JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 25, No. 12: 1809~1816, Dec. 2021

머신러닝을 이용한 R&D과제의 연구분야 추천 서비스

김윤정[1] · 신동구[1] · 정회경[2*]

Recommendation System for Research Field of R&D Project Using Machine Learning

Yunjeong Kim[1] · Donggu Shin[1] · Hoekyung Jung[2*]

[1]Principal Researcher, Korea Institute of Science and Technology Information, Daejeon, 34141 Korea
[2*]Professor, Department of Computer Engineering, Paichai University, Daejeon, 35345 Korea

요 약

국가연구개발사업 관련 데이터를 이용한 최신 연구동향 파악, 의미 있는 정보의 생산과 활용을 위해 국가R&D 정보 서비스에도 자동 분류 기술 적용이 요구되어 R&D과제의 연구분야를 자동 분류하고 추천하기 위한 연구를 진행했다. 2013~2020년 국가R&D 과제 데이터 약 45만 건을 수집하여 학습과 평가에 사용했다. 수집 데이터 중 유효한 데이터를 대상으로 데이터 전처리 및 분석, 실험을 통한 성능 분석 후 모델을 선정했다. 최적의 모델 조합 도출을 목적으로 Word2vec, GloVe, fastText 성능을 비교했다. 실험 결과, 과제정보의 필수 항목으로 사용되는 소분류만의 정확도는 90.11%이다. 이 모델은 국가과학기술표준분류 연구분야와 유사한 계층 구조를 가진 다른 분류체계의 자동분류 연구에 활용 가능할 것으로 기대한다.

ABSTRACT

In order to identify the latest research trends using data related to national R&D projects and to produce and utilize meaningful information, the application of automatic classification technology was also required in the national R&D information service, so we conducted research to automatically classify and recommend research field. About 450,000 cases of national R&D project data from 2013 to 2020 were collected and used for learning and evaluation. A model was selected after data pre-processing, analysis, and performance analysis for valid data among collected data. The performance of Word2vec, GloVe, and fastText was compared for the purpose of deriving the optimal model combination. As a result of the experiment, the accuracy of only the subcategories used as essential items of task information is 90.11%. This model is expected to be applicable to the automatic classification study of other classification systems with a hierarchical structure similar to that of the national science and technology standard classification research field.

키워드 : 자동 분류, 머신러닝, Word2vec, 연구분야 추천, 국가과학기술표준분류

Keywords : Auto classification, Machine Learning, Word2vec, Research field recommendation, National science and technology standard classification

Received 9 September 2021, **Revised** 14 September 2021, **Accepted** 22 October 2021

*** Corresponding Author** Hoekyung Jung(E-mail:hkjung@pcu.ac.kr, Tel:+82-42-520-5640)
Professor, Department of Computer Engineering, Paichai University, Daejeon, 35345 Korea

Open Access http://doi.org/10.6109/jkiice.2021.25.12.1809 print ISSN: 2234-4772 online ISSN: 2288-4165

ⓒThis is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License(http://creativecommons.org/licenses/by-nc/3.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
Copyright ⓒ The Korea Institute of Information and Communication Engineering.

Ⅰ. 서 론

데이터와 인공지능이 국가 발전의 동력으로 인식됨에 따라 정부는 공공·민간 분야의 양질의 데이터가 유통될 수 있는 인프라를 확충하고 있다. 국가연구개발(R&D)사업 관련 정보를 통합하여 서비스하는 NTIS는 약 1억 6,081만 건(2021.9.3. 기준)의 데이터를 공개하고, 437개의 국가R&D 표준 정보 항목 중 81.7%를 개방하고 있다. 공개하는 데이터의 양, 파일 형태로 다운로드할 수 있는 개방 항목이 증가할수록 데이터 분석 수요가 높아졌다.

