Journal of Digital Convergence
Vol. 18. No. 1, pp. 137-146, 2020

ISSN 2713-6434 / eISSN 2713-6442
https://doi.org/10.14400/JDC.2020.18.1.137

국내외 특허데이터 분석을 통한 자연어처리의 의미분석 관련 기술동향 분석에 대한 연구

현영근[1], 한정현[2], 채우리[2], 이기현[3], 이주연[4*]
[1]아주대학교 산업공학과 석박사통합과정, [2]아주대학교 산업공학과 박사과정,
[3]아주대학교 산업공학과 석사과정, [4]아주대학교 산업공학과 교수

A Study On Technical Trend Analysis Related to Semantic Analysis of NLP Through Domestic/Foreign Patent Data

Young-Geun Hyun[1], Jeong-Hyeon Han[2], Uri Chae[2], Gi-Hyun Lee[3], Joo-Yeoun Lee[4*]
[1]Division of Industrial Engineering, Ajou University, M.D. integration process
[2]Department of Industrial Engineering, Ajou University, Ph.D candidate
[3]Department of Industrial Engineering, Ajou University, M.S. candidate
[4]Division of Industrial Engineering, Ajou University, Professor

요 약 자연어처리 기술은 사람이 말하는 언어를 기계적으로 분석해 컴퓨터가 이해할 수 있는 형태로 만드는 것을 의미한다. 이것이 중요한 이유는 인공지능의 기본인 인간과 디바이스 간 커뮤니케이션을 위한 핵심기술이기 때문이다. 본 논문에서는 자연어처리, 특히 의미분석과 관련된 기술동향을 확인하기 위해 미국과 한국의 특허정보에 대해 분석하였으며, 본 연구를 통해 향후 자연어처리 관련 연구에 의미있는 정보제공을 그 목적으로 한다. 결론적으로, 국내 특허 수는 미국 대비 7.9% 수준이며, 주요 Keyword의 상이한 빈도는 기술적 방향성에 국가별로 차이가 있음을 확인하였다. 또한 상향 또는 하향 성향의 Keyword가 한국 대비 미국이 2배로 나타나 시대적 흐름을 상대적으로 더 반영한 것으로 분석되었다. 향후 연구에서는 실질적인 기술예측을 위해 상향 성향의 Keyword가 특허에서 어떻게 기술되고 있는지 구체적으로 분석하고자 한다.

주제어 : 자연어처리, 의미분석, 인공지능, 특허정보, 키워드 네트워킹

Abstract NLP means the technology that mechanically analyzes a language spoken by a human and makes it into a form that can be understood by a computer. This is important because it is a core technology for communication between humans and devices, which is the basis of artificial intelligence. In this paper, I analyzed patent information of US and Korea in order to identify technical trends related to NLP, especially semantic analysis. and the purpose of this study is to provide meaningful information for future research on NLP. In conclusion, the number of Korea patents is 7.9% compared to the USA and the different frequencies of the major keywords were found to differ from country to country in technical direction. In addition, the upward or downward keywords are twice as many in the U.S. as in Korea, and reflect the trend of the times relatively more. Based on these results, in future study, I will analysis how upward trending keywords are described in actual patents for concrete technology prediction.

Key Words : NLP, Semantic analysis, Artificial Intelligence, Patent information, Keyword Network

*This paper was studied with the support of the 2019 Ajou University Academic Research and Development Fund (S-2019-G0001-00522)
*Corresponding Author : Joo-Yeoun Lee(jooyeoun325@ajou.ac.kr)
Received October 11, 2019 Revised November 26, 2019
Accepted January 20, 2020 Published January 28, 2020

1. 서론

인간은 호모사피엔스, 즉 지혜를 가진 사람이라고 하며, 이것은 인간의 정신적인 능력이 일상생활에서 매우 중요하다는 것을 의미한다. '인공지능(Artificial Intelligence)'란 바로 이러한 인간의 지능적인 작용들을 이해해 보려는 것으로 인간의 지능을 기계가 갖출 수 있도록 하려는데 목표가 있다[1].

인공지능이라는 표현은 1956년 미국 다트머스대학의 컴퓨터사이언스 워크숍에서 처음 등장했다. 2009년 이전 인공지능 연구자들은 인공지능이란 지능적인 기계를 만드는 공학 및 과학(McCarthy etal., 1955), 여러 계산 모델을 이용하여 인간의 정신적 기능을 연구하는 것(Chamiak et al., 1985), 컴퓨터가 특정 순간에 사람보다 더 효율적으로 일을 할 수 있도록 하는 연구(Rich etal., 1991), 지능적인 행동의 자동화에 관한 컴퓨터 과학의 한 부문(Luger etal., 1993)으로 정의하였다(ETRI, 2015a). 최근, 이러한 弱인공지능에서 스스로 사고·판단·예측 그리고 학습·진화할 수 있는 强인공지능 기술로 진화될 것으로 전망하고 있다[1].

전 세계적으로, 그리고 모든 분야에서 인공지능이 화두로 떠오름에 따라, 인간과 인공지능 기기 간 원활한 인터페이스가 가능하도록 하는 자연어 처리 기술 또한 각광을 받고 있다[2,3]. 자연어 처리는 사람이 말하는 언어를 기계언어로 분석하여 컴퓨터가 읽어 들어 작동할 수 있는 형태로 만드는 자연어의 이해나 그러한 형태를 반대로 인간이 이해할 수 있는 자연어로 표현하는 기술을 의미한다[4]. 이러한 커뮤니케이션 기술은 형태소 분석, 품사 태깅, 구문 분석 등의 다양한 기술을 기반으로 한다[5].

