 차이가 작을수록, 특정 데이터 분할에 따른 편향이 적고 분류 성능이 안정적으로 유지됨을 의미한다. 이러한 방식을 통해 단일 테스트 세트만으로 최종 성능을 평가할 때 발생할 수 있는 편향을 줄이고 결과의 안정성을 높일 수 있다.

성능 지표로는 AUC-ROC, 정확률(Precision), 재현율(Recall), F1 점수를 사용한다. AUC 값은 ROC(Receiver Operating Characteristic) 곡선 아래 면적으로, 여러 분류 임계값에서의 TPR(True Positive Rate)과 FPR(False Positive Rate) 간 관계를 나타낸다. 이 지표는 클래스

불균형에 덜 민감하며 다중 레이블 분류에서 개별 라벨의 신뢰도를 평가하는 데 사용한다(Kou et al., 2020). 정확률은 모델이 특정 라벨로 예측한 전체 데이터(TP+FP) 중 정답 라벨과 일치한 비율(TP)을 의미하며, 재현율은 실제로 특정 라벨에 해당하는 데이터(TP+FN) 중 모델이 올바르게 예측한 비율(TP)을 뜻한다. F1 점수는 정확률과 재현율의 조화평균으로, 텍스트 분류에서 모델의 성능을 종합적으로 평가한다.

4. 연구결과

4.1 한국소설 영어번역서에 부여된 주제명의 현황 분석

한국소설 영어번역서 492편에 대한 WorldCat의 주제명 데이터를 수집한 결과, 총 477편의 자료에 대해 5,094개의 개별 주제명이 부여된 것으로 나타났다. 저작 당 평균 10.68개의 주제명이 부여된 셈이다. 〈그림 2〉는 개별 주제명의 빈도수를 워드클라우드로 시각화한 것이다. 'Fiction', 'Translations', 'Short Stories', 'Romans'와 같은 주제명의 빈도수가 비교적 높게 나타났다.

작품에 부여된 LCSH 주제명은 총 3,330개로, 전체 주제명 데이터의 약 65%를 차지하였다. 나머지 1,764개의 주제명은 영어 이외의 언어로 표기된 주제명(1,174개), 비통제 영어 주제명(314개), LCSH 이외의 다른 통제어휘집에 속하는 주제명(276개)으로 구분하였다. 세부적으로, 영어 이외의 언어로 표기된 주제명은 프랑스 국가도서관의 주제명표목표(RAMEAU)에 속하는 'Romans'(소설), 'Corée du Sud Romans, nouvelles, etc'(한국 소설, 단편 소설, 기타), 국립중앙도서관의 주제명표목인 '한국 단편 소설[韓國短篇小說]' 등이 조사되었다. 또한 BISAC(Book Industry Standards and Communications)의 주제명표목표에 속한 'FICTION / World Literature / Korea', 네덜란드어 기반의 GOO-trefwoorden 시소러스에 속한 'Anthologies(form)',

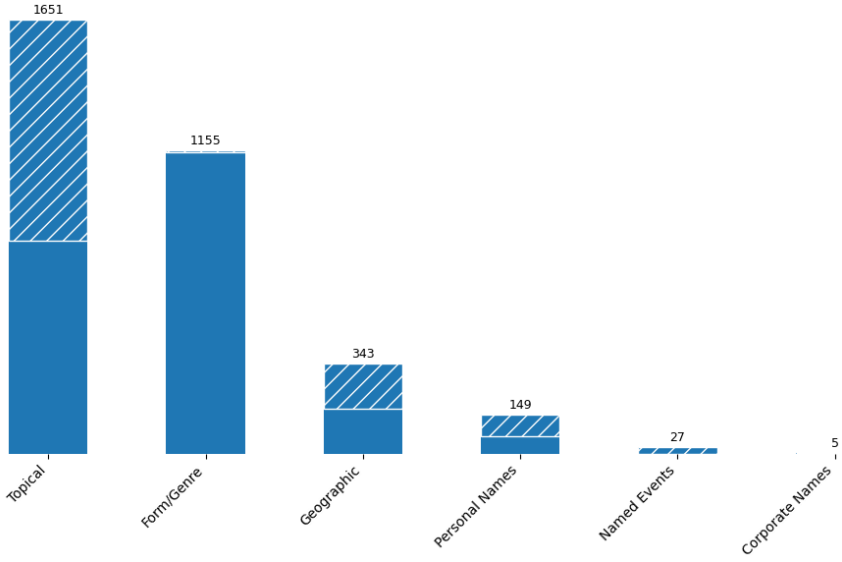
〈그림 2〉 한국소설 영어번역서에 부여된 주제명 빈도 시각화

LGBTQ+ 관련 개념을 다루는 Homosaurus Vocabulary에 속한 'Gay Romance Fiction' 등과 같이 LCSH 이외의 통제어휘집에서 비롯된 주제명도 확인되었다.

LCSH 주제명 중에서 단일 표목으로 사용된 경우가 2,196개로 가장 많았고, 세목을 가진 표목은 총 1,134개로 나타났다. 세목은 'Democratization-Korea (South)-History-20th century-Fiction'와 같이 최대 4개까지 결합되었다. 세목 사용 횟수별로는 1개의 세목을 가진 주제명이 865개로 가장 많았으며, 2개인 경우는 215개, 3개인 경우는 49개, 4개인 경우는 5개였다.

〈그림 3〉은 주표목을 기준으로 FAST 패싯별 LCSH 주제명의 분포를 분석한 것이다. FAST의 전체 9개 패싯 중에서 일반주제명, 형식・장르, 장소, 개인명, 사건, 단체명 패싯 순으로 많은 주제명이 부여되었고, 회의명, 표제, 시대 패싯에 해당하는 주제명은 부여되지 않았다. 일반주제명 패싯에 속하는 주제명이 1,651개로 가장 많았는데, 'Short stories, Korean'(292개), 'Korean fiction'(180개), 'Korean literature'(95개)와 같은 주표목이 주요 비중을 차지하였다.

형식・장르 패싯에 해당하는 것은 1,158개로 'Fiction'(313개), 'Short stories'(265개), 'Translations'(209개), 'Novels'(131개)의 순서로 높은 빈도를 보였다. 〈그림 3〉에서 세목을 가진 표목의 비율은 빗금으로 표시하였다. 일반주제명 패싯에서는 842개(51%)의 주제명이 세목과 결합되었고, 장소 패싯에서는 361개 중 190개(53%)가 세목과 결합되었다. 반면, 형식・장르 패싯에서는 세목과 결합된 주제명이 7개(1%)에 불과했다. 이는 LCSH 주제명 특성상 형식・장르 관련 주제명이 단독으로 부여되거나, 다른 주제명의 세목으로 사용되는 경우가 일반적이기 때문이다.



〈그림 3〉 FAST 패싯별 LCSH 주제명(주표목) 분포

장소 패싯에 해당하는 주제명은 'Korea'(170개), 'Korea(South)'(116개), 'Seoul(Korea)'(21개), 'Korea(North)'(12개) 순으로 빈번하게 나타났다.

개인명 패싯 내에서는 김영하(Kim, Young-ha, 1968-, 8개), 황순원(Hwang, Sun-wŏn, 1915-2000, 6개), 김동리(Kim, Tong-ni, 1913-1995, 6개), 박완서(Pak, Wan-sŏ, 1931-2011, 6개)의 순서로 작가명 관련 주제명이 빈번하게 부여되었다. 선행연구에서 정리한 1948년부터 2023년까지 '15회 이상 번역된 작가 목록'에서는 황순원이 1위를 차지했으며 박완서, 김동리도 포함된 바 있다(유건수 외, 2023).