명확한 분류체계를 갖고 있는 데이터는 분석, 예측에 더 효율적이라 할 수 있다. 국가 R&D사업에는 과학기술예측, R&D정보의 관리·유통, 통계·분석 등의 목적을 위해 다양한 분류체계가 사용되고 있다. 대표적으로 국가과학기술표준분류, 미래유망 신기술(6T), 국가중점과학기술, 국제특허분류를 들 수 있고, 이 가운데 국가과학기술표준분류는 과제의 연구분야, 연구자의 전문분야 값으로 사용되는 필수 분류체계이다. 이 분류체계는 새로운 기술의 출현을 고려하여 5년마다 개정되지만 대-중-소분류의 코드와 이름, 중-소분류의 간략한 설명으로 작성된 해설서만 제공되고 있다. 분류체계에 대한 정보가 부족한 연구자들은 변화하는 분류체계를 이해하기 어렵다. 따라서 데이터 생산자인 연구자의 경험과 감에 의존하지 않고 과학적 방법으로 R&D 데이터의 분류 접근이 필요하다. 분류는 어떤 입력 데이터가 주어졌을 때 특징을 기준으로 해당 데이터의 클래스를 예측하는 기법이다. 최근 R&D 과제, 논문, 특허 등 기술문서의 텍스트를 자동 분류하기 위한 연구가 눈에 띄게 늘고 있다. 하지만 대형 포털서비스를 통해 뉴스, 음악, 쇼핑 등 콘텐츠 추천과 챗봇(chatbot)의 요소로 분류를 체감할 수 있는데 비해 공공서비스에서는 텍스트 분류의 학문적 결과가 실제 서비스에 활용된 사례를 접하기가 쉽지 않다.

본 연구는 국가R&D 과제 데이터를 학습하고 머신러닝을 이용하여 과제의 연구분야 소분류를 자동 분류함으로써 연구자의 과제신청 등 과제관리 지원과 과제 데이터의 활용 편의성을 높이는 것을 목적으로 한다. 본 논문의 구성은 다음과 같다. 2절에서는 텍스트 분류와 관련된 연구들을 간략히 살펴보고, 3절에서는 제안하는 연구분야 분류 모델과 실험 결과를 소개한다. 4절에서는 분류 모델을 적용한 연구분야 추천 서비스를 소개한다. 마지막으로 5절은 결론과 향후 연구 방향을 제시한다.

Ⅱ. 관련 연구

빅데이터 시대에 텍스트 데이터를 효율적으로 분류하는 것은 해결해야할 주요 과제 중 하나이다. 텍스트 데이터를 분류하기 위해 국내 대학, 연구소에서도 연구가 활발해지고 있다. 연구보고서의 제목, 키워드를 중요한 요소로 이용한 TK_CNN(Convolutional Neural Networks) 기법의 국가과학기술표준분류 자동 분류[1], 딥러닝 모델을 이용한 특허 문헌 분류[2], BERT(Bidirectional Encoder Representations from Transformers) 기반 분류 모델을 이용한 R&D과제의 기술분야 분류[3], 벡터 공간 모델을 이용하여 분류 대상 문서와의 유사도를 계산한 한국표준산업분류 기준 문서 자동 분류[4] 연구 등을 들 수 있다.

해외 연구 사례를 살펴보면, Li[5]는 대부분 수동으로 수행되는 특허 분류를 위해 특허 제목과 초록 정보만을 사용하여 CNN과 워드벡터 임베딩을 기반으로 한 DeepPatent 딥러닝 알고리즘을 제안했다. Sharma[6]는 소셜 미디어와 유튜브의 긍정적 및 부정적인 리뷰 분류 예측을 목적으로 사전 훈련된 Word2vec 모델과 CNN layer를 이용한 모델을 제안했다.

