워드 임베딩 기반의 기술 개체명 인식 방법 연구

이유진*, 김세빈*, 김장원*†
*국립군산대학교 소프트웨어융합공학과
e-mail: {u.jin, bhy401, jwgim}@kunsan.ac.kr

A Study on Technology Name Recognition Method based on Word Embedding

Yujin Lee*, Sebin Kim* and Jangwon Gim*†
*Dept. of Software Convergence Engineering, Kunsan National University

요 약

최근 4차 산업 혁명시대에 이르러 다양한 기술이 급속도로 발전함에 따라 지적 재산권 확보가 중요하게 되었다. 따라서 대표적인 지식재산권의 하나인 특허의 발명 또한 급증하고 있다. 본 논문에서는 특허 데이터에 포함된 기술명 식별을 위해 딥러닝 기반 기술명 분류 방법을 제안한다. 그 결과 특정 분야에서 사용되는 전문 용어에 대한 개체명 식별이 가능함을 보인다.

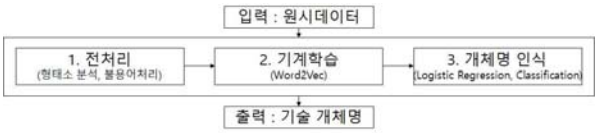
1. 서론

특허 발명은 발명자의 기술에 대해 일정한 법률적 권리나 능력, 포괄적 법률관계를 설정하는 행위를 의미한다. 특허는 창의성, 신규성, 진보성을 가진 기술에 대한 권리를 주장하기 위해 핵심 기술 및 상세 요소 기술에 대한 정보가 포함되어 있다. 따라서 특허에 포함된 기술을 분석하여 최신 기술 동향 파악 및 미래 유망 기술 도출에 활용할 수 있다. 또한 최근 4차 산업 혁명시대에 이르러 다양한 기술이 급속도로 발전함에 따라, 기술에 대한 소유권이 중요해지면서 특허 문헌의 양 또한 함께 증가한다. 이로 인해 많은 수의 특허 문헌으로부터 핵심 기술 분석 및 관련 기술에 대한 신속하고 정확한 분석이 어려워지고 있다. 그러므로 대용량의 특허문헌으로부터 원천 및 핵심 기술명의 식별 및 추출을 위한 연구가 필요하다.

2. 관련 연구

특허 문헌에 포함된 기술의 분석을 위해서는 기술 용어에 대한 개체명 인식이 선행될 필요가 있다. 그러므로 다양한 분야에서 개체명 인식을 위한 연구가 진행되고 있으며 기계학습 기반의 CRF 모델이 주로 적용되고 있다[1,2]. 또한, 최근 딥러닝 기술이 발전하면서 개체명 인식 성능을 향상시킨 하이브리드 방법(LSTM-CRF)등이 이 제안되었다[3,4]. 그 결과 하이브리드 방법을 기반으로 특정한 분야의 전문 용어에 대한 개체명 식별 연구가 진행되어 화학 및 식품 분야 각각의 도메인 용어에 대한 개체명 인식 방법이 연구되었다[5,6]. 본 논문에서는 다양한 분야의 기술이 포함된 특허 문헌으로부터 딥러닝 기술을 이용한 기술 개체명 인식 방법을 제안한다.

3. 제안 방법



(그림 1) 제안 모델의 전체 개요도

본 논문에서 제안하는 모델의 전체 개요도는 그림 1과 같다. 특허 문헌에 포함된 기술 개체명 인식 과정은 다음과 같이 3단계로 구성된다. 첫 번째 단계는 기술 개체명 추출에 사용되는 특허 데이터의 전처리 단계로서 형태소 분석기를 이용하여 토크나이징을 수행하고, 불용어 및 의미상의 불용어를 제거한다. 두 번째 단계는 정제된 데이터를 기반으로 워드 임베딩을 하는 단계로 워드 임베딩에 필요한 Seed set은 특허 분류체계코드(CPC) 및 특허 초록에서 각 분류체계를 대표하는 단어를 이용하여 구축한다. 그 결과 구축된 Seed set을 기반으로 워드 임베딩을 수행한다. 마지막 단계는 어휘들로부터 기술명을 분류하는 단계로서 로지스틱 회귀(Logistic Regression) 기법을 적용

† 교신저자

※ 이 성과는 2018년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2018R1C1B6008624).

하여 기술명에 대한 분류 정확도 점수를 도출하고 분류(classification)를 수행하여 기준치 이상의 값을 가지는 어휘를 기술명으로 분류한다.

4. 실험

본 논문에서는 한국특허정보원(KIPRIS)에서 제공하는 특허 데이터를 수집하였으며, 수집된 데이터에서 높은 빈도수를 가지는 CPC 분류코드(B60L 11/18, G06F 17/30, H01L 21/67)를 포함한 특허들을 실험 데이터로 선정하여 총 7,500(분야별 2,500)개의 특허에 포함된 초록을 최종 실험 데이터 집합으로 구축한다. 데이터 분석을 위해 불용어 제거 및 형태소 분석을 통한 토크나이징을 수행하여 특허 기술명 코퍼스를 구축한다. 또한, 워드 임베딩은 파이썬 라이브러리 Gensim을 이용하고, Seed set 구축을 위해 특허 초록에서 고빈도로 출현하는 어휘 집합과, CPC 분류코드 명세서에서 나오는 CPC 분류체계 설명 어휘 집합으로부터 대표 키워드를 추출한다. 표 1은 각 분류 코드별로 추출된 대표 키워드를 나타낸다.

