ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2020 September: 26(3): 127~147
http://dx.doi.org/10.13088/jiis.2020.26.3.127

국가 감염병 공동R&D전략 수립을 위한 분류체계 및 정보서비스에 대한 연구: 해외 코로나바이러스 R&D과제의 분류모델을 중심으로*

이도연
한국과학기술정보연구원 데이터분석본부
(dylee@kisti.re.kr)

이재성
과학기술연합대학원대학교 과학기술경영정책학과
(jslee@kisti.re.kr)

전승표
한국과학기술정보연구원 데이터분석플랫폼센터
(spjun@kisti.re.kr)

김근환
한국과학기술정보연구원 데이터분석본부
(khkim75@kisti.re.kr)

세계는 신형 코로나바이러스 감염증(COVID-19)으로 수 많은 인명 피해와 경제적 손실을 기록하고 있는 상황이다. 우리나라 정부는 연구개발(Research & Development)을 통해 국가 감염병 위기를 극복하려는 전략을 수립하고 실행하기 위한 투자방향을 수립하였다. 기존 기술분류나 과학기술 표준분류에 따른 통계를 활용하면 특정 R&D 분야의 특이점 및 변화를 발견하기 어렵다는 한계가 존재해왔다. 최근 우리나라 감염병 연구개발 과제를 대상으로 수요자의 목적에 맞게 분류체계를 수립하고 연구비 비교 분석을 통해 투자가 요구되는 연구 분야를 제시하는 연구들이 진행되었다. 하지만 현재 국가 보건 안보와 신성장 산업육성이라는 목표를 달성하기 위한 실행방안으로 요구되고 있는 전염병 연구분야의 국가간 협력전략 수립에 필요한 정보를 체계적으로 제공하고 있지 못한 상황이다. 따라서 국가 공동 연구개발 전략 수립을 위한 분류체계와 분류모델기반의 정보서비스에 대한 연구가 요구되고 있다. 우선 감염병관련 NTIS 과제데이터를 기반으로 정성분석을 통해 7개의 분류체계를 도출하였다. 스코퍼스(Scopus) 데이터와 양방향 RNN모델을 사용하여, 분류체계 모델을 학습시켰다. 최종적인 모델의 분류 성능은 90%이상의 높은 정확도와 강건성을 확보하였다. 실증연구를 위해 주요 국가의 코로나바이러스 연구개발 과제를 대상으로 전염병 분류체계를 적용하였다. 주요 국가의 감염병(코로나바이러스) 연구개발 과제를 분류체계별로 분석한 결과, 세계적으로 유행하는 바이러스의 예상치 못한 창궐이 확산되는 속도에 비해 백신과 치료제 개발이 제대로 이뤄지지 않는 원인의 배경을 간접적으로 확인할 수 있었다. 국가별 비교분석을 통해 미국과 일본은 상대적으로 모든 영역에 골고루 연구개발 투자를 하고 있는 것으로 나타난 반면, 유럽은 상대적으로 특정 연구분야에 많은 투자를 하는 집중화 전략을 취하는 것으로 나타났다. 동시에 주요 국가의 코로나 바이러스 주요 연구조직에 대한 정보를 분류체계별로 제공하여 국제 공동R&D 전략의 기초정보를 제공하였다. 본 연구 결과를 통해 세 가지 정책적 의미를 도출할 수 있다. 첫째, 데이터기반 과학기술정책 관점에서 수요자 관심분야에 대한 국가 R&D사업의 정보를 글로벌 기준으로 문서를 분류하는 방안을 제시하였다. 둘째, 감염병관련 국가 R&D사업 영역에 대한 정보분석 서비스 기획의 기반을 마련하였다. 마지막으로 국가 감염병 R&D 분류체계 수립을 통해 분류 체계의 궁극적 목표인 산업, 기업, 정책 정보를 제공할 수 있는 기반을 마련한 것이다.

주제어 : 국가 연구개발 과제, 감염병, 문서분류, 양방향 RNN, 코로나바이러스

논문접수일 : 2020년 8월 18일　논문수정일 : 2020년 9월 11일　게재확정일 : 2020년 9월 18일
원고유형 : 일반논문(급행)　교신저자 : 김근환

---

* 이 연구는 한국과학기술정보연구원 주요사업(No. K-20-L03-C03-S01)의 지원을 받아 수행되었습니다.