워드임베딩은 자연어 처리에 널리 활용되고 있으며 분류 정확도를 높이기 위해 사용되는 방법 중 하나이다. 단어들 사이 유사성을 표현하기 위해 분산 표상을 학습하는 대표적인 모델은 Word2vec, GloVe, fastText를 들 수 있다. 2013년 Google에서 개발한 Word2vec[7]은 실제값과 예측값에 대한 오차를 줄여나가며 학습하는 예측 기반 방법으로, 빈번한 단어들(frequent words)의 하위 샘플링(subsampling)으로 훈련 속도가 빨라져 빠른 속도로 대용량의 데이터를 학습시킬 수 있다. 2014년에 스탠포드대학이 발표한 GloVe(Global Vectors for Word Representation)[8]는 분산된 단어 표현을 위한 모델로, 코퍼스(corpus)에서 집계된 global word-word 동시 등장 행렬(co-occurrence matrix)의 0이 아닌 항목에 대해 학습하기 때문에 후속 훈련 반복(subsequent training iteration)이 빠르다. 2016년에 Facebook이 발표한 fastText[9]는 워드임베딩 및 텍스트 분류 학습을 위한

라이브러리로, 학습 데이터에 잘 나타나지 않는 저빈도 단어(rare word)에 강하고 정확성 면에서 딥러닝 분류기와 대등하다고 한다.

워드임베딩 방법에 따른 분류 정확도를 비교한 연구들도 다수 수행되었다. Kim[10]은 워드클러스터링과 CNN을 기반으로 한 문장 분류 연구에서 TREC(Text Retrieval Conference) 질의 데이터셋과 영화 리뷰 데이터셋을 대상으로 워드임베딩 방법에 따른 분류 정확도를 비교했다. 실험 결과 TREC-major에서는 fastText, TREC-minor에서는 Word2vec이 좋은 성능을 보였다. Jang[11]은 CNN with CBOW(continuous bag-of-word), CNN with Skip-gram, CNN without Word2vec 모델 성능 비교를 통해 Word2vec 사용 시 분류 모델의 성능이 크게 향상되는 것을 확인했다. 뉴스 기사와 트윗(tweets)을 기반으로 실험한 결과, 뉴스 기사에는 CBOW 알고리즘, 트윗에는 Skip-gram 알고리즘이 사용될 때 더 나은 성능을 보여 데이터 유형에 따라 적절한 워드임베딩 모델을 사용하는 것이 효과적임을 설명했다. Choi[2]의 연구에서도 특허 문헌에 대한 워드임베딩 벡터는 fastText, Skip-Gram, CBOW, GloVe 순으로 좋은 결과를 보였다.

분류의 성능을 높이기 위해 Doc2Vec[12]이 함께 활용된 연구들도 있다. Yuk[13]의 실험에서는 Naïve Bayes, KNN(K-Nearest Neighbor), Ridge로 분류한 경우보다 더 높은 성능을 보였고, Kim[14]의 연구에서도 문서의 분류에 Doc2Vec을 함께 활용하는 것이 효과적임을 검증하였다.

Ⅲ. 연구분야 분류 모델 및 성능 평가

텍스트를 분류하려면 텍스트 전처리 후 특징을 추출하고 학습, 평가하는 일련의 과정이 필요하다. 텍스트 분류를 위한 서비스 도메인의 데이터셋과 특징을 살펴보고, 최적의 모델 조합 도출을 위해 Word2vec, GloVe, fastText 워드임베딩 성능을 비교한 결과를 소개한다. 이어서 우리 시스템에 적용하기 위한 모델의 분류 성능을 소개한다.

3.1. 서비스 도메인의 데이터셋

본 연구의 목적은 과제신청 단계에서 연구자가 제안하려는 과제의 요약정보에 적합한 연구분야 소분류를 자동 분류하여 추천하는 것이므로 과제 데이터 분석이 선행되어야 한다. 한국의 국가R&D 사업 관련 정보는 후속연구, 융합연구 수행 등에 활용될 수 있도록 NTIS (ntis.go.kr)라는 국가R&D정보 지식포털을 통해 제공 중이다. 국가R&D 과제는 협약이 체결되면 NTIS에 상시 연계되는데, 과제의 입력 항목은 표 1과 같이 일반/국방/인문 3개 사업 유형별로 다르다. 일반 연구사업의 과제는 과제명, 연구분야, 적용분야, 과제요약서(연구목표, 연구내용, 기대효과, 한글 키워드, 영문 키워드) 정보를 필수로 입력하지만, 국방 사업은 과제명, 연구분야, 적용분야가 필수이고 인문 사업은 과제명, 적용분야만 필수로 입력하면 된다. 융합기술인 경우 연구분야를 3개까지 입력할 수 있으며, 연구분야 입력 조건은 2015년까지는 중분류 코드, 2016년 이후는 소분류 코드이다.

