기술과학 분야 학술문헌에 대한 학습집합 반자동 구축 및 자동 분류 통합 연구*

Semi-automatic Construction of Learning Set and Integration of Automatic Classification for Academic Literature in Technical Sciences

김선우 (Seon-Wu Kim)**, 고건우 (Gun-Woo Ko)***
최원준 (Won-Jun Choi)****, 정희석 (Hee-Seok Jeong)*****
윤화묵 (Hwa-Mook Yoon)******, 최성필 (Sung-Pil Choi)*******

초 록

최근 학술문헌의 양이 급증하고, 융복합적인 연구가 활발히 이뤄지면서 연구자들은 선행 연구에 대한 동향 분석에 어려움을 겪고 있다. 이를 해결하기 위해 우선적으로 학술논문 단위의 분류 정보가 필요하지만 국내에는 이러한 정보가 제공되는 학술 데이터베이스가 존재하지 않는다. 이에 본 연구에서는 국내 학술문헌에 대해 다중 분류가 가능한 자동 분류 시스템을 제안한다. 먼저 한국어로 기술된 기술과학 분야의 학술문헌을 수집하고 K-Means 클러스터링 기법을 활용하여 DDC 600번 대의 중분류에 맞게 매핑하여 다중 분류가 가능한 학습집합을 구축하였다. 학습집합 구축 결과, 메타데이터가 존재하지 않는 값을 제외한 총 63,915건의 한국어 기술과학 분야의 자동 분류 학습집합이 구축되었다. 이를 활용하여 심층학습 기반의 학술문헌 자동 분류 엔진을 구현하고 학습하였다. 객관적인 검증을 위해 수작업 구축한 실험집합을 통한 실험 결과, 다중 분류에 대해 78.32%의 정확도와 72.45%의 F1 성능을 얻었다.

ABSTRACT

Recently, as the amount of academic literature has increased rapidly and complex researches have been actively conducted, researchers have difficulty in analyzing trends in previous research. In order to solve this problem, it is necessary to classify information in units of academic papers. However, in Korea, there is no academic database in which such information is provided. In this paper, we propose an automatic classification system that can classify domestic academic literature into multiple classes. To this end, first, academic documents in the technical science field described in Korean were collected and mapped according to class 600 of the DDC by using K-Means clustering technique to construct a learning set capable of multiple classification. As a result of the construction of the training set, 63,915 documents in the Korean technical science field were established except for the values in which metadata does not exist. Using this training set, we implemented and learned the automatic classification engine of academic documents based on deep learning. Experimental results obtained by hand-built experimental set-up showed 78.32% accuracy and 72.45% F1 performance for multiple classification.

키워드: 자동 분류, 텍스트 마이닝, 자연어 처리, 심층학습, 준지도 학습
automatic classification, text mining, NLP(Natural Language Processing), deep learning, semi-supervised learning

---

* 본 연구는 2018년도 한국과학기술정보연구원(KISTI) 주요사업 과제로 수행한 것임(K-18-L11-C01-S01). 이 논문은 2017년도 정부(과학기술정보통신부)의 재원으로 한국연구재단 기초연구사업의 지원을 받아 수행된 연구임(No. 2017M3C4A7068188).
** 경기대학교 문헌정보학과 석사과정(kimsw@kgu.ac.kr) (제1저자)
*** 경기대학교 문헌정보학과 석사과정(zellyshu@kgu.ac.kr) (공동저자)
**** 한국과학기술정보연구원 콘텐츠큐레이션센터 연구원(cwj@kisti.re.kr) (공동저자)
***** 한국과학기술정보연구원 콘텐츠큐레이션센터 선임연구원(hsjeong@kisti.re.kr) (공동저자)
****** 한국과학기술정보연구원 콘텐츠큐레이션센터 책임연구원(hmyoon@kisti.re.kr) (공동저자)
******* 경기대학교 문헌정보학과 조교수(spchoi@kgu.ac.kr) (교신저자)
■ 논문접수일자 : 2018년 11월 17일 ■ 최초심사일자 : 2018년 12월 10일 ■ 게재확정일자 : 2018년 12월 17일
■ 정보관리학회지, 35(4), 141-164, 2018. 〔http://dx.doi.org/10.3743/KOSIM.2018.35.4.141〕

1. 서 론

최근 전 세계적으로 학술정보의 생산 및 유통이 폭발적으로 증가하였다. 이에 따라 연구자들은 연구 동향을 파악하기에 어려움을 겪고 있다. 이에 학문분야별로 세부적인 연구의 양상을 구체적으로 파악하기 위해서는 개별 자료의 분류 정보가 필수적이다. 그러나 이러한 정보가 기본 항목으로 제공되는 해외 학술데이터베이스와 달리 국내에서는 학술지와 학술문헌에 대한 분류 정보가 제대로 제공되지 않고 있다(김판준, 이재윤, 2014). 국내의 학술문헌에 대한 분류 정보를 포함하고 있는 학술 데이터베이스가 필요한 시점이지만, 대규모의 학술문헌집단의 모든 문헌을 특정 시점에서 일괄적으로 수작업 분류하는 것은 막대한 시간과 인력, 비용이 소요되므로 사실상 불가능하다(김판준, 2018).

이러한 동향에 따라, 효율적인 대안으로서 문헌에 대한 자동 분류 연구의 중요성은 확대되고 있지만, 국내에서는 학술문헌에 대한 자동 분류 시스템을 구성하기 위한 기초 데이터가 부족한 실정이다. 공개적으로 배포되어 있는 학술문헌 자동 분류 학습집합이 존재하지 않기 때문에 학습집합의 구축이 우선적이다. 그러나 이를 수작업으로 구축하기에는 비효율적이며, 분류 작업과 마찬가지로 시간과 인력, 비용이 소모된다.

한편, 4차 산업혁명 시대의 도래에 따라, 융복합적인 연구가 활발히 진행되며 다중적인 주제 분야를 가진 학술문헌이 적극적으로 생산되기 시작하였다. 그러나 일반적으로 활용되는 표준분류체계를 활용한 분류는 단일 분류를 기준으로 한다. 다중적인 주제 분야의 경우, 별도의 기준을 두어 처리하는 경우가 일반적이다. 융복합적인 연구가 지속적으로 증가하고 있는 현재의 연구 동향을 고려하였을 때, 학술문헌에 대한 단일 주제 분류는 점진적으로 많은 전문적인 지식을 요구하게 될 것이며, 다중 분류의 경우도 별도의 기준을 정함에 있어 보다 높은 전문지식을 요구하게 될 것이다.

본 연구에서는 국내 학술문헌에 대한 자동 분류 시스템을 구현하기 위해, 우선적으로 반자동적으로 학습집합을 구축한다. 이 과정에서, 융복합적인 현재의 연구 동향에 맞게 유연한 분류가 가능한 학습집합을 구성하기 위한 다중 분류 기반의 클러스터링 기법을 활용한다. 구축한 학습집합의 의도에 적합하도록 심층학습 기술을 기반한 학술문헌 자동 분류 시스템을 구현한다. 제안하는 시스템은 다중 분류와 단일 분류에 대해 나누어 구현되며, 각 성능을 측정하고 분류할 수 있다.

2. 관련 연구

학술문헌에 대한 자동 분류 문제는 자동 분류 연구의 한 부분이다. 문헌에 대한 자동 분류 연구는 활발히 이뤄지고 있다. 문헌에 대한 자동 분류 연구는 일반적으로 지도 학습과 비지도 학습 기반의 연구로 나뉜다. 지도 학습 기반의 문헌 자동 분류 연구로는 먼저 이용구(2013)의 문헌빈도와 장서빈도를 활용한 kNN(k-nearest neighbors) 분류기의 자질 선정 연구가 있다. 그는 장서빈도와 문헌빈도의 두 자질 정보를 활용하여, 자질 선정 범위를 바꿨을 때의 효과

적인 방법에 대한 연구를 수행하여 우수한 성능을 보였다. 조현양(2017)은 책 소개 정보를 통한 성격 유형별 도서 자동 분류 시스템을 연구하여 선형 커널 기반 SVM보다 비선형 분류를 지원하는 LibSVM(A Library for Support Vector Machines) 모델이 우수한 성능을 보임을 입증하였다. 최근 심층학습(deep-learning) 기술의 등장으로, 국내외로 문헌 자동 분류에 대한 적용 연구가 적극적으로 이뤄지고 있다. 조휘열, 김진화, 윤상웅, 김경민, 장병탁(2015)은 대용량 텍스트 분류 모델을 CNN(Convolutional Neural Networks) 기반으로 구성하여 우수한 성능을 보인바 있으며, Kowsari, Brown, Heidarysafa, Meimandi, Gerber, Barnes(2017)는 심층학습 모델 구조를 다중으로 쌓아 계측적인 문헌 분류가 가능한 문헌 분류 모델을 구현하였다.

비지도 학습 기반의 문서 분류 연구로는 노대욱, 이수용, 나동열(2007)의 비지도 학습 기반의 문서 분류 시스템 개발 연구가 있다. 원시 말뭉치를 통한 비지도 학습 기법을 활용하여, 중심 단어와 클러스터에 대한 부트스트래핑 기법을 활용하여 가중치를 조정하고 대표 벡터를 생성하여 문서를 분류하는 연구를 수행하였다. 한규열, 안영민(2013)의 LDA(Latent Dirichlet Allocation) 기반 한국어 문서 클러스터에 대한 자동 제목 생성 연구를 수행하여, 토픽 모델링 방법인 LDA를 활용하여 디지털 문서에 대한 주제 분야(제목)을 자동으로 생성하는 모델을 구현하여 후보군을 정할 수 있도록 적용하였다. 박영근, 박수빈, 박노일, 이현아(2017)는 웹 뉴스 분류를 위하여 K-Means 알고리즘을 활용하여 웹 뉴스 내의 내포된 정보를 기반하여 분류를 수행할 수 있는 모델을 구현하였다.