1. 서론

중국 후베이성 우한('19년12월12일)에서 원인을 알 수 없는 호흡기 폐렴 사례가 WHO에 처음 보고('19년12월31일)된 이후, 전 세계로 급속히 확산되면서 세계보건기구(WHO)는 2020년 3월12일 2019년 신형 코로나바이러스 감염증(Coronavirus Disease 2019, COVID-19)을 세계적 대유행(Pandemic)으로 선포하였다(Lee et al., 2020). 2020년 2월 20일 첫 확진자가 발생한 이후 현재(2020년 8월 15일) 우리나라 코로나 감염자는 15,039명(사망:305명), 세계적으로는 21,050,003명(사망:757,003명)에 이르고 있으며, 아직까지 특별한 치료법이 없는 상황이다. 세계보건기구(WHO)는 신종 코로나바이러스감염증(COVID-19)으로 전 세계가 한 달에 3750억달러(약 444조원)가 넘는 경제적 손실을 기록하고 있다고 분석하였고(Hankookilbo, 2020), 우리 경제의 주축인 수출은 2020년 2분기 16.6% 급감하여 1963년 4분기(-24%) 이후 56년 6개월만에 최악을 맞이하고 있다(Joseilbo, 2020).

정부에서는 연구개발(R&D)를 통해 국가 감염병 위기에 대응하려는 전략을 수립하고 있다. 이미 '제4차 과학기술 기본계획('18~'22)'에서는 신·변종 감염병 발생 및 대유행 등 국가적 재난상황에 대응하기 위한 감염병 대응 시스템 강화 전략 제시하였고, '제2차 보건의료기술육성기본계획('18~'22)'에서는 범부처 협업을 통한 국가 방역체계 강화 및 국내 백신 자급화를 위한 백신 개발 인프라 지원 추진해오고 있다(Ministry of Health and Welfare, 2019). COVID-19 사태 이후 정부는 기존에 수립된 투자방향에 포스트 코로나 시대 대응역량 강화를 위해 "2021년도 정부연구개발 투자방향 및 기준(안)('20.3.12.)"을 통해 제시하였다. (Ministry of Science and ICT, 2020). 포스트 코로나 시대 신성장산업으로 육성하기 위해 감염병 대응체계인 방역·예방, 진단·검사, 백신·치료 등 3대 영역별 경쟁력 강화를 적극 지원하기 위한 "감염병 대응 산업 육성 방안('20.5.14.)"을 제시하였다(Ministry of Economy and Finance, 2020).공통적으로 신·변종 감염병 대응을 위해 범부처 차원의 사전대비 R&D 강화, 국가/부처 간 협력 협업 내실화를 통한 국가 보건 안보를 강화하고 신성장 산업을 육성하려 목적을 담고 있다.

세계적으로 데이터를 활용한 증거 기반 R&D 정책을 기획해야 한다는 '과학기술정책의 과학화(Science of Science Policy)'가 추진되고 있으며, 이를 통해 합리적 의사결정에 근간이 되는 유용한 정보를 제공하려는 노력을 지속하고 있다. 물론 정책 효과를 예측하는 것은 인력, 재정, 정책수행 부처의 정치적 영향력, 기존 인프라 역량 등 정책수단의 결정과 수행에 요구되는 다양한 자원에 대한 요인들을 종합적으로 분석해야 가능하다. 그러므로 특정 정보만으로 의사결정자들에게 영향을 주어 정책의 성과를 예측하거나 설명하는 것은 매우 어려운 일이다. 따라서 특정 정책 문제의 영역을 대상으로 데이터를 활용하여 일관된 분류체계를 통해 비편향적이고 신속하게 기초정보를 제공하여 정책수립과정에서 이해관계자들간의 필요한 토론과 협의과정에 시작단계에서부터 추진하는 것이 더 효과적이라고 할 수 있다. 즉, 사용자 중심의 정보를 분류하여 이질적인 정책적 의사결정자들에게 신속하고 비편향적인 맞춤형 정보를 제공하는 것이 필수