Table. 1 Example of National R&D Project Information Input Item (NTIS)

Project information item			Research program		
			G	D	HS
Project name_Korean			●	●	●
Project name_English			○	○	○
National science and technology standard classification	Research field	Category 1	●	●	-
		Category 2	○	○	-
		Category 3	○	○	-
Project type			●	●	-
Project summary information	Research Objectives		●※	-	-
	Research Contents		●※	-	-
	Expectation effectiveness		●※	-	-
	Keywords-Korean		●※	-	-
	Keywords-English		●※	-	-

[Legend]
G: General, D: Defense, HS: Humanities society
●: Required, ○: Required by condition
※: If the project type is 'R&D', the project summary information is required

자동 분류 대상인 국가과학기술표준분류는 연구분야와 적용분야의 2차원 분류체계로 현재 2018년에 개정된 분류체계가 활용 중이다. 연구분야는 표 2와 같이 33개 대분류, 371개 중분류, 2,898개 소분류의 계층적 구조로 되어 있다.

Table. 2 Composition of research fields in National science and technology standard classification

Research field		Major category	Mid category	Sub category
Science Technology	Nature	4	47	339
	Life	3	49	448
	Artifact	9	112	858
Humanities and Social Sciences	Human	5	61	546
	Society	9	88	634
	Human Science and Technology	3	14	73
Total		33	371	2,898

분류를 위해서는 미리 분류되어 정답으로 활용 가능한 많은 양의 학습 데이터가 필요하다. 학습 데이터가 많을수록 단어들의 관계에 대해 더 쉽게 파악할 수 있기 때문이다. 국가연구개발사업 조사·분석을 통해 확정된 국가R&D 과제정보는 검증된 데이터로 인식된다. 우리는 2013년부터 2019년까지의 조사·분석 확정 과제 데이터와 NTIS에 상시 연계·수집되는 당해 연도 과제 데이터를 학습과 테스팅 대상으로 했다. 국가과학기술표준분류가 5년마다 개정되고 있는데, 2018년 개정된 최신 버전과 구 버전 간 매핑표를 활용할 수 있기 때문이다. 수집한 R&D과제 데이터 현황은 표 3과 같다.

수집 데이터 중 유효한 데이터만 활용하기 위해 연구분야 코드, 과제 요약서의 텍스트 입력 현황을 분석했다. 우선 연구분야 코드가 입력되지 않은 과제를 제외하고 연구분야 분포를 살펴보면, 과제 수 기준으로 전체 과제의 약 35%가 상위 5%의 연구분야, 전체 과제의 약 50%가 상위 10%의 연구분야에 집중되어 있다. 연구분야의 소분류 코드가 필수 입력인 2016년 이후의 과제 분포 예시인 그림 1에서도 유사한 형태를 확인할 수 있다. 이러한 분포는 연구분야 코드가 필수 입력이 아닌 인문사업의 과제를 제외하더라도 일부 연구분야에 연구개발이 편중되어 있음을 시사한다. 또한 최신 버전의 연구분야 소분류 코드 입력 현황을 보면, 2018년과 2019년 모두 2,100개 이내의 코드만 사용되었다. 즉 800여개의 소분류는 학습에 활용 가능한 데이터가 없는 상태이다.