사건 패싯에서는 17개의 주제명에서 주표목으로 사용된 'Korean War, 1950-1953'이 가장 높은 비중을 차지했다. 이를 근거로 한국소설 영어번역에서 6.25전쟁(한국전쟁)과 밀접하게 관련된 작품이 자주 다뤄지고 있음을 유추할 수 있다.

단체명 패싯에 속한 주제명 중에는 일본 오미네 강제수용소를 의미하는 'Omine(Concentration camp)'가 4개, 2014년 세월호 참사를 나타내는 'Sewŏrho(Ferry)'가 1개 나타났다.

주표목을 기준으로 중복을 제거한 결과, 총 514개의 서로 다른 주제명이 도출되었다. 자동분류를 위한 주제명 라벨을 선정하기 위해, 〈그림 4〉에 도식화한 순서대로 실험 데이터셋을 구축하였다. 우선, 특정 개별 자료에만 해당되는 주제명을 배제하고자, 10편 미만의 작품에 부여된 저빈도 주제명을 제외하였다. 〈그림 5〉는 저빈도 주제명을 제외하고 남은 주요 주제명이 부여된 작품 수를 시각화한 것이다. 각 주제명에 대해 세목을 가진 표목은 주석에 별도로 명시하였다.

〈그림 5〉에서 23개의 주요 주제명은 다음과 같다. 'Fiction'[1](312개), 'Translations'(209개), 'Short stories, Korean'[2](154개), 'Short stories'[3](134개), 'Korean fiction'[4](106개), 'Korea'[5](96개), 'Novels'(90개), 'Korea (South)'[6](56개), 'Korean literature'[7](50개), 'Manners and customs'

1) **Fiction**-21st century.

2) **Short stories, Korean**-20th century, **Short stories, Korean**-20th century-Translations into English, **Short stories, Korean**-21st century, **Short stories, Korean**-21st century-Translations into English, **Short stories, Korean**-History and criticism, **Short stories, Korean**-Korea (North)-Translations into English, **Short stories, Korean**-Translations into English, **Short stories, Korean**-Women authors, **Short stories, Korean**-Women authors-20th century-Translations into english.

3) **Short stories**-21st century.

4) **Korean fiction**-20th century, **Korean fiction**-21st century, **Korean fiction**-21st century-Translations into english, **Korean fiction**-Translations into English, **Korean fiction**-Women authors, **Korean fiction**-Women authors-20th century-Translations into english, **Korean fiction**-Women authors-Translations into English.

5) **Korea**-Fiction, **Korea**-Foreign relations-Japan-Fiction, **Korea**-History-1637-1864-Fiction, **Korea**-History-1864-1910-Fiction, **Korea**-History-20th century-Fiction, **Korea**-History-Chejudo Rebellion, 1948-Fiction, **Korea**-History-Chosŏn Dynasty, 1392-1910-Fiction, **Korea**-History-Japanese occupation, 1910-1945, **Korea**-History-Japanese occupation, 1910-1945-Fiction, **Korea**-History-Fiction, **Korea**-Juvenile fiction, **Korea**-Kings and Rulers-Fiction, **Korea**-Rural conditions-Fiction, **Korea**-Social conditions-1910-1945-Fiction, **Korea**-Social life and customs, **Korea**-Social life and customs-1910-1945-Fiction, **Korea**-Social conditions-1945-Fiction, **Korea**-Social life and customs-20th century, **Korea**-Social life and customs-20th century-Fiction, **Korea**-Social life and customs-Fiction, **Korea**-Social life and customs-Literary collections.

6) **Korea (South)**-20th century-Fiction, **Korea (South)**-Cheju island, **Korea (South)**-Fiction, **Korea (South)**-Politics and government-1960-1988-Fiction, **Korea (South)**-Politics and government-Fiction, **Korea (South)**-

(38개), 'Historical fiction'(27개), 'Psychological fiction'(26개), 'History'(25개), 'Seoul(Korea)'[8](21개), 'Women'[9](19개), 'Korean War, 1950-1953'[10](17개), 'Families'[11](15개), 'Koreans'[12](15개), 'Domestic fiction'(15개), 'Thrillers(fiction)'(12개), 'Science fiction'(12개), 'Bildungsromans'(12개), 'Electronic books' 순으로 높은 빈도를 보였다.

그 외에도, 'Suspense fiction'(9개), 'Murder'(9개), 'Romance fiction'(8개), 'Man-woman relationships'(8개), 'Love stories'(8개), 'Autobiographical fiction'(6개), 'Biographical fiction'(5개), 'Political fiction'(5개), 'War stories'(5개) 등은 상대적으로 낮은 빈도를 보였으나, 장르적 특성이 반영된 주제명을 통해 한국소설 영어번역서의 주제적 다양성을 확인할 수 있다.

〈그림 5〉의 결과를 바탕으로 주제명 라벨을 선정하기 위한 데이터 필터링 과정에서 다음과 같은 3가지 사항을 반영하였다. 첫째, 자동분류 대상인 한국소설 영어번역서의 일반적 속성을 나타내는 표목 8개(Fiction, Translations, Korean fiction, Korea, Novels, Korea(South), Korean literature, Koreans)를 제외하였다. 이는 자료들의 공통적인 속성일 가능성이 매우 높아서 라벨로써 실효성이 낮다고 판단하였다.

〈그림 5〉는 긴꼬리형 분포를 보인다. 번역문학의 일반적 속성을 나타내는 소수의 주제명은 다수의 작품에 부여된 반면, 특정 장르나 자료의 세부 내용을 반영한 다수의 주제명은 상대적으로 소수의 작품에 부여되었다. 이러한 경향은 〈그림 4〉에서도 확인된다. 491개의 저빈도 주제명만을 가진 작품은 9편에 불과했으나, 자료의 일반적 속성을 나타내는 8개의 주요 주제명만을 가진 작품은 124편이었다.

둘째, 빈도수 상위 주제명 중 유일하게 장소 패싯에 속하는 'Seoul(Korea)'과 매체 유형에 관한 'Electronic books'는 상대적으로 문학적 주제와의 관련성이 낮다고 판단하여 제외하였다.

셋째, 의미적으로 유사한 표목들을 하나의 주제명 라벨로 통합하였다. 'Short stories', 'Short stories, Korean'은 '단편소설'이라는 형태적 특성을 공유하므로, 하나의 라벨로 취급하였다. 'Historical fiction', 'History', 'Korean War, 1950-1953'도 하나의 범주로 보았다. 국내 온라인 및

Seoul, **Korea (South)**-Social conditions-Fiction, **Korea (South)**-Social life and customs-Fiction, **Korea (South)**-Social life and customs-20th century-Fiction.

7) **Korean literature**-20th century, **Korean literature**-20th century-Translations into English, **Korean literature**-21st century, **Korean literature**-21st century-Translations into English, **Korean literature**-Translations into English.

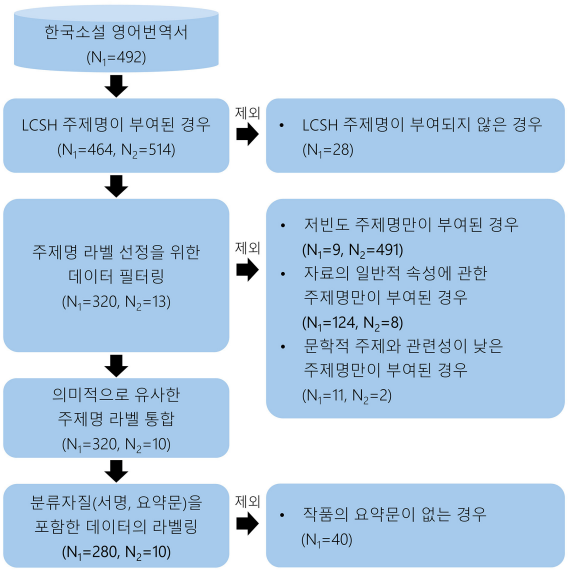
8) **Seoul (Korea)**-Fiction, **Seoul (Korea)**-Social conditions-Fiction.