앞서 설명한 개별적인 연구 외에 비지도 학습과 지도 학습에 대한 통합적인 연구는 드물게 존재하며, Shafiabady, Lee, Rajkumar, Kallimani, Akram, Isa(2017)은 비지도 클러스터 기법을 통하여 구성한 학습집합을 활용하여 SVM 모델을 학습하여 문헌 분류기를 구축한 바있다. 국내에서는 육지희, 송민(2018)이 LDA 토픽 모델링을 통한 자질 집합 탐색과 심층학습 기법을 이용한 분류 성능을 측정하는 생의학 분야의 문헌 자동 분류 연구를 수행하였다.

문헌 자동 분류 연구가 활발한 것에 비하여 학술문헌을 대상으로 한 연구는 비교적 소극적으로 이루어지고 있는데, 특히 국내의 학술문헌 자동 분류 연구는 몇몇 연구자의 주도로 지도 학습 기반의 연구가 진행되고 있다. 김판준(2018)은 문헌정보학 분야의 국내 학술문헌집합을 대상으로 기계학습에 기초한 자동분류의 성능에 영향을 미치는 요소들을 검토하였고, 용어 가중치부여 기법, 학습집합 크기, 분류 알고리즘, 범주 할당 방법 등 주요 요소들의 특성을 다각적인 실험을 통해 살펴보았다. 그 결과 국내 학술문헌의 자동분류에 적합한 분류모형을 제안하였다. 자동 분류의 성능을 높이기 위해서 김판준, 이재윤(2014)은 현재 국내 데이터베이스에서 대부분의 학술문헌에 통제키워드가 부여되어 있지 않아 디스크립터의 제공이 어렵다는 것을 판단하고, 해외 데이터베이스의 학술문헌에 부여된 통제키워드를 학습한 분류기를 사용하여 국내 학술지 논문에 디스크립터를 자동 할당하여 분류기 성능을 높였다. 이 외에도 나동열, 김윤식, 신현주, 이규희, 김태규, 강현규, 최호섭, 윤화묵(2007)은 한국어 문서분류 시스템의 개발을 위해 한국어 문서분류 테스트컬렉

션의 구축을 진행하였고, 이를 위하여 한국과학기술연구원에서 선행적으로 구축한 기본적인 컬렉션인 KRTC의 개선과 재구축 작업에 필요한 요소들에 대하여 기술하였으며, 나동열, 강현규, 김현태, 박경일, 장형일, 염성욱, 김윤식, 김태규, 이규희, 신현주(2007)은 기 배포된 HANTEC 2.0 테스트 컬렉션의 12만건 문서 중 2만건을 대상으로 해당 질의에 대한 적합성 정보의 수동 판정을 행함으로써 적합성을 평가하였고, 기존의 문서분류 테스트컬렉션 KRTC의 범주체계 및 문서 범주 태그의 정제작업을 수행하여 보다 고품질의 한국어 문서분류 테스트 컬렉션의 구축을 진행하였다.

국내에서 반자동 학습집합 구축과 관련한 연구는 그 양이 매우 적다. 최성필, 유석종, 조현양(2016)의 생의학 분야에 대한 분야별 관계 추출 데이터셋을 반자동 구축한 연구를 진행하였는데, 이는 대상 분야의 용어집을 기반으로 용어 간의 연관관계를 생성하고, 연관관계 집합을 데이터베이스에 검색하여 문장을 추출하는 형태의 반자동 기법을 활용한 연구이다. 그러나 문헌 자동 분류에 대한 학습집합 반자동 구축 방법론을 다룬 국내 연구는 거의 존재하지 않는다.

이러한 많은 관련 연구에도 불구하고, 학술문헌 자동 분류를 위한 반자동 학습 구축과 자동 분류 모델을 통합적으로 구현한 연구는 현재 국내에는 존재하지 않는다. 이에 본 연구에서는 학술문헌 자동 분류를 위한 반자동 학습 구축 과정에서 비지도 학습 기반 방법론을 활용하고, 구축한 학습집합을 통해 학술문헌에 대한 자동 분류가 가능한 지도 학습 기반 모델을 구현한다.

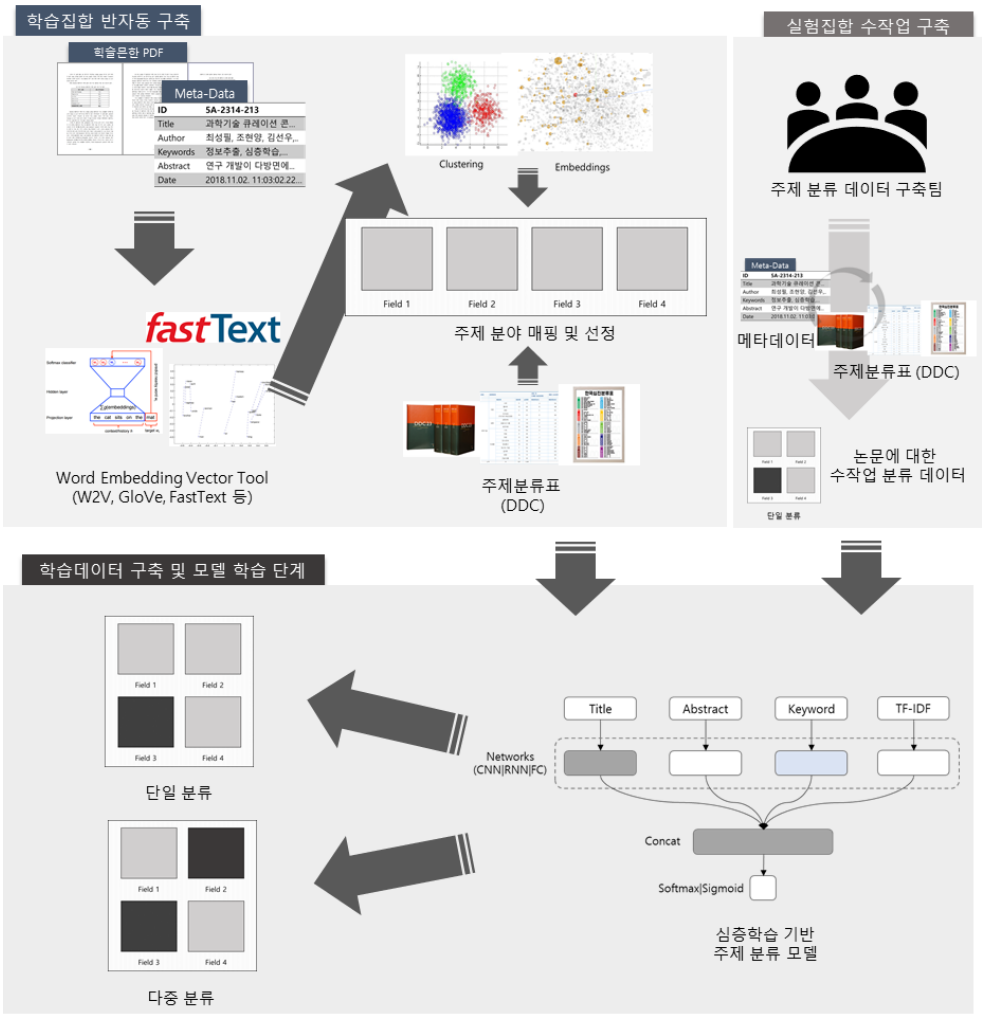
3. 학습집합 반자동 구축 및 자동 분류 방법론

본 연구에서는 정답 학습집합이 없는 상황에서 융복합적인 연구 동향에 적합한 학술문헌 자동 분류를 위해, 비지도 학습 기반의 반자동 학습집합 구축 방법론을 제시하며, 구축한 학습집합을 활용하여 학습한 심층학습 기반의 학술문헌 자동 분류 모델을 제안한다. 〈그림 1〉은 연구 과정 전반을 간략하게 표현한 것이다.

이 장에서는 학습집합 구축을 위한 대상 데이터 범위 선정, 비지도 학습 기반의 반자동 학습집합 구축 과정, 수작업 구축 실험집합 구축 과정, 심층학습 기반의 학술문헌 자동 분류 모델을 각각 나누어 설명한다.

3.1 대상 데이터 범위 선정

본 연구에서는 학술문헌 자동 분류를 위해 2010년 이후의 한국어 기반의 메타데이터가 존재하는 학술지를 수집하였다. 수집 과정에서는 한국과학기술정보연구원(KISTI)의 데이터베이스를 제공 받았으며, 총 718개의 학술지를 통해 144,786건의 학술논문을 수집하였다. 본 연구에서는 비지도 학습 기반의 학습집합 구축 과정에 대한 성능을 고려하여 원문이 한국어로 기술된 학술문헌을 대상으로 범위를 한정하였다. 또한, 수집한 학술지의 분야 분포를 고려하여 대상 분야를 기술과학 분야로 선정하였다. 이후, 기술과학 분야의 학술지와 학술논문을 도출하기 위해 공학, 의약학, 이학과 기타 분야의 복합학에서 각각 공학의 주제를 가지는 학술지만을 별도로 추출하였다.



〈그림 1〉 기술과학 분야 학술문헌에 대한 자동 분류 통합 연구 과정

KISS는 전 주제 분야 총 1,785종/등재지 901종(KCI 등재지 716종, KCI후보지 115종, WOS 27종, SCOPUS 43종) 학술지의 원문제공 서비스이다. KISS의 주제별 간행물 중 DDC 분류 체계에 따른 기술과학 전체(농학, 가정학 제외)의 학술지명을 참고한 뒤 학회명, 학술지, 권-호 정보만 별도로 처리하였다. 그리고 KISS에서 참고한 학술지명과 앞서 추출한 메타데이터의 학술지명이 서로 동일하지 않은 문제가 존재하였는데, 동일하지 않은 학술지명을 하나로 통합해주는 형태로 해결하였다. 서지정보 메타데이터에서 공학을 제외한 복합학, 의약학, 이학과 KISS의 학술지명을 함께 참고한 뒤 전처리한 데이터를 서로 비교하여 기술과학 분야의 공학 학술문헌만을 추출하였다. 추출한 결과는 〈표 1〉과 같다.