자연어 처리 방식으로는 전통적으로 규칙 기반 접근법과 통계기반 접근법이 있으며, 이 둘의 강점을 통합한 하이브리드 방식이 있다. 최근에는 인공신경망 방식이 부상하고 있으며, 딥러닝(Deep Learning)이 이에 해당한다. 딥러닝을 이용한 방식은 입력 문장과 출력 문장을 하나의 쌍으로 두고, 가장 적합한 표현 및 번역 결과를 찾는 방식을 의미한다. 딥러닝이 급부상하게 된 이유는 GPU 성능이 향상되었고, 머신러닝에 활용할 수 있는 데이터가 대량으로 증가하였으며, 인공신경망 알고리즘 또한 개선이 되어 기존 타 알고리즘에 비해 5~10%의 성능 향상을 가져왔기 때문이다[6,7]. 또한 학습, 추론, 인식 등의 복잡한 인공지능 알고리즘을 개발할 수 있는 주요 플랫폼들이 오픈 소스로 공개되면서, 이를 활용한 기술과 서비스들의 개발이 비약적으로 증가하고 있는 것이 주요 요인 중 하나이다[8].

이러한 딥러닝을 기반으로 실생활에서 자연어처리를 적용한 대표적인 사례에는 인공지능 스피커, 텍스트 인식 업무처리 챗봇 등이 있다[5].

과거의 데이터를 사용하여 미래 기술을 예측한다는 것은 쉽지 않은 일이다. 그러나 기술의 발전과 상용화에 근접한 유용하고 효율적인 기술 지표를 나타내는 특허 정보를 이용하면 가능하다. 특허에서 확인할 수 있는 정보를 활용하여 관련 분야의 기술동향을 분석하고, 이를 기술 개발에 적용할 수 있다. 각 특허에는 국제특허분류 코드인 IPC(International Patent Classification)가 부여되며, 그 발명 내용에 따라 1개 또는 그 이상이 될 수 있는데, 기술내용이 여러 개일 경우에는 그 중 가장 중심이 되는 기술 내용을 주분류(Main Category)로 하고, 그 외의 다른 기술내용을 부분류(Sub Category)로 한다[9,10].

본 논문에서는 인간과 다양한 인공지능 디바이스 간 커뮤니케이션의 기본인 자연어처리, 특히 의미분석(Semantic Analysis) 기술에 집중하여 분석을 진행하였으며, 그 이유는 자연어처리의 정확도를 향상시키기 위한 핵심영역이기 때문이다. 연구방법은 자연어처리의 선진국인 미국과 한국의 특허를 비교하여 Keyword 분석, Keyword Network 그리고 시대별 핵심 Keyword의 변화과정을 분석하였다. 이러한 특허정보 분석 및 결과도출을 통해 자연어처리 관련 의미분석 기술 분야의 연구 활성화에 기여하고, 앞으로 진행되어야 할 연구 방향을 위한 중요한 정보를 제공할 수 있으리라 사료된다.

본 논문은 2장에서 자연어처리의 의미분석과 관련한 기존 연구를 분석하고, 3장에서는 특허정보 분석을 위한 연구 프로세스를, 4장에서는 미국과 한국의 특허정보에 대한 Keyword, Keyword Network 그리고 시대별 Keyword 변화추이에 대한 비교분석 결과를 제시함으로써 관련 동향에 대해 고찰 해보고자 한다.

2. 관련 연구

자연어처리는 1940년대 컴퓨터가 등장한 이후 부터 시작되어, 1990년대에 컴퓨터 성능이 크게 발전하게 되면서 대규모 말뭉치(corpus)를 구축하여 다양한 방법을 사용한 통계적 분석이 가능하게 되었다. 2000년대 이후에는 기계학습(Machine Learning)의 급속한 발전과 더불어, 컴퓨터가 지속적으로 추가되는 문서를 바탕으로 스스로 학습하여 숨은 의미패턴을 자동으로 찾고 개선하는 딥러닝(Deep Learning) 기술을 활용하는 방법이 본격

적으로 연구되고 있다[9].

2.1 자연어처리 및 의미분석 연구 동향

이동영(2018)은 자연어 처리의 핵심이라고 할 수 있는 워드 임베딩(Word Embedding)의 5개 알고리즘(CBoW (Continuous Bag-of-Words), RN (Relation Network), CNN (Convolutional Neural Network), Self Attention, RNN(Recurrent Neural Network))의 비교를 통해 자연어처리의 성능향상을 위한 방법(방법론 결합)을 제안하고 있다[4].

이성호, 정윤경(2016)은 영화의 시나리오에 대해 주제(Topic Modeling)외에 감성분석(Opinion Mining)을 위해, SentiWordNet 어휘사전(긍정, 부정, 중립으로 나누어지는 단어의 극성으로 수치로 부여) 및 NLTK (Natural Language Toolkit, 자연어처리의 라이브러리인)을 활용하여 반지도 학습(Semi-Supervised Learning) 알고리즘을 적용하였다. 이를 통해 영화 시나리오의 시간대별(0분~300분) 감정의 긍정도 추이를 분석하는 방법에 대해 제시하였다[11].

김진수(2016)는 SNS를 통해 작성된 짧은 문단 내 함축된 키워드와 키워드들 간의 연관성을 이용하여 문단에 나타난 감정을 예측하는 시스템을 제안하였다. 이 연구의 목적은 예측된 감정에 대해 적절한 답변 혹은 예측된 감정에 부합된 상품/영화를 추천하는 등 활용측면의 가능성을 확인하고자 하였다[12].

한상욱, 김승인(2019)은 다양한 IoT와 연결하여 허브 역할을 수행할 수 있는 대화형 에이전트(인공지능 스피커. 예: 빅스비, 누구, 알렉사 등)가 사용자와의 올바른 상호작용을 위해서 사용자의 감정을 인지하는 것이 핵심이라고 강조하였다. 다만, 해당 논문에서는 감정분석 정확도 향상을 위한 자연어처리 보다는, 대화형 에이전트에 대한 사용자 중심의 설계방향성에 집중하여 분석하였다[13].