9) **Women**-Fiction, **Women**-Korea-Fiction, **Women**-Korea-20th century-Fiction, **Women**-Korea (South)-Fiction, **Women**-Social conditions, **Women**-Social life and customs, **Women**-Korea-Social life and customs-Fiction, **Women**-Suicidal behavior, **Women**-Suicidal behavior-Fiction.

10) **Korean War, 1950-1953**-Campaigns-Korea (North)-Changjin Reservoir-Fiction, **Korean War, 1950-1953**-Fiction, **Korean War, 1950-1953**-Korea (North)-Hwanghae-do-Fiction, **Korean War, 1950-1953**-Veterans-Fiction.

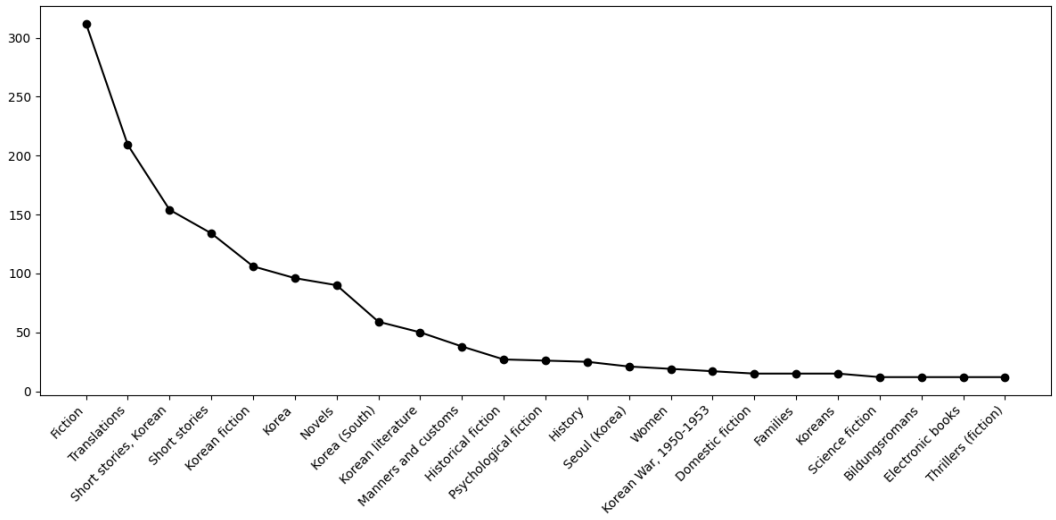
11) **Families**-Fiction, **Families**-Korea-Fiction, **Families**-Korea (North)-Fiction, **Families**-Korea (south)-Fiction.

12) **Koreans**-Australia-Fiction, **Koreans**-Belgium-Fiction, **Koreans**-China-Fiction, **Koreans**-China-Manchuria-Fiction, **Koreans**-England-London-Fiction, **Koreans**-Fiction, Koreans-Germany-Berlin-Fiction, **Koreans**-Hawaii-Fiction, **Koreans**-Migrations, Koreans-Migrations-Fiction, **Koreans**-United states-Fiction.



〈그림 4〉 자동분류 실험 데이터셋 구축 과정도

(N_1, N_2는 각 조건에 따른 '작품 수', '주제명(주표목) 개수'를 표기한 것임.)



〈그림 5〉 주요 주제명(주표목)별 작품 수

오프라인 서점의 소설장서 분류체계와 KDC, DDC의 소설 분류표를 포괄하여 정리한 '소설 장서 장르 종합 및 대표표목'(박은희, 이미화, 2020)에서 '전쟁'이라는 주제에 관해 대표 표목을 '역사'로 지정한 것을 근거로 하였다. 또한, 'Korean War, 1950-1953'은 다른 표목에 비해

그 의미와 표목이 부여된 자료의 범위가 한정적이기 때문에 단독으로 라벨을 설정한다면 충분한 학습 데이터를 확보하기 어려운 점도 고려하였다.

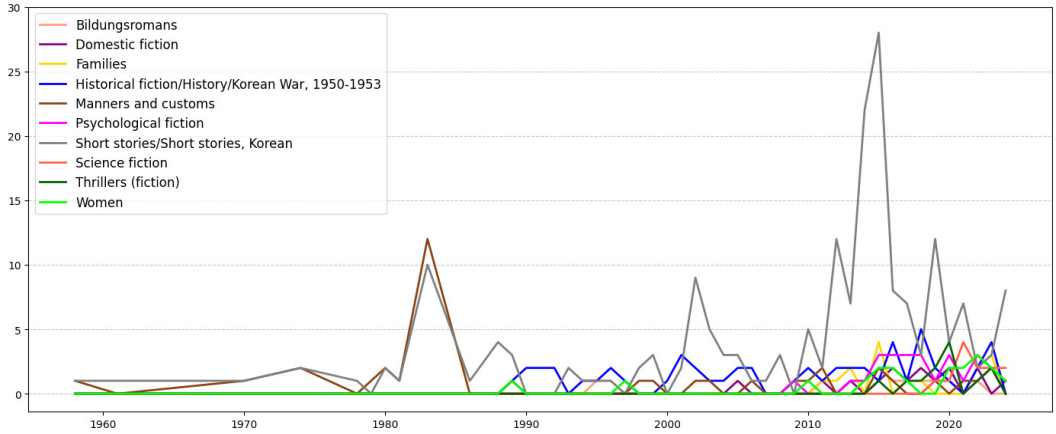
이상의 과정을 통해 주제명 라벨로 'Bildungsromans', 'Domestic fiction', 'Families', 'Historical fiction/History/Korean War, 1950-1953', 'Manners and Customs', 'Psychological fiction', 'Short stories/Short stories, Korean', 'Science fiction', 'Thrillers(fiction)', 'Women'을 선정하였다.

〈그림 6〉은 10개의 주제명 라벨 중 하나 이상을 주표목으로 가진 한국소설 영어번역서 320편을 대상으로 출판연도별 작품 수 추이를 나타낸 것이다.

2010년대 이후 한국소설 번역서의 주제적 다양성이 이전보다 현저히 확대되었음을 확인할 수 있다. 이러한 변화는 해당 시기를 전후로 'Short stories/Short stories, Korean'에 속하는 작품을 비롯해 번역서의 출간이 양적으로 증가한 흐름과 밀접한 관련이 있다. 또한, 2011년에 번역되어 미국에서 베스트셀러가 된 『엄마를 부탁해』와 2016년에 맨부커상을 수상한 『채식주의자』 등 2010년대에 한국소설이 영미권의 큰 주목을 받은 점(유건수 외, 2023)도 이러한 현상에 영향을 주었을 것으로 유추할 수 있다.

'Historical fiction/History/Korean War, 1950-1953'에 해당하는 번역서는 1990년대부터 꾸준히 출간되어 왔다. 해당 주제명의 세목을 고려하면, 이들 작품이 일제강점기, 한국전쟁, 분단 등 한국의 주요 근현대사를 배경으로 하고 있음을 알 수 있다. 강은진과 이용재(2023)의 연구에 따르면 한국 현대시에 관한 국제 연구에서 한국사가 주요 주제로 다뤄지고 있는데, 한국소설 번역에서도 이와 유사한 경향을 확인할 수 있다.