〈표 1〉 수집한 학술문헌과 공학 분야 추출 학술문헌 통계

분야	전체 수집 데이터		추출 대상 데이터	
	학술문헌	학술지	공학 분야 학술문헌	학술지
공학	78,924	312	78,924	312
의약학	24,221	191	1,502	8
이학	21,445	121	2,647	12
복합학	6,530	23	169	1
농수해	13,265	63	0	0
사회	368	6	0	0
인문	33	2	0	0
총	144,786	718	83,242	333

이후, 학습집합 구축 및 심층학습 엔진의 성능을 고려하여 추출한 333종의 학술지 내의 83,242건의 학술논문에 대해, 원문이 한국어로 기술된 학술논문을 선정하였다. 선정 결과는 〈표 2〉와 같다.

〈표 2〉 공학 분야 추출 학술논문의 언어 분포

언어	건
영문	14,770
중문	2
일문	1
한글	**68,469**
전체	83,242

68,469건의 데이터는 학습 데이터와 실험 데이터로 각각 나누었으며, 모델의 객관적인 검증을 위해 전체 데이터의 약 5%에 해당하는 3,400건을 수작업으로 분류하여 실험 데이터로 활용하였고, 나머지 95%인 65,069건을 학습 데이터로 활용하였다. 향후 학습 데이터의 양을 늘리는 것은 물론 검증 데이터의 양도 10% 이상으로 하여 더욱 높은 신뢰도를 얻고자 한다.

3.2 비지도 학습 기반 반자동 학습집합 구축

본 연구에서는 65,069건의 학술문헌 PDF 내에서 Python PDF 라이브러리 중 하나인 PDFMiner를 활용하여, 텍스트 추출 작업을 수행하였다. 텍스트 추출 전문을 기준으로 자체 구축한 한국어 품사 태거(김선우, 최성필, 2018)를 활용하여, 품사 태깅을 수행하고 명사만을 추출하였다. 클러스터링의 입력자질로서 활용하기 위한 추출된 명사 데이터를 기준하여 다양한 방법의 임베딩 벡터를 구성하였다.

단어 임베딩은 컴퓨터가 자연어를 인식할 수 있도록 자연어를 단어 단위로 수치화하는 방법이다. 이는 개체명 인식, 형태소 분석, 품사 태깅, 문장 임베딩 등 다양한 자연어 처리 과제에 활용되고 있다(조현수, 이상구, 2017). 단어 임베딩 생성 방법에는 다양한 방법론들이 존재하지만 가장 많이 활용되는 방법론으로는 Word2Vec의 Skip-gram와 CBOW(Mikolov, Sutskever, Chen, Corrado, & Dean, 2013), GloVe(Pennington, Socher, & Manning, 2014), FastText(Bojanowski,

Grave, Joulin, & Mikolov, 2016: Joulin, Grave Bojanowski, & Mikolov, 2016) 등이 존재한다.

이다빈, 최성필(2018)의 비교 분석 연구를 통해 가장 성능이 좋은 모델로 확인된 FastText 방법론을 활용하여 단어 임베딩을 생성하였고, FastText 모델의 성능을 해당 데이터로 다시 확인하기 위해 FastText 다음으로 성능이 좋은 Word2Vec의 Skip-gram을 적용하여 생성하였다. 이때 변수 설정은 FastText로 생성한 단어 임베딩 중 가장 성능이 좋은 것과 동일하게 설정하였다.

단어 인베딩 성능 평가에는 단어 짝 간의 유사도를 사람들이 평가한 평균값과 단어 임베딩의 유사도 값을 상관분석(correlation analysis)하는 단어 유사도(word similarity) 방법을 사용하였다. 이는 결과 값이 클수록 더욱 두 값과의 유사도가 높다는 것을 보여준다. 단어 유사도 평가에서 사용되는 다양한 평가 테스트 셋 중 가장 활발하게 사용되는 WordSim353을 한국어로 번역하여 사용하였다(이다빈, 최성필, 2018). 도출된 단어 임베딩 성능 값은 백분위로 표현하였고, 소수점 셋째 자리에서 반올림하였다.

〈표 3〉은 임베딩 벡터의 성능 실험 결과를 나타내는 표로 〈표 3〉의 결과를 통해 확인할 수 있듯이 단어 최소등장 횟수를 12로 설정하고 생성한 단어 임베딩이 62.50으로 가장 성능이 높았다. 본 연구에서는 이를 기준하여, 해당 임베딩 벡터를 활용한 클러스터링을 수행하기로 결정하였다. 구성된 임베딩 벡터는 각 명사에 대하여 300개의 자질 벡터가 구성된 형태이며, 해당 벡터는 곧 각 명사를 설명하는 값이 된다. 본 연구에서는 이를 통하여 K-Means 알고리즘을 활용한 클러스터링을 수행하였다. K-Means 알고리즘은 가장 기본적이고 일반적으로 활용되는 클러스터링 알고리즘으로, k개로 설정한 기준 점을 설정하고, 이에 유클리드 거리를 계산하여 가까운 값을 군집으로 이루어 클러스터링을 수행하는 알고리즘이다(Bock, 2007). K-Means 클러스터링 과정에서, 총 클러스터의 수를 정하기 위한 k값 선정은 최가람, 최성필(2018)의 연구 방법을 참고하였다. 50~300의 범주 내에서 50 단위로 늘려가며 클러스터 결과를 연구자들 간의 질적 평가를 수행하였으며, 결과적으로 k값을 200으로 선정하였다. 클러스터 결과는 클러스터 내의 중심에 가까운 명사 순으로 정렬하였다. 〈표 4〉는 클러스터 결과에 대한 예시이다.

각 클러스터 결과는 DDC(듀이 십진 분류표) 주제분류표에 대한 매핑 사전을 구축하였다. 주제분류표 선정은 DDC 외에도 연구재단

〈표 3〉 임베딩 벡터 성능 실험 결과

생성 모델	차원	윈도우 사이즈	학습반복 횟수	단어최소 등장 횟수	성능(%)
FastText	300	5	50	5	62.39
FastText	300	5	50	7	60.54
FastText	300	5	50	10	60.79
FastText	300	5	50	12	62.50
Word2Vec	300	5	50	12	55.58

〈표 4〉 K-Means 클러스터링 결과 예시

클러스터	명사 리스트
0	건강 여자 수면 성인 남자 비만 혈압 음주 흡연 습관 검진 당뇨병 유병 중년 우울증 금연 당뇨 건강관리 심혈관 …
1	엔진 연소 액체 화염 노즐 분사 배기 분무 액적 디젤 로켓 연소기 산화제 점화 연소실 제트 추진공 버너 …
2	만족도 세 연령 미만 성별 소득 남성 학력 나이 남녀 기혼 연령대 가계 주부 더미 고졸 대졸 미혼 주말 …
3	수산 어업 수산물 어선 조업 어획 양식장 어장 어구 어초 어획량 낚시 종묘 어항 그물 수산업 고등어 …
4	년 자료 조사 평균 비율 단위 지수 나라 규모 지표 면적 빈도 현황 건 차지 인구 순위 등급 비중 …
5	강도 하중 변형 응력 파괴 전단 균열 시편 탄성 인장 피로 충격 소성 파손 연성 파단 시험편 층간 시험기 …
6	건설 공사 프로젝트 리스크 자재 물량 공사비 공종 발주 입찰 용역 발주자 감리 준공 건설업 수주 건설업체 …
7	가열 냉동 두부 시판 부패 열풍 냉장 공전 신선 염도 해동 쇠고기 풍미 닭고기 돈육 미생물학 식육 소시지 치즈 신선도 …
8	통신 채널 무선 수신 할당 간섭 링크 송신 대역폭 수신기 통신망 기지국 송수신 상향 심볼 하향 중계 단말 송신기 …
9	공정 가공 성형 이송 금형 연마 절삭 다이 공구 단조 압출 사출 판재 정밀공 드로잉 프레스 밀링 학회지제 펀치 공작 …

의 학문분류체계와 과학기술표준분류 등의 학술지에 적합한 다른 분류 체계 등이 논의되었다. 그 중에서 참고하였던 KISS의 학술 정보와의 연관성, 분류값의 범위 등을 고려하여 DDC를 선정하였다. DDC 분류 중에서도 학술지의 비율과 주제 분류에 적합성을 고려하여, 23판의 이공학 및 기술과학에 대한 주제 분야인 600번 대의 중분류(600~690의 10 단위 분류)로 범위를 축소하여 활용하였다. 실험 데이터를 수작업으로 구축한 3인의 주제 분류 학습집합 구축 연구원이 논의하여 클러스터의 매핑 결과를 결정하였으며, 실제 분류를 수행한 600번 대의 중분류는 〈표 5〉와 같다.

각 클러스터는 주제 분류 경험이 있는 교수급 1인, 석사 2인, 학부생 2인, 석사 이상의 연구원인 외부 전문가 3인이 회의를 통해 기준을 정립하고, 석사 1인과 학부생 2인이 DDC 23판의 600번 대의 중분류에 맞게 매핑 작업을 거쳤

〈표 5〉 DDC 23판의 600번 중분류

대분류	중분류	분류명
600 기술(Technology)	600	기술과학(Technology)
	610	의학(Medicine & Health)
	620	공학(Engineering)
	630	농학(Agriculture)
	640	가정학(Home & Family Management)
	650	경영관리 및 보조 서비스(Management & Relation)
	660	화학공학(Chemical Engineering)
	670	제조업(Manufacturing)
	680	제조업(Manufacture for Specific uses)
	690	건축공학(Construction of Building)

다. 클러스터 매핑 작업의 신뢰도를 높이기 위하여 주마다의 회의를 통해 기준을 재정립하였고, 재정립한 기준을 바탕으로 클러스터 매핑 재작업을 거쳤다. 〈표 6〉은 분류자 3인의 일치와 불일치에 대한 통계를 범주별로 나타낸 표이다.