2.2 특허 데이터 기반 기술동향 연구 동향

특허 데이터는 출원 및 날짜, 등록자, 특허제목, 기술요약, 인용정보, 상세기술, 도면, 절차도 등 다양한 정보를 포함하고 있으며, 특히 전 세계적으로 건수가 많아 그 활용가치가 매우 높다고 할 수 있다. 또한 특허 데이터는 분석 방법에 따라 기술동향이나 관련 산업/시장 동향 등의 전반적인 흐름을 볼 수 있기 때문에 중요한 데이터로 활용되고 있다[14].

정명석, 정소희, 이주연(2018)은 4차 산업혁명의 핵심기술인 '인공지능'에 대해 특허 데이터를 대상으로 Keyword를 추출하여 국내외 기술동향을 비교 분석하였다. 해당 논문은 본 논문의 데이터 추출 및 분석 방향성은 유사하나, 다만 차이점은 해당 논문은 IPC 분류기준 중심으로 사실전달에만 집중하였다[15].

박재용(2018)도 인공지능에 대해 특허 데이터를 활용하여 관련 기술을 분석하였다. 다만 본 논문과의 차이점은 IPC 분류기준 별 특허등록 건수를 기준(Section, Class, Sub Class, Main Group, Sub Group)으로 기술예측에 집중하여 분석을 진행하였다[9].

노승민은(2014) 4차 산업혁명에서 또 다른 핵심기술인 빅데이터와 관련하여 특허 분석을 통해 클라우드 기반의 빅데이터 플랫폼 연구 및 개발 동향을 분석하였다. 본 논문의 분석방향성과 유사하게, 1991년부터 2010년까지 총 7구간으로 구분하고 기술의 성숙도를 5단계(Begin, Growth, Maturity, Decline, Recovery)로 설정하여 한국, 미국, 일본 및 유럽의 특허를 비교 분석하였다. 다만, 앞 사례와 마찬가지로 특허출원 건수를 기준으로 분석하였다[16].

이재환(2017)은 세계 4대 특허 강국인 중국, 미국, 일본, 한국 4개국을 중심으로 SIPO(중국지식산권국), USPTO(미국특허청), EPO(유럽특허청)에서의 특허 출원 동향을 ICT 기술(의료, 광학, 반도체, 컴퓨터, 통신 등)별로 비교분석 하였다. 해당 논문 또한 특허출원 건수를 기준으로 분석을 진행하였다[17].

관련 논문을 분석한 결과, 특허정보 분석을 통해 자연어처리 기반의 의미분석과 관련한 기술동향 연구사례는 찾아보기 어려운 것으로 확인되었다. 특히, 본 논문의 주요 연구관점인 시대별 동향분석은 특허등록 건수만을 제시했을 뿐, 본 논문과 같이 Keyword 중심의 연구사례는 없었다.

3. 연구 방법론

본 연구의 분석 데이터, 즉 한국 및 미국의 특허 정보는 특허정보넷 KIPRIS(www.kipris.or.kr)에서 다운 받았으며, 한국은 "(자연어처리)*(의미분석)"을, 미국은 "(natural*language*processing)*(semantic*analysis)"을 검색어로 하였다. 2000년 1월부터 현재(2019년 8월30일)까지 해당 검색어로 검색한 결과, 한국은 2,411건 그리고 미국은 30,666건으로, 한국의 특허등록 수는 미국 대비 7.9% 수준인 것으로 확인되었다(단, 특허등록

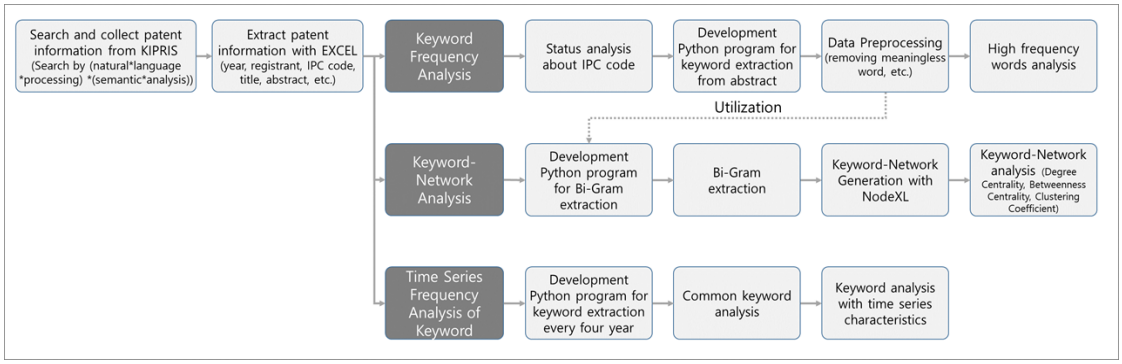

Fig. 1. Methodology of Patent Data Analysis

거절 건 제외). 특히, 2000년 이후의 데이터를 분석대상으로 한 이유는, 한국어의 자연어처리 관련 연구는 정부가 한국어 디지털 언어자원 구축을 목표로 진행한 “21세기 세종 계획”이 그 출발점이라고 할 수 있으며, 2000년부터 초기단계로서의 실질적인 결과인 “세종 말뭉치”가 발표된 시점이기 때문이다.

3.1 분석절차

본 연구는 Fig. 1과 같은 자료수집, 전처리 및 분석절차를 바탕으로 수행하였다. 구체적인 분석과정 전, 데이터를 수집하기 위해 KIPRIS에서 상기 검색어를 검색하여 엑셀로 출원년도, 등록자, IPC 코드, 제목, 초록 등의 특허정보를 다운받았다.