반면, 'Manners and customs'에 해당하는 작품의 비중은 과거에 비해 상대적으로 감소하는 추세이다. 이와 관련하여, 『채식주의자』를 번역한 Deborah Smith는 해당 소설을 홍보할 때 '다른 문화로의 창'이라는 기존의 접근 방식을 지양하고, 오히려 '한국성'을 강조하지 않으려는 전략을 선택했다고 밝힌 바 있다(박다해,



〈그림 6〉 출판연도별 주제명 라벨이 부여된 작품 수 추이

2016). 한국소설 번역이 특정 문화의 관념에서 벗어나 다양한 문학적 가치를 추구하는 방향으로 전환되고 있는 것이다.

'Psychological Fiction', 'Thrillers(fiction)', 'Women', 'Science fiction' 등의 주제명이 부여된 작품의 번역 빈도는 최근 들어 증가하는 경향을 보인다. 특히 'Women' 관련 소설의 번역이 늘어난 점은 2010년대 이후 한국문학 번역에서 여성 작가의 작품 비율이 현저히 증가한 흐름과 맞닿아 있다(유건수 외, 2023). 또한 2014년에서 2020년 사이에는 가부장제, 여성 차별 등을 주제로 한 작품의 번역이 활발하게 이루어졌는데(신지선, 2023), 〈그림 6〉의 결과는 이러한 선행연구의 결과와 유사하였다.

4.2 자동분류 모델의 학습 및 성능 검증

자동분류 실험은 280편의 한국소설 영어번역서를 대상으로 진행하였다(〈그림 5〉 참조). 이 데이터셋에서 라벨별 분포 비율을 유지하면서 약 5:5로 분할하여, 137건을 학습・검증 세트로 사용하고 나머지 143건은 테스트 세트로 할당하였다.

추가로, 10개 라벨 중 최소 1개 이상이 주표목으로 부여된 10,993건의 영문소설 서지데이터를 WorldCat에서 수집하였다. 〈표 1〉과 같이 학습・검증 세트의 규모가 137건에서 11,130건으로 약 80배 증가하였다. 라벨별 데이터 불균형 문제도 크게 개선되었다. 기존 학습・검증 세트에서 가장 많은 비중을 차지하는 라벨과 가장 적은 비중을 차지하는 라벨의 비율은 약 20:1이었으나, 개선된 학습・검증 세트에서는 그 비율이 약 4:1로 조정되었다.

분류 자질과 모델 아키텍처에 따른 모델의 학습 및 성능 검증에는 8,904건의 학습 세트와 2,226건의 검증 세트를 무작위로 추출하여 사용하였다. 4가지 모델 유형별 성능을 검증한 결과는 〈표 2〉와 같다.

spaCy의 TextCategorizer는 배치 크기(batch size), 기울기 누적(gradient accumulation), 학

〈표 1〉 라벨별 학습・검증 세트의 구성

라벨명	데이터 수(건)
Bildungsromans	1,034(6)
Domestic fiction	1,975(7)
Families	737(8)
Historical fiction/History/Korean War, 1950-1953	2,943(23)
Manners and Customs	1,258(10)
Psychological fiction	2,164(12)
Short stories/Short stories, Korean	1,258(83)
Science fiction	1,433(5)
Thrillers(fiction)	1,286(4)
Women	867(10)
합계	11,130(137)

*( )는 한국소설 영어번역서에 해당하는 데이터의 수를 별도 표기한 것임.

〈표 2〉 4가지 모델 유형별 성능 검증 비교

분류 자질		서명		요약문	
모델 아키텍처		TextCatBOW	TextCatCNN	TextCatBOW	TextCatCNN
AUC-ROC		0.7227	0.755	0.8232	**0.8909**
마이크로	F1	0.2814	0.3887	0.5217	**0.6157**
	Precision	0.6536	0.5559	0.6856	**0.6951**
	Recall	0.1793	0.2988	0.421	**0.5527**
매크로	F1	0.228	0.3097	0.4612	**0.5683**
	Precision	0.5673	0.5513	0.6307	**0.6689**
	Recall	0.1596	0.2589	0.3818	**0.5175**

습률 스케줄링(learning rate scheduling)과 같은 주요 하이퍼파라미터를 자동으로 최적화하는 기능을 가진다. 또한, 200회마다 검증 세트로 성능을 평가하고, 최고 성능이 갱신될 때 해당 모델을 저장하도록 설계하였다.

서명을 분류 자질로 한 전통적인 기계학습 모델은 262,144차원의 단어 벡터 공간에서 단일 어절 단위로 텍스트를 표현하였다. 학습 시 배치 크기는 100에서 시작하여 1,000까지 복합률 1.001로 점진적으로 증가하도록 설정하였다. 학습률은 0.001로 고정되었으며, 4,200회의 학습 단계까지 진행한 결과, 분류 손실값이 7.6까지 감소하고 최고 AUC 값은 0.7227을 기록하였다.

서명을 분류 자질로 한 딥러닝 모델은 입력 텍스트를 128개의 토큰 단위로 구분하였으며, 96토큰씩 중첩되도록 분석 구간을 이동하며 처리하였다. 학습 초기에 학습률을 0.00005로 설정하고, 250회까지 점진적으로 증가시킨 뒤 감소시켰다. 2,000회의 학습 단계에서 최저 분류 손실값은 5.36, 최고 AUC 값은 0.755로 나타났다.

요약문을 분류 자질로 한 전통적인 기계학습 모델은 서명보다 더 긴 텍스트인 요약문을 처리하도록 하이퍼파라미터가 조정되었다. 5,400회의 학습 단계까지 진행한 결과, 분류 손실값은 3.67까지 감소하였으며 최고 AUC 값은 0.8232로 나타났다.

요약문을 분류 자질로 한 딥러닝 모델은 최대 배치 크기를 4,096으로 설정하고, 기울기 누적 단계를 3으로 설정하여 학습 효율성을 높였다. 학습률은 초기 250회 동안 점진적으로 증가 후 감소하도록 설정되었다. 3,200회의 학습 단계에서 최저 분류 손실값은 22.81, 최고 AUC 값은 0.8909로 나타났다.

요약문을 분류 자질로 한 딥러닝 모델이 모든 성능 지표에서 높은 값을 기록하였다. 소설의 주제명을 자동분류할 때, 서명보다 요약문이 분류 성능을 더 효과적으로 향상시켰다. 전통적인 기계학습 모델에서 서명을 분류 자질로 사용했을 때 마이크로 F1 점수는 0.2814로 낮은 성능을 보였으나, 같은 조건에서 요약문을 분류 자질로 사용했을 때의 점수는 0.5217로 약 24%p 상승했다. 딥러닝 모델에서도 요약문을 활용한 경우에 F1 점수가 0.6157로 서명을 사용했을 때보다 약 27%p 더 높았다.

모델 아키텍처의 측면에서는 RoBERTa와 CNN을 결합한 TextCatCNN을 적용했을 때 분류 성능이 개선됨을 검증하였다. 요약문을 분류 자질로 사용한 두 유형을 비교했을 때,

TextCatCNN을 사용한 모델은 TextCatBOW을 사용한 것보다 마이크로 F1 점수가 약 9%p 이상 높았다. CNN은 텍스트의 계층적 특징을 효과적으로 포착하며, 특히 트랜스포머와 결합하면 문서의 국소적 문맥 정보(local contextual information)뿐 아니라 전역적 의미 구조(global semantic structure)를 동시에 학습할 수 있다(Liu et al., 2021). 이를 통해 딥러닝 모델이 요약문에 담긴 맥락적 의미를 효과적으로 포착하여 분류 성능이 향상되었다.