<표 1> Seed set 구축을 위한 CPC별 대표 키워드

CPC codes / Methods	B06L 11/18	G06F 17/30	H01L 21/67
CPC descriptions	'전지', '연료', '전력'	'데이터베이스', '정보', '검색'	'반도체', '웨이퍼', '전기'
High frequency	'배터리', '충전', '전기'	'데이터', '검색', '인터넷'	'기판', '웨이퍼', '챔버'

4.1 실험 평가

표 2는 특허 기술 용어 코퍼스에 대한 Seed set의 긍정 값을 각 CPC 코드별로 나타낸 것이다. 긍정 값은 Seed set의 단어들과 코퍼스의 단어들에 대한 유사도를 계산하며, CPC 분류코드 명세서의 키워드를 이용해 Seed set을 구축한 방법과 고빈도의 키워드를 이용해 Seed set을 구축한 방법을 비교한다. 이를 위해 Jaccard 유사도 방법을 이용하고 결과를 표 2에서 보인다.

<표 2> Jaccard 유사도에 대한 긍정 값

CPC codes / Methods	B06L 11/18	G06F 17/30	H01L 21/67
CPC descriptions	0.61	0.57	0.58
High frequency	0.63	0.62	0.61

실험 결과 CPC 분류코드 명세서의 키워드를 이용한 3개의 CPC 평균 긍정 값은 58%이고, 고빈도 키워드를 이용한 3개의 CPC 평균 긍정 값은 62%로 고빈도 키워드를 이용한 경우가 약 4%높다. 이 결과로 CPC 분류코드 명세서의 키워드로 Seed set을 구축하는 경우보다 고빈도의 키워드로 Seed set을 구축하는 경우가 긍정 값이 더 높은 것을 알 수 있다. 비교 결과를 토대로 고빈도의 키워드를 이용한 Seed set을 사용하여 로지스틱 회귀와 분류 기법을 이용해 기술명 분류를 수행한다. 표 3은 로지스틱 회귀를 통해 도출된 기술명과 함께 출현한 문장 패턴을 나타낸다. 각 CPC별 기술 개체명 확률이 높은 기술명과 해당 기술 용어가 등장한 문장 패턴의 일부 예를 보인다.

<표 3> CPC별 문장 패턴 및 일부 예

	Pattern	Sentences(Abstract)
B60L 11/18	**기술명** + "전기" + "자동차"	"하이브리드 전기자동차의"
	"을" + "으로" + **기술명**	"전력을 무선으로 수신하도록"
	"는" + "을" + **기술명** + "한다"	"는 전자력을 제공한다"
G06F 17/30	**기술명** + "인터넷" + "을"	"클라이언트와 인터넷을"
	"를" + "는" + **기술명** + "으로" + "한다"	"를 저장하는 수단으로 한다"
	"에" + "관한" + **기술명** + "에" + "따른"	"에 관한 정보에 따른"
H01L 21/67	"을" + "으로" + **기술명**	"모듈을 진공압으로 흡착하는"
	기술명 + "반도체" + "제조"	"플라즈마 반도체 제조"
	"는" + "을" + **기술명** + "한다"	"는 기판코팅시스템을 개시한다"

5. 결론

본 논문에서는 특허 데이터를 대상으로 기술 개체명 인식 방법을 제안한다. 기술명 식별을 위해 워드 임베딩을 수행하고, 로지스틱 회귀와 분류 기법을 이용해 기술 개체명 분류를 수행하였다. 기술 개체명 정확도 향상을 위해 기술 용어 사전을 확장하여 구축하고 다양한 도메인을 대상으로 한 딥러닝 기반 기술명 인식을 향후 연구로 한다.

참고문헌

[1] Yongmin Park and Jaesung Lee, "Named Entity Recognition and Dictionary Construction for Korean Title: Books, Movies, Music and TV Programs.", KIPS Transactions on Software and Data Engineering, vol. 3, no. 7, pp. 285-292, 2014.

[2] Chanki Lee et. al. "Fine-Grained Named Entity Recognition using Conditional Random Fields for Question Answering.", Asia Information Retrieval Symposium. Springer, Berlin, Heidelberg, pp. 268-272, 2006.

[3] Seunghoon Na and Jinwoo Min, "Character-Based LSTM CRFs for Named Entity Recognition.", Journal of Computing Science and Engineering, vol. 11, no. 1, pp. 32-38, 2016.

[4] Hayoung Lee et al. "A study on the Algorithm for automated extraction for chemical term in Korean patents.", KCI, vol. 27, no. 2, pp. 273-276, 2019.

[5] Jinwoo Min et al. "Character-Based LSTM CRFs for Food Domain Named Entity Recognition.", kiise, pp. 500-502, 2016.