먼저, ‘Keyword 빈도분석’에서는 수집된 초록에서 Keyword(명사)를 추출하기 위해 Python 프로그램을 개발하였으며, “상기”, “발명”, “대한”, “input”, “re” 등 무의미한 단어는 수작업으로 전처리 과정을 수행하였다. 다만 “발명”이라는 단어를 제외한 것은, 예를 들어, 문장상 “본 발명은 자연어처리를 위한...”와 같은 경우가 대부분이여서 삭제처리 하였다.

‘Keyword Network 분석’에서는 ‘Keyword 빈도분석’에서 전처리 과정을 거친 데이터를 활용하여 Bi-Gram을 생성하였으며, 이를 위해 Python 프로그램을 개발하였다. Keyword Network 분석을 위해 Microsoft Office의 Excel 프로그램에 Add-on 프로그램인 NodeXL을 활용하여 Degree Centrality, Betweenness Centrality 그리고 Clustering Coefficient을 분석하였다.

마지막으로 ‘시계열 Keyword 빈도분석’에서는 4년 단위로 키워드 상위 100개를 추출하는 Python 프로그램을 개발하였다. 이를 통해 각 단위별 공통적으로 사용되는 단어와 시계열적으로 그 출현빈도가 높아지는 Keyword, 그리고 낮아지는 Keyword를 분석하여 그 의미를 추론해 보고자 하였다.

4. 분석 결론

특허정보넷 KIPRIS에서 (자연어처리)*(의미분석)를 검색어(미국 검색어는 (natural*language*processing)*(semantic *analysis))로 추출된 특허정보를 기반으로 한국과 미국 특허에 대한 IPC별 빈도, Keyword Network 그리고 Keyword의 시계열적 흐름은 하기와 같이 분석되었다.

4.1 IPC & Keyword Frequency Analysis

국제적으로 특허문헌에 대해 통일된 분류 및 검색체계를 갖추고 있으며 IPC를 그 기준으로 하며, 한국과 미국의 IPC별 특허등록 현황은 Table 1과 같다.

Table 1. Status of Korea and USA patent

IPC Code & Content		Korea	USA
A section	Human Necessities	20	588
B section	Performing Operations; Transporting	48	168
C section	Chemistry; Metallurgy	2	170
D section	Textiles; Paper	3	2
E section	Fixed Constructions	0	7
F section	Lighting; Heating; Weapons; Mechanical Engineering; Blasting	22	22
G section	Physics	2,009	26,182
H section	Electricity	307	3,527
Total	-	2,411	30,666

한국 특허의 경우 G Section(2,009건, 83%)이 가장

많았으며, H Section(307건, 13%), B Section(48건,2%), F Section(22건, 1%), A Section(20건,1%) 순으로 나타났다. 나머지 C Section, D Section 그리고 E Section은 0~3건 정도로 그 비율은 0%(반올림)으로 나타났다.

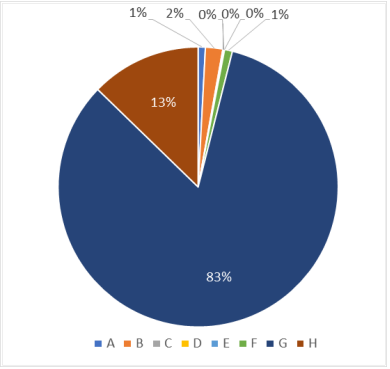

Fig. 2. IPC Ratio of Korea patent

미국 특허 또한 한국 특허와 유사한 형태를 보이고 있으며, G Section(26,182건, 85%)이 가장 많았으며, H Section(3,527건,11%), A Section(588건,2%), C Section (170건,1%), B Section(168건, 1%), 순으로 나타났다. 나머지 F Section(22건), E Section(7건) 그리고 D Section(2건)은 0%(반올림)으로 나타났다.

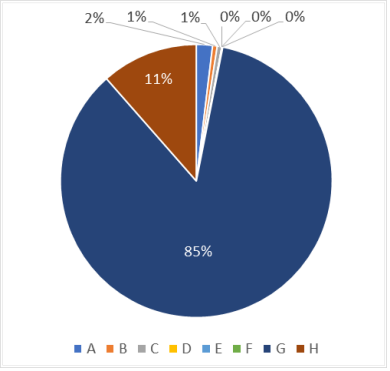

Fig. 3. IPC Ratio of USA patent

상위 20위의 빈도를 보이는 Keyword는 Table 2와 같으며, 그중 공통적으로 높은 빈도수를 보이는 것은 6개(Information, User, Method, System, Device, Search)로 분석되었다. 특이한 것은, 어떤 한 Keyword가 각 국가별 빈도순위에서 상이할 뿐 대부분 공통적으로 사용되고 있다는 것이다. 예를 들어, 한국 특허에서 높은 빈도를 보이는 Mobile(13위)은 미국 특허에서는 901위, 미국에서 높은 빈도를 보이는 Computer(7위)는 한국 특허에서는 135위를 나타내고 있으며, 이것은 자연어처리 기반 의미분석과 관련한 기술적 방향성에서 각 국가별 차이가 있음을 의미할 수 있다.

Table 2. Comparison high frequency words of Korea and USA patent

No.	Korea		USA	
1	information	11,889	system	26,419
2	user	4,048	information	19,524
3	method	3,156	method	18,222
4	phase	3,003	user	15,629
5	data	2,858	plurality	12,701
6	system	2,496	device	11,275
7	voice	2,459	computer	10,941
8	offer	2,338	content	10,525
9	device	2,246	language	10,387
10	terminal	2,207	set	9,807
11	member	2,193	network	8,164
12	search	2,115	model	8,119
13	mobile	2,087	search	7,845
14	analysis	1,949	application	7,353
15	ratio	1,897	query	7,316
16	order	1,832	document	7,089
17	use	1,775	image	6,305
18	extraction	1,671	text	6,192
19	service	1,532	program	6,047
20	product	1,502	processing	5,016

4.2 Keyword Network Analysis

Python으로 추출된 Bi-Gram을 활용하여 Microsoft Office Excel의 Add-on 프로그램인 NodeXL을 통해 Keyword Network를 도출하였다. 다만, 추출된 Bi-Gram이 약 12만개가 되어 NodeXL에서 Keyword Network가 원활히 추출되지 않아 전처리 과정을 통해 Frequency 상위 3,000개를 대상으로 수행하였다.