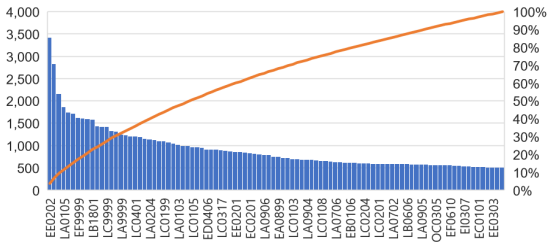

Fig. 1 Example of project distribution by sub-category code in the research field (From 2016 to 2019, the frequency of projects per code is 500 or more)

높은 단어 수(term count)와 라인 수를 갖는 데이터일수록 학습에 활용하기 좋은 데이터라 할 수 있다. 과제 요약서 정보가 입력되지 않은 과제를 제외하면 전체 과제의 약 85%를 학습에 활용할 수 있는데, 이 가운데 라인 수가 6 이하, term count가 32 이하인 과제가 약 10% 분포를 보였다. 우리는 이들을 제외한 약 75%의 데이터를 학습에 사용할 수 있는 데이터로 정의했다. 마찬가지로 이용자가 분류를 요청하기 위해 입력하는 유효한 텍스트 길이도 term count 32 이상으로 설정했다.

Table. 3 National R&D project information construction status by year of NTIS (As of 2021.8.18.)

Year	Number of projects			Input status of research field code				Input status of summary information	
	New project	Continuous project	Sum	1	2	3	None	All (5 items)	None
2013	26,286	24,579	50,865	32,381	6,680	4,587	7,217	42,517	8,114
2014	27,299	26,194	53,493	32,738	8,051	5,076	7,628	44,217	1,522
2015	27,871	26,562	54,433	33,325	8,453	5,227	7,428	49,681	4,626
2016	26,947	27,880	54,827	31,126	8,380	8,283	7,038	46,413	3,432
2017	31,500	29,780	61,280	32,740	9,458	11,797	7,285	53,373	4,429
2018	26,894	36,803	63,697	33,827	9,171	13,771	6,928	55,603	7,404
2019	28,140	42,041	70,181	34,517	10,808	16,343	8,513	60,315	6,466
2020	13,130	27,741	40,871	19,060	8,261	12,158	1,392	39,487	55

3.2. 워드임베딩 성능 평가

텍스트 처리와 워드임베딩 모델의 적절한 조합을 찾으면 분류 모델 성능 향상을 도모할 수 있다[15]. 우리는 최적의 모델을 도출하기 위해 Word2vec, GloVe, fastText 워드임베딩 알고리즘을 비교하였다. Word2vec (CBOW (Continuous Bag of Words)와 Skip-gram), GloVe, fastText 정확도 비교 결과는 그림 2와 같다. dimension이 300일 때 Skip-gram과 fastText가 가장 높은 정확도를 보였다.

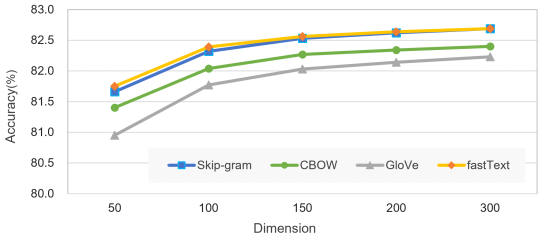

Fig. 2 Skip-gram, CBOW, GloVe, fastText accuracy comparison

파라미터에 따른 데이터 특성을 파악하기 위한 모델 생성 및 테스트 결과는 표 4와 같다. 네거티브 샘플링(negative sampling)은 빈번한 단어에 대한 정확한 표현을 학습하는 훈련 방법으로, Word2vec의 성능은 negative sample의 수가 약 10개 이상으로 증가하면 실제로 감소한다고 한다[8, 16-18]. 우리는 negative sample의 수를 10 이내로 측정하였다.

Table. 4 Model creation and test results (Word2vec vs. fastText)

		Word2vec	fastText
Dimension		300	300
Iteration		3~8	4,6,8,10,12,14
Window size		3~8	4,6,8,10,12,14
Negative sample		3~8	1,3,5,7,9,11
Number of Models		216	216
File size		47.00G	543.67G
Average Accuracy (%)	Top3	91.64	91.70
	Top6	89.42	89.42
	Top9	88.84	88.84
	Total	82.69	82.69

그림 2에서 정확도가 높았던 Word2vec, fastText의 모델 생성 시간과 성능 측정 시간(질의 성능)을 비교했으며, 전체 평균 정확도가 가장 높게 나타난 파라미터별 결과는 표 5와 같다. fastText는 저차원에서는 만족할만한 성능을 보였지만 분석시간과 모델 사이즈가 크게 나타났다. 질의 성능은 Word2vec에 비해 눈에 띄게 낮게 나타났다.