선결과제이다(Hong et al., 2016). 기존 기술분류나 과학기술 표준분류에 따른 통계를 활용하면, 각 사업의 성격, 특징이 정확하게 드러나지 않는 등 특정 R&D 분야의 투입 대비 성과, 특이점 및 변화를 발견하기 어렵다는 한계가 존재해왔다. 따라서 R&D 사업의 목적별, 속성별, 단계별 등 전략적인 기술분야의 집중 및 정부 R&D 투자 이후 향후 성과를 예측하거나 전략적 예산 배분을 위한 미래 투자 포트폴리오 조정을 위해서 특정 분야의 특성에 맞는 보다 세분화된 분류기준 및 분류체계 연구의 필요성과 중요성이 지속적으로 요구되고 있다. 따라서 본 연구는 현재 국가 정책 및 관련 연구자들의 최대 관심분야인 감염병을 대상으로 국가 보건 안보와 신성장 산업 육성이라는 목표를 달성하기 위한 실행방안으로 요구되고 있는 국가간 협력전략 수립에 필요한 기초정보를 제공할 목적으로 분류체계 및 문서분류에 대한 방법론을 제안하고, 해외 코로나바이러스관련 국가 R&D 과제를 대상으로 실증분석을 수행하였다. 이를 통해 관련 이해관계자들이 정책 효과를 예측하기 위해 요구되는 다양한 요인들과 함께, 해외 공동연구가 요구되는 정책적 특성이 필요한 감염병(코로나바이러스)관련 국가 R&D 현황 분석의 기반을 마련하고, 해외 주요 국가의 정부R&D사업을 분류체계에 따라 적용하여 분류체계별 해외 공동연구 조직에 대한 기초정보를 제공하여 감염병(코로나바이러스) R&D관련 이해관계자간 토론과 협의에 필요한 일관성 확보 및 체계적 관리의 초석을 제공하고자 하였다.

2. 이론적 배경과 연구 모델

2.1. 국가 과학기술분야 분류체계 선행 연구

자료 분류란 “사물이나 현상, 개념 등을 유사한 것은 모으고, 상이한 것은 구분하여 체계화하고, 그 결과 분류된 사상의 명칭이 체계적으로 배열된 표”로 정의된다 (Seo and Kim, 2015). 특히, 국가 R&D사업을 특정 분야 및 정책적 목적에 맞게 분류하고, 분류된 정보를 활용하여 연구개발 투자의 효율화와 과학기술정책의 효과를 증진시키기 위한 많은 연구들이 진행되어져 왔다.

김상균 외(Kim et al., 2008)는 일반적인 한의학분류체계로는 한의학연구원에서 수행하고 있는 한의학 연구개발과제의 특성이 반영된 분류로 적합하지 않은 문제가 발생하여 연구개발의 효율적 운영을 어렵게 하는 문제가 있어서, 한의학 연구개발과제를 세부적으로 분류할 수 있는 특징을 중심으로 총 9개 대분류와 33개 중분류 그리고 12개 소분류를 제시하였다. 곽승진&이혜영(Kwak & Lee, 2013)은 연간 7,500억원의 에너지기술 R&D 연구비를 투여하는 한국에너지기술평가원의 연구과제에서 생성되는 연구보고서의 수집, 관리, 서비스 및 공유를 위해 에너지기술분야의 고유분류체계를 제안하여 연구자와 산업현장의 이해당사자들에게 효율적인 정보를 제공하고자 하였다. 문세영(Moon, 2015)은 생명·보건의료 분야의 대표적인 산출물인 신약개발분야 정부 R&D 과제들을 대상으로 과학기술표준분류, 연구개발단계, 질환에 대한 분류기준을 추가하여 국가연구개발투자 효율성과 효과성 제고를 위한 분류기준을 제안하였다. 정명진 외

(Jung et al., 2015)는 고령화에 따른 건강한 삶과 질병 퇴치라는 사회적 수요로 인해 주목받고 있는 HT(Health Technology: 보건기술)은 다른 요소기술과의 융합화로 인해 분류군간 배타성이 충분하지 않은 문제를 안고 있는 국가과학기술분류체 대신 HT융합기술을 119개로 분류하는 방안을 제시하였다. 유진석 외(Yoo et al., 2016)는 정부차원에서 온실가스 감축을 위해 진행해야 할 기술적 범위에 대한 논의를 하는데 중요한 정책적 방향을 제공하고, 기술별·분야별로 정부 R&D 투자현황을 분석할 수 있는 기후기술 분류체계를 제시하였다.