Fig. 4는 한국 특허정보를 대상으로 Keyword Network를 시각화하여 표현한 것이다.

Keyword Network에서 Keyword간 중심성, 연관성 및 밀집도를 표현하기 위해 Degree Centrality, Betweenness Centrality 그리고 Clustering Coefficient의 정도를 분석하였다.

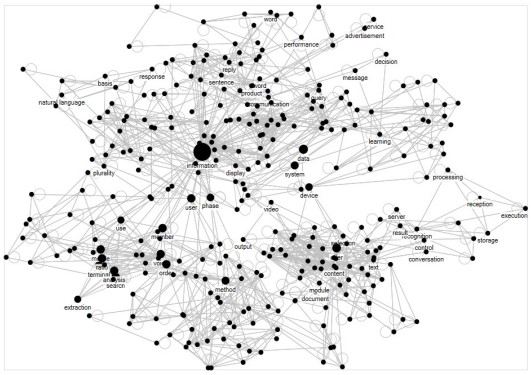

Fig. 4. Fruchterman-Reingo Layout of Korea patent

Table 3. Keyword Network analysis of Korea patent

keyword	Degree Centrality	Betweenness Centrality	Clustering Coefficient
information	393	10,583.804	0.140
user	371	9,860.029	0.152
method	377	9,423.735	0.149
phase	351	7,156.641	0.168
data	330	5,748.033	0.177
system	328	5,518.669	0.180
offer	303	4,807.321	0.192
device	287	4,391.548	0.165
terminal	296	4,365.636	0.208
member	275	3,349.651	0.223
search	258	3,115.361	0.237
mobile	277	2,962.472	0.221
analysis	273	2,877.341	0.235
voice	263	2,812.466	0.215
ratio	264	2,744.130	0.178
order	248	2,340.231	0.247
use	223	2,132.632	0.267
extraction	232	2,113.718	0.261
service	243	2,096.290	0.258
product	193	2,067.721	0.307

먼저, Degree Centrality는 중심성(Centrality) 분석의 가장 기본적인 측정방법으로 각 Keyword가 구성하는 Network에서 각 Keyword가 직접 연결된 다른 Keyword와의 연결(Edge)정도를 측정하여 Network상에서 얼마나 중심에 위치하는지를 알아보는 것으로, 해당 Keyword의 중요도 또는 허브(Hub) 역할을 수행하는 Keyword를 파악할 수 있다. 즉, 연결된 Keyword의 수에 따라 그 중요도와 허브 역할 정도를 판단하는 방법이다. 따라서, 단지 오류를 최소화하기 위해서는 연결정도가 많으나(High Frequency) 의미가 없는 Keyword의 전처리과정이 반드시 필요하다.

Betweenness Centrality는 Degree Centrality의 단점을 보완한 측정방법이며, 단순히 다른 Keyword와 얼마나 연결되었는지를 분석하는 것이 아닌, 전체 Network상에서 해당 Keyword가 얼마나 다른 Keyword들과 잘 연결되어있는가를 분석하는 기법이다. 연관성을 측정할 때 가장 일반적인 방법이 최단경로(Shortest Path)를 이용하는 방법으로, A, B 두 Keyword간에 거리가 얼마나 되는가를 통해 Keyword간의 관계를 분석한다. Betweenness Centrality는 이를 이용하여 A, B간의 영향력을 조사할 때 V라는 Keyword를 꼭 지나가야 한다면 V가 "두 관계를 정의하는데 중심이 되는 역할을 한다" 라는 의미를 수식화하여 중요도를 표현한다. 즉, V Keyword를 거쳐서 가는 경우가 많을수록 V가 중요하다고 표현하는 것이다.

Clustering Coefficient는 특정 Keyword와 이웃한 Keyword들이 서로 연결되어 있을 확률(0부터 1 사이의 값)을 의미하는 것으로, Network상에서 얼마나 뭉쳐져 있는지, 얼마나 밀도가 높은지를 분석하는 것이다. 즉, 특정 Keyword와 이 Keyword와 연결된 주변의 Keyword가 있다고 가정할 때, 이 주변의 Keyword들 간의 연결되어 있을 확률을 분석하여 밀집도 정도를 분석한다.

Fig. 5는 미국 특허의 Key Network에 대해 시각화하여 표현한 것이며, Table 7은 Degree Centrality, Betweenness Centrality 그리고 Clustering Coefficient를 분석한 결과이다.

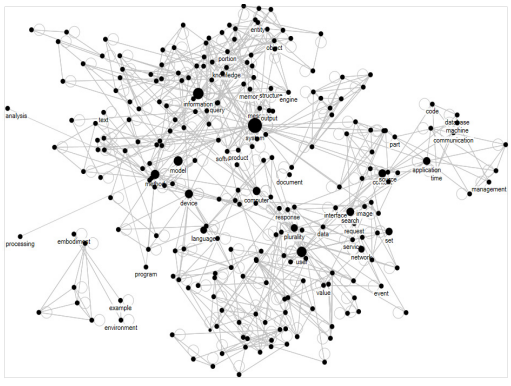

Fig. 5. Fruchterman-Reingo Layout of USA patent

Table 4. Keyword Network analysis of USA patent

keyword	Degree Centrality	Betweenness Centrality	Clustering Coefficient
system	1,008	203,812.715	0.046
information	844	130,618.278	0.058
method	806	110,931.195	0.063
user	650	76,946.715	0.079
plurality	690	76,748.321	0.078

device	621	67,060.131	0.089
computer	654	66,314.312	0.084
content	587	58,002.419	0.097
set	605	52,953.452	0.095
language	502	39,505.824	0.119
network	495	39,234.084	0.120
model	466	36,775.431	0.125
search	497	35,543.747	0.119
application	455	28,248.635	0.137
query	416	25,561.063	0.147
document	406	24,414.650	0.147
image	413	23,827.440	0.153
text	406	21,941.791	0.154
program	420	20,998.279	0.146
processing	381	20,151.400	0.164

4.3 Time Series Frequency Analysis

Table 5는 자연어처리 기반 의미분석과 관련된 한국 및 미국 특허를 2000년부터 2019년(8월)까지 4년 단위, 총 5단계로 나누어 Keyword가 어떤 흐름을 보이는지 비교 분석한 결과이다.