Table. 5 Speed of model creation and performance measurement (Word2vec vs. fastText)

		Word2vec		fastText	
Dimension		300	300	300	300
Iteration		5	4	4	6
Window size		6	4	12	10
Negative sample		3	3	3	3
Model creation time		2:38:11	1:19:20	1:43:19	2:07:43
Average Accuracy (%)	Top3	91.76	91.68	91.83	91.87
	Top6	89.55	89.49	89.54	89.55
	Top9	88.96	88.89	88.96	88.96
	Overall average	82.81	82.78	82.82	82.80
Performance measurement time		0:21:37	0:21:36	1:04:27	1:05:04

모델 질의 성능은 엔진의 응답 성능과 직결된다. 모델 생성과 질의 처리 성능 관점에서 우리 어플리케이션에는 Word2vec 모델이 적합하다는 결론을 도출했다.

3.3. 모델 성능

모델의 적절한 중요한 특징 값을 선택하기 위해서 전처리 단계에서 언어적으로 잘못된 부분을 교정(오탈자 및 띄어쓰기 교정, 불용어 제거 등)하고, 형태소 분석기를 사용해 명사에 해당하는 단어만 추출했다.

정확도, 학습 시간, 모델 생성 시간, 빠른 응답 속도 등 도메인과 상황에 따라 요구하는 능력치가 다르다. 모델 아키텍처, 벡터 사이즈, 서브샘플링 비율, training window 사이즈가 실험에서 성능에 영향을 미치는 가장 중요한 결정 대상이다[9]. 우리는 수집 데이터 중 2013년부터 2019년까지의 과제 데이터를 대상으로 Word2vec 모델에 파라미터(dimension, iteration, window size, negative sample)를 변경하면서 반복 학습하였다. dimension은 학습 시간에 영향을 미치기 때문에 50, 100으로 학습하였다. 학습된 분류 모델의 성능은 테스트 데이터셋(2020년 과제 데이터 약 3.9만개) 질의를 통해 연구분야 대분류, 중분류, 소분류 단위로 평가했다. 분류

모델의 정확도(F1 Score)는 표 6과 같고, 대-중-소분류 전체 평균 정확도가 79.07%로 가장 높은 Test 5 모델을 최종 선정했다.

Table. 6 Classification model accuracy

Query result	Test 1	Test 2	Test 3	Test 4	Test 5
Major category	94.95%	94.10%	94.11%	94.08%	95.04%
Mid category	91.79%	90.38%	89.99%	90.31%	92.32%
Sub category	90.02%	89.11%	89.07%	89.03%	90.11%
Average	78.46%	75.78%	75.43%	75.65%	79.07%

2013년부터 2018년까지 과제 데이터로 학습한 모델과 2013년부터 2019년까지 과제 데이터로 학습한 Test 5 모델 성능을 비교했을 때 모델의 평균 정확도가 2.48% 향상되어 학습 데이터셋이 늘어나면서 실 예측 성능이 좋아짐을 확인할 수 있었다.

Ⅳ. 연구분야 추천 시스템

학습을 통한 최적의 모델을 생성하고 머신러닝 엔진에 탑재된 엔진을 개발했다. 연구분야 추천 시스템은 입력되는 과제정보(sentence)에 대해 Word2vec 모델을 통하여 단어를 벡터화하고 과제정보의 단어 벡터들을 기반으로 Doc2Vec 모델을 통해 문장 벡터를 생성한다. K-NN을 사용하여 가장 근접한 추천 후보 5개를 선정하고, Naïve Bayes를 이용하여 소분류에 대한 최종 점수를 계산한다.