이용구(2023)의 연구에서는 데이터셋에 따라 주제명 분류 모델의 마이크로 F1 점수가 0.61에서 0.8184 사이의 범위를 기록하였다. 연구 대상과 데이터셋 구성이 다르므로 절대적인 비교에는 한계가 있으나, 검증 세트에 대한 최적 모델의 마이크로 F1 점수(0.6157)는 해당 범위 내에 위치하여 유의미한 수준으로 평가할 수 있다.

한편, 모든 모델 유형에서 정확률보다 재현율이 상대적으로 낮게 나타났다. 이는 모델이 특정 라벨에 속하는 데이터를 온전히 탐지하지 못했음을 의미한다. 전체 데이터 수가 적고, 특정 라벨의 데이터 비중이 낮아 모델이 라벨별 특징을 충분히 학습하지 못했을 가능성이 있다. 향후 성능 향상을 위해 데이터 증강 기법을 적용하거나, 소수 클래스 탐지율 개선을 위해 예측 임계값을 조정할 수 있다.

4.3 자동분류 모델의 성능 평가 및 분석

분류 자질로 요약문을 사용한 딥러닝 모델의 최종 성능을 평가하기 위해 테스트 세트에 대한 자동분류 결과를 분석하였다. 소규모 테스트 세트에서 얻은 결과의 안정성을 높이기 위해 〈표 3〉과 같이 서로 다른 테스트 세트에서 두 모델의 최종 성능 차이를 비교하였다. 테스트 세트①, ②는 280편의 한국소설 영어번역서로 구성된 데이터셋을 앞서 5:5로 분할한 그대

〈표 3〉 테스트 세트와 라벨에 따른 성능 평가 비교

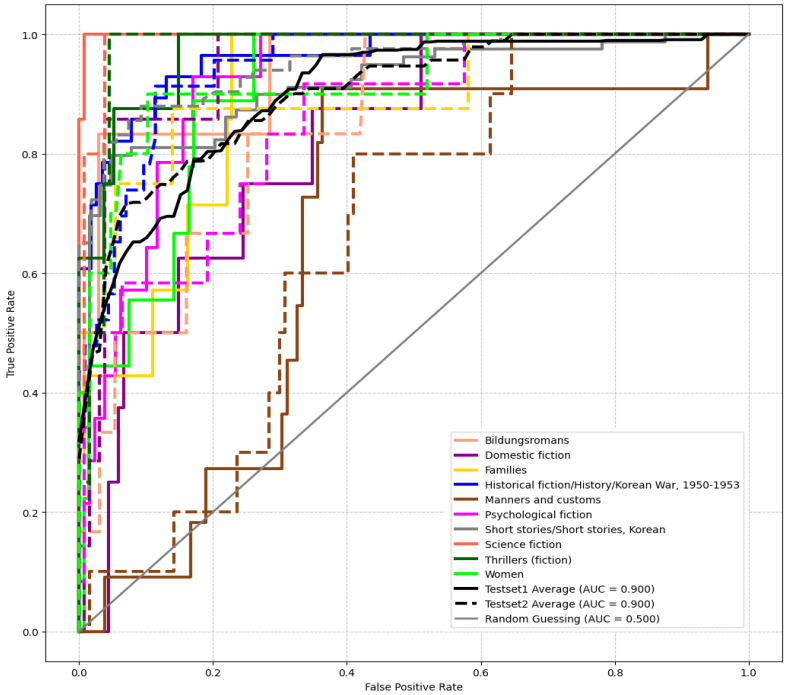
	테스트 세트①				테스트 세트②			
	데이터 수	F1	Precision	Recall	데이터 수	F1	Precision	Recall
라벨명	143	0.6726	0.7106	0.6384	137	0.6483	0.6666	0.6309
Bildungsromans	6	0.4347	0.2941	0.8333	6	0.2857	0.2	0.5
Domestic fiction	8	0.3478	0.2666	0.5	7	0.4	0.375	0.4285
Families	7	0.5	0.6	0.4285	8	0.6153	0.8	0.5
Historical fiction/History/Korean War, 1950-1953	28	0.7857	0.7857	0.7857	23	0.619	0.6842	0.5652
Manners and Customs	11	0	0	0	10	0.1	0.1	0.1
Psychological fiction	14	0.4545	0.625	0.3571	12	0.5	0.5	0.5
Short stories/Short stories, Korean	79	0.8368	0.9516	0.7468	83	0.859	0.9696	0.771
Science fiction	7	0.923	1.0	0.8571	5	0.6666	0.5714	0.8
Thrillers(fiction)	8	0.5882	0.5555	0.625	4	0.4444	0.4	0.5
Women	9	0.5333	0.6666	0.4444	10	0.5454	0.5	0.6

로를 반영한 것으로 각각 143건과 137건으로 나누어진다. 두 모델의 성능을 비교한 결과, 마이크로 F1 점수는 각각 0.6726과 0.6483으로 나타났다. 점수 차이가 0.05 이내로 근소하여, 테스트 세트에 대한 분류 결과의 안정성을 확인하였다.

10개의 라벨, 'Bildungsromans(이하 B)', 'Domestic fiction(이하 D)', 'Families(이하 F)', 'Historical fiction/History/Korean War, 1950-1953(이하 H/K)', 'Manners and Customs(이하 M)', 'Psychological fiction(이하 P)', 'Short stories/Short stories, Korean(이하 S)', 'Science fiction(이하 SF)', 'Thrillers (fiction)(이하 T)', 'Women(이하 W)'별로 분류 성능을 비교한 결과는 다음과 같다. D, F, M, P, S, W, 6개 라벨에서 두 모델 간 F1 점수 차이는 0.1 이내로 유지되었다. 특히 라벨 S에 대한 F1 점수는 0.8368과 0.859로, 일관되게 높은 성능을 보였다.

이 외에도 라벨 SF의 F1 점수는 두 모델에서 모두 0.65 이상을 기록하며 높은 성능을 보였다. 다만 해당 라벨은 모델 간 F1 점수 차이가 0.25 이상으로 성능 차이가 컸다. 두 테스트 세트에서 이 라벨에 할당된 텍스트는 모두 10건 미만이기 때문에, 분류 자질인 요약문의 특징에 따라서 분류 성능이 크게 변동되었을 가능성이 있다.

〈그림 7〉은 테스트 세트에 대한 모델의 성능



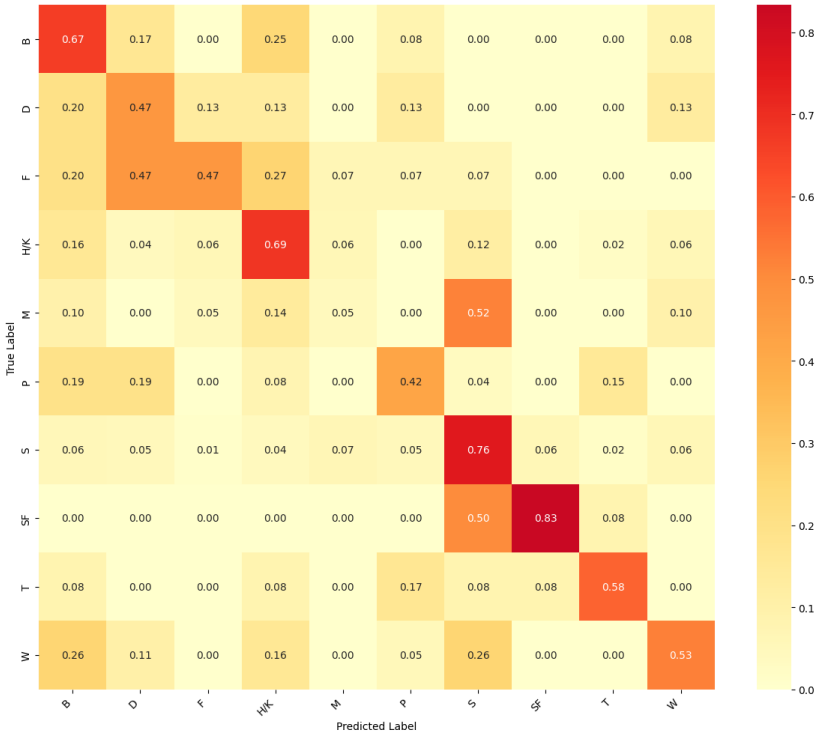
〈그림 7〉 테스트 세트에 대한 모델의 ROC 커브

* 테스트 세트①, ②에 대한 결과는 각각 실선, 점선으로 구분함.