김주원, 여창민(Kim and Yeo, 2020)은 감염병 8대 주요 연구분야인 감시/예측, 임상/정책, 기초/기전, 진단기술, 치료기술, 백신, 인프라, 방역/방제로 분류하여 우리나라 연구개발과제를 대상으로 연구비 분석을 수행하였다. 질병관리본부(Korea Centers for Disease Control and Prevention, 2020)은 9대 중점분야(신변종 감염병, 기후변화 및 인수공통 감염병, 인플루엔자, 세균감염 및 항생제 내성, 결핵, 만성감염, 감염병 재난 대비 및 관리, 예방접종질환 및 백신, 국제협력)를 대상으로 국가 연구개발과제를 분석하여 국가-부처-중점분야간 감염병 R&D 예산 비교를 통해 투자가 요구되는 분야를 제시하였다. 지금까지 국가 R&D사업 투자 목적에 맞는 분류체계를 많은 연구에서 제시하였고, 감염병분야에서도 국내 연구개발 투자 자료를 바탕으로 목적에 맞는 분류체계를 수립하고 분석을 수행하여 의사결정에 필요한 정보를 제공하고자 하였다. 하지만, 감염병분야은 국가간 협력이 강하게 요구되는 분야로 국제 협력을 위한 분류체계와 분류체계별 협력조직에 관한 정보를 제공해야하는 필요성이 절실히 요구되고 있다.

2.2. 과학기술분야 분류모델에 대한 선행 연구

디지털 형태로 작성된 정보들의 양과 접근성이 증가함에 따라 관련 있는 정보를 걸러내는 기술이 중요하다(Folts and Dumais, 1992). 이와 관련된 연구는 강력한 컴퓨팅 파워를 기반으로 하는 기계학습에 기반을 두고 이루어지고 있다(Yang, 1999; Sebastiani, 2002). 이러한 기계학습 기반의 분류 모델은 전문가의 정성적 판단에 근거한 도메인 지식에 상대적으로 독립적이고, 보다 많은 양의 정보를 분류해 낼 수 있다는 장점이 있다.

과학기술분야에서도 이러한 시도가 많이 이뤄졌는데, 이상의 분류 기능을 지원하는 오늘날 유사 시스템들의 기원은 SDI(Selective Dissemination of Information)에서 유래한다고 볼 수 있다. 해당 시스템은 연구자들에게 그들의 전문화된 영역에서 새로 출판되는 문서들에 대한 정보를 알려주도록 고안됐다. 연구자들은 그들의 관심사에 맞는 정보를 보다 잘 받기 위해서 자신의 관심 분야 영역을 새롭게 설정하거나 수정했다. SDI는 이러한 정보를 바탕으로 연구자의 관심사와 가장 관련성이 높은 연구논문 자료를 제공했다(Houseman et al., 1970). 이러한 분류 모델의 기능은 연구자에게 정보 비대칭성으로 인해 인지하지 못하고 있었던 유용한 정보를 제공할 수 있다(Chen et al., 2013).

이상의 기술은 다음과 같이도 적용할 수도 있

다. 예컨대 고기술 수준을 가진 기업이나 선도적인 제조업체들이 협력을 구하기 위해 공공 웹사이트에 그들의 어려운 사안을 포스팅할 수 있다. 하지만 관련이 높은 연구자들이나 실무자들이 해당 과제 기회 정보를 접하는 채널을 찾지 못해 이러한 포스팅에 노출되지 못할 수도 있다. 이때 분류 모델을 응용하면 적절한 R&D 과제 후보 집합을 찾아주는 단계와 그 집합에서 과제 신청자를 위해 적절한 R&D 과제를 분류하는 기능을 수행할 수 있다(Xu et al., 2016).

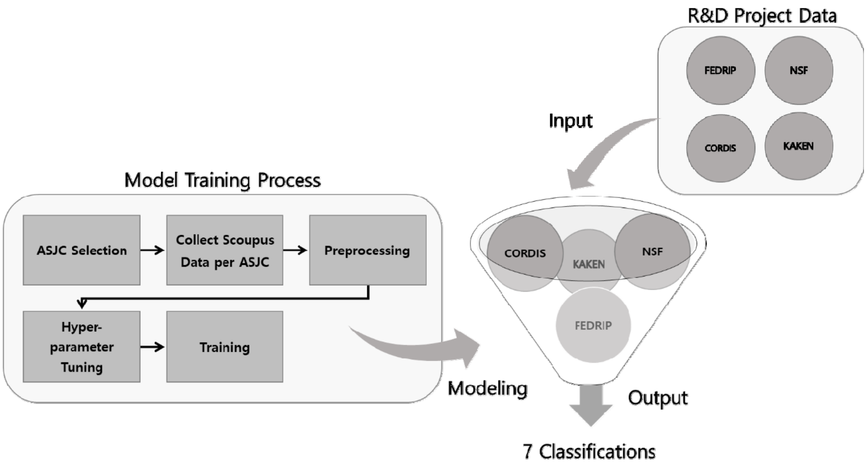
한편 Tian 등은 과학기술 분류 모델을 조직의 업무수행 성격에 가장 적절한 R&D 과제를 찾는데 활용했다(Tian et al., 2005). Trappey 등은 기업제품의 혁신적인 디자인 협력을 위해 비슷한 기술내용을 보유한 특허를 분류했다(Trappey et al., 2013). 그리고 Jeong 등은 국가 R&D 과제들에 대한 위원회를 구성하기 위해 연구자들의 연구논문들을 토대로 위원회 분과에 맞는 전문가들을 분류하는 모델을 구축했다(Jeong et al., 2016).