전체 5단계의 상위 20위 Keyword는 총 41개 단어만 사용되고 있으며, 이중 22. 0%인 9개(Information, System, Method, Data, Phase, User, Analysis, Offer, Voice)가 공통적으로 노출되고 있다. 상세내용(노출순위 및 노출빈도 수)은 Table 6과 같다.

Table 6. High frequency words of Korea patent

keyword	'00~'03	'04~'07	'08~'11	'12~'15	'16~'19
information	1(240)	1(430)	1(826)	1(7,141)	1(3,252)
system	3(184)	5(275)	7(344)	12(595)	12(1,098)
method	5(180)	4(356)	4(434)	9(825)	6(1,361)
data	6(156)	8(218)	14(227)	11(601)	3(1,656)
phase	7(155)	10(203)	13(253)	10(749)	4(1,643)
user	8(151)	2(367)	3(486)	4(1,108)	2(1,936)
analysis	10(104)	12(172)	11(274)	19(428)	17(971)
offer	11(102)	7(226)	10(282)	17(445)	8(1,283)
voice	19(81)	20(103)	17(181)	8(890)	9(1,204)

반면, 미국 특허정보에서는 총 28개 Keyword만 사용되고 있고, 그중 50. 0%인 14개(System, Method,

Table 5. Comparison of patent keyword in Korea and USA

No.	'00~'03		'04~'07		'08~'11		'12~'15		'16~'19	
	Korea	USA	Korea	USA	Korea	USA	Korea	USA	Korea	USA
1	information	system	information	system	information	system	information	system	information	system
2	system	information	user	information	content	information	member	information	user	user
3	search	method	search	method	user	method	terminal	method	data	method
4	method	user	method	user	method	user	user	user	phase	plurality
5	data	language	system	computer	advertisement	content	mobile	plurality	device	device
6	phase	network	offer	language	system	plurality	ratio	device	method	information
7	user	computer	data	content	offer	computer	order	set	offer	computer
8	analysis	program	phase	plurality	search	model	product	content	voice	set
9	offer	application	service	search	offer	search	method	computer	system	content
10	processing	model	analysis	model	analysis	device	phase	search	reception	network
11	use	document	use	document	phase	network	data	model	use	image
12	natural language	database	extraction	device	use	set	system	query	recognition	model
13	document	plurality	conversation	network	data	query	device	network	plurality	application
14	storage	text	advertisement	query	extraction	language	advertisement	language	server	language
15	extraction	content	storage	application	word	application	offer	application	execution	search
16	database	code	query	program	voice	document	analysis	text	analysis	query
17	sentence	search	knowledge	set	device	program	search	image	service	response
18	recognition	query	device	database	multimedia	communicatio n	voice	document	communicatio n	processing
19	voice	set	processing	interface'	storage	image	control	service	basis	analysis
20	word	processing	voice	analysis	natural language	text	query	processing	conversation	data

Information, Data, User, Language, Network, Computer, Application, Model, Plurality, Content, Set, Search, Query)가 공통적으로 노출되고 있다. 상세 내용(노출순위 및 노출빈도 수)은 Table 7과 같다.

Table 7. High frequency words of USA patent

keyword	'00~'03	'04~'07	'08~'11	'12~'15	'16~'19
system	1(2840)	1(3,734)	1(4,214)	1(7,380)	1(8,251)
Information	2(2161)	2(2,880)	2(3,652)	2(5,838)	6(4,993)
method	3(1743)	3(2,335)	3(3,157)	3(5,568)	3(5,419)
user	4(1251)	4(1,665)	4(2,310)	4(4,736)	2(5,667)
language	5(1018)	6(1,268)	14(1,198)	14(1,910)	14(4,993)
network	6(841)	13(908)	11(1,326)	13(2,011)	10(3,078)
computer	7(835)	5(1,331)	7(1,769)	9(2,860)	7(4,146)
application	9(680)	15(853)	15(1,193)	15(1,830)	13(2,797)
model	10(670)	10(1,041)	8(1,479)	11(2,117)	12(2,812)
plurality	13(616)	8(1,154)	6(1,839)	5(3,758)	4(5,334)
content	15(514)	7(1,175)	5(2,062)	8(3,202)	9(3,572)
search	17(513)	9(1,094)	9(1,358)	10(2,180)	15(2,700)
query	18(506)	14(887)	13(1,255)	12(2,019)	16(2,649)
set	19(502)	17(819)	12(1,306)	7(3,212)	8(3,968)

한국과 미국 모두에게서 사용되는 공통 Keyword는 4개(Information, System, Method, User)이며, 특히 그 노출빈도가 가장 높은 1위 ~ 3위의 단어가 한국과 미국이 동일하다는 것이 가장 큰 특징이라고 할 수 있다.

상기 Keyword에 대해 상위 50위까지 확대하여 시계열적 Keyword의 변화추이를 분석한 결과, 상승 및 하강하는 Keyword가 확인되었으며, 한국 특허정보의 경우, 출현 빈도가 하강하는 Keyword는 4개(System, Search, Storage, Extraction), 상승하는 Keyword는 2개(Voice, Device)로 확인되었다.