그림 3과 같이 이용자가 확인하고자 하는 텍스트를 입력하면 연구분야를 자동 분류하여 최대 5개의 대-중-소분류를 추천한다. 과제신청과 협약 단계에서 사용되는 연구분야 코드가 소분류이기 때문에 추천결과 목록의 정확도 값은 평균 정확도가 아닌 소분류에 대한 정확도를 나타낸다. 이용자가 확인하고자 하는 요청정보가 많을 경우에는 한 번에 연구분야 매칭 결과를 확인할 수 있도록 '파일 등록'(최대 300건) 기능도 부가적으로 제공한다.

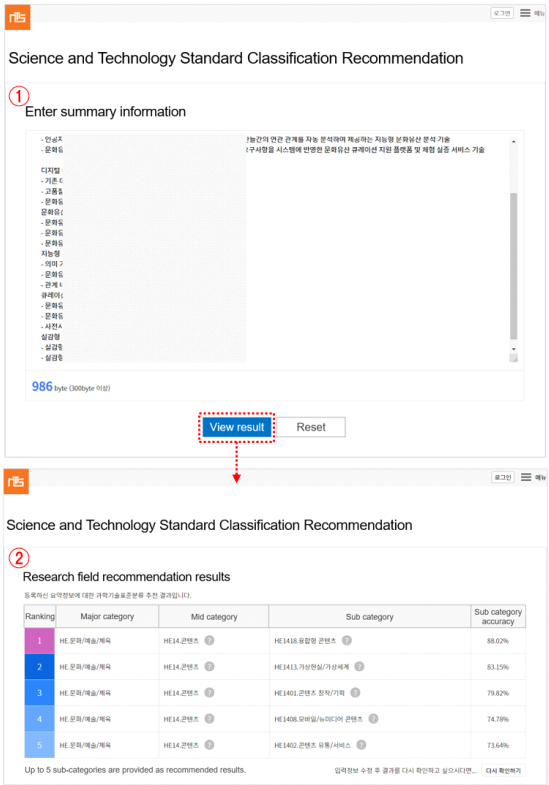

Fig. 3 Research field recommendation service

Ⅴ. 결론 및 향후 연구

본 연구에서는 Word2vec을 이용한 분류 모델의 연구분야 분류 예측 결과를 확인했다. 약 45만건의 국가 R&D 과제를 데이터로 활용했고, 당해 연도 과제가 해당되는 33개 대분류, 371개 중분류, 2,898개 소분류를 예측하도록 했다. 학습된 모델은 대-중-소분류 평균 79.07%의 정확도 값을 보였고, 연구자가 필수 항목으로 사용하는 소분류의 정확도는 90.11%로 나타나 의미 있는 결과를 거두었다. 학습에 활용 가능한 데이터가 부족한 중-소분류에 대해서는 학습데이터 수가 적을 때 효과적인 전이학습 기술 등의 고려가 필요하다.

정확도를 높이기 위해 앞으로 더 많은 학습 데이터셋을 주기적으로 반영할 계획이다. 본 연구 결과는 국가과학기술표준분류 연구분야와 유사한 계층 구조를 가진 다른 분류체계(산업기술분류, 특허 분류체계 등)의 자

동 분류 연구에도 활용 가능할 것으로 기대한다. 향후 과학기술표준분류와 타 분류체계 간 매핑도 일부 가능할 것으로 보인다.

ACKNOWLEDGEMENT

This research was supported by Construction of NTIS funded by the Ministry of Science and ICT.

References

[1] J. Y. Choi, H. Hahn, and Y. C. Jung, "Research on Text Classification of Research Reports using Korea National Science and Technology Standards Classification Codes," *Journal of the Korea Academia-Industrial cooperation Society*, vol. 21, no. 1, pp. 169-177, 2020.

[2] Y. Choi and S. P. Choi, "A Study on Patent Literature Classification Using Distributed Representation of Technical Terms," *Journal of the Korean Society for Library and Information Science*, vol. 53, no. 2, pp. 179 - 199, May 2019.

[3] S. Hwang and D. Kim, "BERT-based Classification Model for Korean Documents," *The Journal of Society for e-Business Studies*, vol. 25, no. 1, pp. 203-214, Feb. 2020.