〈표 6〉 분류자 3인의 일치와 불일치에 대한 통계별 범주

일치 개수	Count	%
3개 일치	69	34.5%
2개 일치	94	47%
1개 일치	13	6.5%
분류 제외	24	12%
계	200	100%

〈표 6〉에서 '분류 제외'는 클러스터가 학회지만을 나열하는 경우, 이름, 지명, 회사명 등의 일반명사만을 나열하는 경우, 공학과 거리가 있는 전문 용어를 나열한 경우의 클러스터를 나타낸다. 분류자 3인의 결과가 모두 달랐던 13개의 '1개 일치'에 대하여 별도의 회의를 진행하였고, 기준을 재정립하여 1개 일치 데이터에 대한 재분류를 진행하였다. 그 과정에서, 〈표 4〉의 클러스터 4와 같이 일반적인 연구 전반에 활용되어 분류 자체에 대한 특징성이 적은 클러스터는 이에서 제외하였다. 결과적으로, 200개의 클러스터 중, 63개의 클러스터가 제외 되었으며 137개의 클러스터가 매핑되었다. 제외된 클러스터는 대체로 연구 전반에 흔히 쓰이는 용어들, "초록", "키워드" 등의 메타데이터에 대한 필드명, 인명 등이 해당되었다. 〈표 7〉은 150개의 클러스터의 매핑 결과이다.

〈표 7〉 각 클러스터의 DDC 매핑 결과

DDC	클러스터 넘버	개수
600	16 58	2
610	0 10 21 25 31 33 35 47 53 54 78 86 89 98 114 128 153 155 159 177 187 188 190	23
620	1 5 8 13 14 17 18 19 20 22 26 28 30 34 36 39 40 44 45 46 48 52 61 63 65 69 70 82 83 94 97 101 103 104 108 110 112 115 116 117 118 120 124 126 130 133 139 140 141 142 148 149 152 154 157 158 160 161 164 167 168 173 175 176 180 183 191 192 193 199	70
630	3 51 68 71 76 80 91 129 137 169 178 181 182 185 194	15
640	2 7 32 43 73 113 147 150 151 171 179	11
650	37 42 64 106 119 146	6
660	23 27 57 127 172	5
670	12	1
680		0
690	6 62 123 131	4
제외	4 9 11 12 15 24 29 38 41 49 50 55 56 59 60 66 67 72 74 75 77 79 81 84 87 85 88 90 92 93 95 96 99 100 102 105 107 109 111 122 125 132 134 135 136 138 143 144 145 156 162 163 165 166 170 174 184 186 189 195 196 197 198	63

〈표 7〉의 클러스터 매핑 결과를 통해 알 수 있듯이, DDC 620 분류에 대한 클러스터 편중 현상이 발생하였는데 이는 620이 "공학" 분야로, 보유한 데이터에 대한 연관성이 가장 높은 편이기 때문으로 보인다. 그 외에는 600, 670, 680의 클러스터가 각각 2개, 1개, 0개로 현저히 적거나 없는 경우도 존재한다. 이는 600이 전반 기술 과학으로서 기술철학, 기술심리학 등의 기초적이고 근본적인 내용을 다루는데, 최근 이러한 연구가 부족하기 때문에 발생한 형상으로 보인다. 한편, 670과 680의 경우, 670이 제조업 분야의 분류인데, 680의 경우 특정 용도 제조업으로 서로의 분야가 겹치며, 관련 연구에 대한 클러스터도 1개 밖에 존재하지 않았는데, 이 경우 범위가 조금 더 넓고 일반적인 670으로 분류하는 것이 옳다 판단한 것이다. 결과적으로, 학습 데이터는 DDC 23판의 600번대의 중분류를 소화하면서도 680 분류를 제외한 9종의 분류값을 갖게 되었다.

이후, 본 연구에서는 수작업 구축 학습집합으로부터 실험데이터로 활용하기 위해 무작위로 선정한 학술문헌 3,400건을 제외하고, 각 학술문헌에 대한 제목, 초록, 키워드를 추출하였다. 클러스터링 과정과 마찬가지로, 자체 구축한 한국어 품사 태거 ICLKT(김선우, 최성필, 2018)를 활용하여, 추출한 제목, 초록, 키워드에 대한 명사 리스트를 추출하였다. 추출한 명사 리스트를 활용하여, TF-IDF 계산을 통한 키워드 가중치 부여 작업을 진행하였다. TF-IDF란, 문서 내의 단어 빈도(Term Frequency)와 전체 문서 내의 역문헌 빈도(Inverse Document Frequency)를 계수하여, 이를 곱하는 방식의 가장 일반적으로 활용되는 키워드 가중치 부여 알고리즘이다. 명사 리스트에 대한 가중치 부여 과정에서, 메타데이터의 특징에 따라 제목과 키워드의 중요성이 높다고 판단하여 가중치를 초록에 비하여 1.5배 높게 활용하였으며, TF-IDF 가중치를 통하여 랭킹된 명사들 중에서 상위 10개를 추출하고, 이를 메타데이터 내의 키워드에 대한 명사 추출 리스트와 함께 클러스터 및 DDC 매핑에 활용하였다.

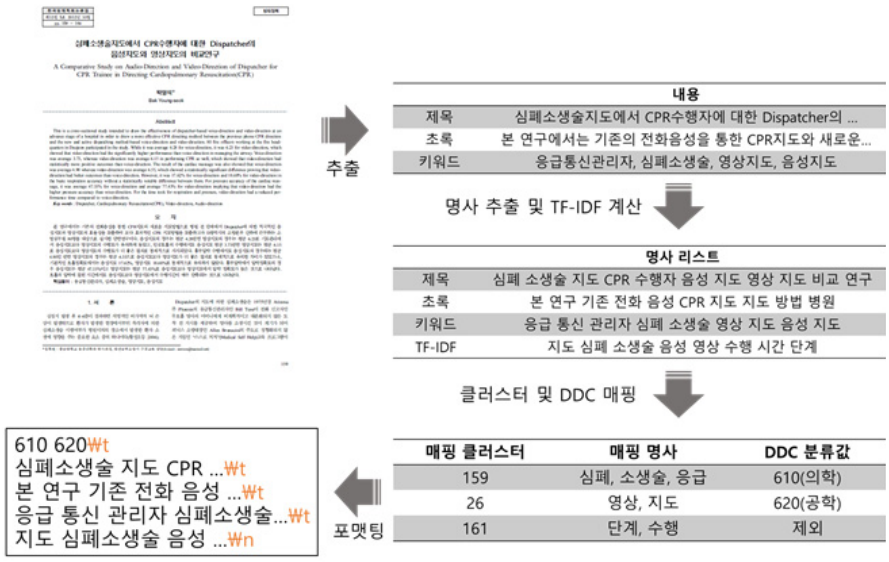
준비된 학술문헌별 추출 명사 리스트, 클러스터, 클러스터 DDC 매핑 사전을 활용한 학습데이터 구축은 다음과 같다. 먼저 학술문헌별 추출 명사 리스트를 클러스터 내의 키워드와 매핑하였다. 매핑 과정에서 키워드 추출 명사 리스트와 TF-IDF 적용 명사 리스트를 별도로 매핑한 후, TF-IDF 적용 명사 리스트의 매핑 결과에 가중치를 1.5배 추가 부여하고 이를 합산하여 상위 클러스터 종을 추출하였다. 각 클러스터는 DDC 매핑 사전을 통하여 DDC 분류에 매핑되어 학술문헌을 DDC 분류에 맞게 매핑할 수 있었다. 이 과정에서, 클러스터가 더 많이 매핑된 DDC 분류에 학습데이터를 분류하였다. 만일 매핑된 클러스터의 수가 같을 경우에는 클러스터에 매핑된 각 명사가 클러스터 중심점으로부터 가까운지를 판단하여 더 클러스터에 가까운 명사가 있는 쪽에 DDC 분류를 적용하였다. 중심점에 대한 거리는 〈표 4〉의 명사 정렬을 활용하여, 그 위치를 기준으로 적용하였다. DDC 매핑은 최소 1개에서 최대 3개의 분류를 갖도록 다중 분류 형식으로 매핑하였으며, 그중 가중치가 가장 높은 것을 왼쪽에 부여하여 랭킹하였다. 단일 분류 형식으로 활용할 경우에는 가장 랭킹이 높은 DDC 분류를 활용하였다. 결과적으로 각 학술문헌을 매핑한 결과는 〈표 8〉과 같다.

〈표 8〉 학습집합 구축 결과

DDC 분류	다중 분류 학술문헌 수	단일 분류 학술문헌 수
600	533	219
610	8,264	5,435
620	51,231	48,066
630	4,671	1,082
640	4,949	1,724
650	10,073	4,829
660	4,486	1,369
670	2,266	387
680	0	0
690	4,747	804
Total	91,220	63,915

〈표 7〉과 비교해 보았을 때, 매핑된 클러스터 수에 맞게 대부분의 분류 학술문헌 분포가 일치하는 것을 확인할 수 있다. 매핑된 클러스터가 가장 많은 620의 경우, 다중 분류 시 51,231건, 단일 분류 시 48,066건의 학술문헌이 분류되어 가장 수가 많은 분류가 되었다. 한편, 680을 제외하고, 600과 670의 경우 매핑된 클러스터 적은 만큼, 매핑 된 학술문헌 수가 적은 것을 확인할 수 있다. 그러나 650의 경우, 매핑된 클러스터가 6개뿐임에도 불구하고 다른 분류에 비하여 다중 분류 시 10,073건, 단일 분류 시 4,829건의 많은 학술문헌이 분류된 것을 확인할 수 있다. 이는 대상 데이터인 학술문헌의 주요적인 주제가 공학적인 측면이 많기 때문으로 보인다.

이후, 각 학술문헌에 대해서 DDC 분류 정보, 제목, 초록, 키워드에서 추출한 명사들과 TF-IDF 기반 명사 추출 리스트를 열 하나에 탭(Tap)으로 나누어 출력하였다. 서술한 클러스터링 기술을 활용한 반자동 학습집합 구축 과정에 대한 클러스터, DDC 매핑 사전 등의 데이터 구성이 완료되면 각 학술문헌에 대한 매핑 작업을 수행한다. 개별 학술문헌이 학습집합의 인스턴스로서 구성되는 과정은 〈그림 2〉와 같다.