Table 8. Upward trend words of Korea patent

keyword	'00~'03	'04~'07	'08~'11	'12~'15	'16~'19
voice	19(81)	20(103)	17(181)	8(890)	9(1,204)
device	-	19(120)	18(174)	13(495)	5(1,457)

Table 9. Downward trend words of Korea patent

keyword	'00~'03	'04~'07	'08~'11	'12~'15	'16~'19
system	3(184)	5(275)	7(344)	12(595)	12(1,098)
search	4(183)	3(356)	9(283)	20(409)	30(884)
storage	14(87)	16(129)	20(154)	-	22(928)
extraction	15(86)	13(152)	15(219)	26(317)	27(897)

반면, 미국 특허정보의 경우, 하강하는 Keyword는 6개(Program, Database, Code, Interface, Structure, Software), 상승하는 Keyword 또한 6개(Plurality, Set, Device, Time, Image, Response)로 분석되었다.

Table 10. Upward trend words of USA patent

keyword	'00~'03	'04~'07	'08~'11	'12~'15	'16~'19
plurality	13(616)	8(1154)	6(1839)	5(3758)	4(5,334)
set	19(502)	17(819)	12(1306)	7(3212)	8(3,968)
device	27(416)	12(916)	10(1333)	6(3491)	5(5,119)
time	33(330)	34(481)	33(603)	24(1436)	21(2,310)
image	34(325)	31(530)	19(898)	17(1734)	11(2,818)
response	39(307)	35(456)	29(648)	23(1482)	17(2,642)

Table 11. Downward trend words of USA patent

keyword	'00~'03	'04~'07	'08~'11	'12~'15	'16~'19
program	8(776)	16(847)	17(984)	27(1,260)	26(2,180)
database	12(644)	18(793)	23(823)	29(1,226)	35(1,912)
code	16(513)	24(618)	25(784)	33(1,149)	43(1,814)
interface	21(457)	19(740)	24(793)	25(1,403)	27(2,161)
structure	24(430)	32(520)	40(509)	-	-
software	26(421)	29(562)	30(647)	45(904)	-

한국과 미국 공통적으로 상승 및 하강하는 경향의 Keyword는 Device가 유일하며, 이것은 Device를 통해 의미분석을 위한 자연어처리 연구가 지난 20년간 지속적으로 향상되고 있음을 의미할 수 있다.

5. 결론

한국과 미국의 특허정보의 IPC 분류는 그 비율만 다소 상이할 뿐 거의 유사한 것으로 확인되었다. 공통적으로 높은 빈도를 보이는 Keyword는 6개(Information, User, Method, System, Device, Search)이나, 범위를 확대하여 비교해보면 그 순위만 상이할 뿐 사용되는 Keyword는 매우 유사함을 확인할 수 있었다(예: Mobile: 한국 13위, 미국 901위 / Computer: 한국 135위, 미국 7위). 하지만, 이것이 각 국가별 동일한 기술적 특성을 보인다라기 보다는, 오히려 기술적 방향성 및 트랜드에서 차이가 있음을 의미한다고 할 수 있다.

Keyword Network 분석에서는 Degree Centrality, Betweenness Centrality 그리고 Clustering

Coefficient를 분석하였다. 미국 특허정보의 경우 Degree Centrality와 Betweenness Centrality가 높은 반면, 한국 특허정보는 Clustering Coefficient가 높은 것으로 분석되었다. 즉 미국 특허정보의 경우 중심에 위치한 Keyword와 다른 Keyword간의 연결정도가 상대적으로 높은 반면, 한국 특허정보는 이웃한 Keyword들 간의 상호연결 정도, 즉 Network상에서 상대적으로 뭉쳐져 있는 정도가 높다는 것을 의미한다.

마지막으로 Keyword의 시계열적 특성을 분석한 결과, 지속적으로 노출된 단어의 수가 한국보다 미국이 2.5배(한국 20.0%, 미국 50.0%)로 나타났으며, 이것이 곧 특허정보가 경직되었다는 것을 의미하지는 않는 것으로 예상된다. 왜냐하면 상승 또는 하강하는 Keyword의 수가 한국은 6개(Voice, Device / System, Search, Storage, Extraction), 미국은 12개(Plurality, Set, Device, Time, Image, Response / Program, Database, Code, Interface, Structure, Software)로 분석되었기 때문이다.

향후 연구에서는, 자연어처리 기반의 의미분석에 대한 실질적인 기술예측을 위해, 상향 성향의 키워드, 그리고 하향 성향의 키워드가 실제 특허에서 어떻게 기술되었는지를 살펴보고 그 의미를 분석하고자 한다. 특히 국가별 차이를 함께 분석한다면 보다 의미있는 기술예측이 가능하리라 예상된다.

REFERENCES

[1] J. K. Hong. (2017). Artificial Intelligence and Natural Language Processing. *narasarang Publisher, 126,* 128-148.

[2] H. J. Lee & J. W. Kim. (2017). A Study on the Natural Language Processing(NLP) Technical and Standardization Trend. *Telecommunications Technology Association,* 876-877.

[3] H. C. Lim & H. S. Lim & B. H. Yoon. (1994). Natural Language Processing research trend. *The Korean Institute of Information Scientists and Engineers, 12(9),* 20-30.

[4] D. Y. Lee. (2018). Natural Language Processing Research. *KOREA INFORMATION SCIENCE SOCIETY.* 771-1773.

[5] T. K. Lee & K. S. Shin. (2019) Performance Comparison of Natural Language Processing Model Based on Deep Neural Networks. *Korea Institute Of Communication Sciences, 44(7),* 1344-1350. DOI :10.7840/kics.2019.44.7.1344

[6] K. H. Park, S. H. Na, J. H. Shin & Y. K. Kim. (2019). BERT for Korean Natural Language Processing: Named Entity Tagging, Sentiment Analysis, Dependency Parsing and Semantic Role Labeling. *The Korean Institute of Information Scientists and Engineers,* 584-586.