[4] J. S. Lee, S. P. Jun, and H. S. Yoo, "A Study on Automatic Classification Model of Documents Based on Korean Standard Industrial Classification," *Journal of Intelligent Information Systems*, vol. 24, no. 3, pp. 221-241, Sep. 2018.

[5] S. Li, J. Hu, Y. Cui, and J. Hu, "DeepPatent: patent classification with convolutional neural networks and word embedding," *Scientometrics*, vol. 117, no. 2, pp. 721-744, 2018.

[6] A. K. Sharma, S. Chaurasia, and D. K. Srivastava, "Sentimental Short Sentences Classification by Using CNN Deep Learning Model with Fine Tuned Word2Vec," *Procedia Computer Science*, vol. 167, pp. 1139-1147, 2020.

[7] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their Compositionality," in *Advances in neural information processing systems*, pp. 3111-3119, 2013.

[8] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing* (EMNLP), pp. 1532-1543, 2014.

[9] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of Tricks for Efficient Text Classification," *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, vol. 2, pp. 427-431, 2017.

[10] H. Kim, J. Lee, N. Y. Yeo, M. Astrid, S. Lee, and Y. Kim, "CNN based Sentence Classification with Semantic Features using Word Clustering," *International Conference on Information and Communication Technology Convergence* (ICTC), pp. 484-488, 2018.

[11] B. Jang, I. Kim, and J. W. Kim, "Word2vec Convolutional Neural Networks for Classification of News Articles and Tweets," *PloS one*, vol. 14, no. 8, pp. e0220976, 2019.doi: 10.1371/journal.pone.0220976.

[12] Q. Le and T. Mikolov, "Distributed Representations of Sentences and Documents," *International conference on machine learning, PMLR*, 2014.

[13] J. Yuk and M. Song, "A Study of Research on Methods of Automated Biomedical Document Classification Using Topic Modeling and Deep Learning," *Journal of the Korean Society for information*, vol. 35, no. 2, pp. 63-88, Jun. 2018.

[14] D. W. Kim and M. W. Koo, "Categorization of Korean News Articles Based on Convolutional Neural Network Using Doc2Vec and Word2Vec," *Journal of KIISE*, vol. 44, no. 7, pp. 742-747, 2017.

[15] Y. S. Kim and S. W. Lee, "Combinations of Text Preprocessing and Word Embedding Suitable for Neural Network Models for Document Classification," *Korea Information Science Society*, vol. 45, no. 7, pp. 690-700, July. 2018.

[16] T. Mikolov, K. ChenK, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," arXiv preprint arXiv:1301.3781, 2013.

[17] K. M. Jang, E. S. Kim, and H. K. Jung, "A Study on the Standardization of Information for the Integrated Management of Researcher's Information," *Journal of the Korea Institute of Information and communication Engineering*, vol. 25, no. 5, pp. 741-747, May. 2021.

[18] J. S. Bang, D. Y. Hwang, and H. K. Jung, "Product Recommendation System based on User Purchase Priority," *Journal of Information and Communication Convergence Engineering*, vol. 18, no. 1, pp. 55-60, Mar. 2020.

김윤정(Yunjeong Kim)

2005년~현재 한국과학기술정보연구원 재직
1998년 홍익대학교 전자전산공학과(공학사)
2005년 홍익대학교 전자전산공학과(공학석사)
2019년~현재 배재대학교 컴퓨터공학과(박사과정)
※관심분야 : 정보분석, 빅데이터, 머신러닝

신동구(Donggu Shin)

1997년~현재 한국과학기술정보연구원 재직
1998년 홍익대학교 컴퓨터공학과(공학사)
2006년 서울산업대학교 컴퓨터공학과(공학석사)
2019년 건국대학교 컴퓨터공학과(공학박사)
※관심분야 : AI, 데이터분석

정회경(Hoekyung Jung)

1985년 광운대학교 컴퓨터공학과(공학사)
1987년 광운대학교 컴퓨터공학과(공학석사)
1993년 광운대학교 컴퓨터공학과(공학박사)
1994년~현재 배재대학교 컴퓨터공학과 교수
※관심분야 : Machine learning, Big data, Embedded system, U-Healthcare, IoT