〈그림 2〉 학술문헌에 대한 학습집합 반자동 구축 과정

결과적으로 반자동 학습집합의 대상이 되는 65,069건의 학술문헌에 대하여, 한국어 메타데이터(제목, 초록, 키워드)가 하나도 존재하지 않았던 1,154건을 제외한 63,915건의 학술문헌 분류가 완료되었다. 한국어 제목, 초록, 키워드가 존재하지 않는 경우에는 클러스터 매핑 및 DDC 분류 매핑이 이뤄질 수 없기 때문에 학습집합으로서 활용할 수 없다.

3.3 실험집합 수작업 구축

본 연구에서는 실험 결과의 정확도를 위하여 총 68,469건의 학술문헌을 대상으로 구축한 학습집합 중 약 5%에 해당하는 3,400건을 실험집합으로 나누었고, 나머지에 해당하는 65,059건은 학습집합으로 나누었다. 실험집합은 기사 제목, 초록, 키워드 등의 학술문헌 서지 정보가 학술문헌 1건마다 테이블의 형태로 구성되어 있으며, 수작업 구축을 위하여 DDC 필드를 별도로 추가하였다. 수작업 구축은 3,400건의 실험집합을 동일하게 나눈 뒤 DDC(Dewey Decimal Classification) 23판 분류 체계에 맞추어 〈표 5〉의 분류표와 같이 600번대 중분류로 분류하였다. 실험집합 수작업 구축 작업은 앞서 클러스터 매핑을 진행한 인원 중 교수와 외부 전문가, 시스템 구축을 위한 석사급 1인을 제외한 가용인원인 석사급 1인, 학부생 2인이 총 4주간에 걸쳐 진행하였다. 매주 회의를 통하여 분류기준이 모호하거나 분류가 애매한 사항들의 기준을 재정립하는 과정을 통해 신뢰도를 높이고자 하였다. 예를 들어, 식물에 관한 학술문헌의 경우 610번대의 의학(Medicine & health)과 630번대의 농학(Agriculture)으로 분류가 가능하기 때문에 회의를 거쳐 의학적 키워드가 들어있는 경우에만 610번대로 분류하고 없는 경우에는 630번대로 분류하는 등의 기준을 별도로 정하였다. 수작업 구축은 기사 제목을 국립 중앙 도서관, KISS, DBPIA의 학술문헌 및 학회지 검색의 3가지 검증 절차를 통해 이 중 1가지라도 검색 결과가 존재하는 경우 1차적으로 기사 제목에 DDC 필드를 추가하였고, 검색 결과가 존재하지 않거나 결과가 DDC 23판 분류에 적합하지 않을 경우에는 기사 제목과 초록, 키워드 위주로 3인의 주제 분류 전문가가 직접 분류하였다. 모든 실험집합은 분류된 결과 값이 존재하여야 하므로, 분류 결과가 존재하지 않는 경우가 없도록 분류가 애매한 경우 주제 분류 전문가 3인이 추가적인 회의를 거쳐 별도로 재분류하였다.

〈표 9〉는 실험집합 수작업 구축에 대한 결과를 나타내는 표이다. 구축 결과를 살펴보면 620번대에 데이터가 과하게 치중되어 있는 경향이 보이는데 이는 620번대의 공학(Engineering)이 채광공학, 군사 및 항해공학, 토목공학, 도로공학, 수리공학 등 기타 공학분야 전반을 아우르기 때문이다. 반대로, 630번대 농학의 경우 분류된 학술문헌의 건수가 가장 적으며 이는 630번대로 분류가 가능한 농학 분야 특정 기술이나 설비와 같은 내용이 690번대의 건축공학(Construction of building)이나 620번대의 공학 등에도 분류가 가능하기에 최대한 농학이 학술문헌의 주가 되는 학술문헌만을 630번대의 농학으로 분류하였기 때문이다. 680번대의 제조업(Manufacture for specific uses)의 경우에도 다른 번대와 비교하여 상당히 낮은 비율로 분류된 것을 알 수 있다. 이는 670번대의 제조업(Manufacturing)

이 680번대의 제조업의 내용을 포함하는 부분이 많고, 680번대보다 일반적인 범위를 다루기 때문이다.

〈표 9〉 실험집합 수작업 구축 결과

DDC	건
600	194
610	152
620	2,612
630	8
640	67
650	49
660	139
670	33
680	15
690	131
Total	3,400

670번대의 제조업과 680번대의 제조업은 기준 간의 비슷함이나 범위의 좁고 넓음에 대한 수준을 정확히 정하기가 어렵다는 점에서 문제가 존재하였다. 670번대와 680번대를 나누는 기준은 제조업 중에서도 제조를 하는 것의 용도가 특정한 것이냐 아니냐는 것인데 이를 나누는 범위가 애매하기에 주제 분류 전문가 3인의 의견도 부분적으로 일치하지 않았다. 또한, 680번대 제조업의 경우 클러스터링 결과에는 존재하지 않기 때문에 학습 데이터 내에는 존재하지 않는 데이터이다. 그러나 실험에 대한 객관성과 공정성, 향후 연구에서 680번대의 데이터가 구성될 가능성 등을 고려해 680번대로 분류된 그대로 데이터를 활용하였다.

이 외에도 630번대 농학의 경우에는 다른 분류 값 데이터와 비교해보았을 때 매우 적은 비중을 보이고 있는 반면, 620번대 공학에 극도로 데이터가 치중되어 있다. 이러한 분포는 데이터의 과적합 현상 문제로 이어져 모델의 성능이 저하될 수 있다.

3.4 심층학습 기반 학술문헌 자동 분류 모델

본 연구에서는 구축한 비지도 학습 기반의 자동 분류 학습집합을 활용하여, 향후 입력될 학술문헌에 대한 주제 분류를 효과적으로 수행하기 위하여 지도 학습 기반의 학술문헌 자동 분류 모델을 구축하기로 하였다. 학술문헌 자동 분류 문제에 대해, 본 연구에서는 문헌 분류 과정으로 인식하고 접근하였다. 일반적으로 기계학습 기반의 문헌 분류 문제는 키워드 정보, 메타데이터 정보 등의 문헌 내의 다양한 자질 정보를 추출하여, 이를 SVM(Support Vector Machine) 등을 통하여 학습하고 분류하는 것이 일반적이다. 최근 심층학습(Deep-Learning) 기법이 다양하게 적용되면서, 문헌 분류에 대해서도 다양한 접근이 이뤄지고 있다. 현재 문헌 분류에 대해서는 다양한 심층학습 모델 구조가 활용되고 있는데, 일반적으로 활용되는 모델 구조는 다중필터 CNN(Convolutional Neural Networks), 양방향 LSTM(Long-Short Term Memory unit), 다중 층으로 쌓은 완전연결(Fully-Connected) 층 등이 있다. 이러한 각 구조는 문헌의 종류와 자질 특성에 따라 그 성능 차이가 유효하게 나타나는 특징이 있다. 본 연구에서는 각 학술문헌이 가지고 있는 정보를 제목, 초록, 키워드, TF-IDF 추출 명사 리스트

로 나누고, 각 자질 정보가 가지고 있는 특징에 따라 적합한 심층학습 구조가 있을 것이라 판단하고, 이를 경험적으로 선택하여 융복합적으로 활용할 수 있는 메타데이터 자동 분류 모델을 구현하였다. 구현은 Python의 심층학습 라이브러리인 Tensorflow(https://tensorflow.org/)를 활용하였다.

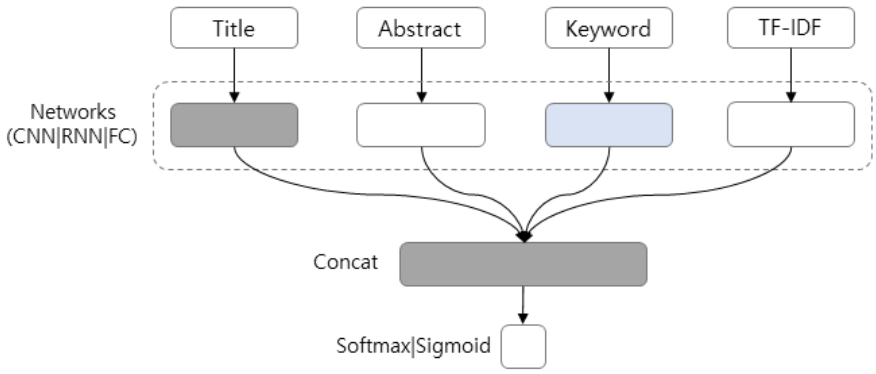
〈그림 3〉은 본 연구에서 제안하는 모델의 전반적인 구조이다. 각 자질 별로 임베딩 등의 벡터화 작업 이후, 적합한 심층학습 구조를 경험적으로 찾아 적용하여 자질 분석을 수행하며, 그 결과를 각각 연결하여 통합 분석을 수행하는 구조를 간략하게 표현하였다.

각 입력 자질 정보는 명사 리스트 형태이다. 이를 심층학습 모델 내부에서 활용하기 위해서는 벡터 형태로 치환하는 작업이 필요한데, 본 연구에서는 언어 정보에 대한 자질 추출에 일반적으로 활용되는 임베딩 벡터를 활용하였다. 임베딩에는 클러스터링 과정에서 활용하였던 FastText 기반의 임베딩 벡터를 학습 과정에서 같이 학습할 수 있도록 동적 활용하였다. 이는 학습 데이터에 대한 명사 정보를 모두 포함하고 있을 뿐 아니라, 학술문헌 자체로 구성된 임베딩 벡터이기 때문에, 일반적으로 활용되는 일반 언어자원에 대한 임베딩 벡터보다 제안하는 모델에 더욱 어울리는 형태이기 때문이다. 적용한 임베딩 벡터의 출력 벡터 차원은 300이기 때문에 각 명사마다 300개의 수치로 구성된 자질 벡터가 추출된다. 결과적으로 각 입력 자질 정보는 명사 리스트의 길이만큼의 300차원의 임베딩 벡터 치환 값을 얻는다.