평가 결과를 ROC 커브로 시각화한 것으로, 서로 다른 테스트 세트를 각각 실선과 점선으로 구분하였다. 두 모델의 평균 AUC 값은 모두 0.9이고 9개 라벨에서 AUC 값이 0.8 이상으로 나타났다. 이에 따라 라벨별 ROC 커브가 대체로 좌측 상단에 근접하게 위치하였다. 반면 라벨 M의 AUC 값이 각각 0.6673과 0.6645로 현저히 낮은 성능을 보였다. 해당 라벨에 대한 검증 세트의 F1 점수는 0.4569로 일관되게 낮게 나타났다. 이는 한국소설 영어번역서뿐 아니라 영문소설을 포함한 학습·검증 세트 전반에서 해당 라벨에 대한 모델의 분류 기준이 모호하여 성능이 제한적임을 시사한다.

〈그림 8〉은 다중 레이블 분류의 혼동행렬로, 임계값 0.5를 기준으로 각 라벨별 TP, FP, FN, TN을 독립적으로 계산하여 시각화한 것이다. 이를 통해 모델이 예측한 라벨과 정답 라벨 간의 상관관계 및 라벨 간 연관성을 분석하였다. 서로 다른 모델의 테스트 세트를 종합했기 때문에 개별 모델의 예측 확률 분포나 성능 차이가 상쇄되는 한계가 있다. 그러나 두 모델이 동일한 분류 자질과 아키텍처를 사용했고 F1 점수와 AUC 값에서 매우 근소한 차이를 보이므로, 이를 통합적으로 분석하는 것이 타당하다고 판단하였다.

〈그림 8〉을 분석한 결과 라벨 B, H/K, S, SF에 대한 재현율은 0.65를 상회하는 것으로 나타났다. 반면 라벨 M의 재현율은 0.05로 가장



〈그림 8〉 테스트 세트(종합)에 대한 혼동행렬

낮았으며, 라벨 M에 해당하는 텍스트 중 절반 이상이 S로 잘못 분류되었다. 라벨 SF에서도 라벨 S와의 혼동이 관찰되었다. 하지만 이 경우에는 모델이 라벨 SF를 올바르게 예측하면서 동시에 라벨 S로 중복 분류하는 경향을 보였다는 점에서 라벨 M의 분류 성능과는 차이가 있다.

소설은 하나의 작품 안에 다층적인 주제 요소가 내포되어 있다. 이러한 특성을 고려하여 아래와 같이 세 가지 측면에서 모델의 예측 오류 패턴에 주목하였다. 이를 근거로 라벨 간의 연관성을 분석하고 한국소설 영어번역서의 주제적 연관성을 유추하고자 하였다. 모델의 성능이 완벽하지 않다는 한계가 있으나, 주제명 간 반복적으로 나타나는 오분류 패턴은 해당 주제명 사이의 주제적 연관성을 시사한다.

첫째, 성장 소설(Bildungsromans, B), 가정 소설(Domestic fiction, D), 가족 관련 소설(Families, F) 간의 연관성이 관찰되었다. 〈그림 8〉에 따르면, 정답 라벨이 D인 텍스트 중 라벨 B와 F로 예측된 비율은 각각 0.2와 0.13이었으며, 정답 라벨이 F인 경우 라벨 B와 D로 예측된 비율은 각각 0.2와 0.47로 나타났다. 이를 통해 가정 소설과 가족 관련 소설의 라벨 간 연관성을 확인하였다. 또한 한국소설 영어번역서에서 성장 서사가 가정을 배경으로 하거나 가족을 중심으로 전개되는 경향이 있음을 추론할 수 있다.

둘째, 여성(Women, W) 관련 소설은 다양한 장르와 복잡하게 얽혀 있는 양상을 보였다. 정답 라벨이 W인 텍스트 중 성장 소설(B), 역사 및 전쟁 소설(Historical fiction/History/Korean War, 1950-1953, H/K), 심리 소설(Psychological fiction, P)로 분류된 비율은 각각 0.26, 0.16, 0.26으로 나타났다. 이를 통해 여성 인물의 성장기를 다룬 소설, 역사적 맥락 속에서 여성의 생활상이나 역할을 조명한 소설, 여성 주인공의 내적 갈등과 심리를 세밀하게 묘사한 소설 등이 번역되었을 것으로 유추할 수 있다. 이는 한국소설 번역서에서 'Women'이라는 주제명이 독립적인 서사적 요소가 아니라 다양한 장르와 상호작용함을 시사한다.

셋째, 심리 소설(P)과 스릴러 장르(Thrillers (fiction), T)는 밀접한 연관성을 보였다. 정답 라벨이 T인 텍스트 중 라벨 P로 분류된 비율은 0.17이었으며, 반대로 라벨 P가 라벨 T로 분류된 비율은 0.15로 나타났다. 이는 '심리 스릴러'에 해당하는 한국소설이 번역되었을 가능성을 시사한다.

다음 〈표 4〉는 5편의 한국소설 영어번역서에 대해 한국문학 디지털도서관의 분류항과 자동분류 모델의 예측 결과를 비교한 예시이다.

한국문학 디지털도서관은 원작의 출판 시기나 문학 형식을 중심으로 번역작품을 분류한 반면, 자동분류 모델은 10개의 주제명 라벨 가운데 작품에 해당하는 주제적 특징을 예측한다. 예를 들어, 최윤의 장편소설 『마네킹』의 번역서인 〈Mannequin〉에 대해서 디지털 도서관은 '21세기 한국소설'로, 모델은 '성장 소설', '여성' 관련 소설로 분류하였다. 이효석의 장편소설 『벽공무한』의 번역서인 〈Endless Blue Sky〉에 대해서는 디지털도서관이 '20세기 근현대소설'로 분류하였고, 모델은 '역사/전쟁 소설', '풍습 및 관습' 관련 소설로 예측하였다. 이처럼 모델의 예측 결과를 통해 번역작품의 장르와 주요 소재를 효과적으로 파악할 수 있다.