[7] D. M. Park. (2016). Natural Language Processing of News Articles : A Case of NewsSource beta. *Korean Society For Journalism And Communication Studies,* 4-52.

[8] J. S. Chong, D. S. Kim, H. J. Lee & J. W. Kim. (2019). A Study on the Development Trend of Artificial Intelligence Using Text Mining Technique : Focused on Open Source Software Projects on Github. *Korea Intelligent Information Systems Society, 25(1),* 1-19. DOI : 10.13088/jiis.2019.25.1.001

[9] J. Y. Park. (2018). Trend Analysis of Artificial Intelligence Technology Using Patent Information. *The Korean Society Of Computer And Information, 23(4),* 9-16. DOI :10.9708/jksci.2018.23.04.009

[10] S. M. Han, Y. W. Kim, S. Y. Yim, S. M. Jeong & Y. S. Shin. (2018). Technology & Industry Evaluation and Prediction Using Patent Data. *KOREA INFORMATION SCIENCE SOCIETY,* 426-428.

[11] S. H. Lee & Y. G. Cheong. (2016). Natural language Processing of Scenarios for Sentiment Analysis, *The Korean Institute of Information Scientists and Engineers,* 1671-1673.

[12] J. S. Kim. (2016). Emotion Prediction of Paragraph using Big Data Analysis. *Journal of Digital Convergence, 14(11),* 267-273. . DOI : 10.14400/JDC.2016.14.11.267

[13] S. W. Han & S. I. Kim. (2019). Suggestion of a Social Significance Research Model for User Emotion. *Journal of the Korea Convergence Society, 10(3),* 167-176. DOI :10.15207/JKCS.2019.10.3.167

[14] J. C. Choi. (2018). Big Data Patent Analysis Using Social Network Analysis. *Journal of the Korea Convergence Society, 9(2),* 251-257. DOI :10.15207/JKCS.2018.9.2.251

[15] M. S. Chung, S. H. Jeong & J. Y. Lee. (2018). Analysis of major research trends in artificial intelligence based on domestic/international patent data. *Journal of Digital Convergence, 16(6),* 1-9. DOI :10.14400/JDC.2018.16.6.000

[16] N. S. Min. (2014). Big Data Analysis Platform Technology R&D Trend through Patent Analysis. *Journal of Digital Convergence, 12(9),* 169-175. DOI : 10.14400/JDC.2014.12.9.169

[17] J. H. Lee. (2017). An Analysis on ICT-related Patent Trend in Leading Countries. *Korea Institute Of Communication Sciences,* 711-712.

현 영 근(Young-Geun Hyun) [정회원]

· 2018년 8월 ~ 현재 : 아주대학교 공과대학 산업공학과 석박사통합과정
· 2017년 8월 ~ 2018년 7월 : 아주대학교 공과대학 산업공학과 석사과정
· 2018년 1월 ~ 현재 : SK 주식회사 C&C DT 전략 Marketing 그룹
· 2004년 12월 ~ 2017년 12월 : SK 주식회사 C&C 제안전략 Consultant
· 1999년 8월 ~ 2004년 11월 : SI Computer programer
· 관심분야 : 융합기술연구, AI, Business Automation
· 저서 : 고객의 마음을 움직이는 제안전략
· E-Mail : hyunyg@ajou.ac.kr

한 정 현(Jeong-Hyeon Han) [정회원]

· 2018년 9월 ~ 현재 : 아주대학교 공과대학 산업공학과 박사과정
· 2017년 7월 ~ 현재 : SK 주식회사 C&C Digital 총괄 Vitality 그룹
· 2007년 1월 ~ 2017년 6월 : SK 주식회사 C&C Business Management
· 1997년 7월 ~ 2006년 12월 : SK 주식회사 C&C Software Engineering
· 관심분야 : 융합기술연구, 디지털 헬스케어, 빅데이터
· E-Mail : hann@ajou.ac.kr

채 우 리(U-ri Chae) [정회원]

· 2017년 3월 ~ 현재 : 아주대학교 공과대학 산업공학과 석박사통합과정
· 2015년 3월 ~ 2017년 2월 : 아주대학교 공과대학 산업공학과 석사과정
· 관심분야 : 융합기술연구, 직업병·직업성질환, 데이터분석
· E-Mail : chaeuri@ajou.ac.kr

이 기 현(Gi-Hyun Lee) [학생회원]

· 2018년 2월 : 남서울대학교 산업공학과(이학사)
· 2018년 3월 ~ 현재 : 아주대학교 산업공학과 석사과정
· 관심분야 : 융합기술연구, 데이터분석, 신재생에너지
· E-Mail : black9255@ajou.ac.kr

이 주 연(Joo-Yeoun Lee) [종신회원]

· 2002년 2월 : 인하대학교 대학원경영학박사
· 2014년 9월 ~ 현재 : 아주대학교 공과대학 산업공학과 교수
· 2015년 2월 ~ 2018년 1월 : 산업통상자원부 산업융합촉진 국가옴부즈만(차관급)
· 2016년 7월 ~ 2019년 6월 : 한국빅데이터서비스학회 학회장
· 2007년 7월 ~ 2011년 6월 : 한국산업정보학회 회장
· 2011년 12월 ~ 2014년 3월 : POSCO ICT 그린사업부문장(전무)
· 2005년 2월 ~ 2011년 11월 : SK C&C 전략마케팅본부장(상무)
· 1999년 12월 ~ 2005년 1월 : Oracle 전략솔루션실장(Director)
· 관심분야 : 산업융합기술, 비즈니스인텔리전스, 서비타이제이션, 스마트그리드
· E-Mail : jooyeoun325@ajou.ac.kr