치환된 벡터를 입력 받아, 각 입력 자질에 대한 자질 분석으로 활용될 수 있는 심층학습 구조의 종류는 완전연결(Fully-Conneced) 층, 다중 필터 CNN, 양방향 GRU로 총 3가지이다. 각 구조에 대한 설명은 다음과 같다.

3.4.1 완전연결(Fully-Connected) 층

완전연결 층은 심층학습의 가장 기본이 되는 네트워크 구조이다. 이는 입력 벡터와 출력 벡터의 노드가 완전 연결을 이루는 구조로, 데이터가 가진 전반의 일반적인 자질 정보를 추출하는데 활용된다. 먼저 입력된 자질 벡터는 명사 리스트의 길이 값 l만큼의 임베딩 차원 깊이



〈그림 3〉 심층학습 기반 학술문헌 자동 분류 모델의 구조

e의 자질 정보를 가지고 있는데, 이를 일렬로 정렬하여 $l*d$의 길이를 가진 형태의 입력 벡터 x_{fc}로 변환시킨다. 이후, 완전연결 층의 가중치 벡터 W_{fc}와 출력 차원 수만큼의 바이어스 값 b_{fc}를 활용하여, 다음 (1)과 같은 연산을 수행한다.

$$f = g(W_{fc} \cdot x_{fc} + b_{fc}) \quad (1)$$

g는 활성함수(Activation)로, 본 연구에서는 완전연결 층에 대하여 ReLU(Rectified Linear Unit)를 활용하였다.

3.4.2 다중필터 CNN(Convolutional Neural Networks)

CNN은 주로 이미지 분석 등에서 활용되는 심층학습 구조로서, 일정한 크기의 가중치 필터를 두고, 이를 움직여가며 연산한 결과를 합하는 형식의 연산을 수행한다. 최근 자연어 처리에도 CNN이 활용되면서, 언어 정보의 구조 분석 등에서 높은 성능을 보였다. 본 연구에서는 선행 연구(Choi, S., 2016)를 참고하여 그 중에서도 필터 크기를 다중으로 두고 연산한 결과를 추후에 연결하여 합치는 형식의 다중 필터 CNN을 활용하였다. 명사 리스트의 길이 l에 대한 임베딩 차원 깊이 e의 형태를 가진 2차원 형태의 입력 자질 x_{cnn}에 대해, 일정한 길이 d에 대한 임베딩 차원 깊이 e의 형태를 가진 2차원 필터를 활용하여, 명사 단어를 d만큼씩 연산하고, 각 연산 결과를 연결하여서 벡터를 추출하는 형태이다. 결과적으로 수식 (2)와 같은 연산을 수행한다.

$$c = \sum_{i=0}^{l-(d-1)} g(W_d \cdot x_i + b_d) \quad (2)$$

g는 활성함수(Activation)로, 본 연구에서는 완전연결 층에 대하여 쌍곡 탄젠트를 활용하였다. 이후, 추출된 c에 대하여, 최대값 풀링(Max Pooling)을 통하여 벡터를 정리하여 준다. 이는 크기 (2, 2)의 필터를 두고 필터 내의 계산된 가중치 중에 가장 높은 값만을 뽑아 취하고, 나머지는 버리는 형식의 가중치 버리기 연산으로, 자질 분석 속도와 과적합 방지 등에 탁월한 성능을 보이는 방법이다. 풀링 작업 이후에 도출된 각 필터 크기 d에 대한 CNN 연산 결과 cnn_d를 각각 병합(Concatenation)하여, 다중필터 CNN 연산 결과 f를 도출한다. 본 연구에서는 선행 연구(김선우, 유석종, 이민호, 최성필, 2017)를 참고하여 다중 필터의 범주를 2, 3, 5로 3가지를 활용하였다.

3.4.3 양방향 GRU(Gated Recurrent Unit)

RNN(Recurrent Neural Network) 구조는 연속적인 자질 분석에 대하여 탁월한 성능을 보이고 있는 심층학습 구조로, 자연어 처리 전반에서 가장 우수한 성능을 보이고 있는 네트워크 구조이기도 하다. 본 연구에서는 RNN을 정방향과 역방향 연산을 동시에 수행하여, 결과값을 연결하여 활용하는 양방향 RNN 구조를 선택하였고, 그 중에서도 GRU(Gated Recurrent Units) 셀을 활용하였다. GRU은 LSTM(Long-Short Term Memory units) 셀에 비하여, 적은 파라미터와 높은 분석 속도를 보이면서도 성능 자체는 큰 차이가 없어 실제 서비스되는 시스템에 주로 활용되는 RNN 셀 중 하나이다. 본

연구에서도 분석을 신속히 처리하기 위해 GRU 셀을 선택하였다. 결과적으로 본 연구에서 활용하는 양방향 RNN 구조는 양방향 GRU 구조라고 할 수 있다.

입력 벡터 X에 대한 양방향 GRU 층의 분석은 각 토큰 단위로 이뤄진다. 리스트 길이 l의 입력 벡터 중에서 t번째에 해당하는 입력 벡터 X의 토큰 x_t는 GRU 셀 내에서 3가지의 연산을 수행한다. 먼저 이전 셀에서의 자질 분석 값 h_{t-1}과 연결된 이후, 가중치 W_z에 대한 곱 연산 이후, 시그모이드(Sigmoid) 연산을 수행하여, z_t를 구한다. 같은 형식으로, r_t를 별도의 가중치를 통하여 구한 이후, 다음 수식 (3)을 수행한다.

$$\begin{aligned} \tilde{h}_t &= \tanh(W \bullet [r_t * h_{t-1}, x_t]) \\ h_t &= (1-z_t) * h_{t-1} + z_t * \tilde{h}_t \end{aligned} \quad (3)$$

이런 식으로 계산된 각 셀은 입력된 토큰 리스트의 길이에 맞게 각 연산을 계속 진행하여 토큰 간의 연속적인 자질 분석을 수행한다. 이를 토큰 리스트에 대해 정방향과 역방향으로 수행하여, 각 토큰 단위의 h값을 도출한다. 이후, 제안하는 모델에서는 각 셀에 대한 분석 결과가 아닌 마지막 셀의 분석 결과만을 추출하여 활용한다. 각 셀의 출력 결과를 전체 활용할 경우, 자질 분석에 좀 더 많은 자질을 활용할 수 있지만 문헌 분류 과정에서는 파라미터의 수와 분석 속도를 고려하여 이를 풀링(Pooling)하여 활용하거나 마지막 셀의 결과 값만을 취하여 전체 분석 결과가 반영된 자질 벡터를 활용하는 것이 일반적이다. 본 연구에서는 풀링 사용 시의 최종 분석 결과가 훼손되는 것을 방지하기 위해, 마지막 셀을 취하는 형식을 선택하였다. 결과적으로 양방향 GRU 층의 분석 결과 값 f는 정방향 GRU 층의 마지막 셀 연산 결과 h_l^{fw}와 역방향 GRU 층의 마지막 셀 연산 결과 h_l^{bw}를 병합(Concatenation)한 결과가 된다.

이러한 세 가지의 심층학습 구조 중에 각 입력 자질은 경험적으로 선별하여 각 연산을 수행한다. 각 연산 수행 결과는 병합되어 통합 분석을 위한 형태로 다음과 같이 정리된다.

$$v = [f_{title}, f_{abstact}, f_{keyword}, f_{tfidf}] \quad (4)$$

이후, 통합 벡터 v를 활용하여 전체적인 자질 분석 및 결과 도출을 위한 완전연결 연산을 수행한다. 이는 수식 (5)와 같은 형식으로 연산된다.

$$y = g(W_y \bullet v + b_y) \quad (5)$$

g는 활성함수(Activation)로, 결과 도출 과정에서의 활성함수는 분류 기법에 따라 나뉘도록 구성하였다. 다중 분류 학습집합에 대한 학습 및 예측의 경우에는 해당 분류를 각 차원에 대한 시그모이드(Sigmoid)를 적용하여 바이너리 리스트 형식의 예측을 수행하도록 하였으며, 단일 분류 학습집합에 대한 학습 및 예측은 소프트맥스(Softmax)를 활용하여 분류하였다. 각각의 분류 방식에 따라 예측한 결과와 실제 정답 간의 거리 값을 구하여, 이를 최소화 하는 형태로 전반 모델을 학습하며 각 층의 가중치 값을 변화시키도록 설계하였다.

4. 실험 진행 및 환경

본 연구에서 구축한 학습집합을 통하여 심층학습 기반의 학술문헌 자동 분류 모델의 구조 확정과 확정된 모델의 성능 확인을 위한 성능 최적화 실험을 수행하였다. 구현한 심층학습 기반 학술문헌 자동 분류 모델은 경험적으로 각 자질에 적합한 모델 형태를 선정하고, 그 외의 심층학습 전반에 대한 파라미터를 확정하는 과정이 필요하기 때문이다. 성능 최적화 실험은 반자동 구축한 학습집합을 활용하여 학습하고, 수작업 구축한 실험집합에 대한 성능을 기준 삼아 실험을 수행한다. 또한, 다중 분류와 단일 분류를 나누어 실험한다.