〈표 4〉 작품별 한국문학 디지털도서관의 분류항과 모델 예측 결과 비교 예시

작품명	한국문학 디지털도서관	자동분류 모델
I'm Waiting For You (Kim Bo-Yong, 2021)	Korean Fiction 〉 21st century 〉 SF 〉 Fantasy	• Short stories/Short stories, Korean • Science fiction
Mannequin (Ch'oe Yun, 2016)	Korean Fiction 〉 21st century	• Bildungsromans • Women
My Brilliant Life (Kim Ae-Ran, 2022)	Korean Fiction 〉 21st century	• Families • Domestic fiction
Endless Blue Sky (Lee Hyo-Seok, 2018)	Korean Fiction 〉 20th century 〉 1945-1999	• Historical fiction/History/Korean War, 1950-1953 • Manners and Customs
The Law Of Lines (Pyun Hye-Young, 2020)	Korean Fiction 〉 21st century	• Thrillers (fiction) • Psychological fiction

5. 결 론

이 연구의 목적은 한국소설 영어번역서에 부여된 주제명의 현황을 파악하고, 기계학습을 기반으로 주제명 라벨을 분류한 모델의 성능을 평가하는 것이다. 이를 위해 한국문학 디지털도서관과 WorldCat에서 한국소설 영어번역서의 서지데이터를 수집하고, 개별 작품을 확인하여 데이터를 정제하였다. 492편의 한국소설 영어번역서에 부여된 주제명의 현황을 분석하고, 자동분류 모델의 성능을 검증 및 평가한 결과는 다음과 같다.

한국소설 영어번역서에 부여된 주제명을 정량적으로 분석한 결과, 개별 주제명은 총 5,094개로, 작품당 평균 10.68개의 주제명이 부여되었다. 한국문학 디지털도서관에서 개별 작품에 할당된 분류항의 개수보다 10배 정도 많은 것으로 나타났다. 전체 주제명 중 65%가 LCSH에 속했으며, FAST 패싯 중에는 일반주제명과 형식·장르 패싯에 해당하는 주제명의 비중이 높았다. 각 주제명이 부여된 작품 수와 표목의 속성을 근거로 10개의 주제명 라벨('Bildungsromans', 'Domestic fiction', 'Families', 'Historical fiction/History/Korean War, 1950-1953', 'Manners and Customs', 'Psychological fiction', 'Short stories/Short stories, Korean', 'Science fiction', 'Thrillers(fiction)', 'Women')을 선정하였다. 또한 출판연도별로 주제명 라벨이 부여된 작품 수 추이를 근거로, 2010년대 이후부터 한국소설 영어번역서의 주제적 다양성이 전반적으로 높아지고 있음을 밝혔다.

분류 자질과 모델 아키텍처에 따른 모델의 성능을 검증한 결과, 요약문을 분류 자질로 사용한 딥러닝 모델이 가장 우수한 성능을 보였다. 검증 세트에서 최적 모델의 마이크로 F1 점수는 0.62, AUC 값은 0.89로 나타났다.

테스트 세트에서 라벨별 분류 성능을 분석한 결과, 9개의 라벨에서 AUC 값이 0.8 이상으로 양호한 성능을 보였다. 특히, 'Bildungsromans', 'Historical fiction/History/Korean War, 1950-1953', 'Short stories/Short stories, Korean', 'Science fiction'에 속하는 텍스트 중 모델이 각

라벨로 올바르게 예측한 비율이 모두 65% 이상으로 나타났다.

ROC 커브와 혼동행렬을 근거로 라벨별 성능 차이와 예측 오류 패턴을 종합하면, 자동분류 모델의 성능 개선을 위한 다음과 같은 시사점을 도출할 수 있다. 첫째, 분류 성능이 높은 라벨은 자동분류에 의존할 수 있는 가능성이 높지만, 성능이 낮은 라벨의 경우에는 인간 작업자의 개입이 불가피하다. 모든 평가 지표에서 낮은 성능을 보인 주제명 'Manners and customs'에 대해서는 모델의 신뢰성이 낮으므로 수작업의 필요성이 특히 높다. 둘째, 일부 라벨의 분류 성능이 낮게 나타난 원인 중 하나로 요약문 내에 특정 주제명과 관련된 정보가 충분히 포함되지 않은 것을 유추할 수 있다. 향후 번역작품의 요약문이 주제명 라벨을 중심으로 주제 정보를 체계적으로 반영하도록 보완된다면, 소설의 요약문이 이용자에게 유용한 자료로 활용될 수 있을 뿐 아니라 이를 학습한 모델의 성능 또한 크게 향상될 것이다. 셋째, 'Domestic fiction', 'Families'와 같이 혼동행렬에서 주제적 연관성이 높게 나타난 라벨의 경우에는 모델 학습 과정에서 세부 조정이 요구된다. 이러한 라벨들을 하나로 통합하면 분류 경계를 명확히 설정할 수 있고, 궁극적으로 모델의 성능이 개선될 것으로 예상한다.

이 연구는 한국소설 영어번역서에 부여된 주제명을 분석한 결과를 시각화하고, 번역작품의 주제를 분류할 때 모델이 보조 도구로 사용될 수 있음을 검증하였다. 다만 이 연구에서 활용한 자동분류 대상의 데이터 수가 적고 라벨의 개수가 10개로 제한되어 있어서, 문학작품의 다양한 주제적 요소를 포괄하지 못하는 한계를 지닌다. 또한, 라벨의 개수를 늘릴 경우 분류 성능이 저하될 가능성이 있다는 점 역시 주요한 제한점이다.

후속 연구에서는 자동분류 실험 데이터셋의 규모를 확장하는 방안을 구체적으로 모색할 필요가 있다. 중국어, 스페인어 등 다양한 언어로 번역된 문학작품을 대상으로 WorldCat의 주제명 데이터를 수집하고 자동분류를 수행하는 방안을 고려할 수 있다. 이를 통해 한국문학 번역작품에 부여된 주제명의 현황을 광범위하게 분석하고, 자동분류 모델의 신뢰성과 일반화 가능성을 높일 수 있다. 또한 요약문을 분류 자질로 사용할 때 요약문이 제공되지 않은 작품이 실험 데이터셋에서 제외되는 문제를 해결하기 위해, 한국어 등 다국어 요약문을 사용하는 방안도 고려해 볼 수 있다. 모델이 언어적 다양성을 반영하여 작품의 주제명을 분류한다면, 이 모델의 활용 범위는 더욱 확대될 것이다.

참 고 문 헌

강은진, 이용재 (2023). 토픽 모델링과 N-gram을 활용한 한국 현대시 연구 동향 분석: 2010년~2023년 WOS 및 SCOPUS DB를 중심으로. 한국어문학국제학술포럼, 61, 169-201. https://doi.org/10.35821/jkc.2023.05.61.169

곽효환 (2022). 세계문학으로서의 한국문학 현황과 전망. 한국문예창작, 21(1), 13-40. https://doi.org/10.47057/jklcw.2022.54.01

한국문학번역금고, 고려대민족문화연구원 공편 (1998). 한국문학 번역서지 목록. 서울: 한국문학번역금고, 고려대민족문화연구원.

노지현 (2010). 장르 분류의 사례를 통해 본 도서관 분류의 의미: 북미 공공도서관을 중심으로. 한국도서관·정보학회지, 41(4), 151-170. https://doi.org/10.16981/kliss.41.4.201012.151

노지현 (2015). 주제명 데이터를 통해 본 현행 목록의 품질과 과제. 한국도서관·정보학회지, 46(4), 379-402. https://doi.org/10.16981/kliss.46.4.201512.379

박다해 (2016.06.09.). 데보라 스미스 "'K-문학' 표현은 쓰지말자". 머니투데이. 출처: https://news.mt.co.kr/mtview.php?no=2016061910172102314

박은희, 이미화 (2020). 학교도서관을 위한 소설장서의 장르 분류 방안에 관한 연구. 한국비블리아학회지, 31(1), 115-136. https://doi.org/10.14699/kbiblia.2020.31.1.115

신지선 (2023). 한국문학 번역서에서 표출되는 한국의 이미지 변화. 통번역교육연구, 21(2), 119-135. https://doi.org/10.23903/kaited.2023.21.2.006

유건수, 김보경, 김지윤, 전세진, 정기인, 정성훈, Chandler, S. (2023). 한국문학 영어번역 양상 멀리서 읽기(1): 현대문학 작품을 중심으로. 한국근대문학연구, 24(1), 7-42.