기준이 되는 성능 지표는 다중 분류와 단일 분류에 각각 다르게 활용한다. 일반적으로 분류 문제에는 정확도(Accuracy)와 F1 성능을 활용하는데 수작업으로 구축한 실험집합은 일단 단일 정답 값을 가지고 있기 때문에, 다중 분류 시의 성능은 일반적인 정확도와 F1 성능 측정이 어려운 측면이 있다. 이에 다중 분류의 경우, 예측한 분류가 1개에서 최대 3개가 되는데, 해당 정답 값이 예측한 결과 내에 존재한다면 일치하는 것으로 판단하는 형식의 정확도를 측정한다. 다중 분류의 F1 성능은 다중 예측된 결과와 실제 정답 값의 재현율(Recall), 정확율(Precision) 성능을 측정하여 F1 성능을 구하였다. 단일 분류의 경우에는 일반적인 정확도와 F1 성능을 측정한다. 본 연구에서 측정한 F1 성능은 Micro F1으로, 전체 분류 값에 대한 개별 값을 일일이 다 더하여 계산한 재현율과 정확율을 기반으로 F1 성능을 측정하는 방식이다. F1 성능을 측정하는 수식은 다음 수식 (6)과 같다. r은 전체 단위로 계산한 재현율을, p는 전체 단위로 계산한 정확율을 의미한다.

$$f = 2\frac{r \bullet p}{r+p} \quad (6)$$

성능 최적화 실험에 활용한 파라미터의 종류와 범위는 〈표 10〉과 같다. 각 Dims와 Model은 각 입력 자질이 가질 네트워크의 출력 차원과 구조를 의미한다. 이에 키워드와 TF-IDF 자질 정보를 제목과 초록에 비하여 더욱 주요한 자질이라 판단하여, 이에 출력 차원의 범주를 조금 더 넓게 두고 실험한다. Learning rate는 학습률을 의미하며, 학습 과정에서 Loss 값을 최소화할 때의 각 가중치를 변경하는 비율을 의미한다. 이는 학습 속도와 최소 Loss 값을 찾는 변화에 대한 연관성이 있는 정보이다. 한편, Dropout은 학습집합에 대한 고착 현상을 방지하기 위해, 각 네트워크의 연산에서 자질 정보를 버리는 비율을 의미한다. 그 외의 Optimizer의 경우, 선행 연구를 참고하여 Adam 알고리즘으로 고정하여 실험하였다. 성능 최적화 실

〈표 10〉 성능 최적화 실험에 활용한 파라미터의 종류와 범위

파라미터 유형	파라미터 범위
Title Dims	10, 30
Title Model	FC, CNN, GRU
Abstract Dims	10, 30
Abstract Model	FC, CNN, GRU
Keyword Dims	10, 30, 50
Keyword Model	FC, CNN, GRU
TF-IDF Dims	10, 30, 50
TF-IDF Model	FC, CNN, GRU
Learning rate	0.01, 0.001
Dropout	0.3, 0.5, 0.7

험은 대해 연구의 목적에 맞도록 다중 분류 모델을 기준으로 한다.

5. 분석 및 결과

본 연구에서 계획한 실험에 따라 실험한 결과를 도출하고 분석하였다. 먼저 〈표 10〉의 내용을 기준으로 각 파라미터에 대한 세트를 전부 구성하여 총 17,496회의 세트를 통해 실험을 수행하였다. 그러나 실험 과정에서 유의미한 성능을 보이지 않는 파라미터가 포함된 세트를 제외해가며 실험하여 그 횟수를 줄여가며 보다 효율적으로 실험하였다. 실험 결과 선정된 파라미터와 성능은 〈표 11〉과 같다.

실험 결과, 도출된 모델의 형태는 제목과 초록 자질에 대해 양방향 GRU 구조를 활용하고, 키워드 자질은 다중필터 CNN 구조를 적용하고, TF-IDF 기반의 추출 명사 리스트를 완전연결 층을 통해 분석하는 모델이다. 먼저 문맥과 같은 자질 정보가 비교적 확실하게 존재하는 있는 제목과 초록 자질에 대해 양방향 GRU 구조가 효과적인 자질 분석을 수행하기 때문으로 보인다. 초록 자질의 경우에는 추출된 명사들의 문맥적 정보보다는 개별적인 특성이 강하고 학술문헌의 저자 선정 키워드의 특성상 먼저 구체적인 주제가 등장하는 등의 구조적인 특징이 있기 때문에, 구조 분석에 탁월한 다중필터 CNN 구조가 높은 분석 성능을 보인 것으로 판단된다. 마지막으로 TF-IDF 자질의 경우에는 명사 간의 어떠한 연관성이 존재하지 않기 때문에, 전반 명사 리스트의 일반적 자질 분석이 가장 높은 성능을 보인 것으로 판단된다.

확정된 모델의 실험집합 활용 실험 결과, 다중 분류의 성능은 78.32%의 정확도와 72.45%의 F1 성능을 보였다. 단일 분류의 경우에는 74.98%의 정확도와 72.20%의 F1 성능을 보였다. 전반적인 성능 면에서 다중 분류 쪽의 성능

〈표 11〉 성능 최적화 실험 결과 선정된 파라미터 세트와 실험집합 성능

파라미터 및 성능 유형		파라미터 및 성능
Title Dims		30
Title Model		GRU
Abstract Dims		30
Abstract Model		GRU
Keyword Dims		30
Keyword Model		CNN
TF-IDF Dims		30
TF-IDF Model		FC
Learning rate		0.001
Dropout		0.3
단일 분류	Accuracy	74.98
	F1-Score	72.20
다중 분류	Accuracy	**78.32**
	F1-Score	**72.45**

이 높은 것을 확인할 수 있다. 이는 성능 측정 방식에서도 그 이유를 찾을 수 있는데, 정확도의 경우에는 측정 방식 자체가 다중으로 예측한 값 내에 정답 값이 존재하기만 해도 일치하는 형태로 성능이 높게 측정될 가능성이 높기 때문이다. 한편, F1 성능의 경우에는 일반적으로 재현율 성능이 단일 분류에 비하여 낮을 가능성이 높으나, 정확율 성능이 단일 분류에 비하여 높게 나타날 가능성이 동일하게 높다. 때문에 다중 분류 모델의 분류 값이 비교적 정확하게 1~2개로 분류한 값이 많다면, 다중 분류 쪽의 F1 성능이 높게 나타난 것도 설명이 된다. 실제로 학습집합 내의 41,795개의 학술문헌이 분류된 DDC 값이 1개로 존재하고, 3개로 분류된 학술문헌은 5,815건 뿐이다. 이러한 학습 데이터를 기반으로 한 모델은 최대 3개의 분류 값을 가질 수 있지만, 대체로 1~2개의 분류 값을 도출할 것이다. 이러한 성능 측정 외에도 기본적으로 전반 파라미터 세트의 도출을 다중 분류를 중심으로 진행하였기 때문이다.

이후, 실험집합에 대한 성능을 자세히 분석하기 위해 Confusion Matrix를 도출하였다. 다중 분류의 경우, 전반적인 예측의 수가 각기 다르기 때문에 정확히 1:1 매핑을 수행할 수 없다. 때문에 본 연구에서는 정확한 Confusion Matrix 분석을 위하여, 1:1 매핑이 이뤄질 수 있는 단일 분류를 성능을 기준으로 하여 도출하였다. 도출 결과는 〈그림 4〉와 같다.

〈그림 4〉를 보면, 전반적인 예측과 정답 값의 분포가 620에 집중되어 있다. 심지어 600, 630, 660 등 다수의 분류 값이 정답인 경우에도, 620으로 예측된 결과가 더욱 많음을 확인할 수 있다. 이러한 현상은 전체적인 학습집합의 데이터가 620에 편중되어 구축되어 있기 때문으로 보인다. 한편, 〈그림 4〉에서 추가적으로 확인할 수 있는 680에 대한 예측값이 전혀 존재하지 않는 점도 그 이유를 학습집합 구성에서 찾을 수 있다. 구성한 학습집합 내에는 680에 매핑되어 있는 데이터가 존재하지 않기 때문이다. 그럼에도 학습집합 내의 620의 비율이 매우 높

Truth \ Prediction	600	610	620	630	640	650	660	670	680	690
600	1	21	127	0	6	30	3	0	0	0
610	0	115	26	1	3	5	0	0	0	0
620	0	94	2284	32	21	101	47	6	0	13
630	0	3	3	1	0	0	1	0	0	0
640	0	21	14	2	23	7	0	0	0	0
650	0	0	16	0	3	30	0	0	0	0
660	0	6	91	2	11	1	21	5	0	1
670	0	2	24	0	1	1	1	4	0	0
680	0	1	11	1	1	1	0	0	0	0
690	0	0	100	2	2	9	1	1	0	16

〈그림 4〉 실험 집합에 대한 단일 분류 Confusion Matrix

기 때문에, 610, 620, 640, 650 등의 분류 값에 대해 비교적 좋은 성능을 보인 것이 정확도와 F1 성능을 높이는데 기여한 것으로 보인다.

6. 결론 및 향후 연구

본 연구에서는 한국어로 기술된 기술과학 분야의 학술문헌 자동 분류를 위해 비지도 학습을 기반으로 학습집합을 구축하고, 이를 통해 심층학습 기반의 학술문헌 자동 분류 모델을 구현하였다. 먼저 워드 임베딩 벡터를 기반으로 K-means 클러스터링을 활용하여 명사 클러스터를 확보하고, 이를 각 학술문헌의 서지 정보를 기준으로 TF-IDF 가중치를 통한 핵심 명사 리스트를 도출하여 클러스터와 DDC 600번 대의 중분류에 매핑하여 학습집합을 반자동 구축하였다. 한편, 객관성 있는 성능 평가를 위하여, 주제 분류 학습집합 구축 경험이 있는 연구원 3인을 통하여 수작업으로 실험집합을 구축하였다. 이후, 각 자질 정보마다의 심층학습 구조를 경험적으로 선정할 수 있도록 심층학습 기반의 학술문헌 자동 분류 학습집합을 구현하였다. 성능 최적화 실험 과정에서 모델 구조를 선정한 결과, 제목과 초록 자질에 양방향 GRU, 키워드 자질에 다중필터 CNN, TF-IDF 자질에 완전연결층을 활용한 모델이 확정되었다. 확정된 모델은 다중 분류에 78.32%의 정확도와 72.45%의 F1 성능을 보였으며, 단일 분류에 74.98%의 정확도와 72.20%의 F1 성능을 보였다.

그러나 해당 결과는 한국어로 기술된 기술과학 분야에 한정한 성능이며, 다른 방법론을 통한 성능 비교 등의 추가 실험의 여지가 있다. 향후에는 성능 최적화 실험에서의 파라미터 범주를 더욱 늘리거나, 학습 데이터 구성 과정에서의 클러스터 방법론과 함께 KDC, 연구재단 학문분류, 과학기술표준분류 등의 주제분류표, 심층학습 기반의 자동 분류 모델 등을 각각 변경하여 실험하는 추가적인 연구를 수행할 것이다. 또한, 더욱 많은 학술문헌을 수집하여 DDC의 분류 범주를 확장하는 연구와 함께, 검증 데이터의 양을 전체 데이터의 10% 이상으로 늘려 실험 결과의 신뢰도를 높일 것이다. 또한, 추가 구축한 데이터는 향후 연구의 객관성을 위하여 추후 공개할 예정이다.

참 고 문 헌

김선우, 유석종, 이민호, 최성필 (2017). 생의학 분야 학술 문헌에서의 이벤트 추출을 위한 심층 학습 모델 구조 비교 분석 연구. 한국문헌정보학회지, 51(4), 77-97. https://doi.org/10.4275/KSLIS.2017.51.4.077

김선우, 최성필 (2018). Bidirectional LSTM-CRF 기반의 음절 단위 한국어 품사 태깅 및 띄어쓰기 통합 모델 연구. 정보과학회학술문헌지, 45(8), 792-800.

김판준 (2018). 기계학습에 기초한 국내 학술지 논문의 자동분류에 관한 연구. 정보관리학회지, 35(2), 37-62. https://doi.org/10.3743/KOSIM.2018.35.2.037

김판준, 이재윤 (2014). 해외 데이터베이스의 통제키워드에 기초한 국내 학술지 논문의 자동분류 성능 향상에 관한 실험적 연구. 한국문헌정보학회지, 48(3), 491-510. https://doi.org/10.4275/KSLIS.2014.48.3.491

나동열, 강현규, 김현태, 박경일, 장형일, 염성욱, … 신현주 (2007). 정보검색 관리 서비스 평가용 테스트컬렉션 구축. 보고서 번호 K-07-IP-02-03S-7. 한국과학기술정보연구원.

나동열, 김윤식, 신현주, 이규희, 김태규, 강현규, … 윤화묵 (2007). 한국어 문서분류 테스트컬렉션 개발. 한국콘텐츠학회 종합학술대회 논문집, 5(1), 435-439.

노대욱, 이수용, 나동열 (2007). 정보검색 기술을 이용한 비지도 학습 기반 문서 분류 시스템 개발. 정보과학회논문지: 소프트웨어 및 응용, 34(2), 160-168.

박영근, 박수빈, 박노일, 이현아 (2017). 잠재 의미 분석을 활용한 웹 뉴스 분류. 한국정보과학회 학술발표논문집, 1828-1830.

육지희, 송민 (2018). 토픽모델링과 딥 러닝을 활용한 생의학 문헌 자동 분류 기법 연구. 정보관리학회지, 35(2), 63-88. https://doi.org/10.3743/KOSIM.2018.35.2.063

이다빈, 최성필 (2018). 대용량 텍스트 자원을 활용한 한국어 형태소 임베딩의 모델별 성능 심층 비교 분석. 한국정보과학회 학술발표학술문헌집, 613-615.

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

조현수, 이상구 (2017). FastText를 적용한 한국어 단어 임베딩. 한국정보과학회 학술발표학술문헌집, 705-707.

조현양 (2017). 자동분류기반 성격 유형별 도서추천시스템 개발을 위한 실험적 연구. 한국도서관·정보학회지, 48(2), 215-236. http://doi.org/10.16981/kliss.48.2.201706.215

조휘열, 김진화, 윤상웅, 김경민, 장병탁 (2015). 컨볼루션 신경망 기반 대용량 텍스트 데이터 분류 기술. 한국정보과학회 학술발표논문집, 792-794.

최가람, 최성필 (2018). 단어 임베딩 (Word Embedding) 기법을 적용한 키워드 중심의 사회적 이슈 도출 연구. 정보관리학회지, 35(1), 231-250. https://doi.org/10.3743/KOSIM.2018.35.1.231

최성필, 유석종, 조현양 (2016). 바이오 분야 학술 문헌에서의 분야별 관계 추출 데이터셋 반자동 구축에 관한 연구. 한국도서관·정보학회지, 47(4), 289-307. https://doi.org/10.16981/kliss.47.4.201612.289

한규열, 안영민 (2013). LDA로 형성된 한국어 문서 클러스터의 자동 제목 생성. 한국정보과학회 학술발표논문집, 616-618.

Bock, H. H. (2007). Clustering methods: a history of k-means algorithms. In Selected contributions

in data analysis and classification, 161-172. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-73560-1_15

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606.

Choi, S. P. (2018). Extraction of protein-protein interactions (PPIs) from the literature by deep convolutional neural networks with various feature embeddings. Journal of Information Science, 44(1), 60-73. https://doi.org/10.1177/0165551516673485

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759.

Kowsari, K., Brown, D. E., Heidarysafa, M., Meimandi, K. J., Gerber, M. S., & Barnes, L. E. (2017, December). Hdltex: Hierarchical deep learning for text classification. In Machine Learning and Applications (ICMLA), 2017 16th IEEE International Conference on, 364-371. https://doi.org/10.1109/ICMLA.2017.0-134

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems, 3111-3119.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532-1543. http://dx.doi.org/10.3115/v1/D14-1162

Shafiabady, N., Lee, L. H., Rajkumar, R., Kallimani, V. P., Akram, N. A., & Isa, D. (2016). Using unsupervised clustering approach to train the support vector machine for text classification. Neurocomputing, 211, 4-10. https://doi.org/10.1016/j.neucom.2015.10.137

Shinyama, Y. (2004). PDFMiner. Retrieved from https://euske.github.io/pdfminer/

• 국문 참고문헌에 대한 영문 표기
(English translation of references written in Korean)

Cho, Hui-Yeol, Kim, Jin-Hwa, Yoon, Sang-Woong, Kim, Kyung-Min, & Zhang, Byung-Tak (2015). Large-scale text classification methodology with convolutional neural network. Korea Information Science Society Academic Conference Academic Literature, 792-794.

Cho, Hyun-Soo, & Lee, Sang-Goo (2017). Korean word embedding using fasttext. Korea Information Science Society Academic Conference Academic Literature, 705-707.

Cho, Hyun-Yang (2017). A experimental study on the development of a book recommendation

system using automatic classification, Based on the Personality Type. Journal of Korean Library and Information Science Society, 48(2), 215-236. http://doi.org/10.16981/kliss.48.2.201706.215

Choi, Ga-Ram, & Choi, Sung-Pil (2018). A study on the deduction of social issues applying word embedding: With an empasis on news articles related to the disables. The Journal of Information Management, 35(1), 231-250. https://doi.org/10.3743/KOSIM.2018.35.1.231

Choi, Sung-Pil, Yoo, Suk-Jong, & Cho, Hyun-Yang (2016). A study on the semiautomatic construction of domain-specific relation extraction datasets from biomedical abstracts - Mainly focusing on a genic interaction dataset in alzheimer's disease domain -. Journal of Korean Library and Information Science Society, 47(4), 289-307. https://doi.org/10.16981/kliss.47.4.201612.289

Han, Kyu-Yeol, & Ahn, Young-Min (2013). Automatic labeling of korean document clusters created by LDA. Journal of Korean Society of Information Science. Korea Information Science Society Academic Conference Academic Literature, 616-618.

Kim, Pan-Jun (2018). An analytical study on automatic classification of domestic journal articles based on machine learning. Information Management Journal, 35(2), 37-62. https://doi.org/10.3743/KOSIM.2018.35.2.037

Kim, Pan-Jun, & Lee, Jae-Yun (2014). An experimental study on the performance improvement of automatic classification for the articles of korean journals based on controlled keywords in international database. Journal of the Korean Society for Library and Information Science, 48(3), 491-510. https://doi.org/10.4275/KSLIS.2014.48.3.491

Kim, Seon-Wu, & Choi, Sung-Pil (2018). Research on joint models for korean word spacing and POS tagging based on bidirectional LSTM-CRF. Journal of Information Science, 45(8), 792-800.

Kim, Seon-Wu, Yu, Seok-Jong, Lee, Min-Ho, & Choi, Sung-Pil (2017). A comparative study on deep learning topology for event extraction from biomedical literature. The Journal of Korean Literature Information, 51(4), 77-97. https://doi.org/10.4275/KSLIS.2017.51.4.077

Lee, Da-Bin, & Choi, Sung-Pil (2018). In-depth comparative analysis of various korean morpheme embedding models using massive textual resource. Korea Information Science Society Academic Conference Academic Literature, 613-615.

Lee, Yong-Gu (2013). A study on the quality selection of KNN classifiers using frequency of documents and frequency of collections. Journal of Korean Library and Information Science Society, 44(1), 27-47. http://doi.org/10.16981/kliss.44.1.201303.27

Noh, Dae-Wook, Lee, Soo-Yong, & Ra, Dong-Yul (2007). Developing a text categorization system based on unsupervised learning using an information retrieval technique. Information Science Journal: Software and Application, 34(2), 160-168.

Park, Young-Keun, Park, Su-Bin, Park, No-il, & Lee, Hyun-Ah (2017). Web news classification using latent semantic analysis. Korea Information Science Society Academic Conference Academic Literature, 1828-1830.

Ra, Dong-Yul, Kim, Yun-Sik, Shin, Hyun-Joo, Lee, Kyu-Hee, Kim, Tae-Kyu, Kang, Hyun-Kyu, … & Yoon, Hwa-Mook (2007). Developing a test collection for korean text categorization. Proceedings of the Korea Contents Association Conference, 5(1), 435-439.

Ra. Dong-Yul, Kang, Hyun-Kyu, Kim, Hyun-Tae, Park, Kyung-Il, Jang, Hyeong-Il, Yeom, Sung-Wook, … & Shin, Hyun-Ju (2007). Development of a test collection HANTEC for evaluating information retrieval · management · service. (report no. K-07-IP-02-03S-7). Korea Institute of Science and Technology Information.

Yuk, Jee-Hee, & Song, Min (2018). A study of research on methods of automated biomedical document classification using topic modeling and deep learning. The Journal of Information Management, 35(2), 63-88. https://doi.org/10.3743/KOSIM.2018.35.2.063