윤지관 (2006.08.08.). 한국문학의 세계화: 무엇을 번역할 것인가. 창작과 비평. 출처: https://magazine.changbi.com/MCWC/WeeklyItem?id=20

이유식 (2000). 한국문학 영어권 번역 소개 연구. 번역학연구, 1(1), 169-202.

이용구 (2013). 문헌빈도와 장서빈도를 이용한 kNN 분류기의 자질선정에 관한 연구. 한국도서관·정보학회지, 44(1), 27-47. https://doi.org/10.16981/kliss.44.1.201303.27

이용구 (2023). BERT 모형을 이용한 주제명 자동분류 연구. 한국문헌정보학회지, 57(2), 435-452. https://doi.org/10.4275/KSLIS.2023.57.2.435

이종호 (2022). 한국문학번역장의 형성과 세계문학을 향한 열망: 유네스코 한국위원회의 Korea Journal을 중심으로. 구보학보, 32, 289-330. https://doi.org/10.35153/gubokr.2022..32.007

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

한국문학번역원 (2023). 번역전문도서관 운영 및 자료 관리지침 제정(안).

한국문학번역원 (2024). 2023 Annual Report.

CEAL (2024). 2024 CEAL Annual Meeting Program. Available: https://www.eastasianlib.org/newsite/meetings/past-meetings/ceal2024/

Cho, H. (Ed.). (2022). The Routledge Companion to Korean Literature. New York: Routledge.

Hur, K. & Yi, H. (2017). Using data visualization to examine translated Korean literature. Journal

of East Asian Libraries, 2017(165).

Jacob, E. (2004). Classification and categorization: a difference that makes a difference. Library Trends, 52.

Kazi, N., Lane, N., & Kahanda, I. (2021). Automatically cataloging scholarly articles using library of congress subject headings. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, 43-49.

Kou, G., Yang, P., Peng, Y., Xiao, F., Chen, Y., & Alsaadi, F. E. (2020). Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods. Applied Soft Computing, 86. https://doi.org/10.1016/j.asoc.2019.105836.

Liu, W., Pang, J., Li, N., Zhou, X., & Yue, F. (2021). Research on multi-label text classification method based on tALBERT-CNN. International Journal of Computational Intelligence Systems, 14(1). https://doi.org/10.1007/s44196-021-00055-4.

Muraina, I. (2022). Ideal dataset splitting ratios in machine learning algorithms: general concerns for data scientists and data analysts. In 7th international Mardin Artuklu Scientific Research Conference, 496-504.

Matykiewicz, P. & Pestian, J. (2012). Effect of small sample size on text categorization with support vector machines. Proceedings of the 2012 Workshop on Biomedical Natural Language Processing, 193-201.

Mitchell, T. M. (1997). Machine learning. New York: McGraw-Hill Science.

OCLC (2024). FAST Quick Start Guide. Available: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.oclc.org/content/dam/oclc/fast/FAST-quick-start-guide-2022.pdf

Park, On-Za (1993). A Bibliography of Korean literature in English or translated into English. Seoul: Hanshin Publishing Co.

Rajput, D., Wang, W. J., & Chen, C. C. (2023). Evaluation of a decided sample size in machine learning applications. BMC bioinformatics, 24(1), 48. https://doi.org/10.1186/s12859-023-05156-9

Yang, L., Huang, B., Guo, S., Lin, Y., & Zhao, T. (2023). A small-sample text classification model based on pseudo-label fusion clustering algorithm. Applied Sciences, 13(8), 4716. https://doi.org/10.3390/app13084716

Yulianti, E. & Rahadianti, L. (2021). Determining subject headings of documents using information retrieval models. Indonesian Journal of Electrical Engineering and Computer Science, 23(2),

1049-1058. https://doi.org/10.11591/ijeecs.v23.i2.pp1049-1058

Zhu, Y., Zhu, T., Li, J., Cao, W., Yong, P., Jiang, F., & Liu, J. (2023). Classify text-based email using naive bayes method with small sample. Journal of Information Science & Engineering, 39(4). https://doi.org/10.6688/JISE.202307_39(4).0010

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

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

Kang, Eun-Jin & Lee, Yong-Jae (2023). A trend analysis of Korean contemporary poetry research using topic modeling and n-gram analysis: focusing on WOS and SCOPUS DB from 2010 to 2023. Journal of Korean Culture, 61, 169-201. http://doi.org/10.35821/jkc.2023.05.61.169

Korea Literature Translation Fund & Institute of Korean Culture at Korea University (Eds.). (1998). Bibliography of Translated Korean Literature. Seoul: Korea Literature Translation Fund & Institute of Korean Culture at Korea University.

Kwak, Hyo-Hwan (2022). Current status and prospects of Korean literature as world literature. The Journal of Literary Creative Writing, 21(1), 13-40. https://doi.org/10.47057/jklcw.2022.54.01

Lee, Jongho (2022). The formation of the Korean literary translation field & the desire for world literature: focusing on Korea Journal published by the Korean National Commission for UNESCO. The Korean Association of Kubo Studies, 32, 289-330. https://doi.org/10.35153/gubokr.2022..32.007

Lee, Yong-Gu (2013). A study on feature selection for kNN classifier using document frequency and collection frequency. Journal of Korean Library and Information Science Society, 44(1), 27-47. https://doi.org/10.16981/kliss.44.1.201303.27

Lee, Yong-Gu (2023). A study on automatic classification of subject headings using BERT model. Journal of the Korean Society for Library and Information Science, 57(2), 435-452. https://doi.org/10.4275/KSLIS.2023.57.2.435

Lee, You-Sik (2000). English translation of Korean literature and its introduction to English-block countries: chiefly on present conditions and problems. The Journal of Translation Studies, 1(1), 169-202.

Literature Translation Institute of Korea (2023). Operational Regulations of the LTI Korea Library.

Literature Translation Institute of Korea (2024). 2023 Annual Report.

Park, Dahae (2016, June 9). Deborah Smith: “Let’s not use the term ‘K-literature’.” Money Today. Available: https://news.mt.co.kr/mtview.php?no=2016061910172102314

Park, Eunhee & Lee, Mihwa (2020). A study on genre classification for fictions in school libraries. Journal of the Korean Biblia Society for Library and Information Science, 31(1), 115-136. https://doi.org/10.14699/kbiblia.2020.31.1.115

Rho, Jee-Hyun (2010). The meanings of genre classification in library classification: the case of American public libraries. Journal of Korean Library and Information Science Society, 41(4), 151-170. https://doi.org/10.16981/kliss.41.4.201012.151

Rho, Jee-Hyun (2015). A study on the quality of subject data in library catalogs. Journal of Korean Library and Information Science Society, 46(4), 379-402. https://doi.org/10.16981/kliss.46.4.201512.379

Shin, Ji-Sun (2023). Evolution of Korea’s representation in translated Korean literature. The Journal of Interpretation and Translation Education, 21(2), 119-135. https://doi.org/10.23903/kaited.2023.21.2.006

Yoo, Geonsu, Kim, Bokyung, Kim, Jiyoon, Jeon, Se-jin, Chong, Ki In, Jung, Seong-hoon, & Chandler, S. (2023). Reading Korean literature in English translation from a distance (1): Modern literature. Journal of Modern Korean Literature, 24(1), 7-42.

Yoon, Ji-Kwan (2006, August 08). Globalizing Korean literature: What should be translated? The Quarterly Changbi. Available: https://magazine.changbi.com/MCWC/WeeklyItem?id=20