이에 본 연구는 이러한 과학기술 분류기술을 응용해 국가 감염병 R&D와 관련된 연구분야를 분류하고자 하였다. 이러한 접근과 가장 비슷한 선행연구로는 Tang 등의 연구가 있다(Tang et al., 2012). 그는 연구자들의 다학제적 연구협력 구축의 어려움을 개선하기 위해 Cross-domain Topic Learning(CTL) 모델을 연구했다. 해당 모델은 방대한 양의 연구논문에서 분류된 연구 도메인 간의 독특한 패턴을 고려해 주제에 맞는 적절한 연구자들을 분류하는데 활용했다. 이에 본 연구는 앞서 선행연구 2.1에서 주장했던 감염병분야의 국제적 협력을 이끌어낼 수 있는 방법을 제시함으로써 전략적 의사결정의 기초자료로 활용될 수 있을 것이다.

2.3. 연구방법론

국가 감염병 R&D와 관련된 연구개발 과제를 분류하기 위해 본 연구는 네델란드의 세계적인 학술논문 출판사인 엘스비어(Elsevier)가 관리하고 있는 스코퍼스(Scopus)의 DB (Database)에서 연구논문 데이터를 학습하는 분류 모델을 고안했다. 본 연구에서 연구논문 데이터를 학습데이터로 선정하고, 수집 및 학습모델 구축에 활용한 이유는 다음과 같다. 본 연구의 목표는 우리나라 국가 감염병 R&D 관련 연구분야의 정책, 기획, 공동연구 전략, 예산 분배, 관리 및 운영의 효율성을 제고하기 위한 기초정보로 활용 가능한 정보의 제공이므로, 신뢰성 있는 감염병 관련 분류체계의 기준을 설정하는 단계와 이러한 분류체계를 과학적 관점에서 체계적으로 잘 반영하는 데이터의 수집이 무엇보다 중요했다. 먼저, 과학적 관점의 체계적인 분류체계의 기준 및 설정을 위해 질병관리본부에서 감염병 분야 전문가들의 의견을 수렴하여 국가 감염병 위기대응 추진전략을 이행을 위한 범부처 R&D 시행계획 및 신규 사업 기획을 위해 발간한 정책 연구보고서를 참고하였다(Ministry of Health and Welfare, 2019). 상기 정책연구보고서에 의하면 감염병 연구분야별 중점 기술을 (1)기전연구(기초), (2)감시/예측, (3)역학, (4)진단, (5)백신, (6)치료제, (7)방역/방제, (8)정책, (9)인프라 등 총 9개 분야로 구성하고 있다. 이에 본 연구는 상기 중점 기술 분야에 기초하여 감염병 관련 R&D 과제의 기술분류를

유형화하는 과정에서 감염병 전문가 집단(대학 및 민간 기업의 바이오신약개발 전문가, 한국생명공학연구원 감염병연구센터, 바이오나노연구센터 소속의 전문가)을 구성하여 전문가 검토를 거친 후, 비R&D 영역을 제외한 (1)기전연구(기초), (2)감시/예측, (3)역학, (4)진단, (5)백신, (6)치료제, (7)방역/방제에 해당하는 7개의 분류체계를 최종적으로 설정하였다. 그리고 상기 분류체계의 속성과 감염병 연구분야별 중점기술 내용을 과학적 관점에서 체계적으로 가장 잘 반영하는 데이터로 논문 데이터를 선정하여 학습모델에 사용하였다. 정부 R&D 연구과제는 국가적 차원의 당면 과제 및 사회 현안을 해결하거나 전략기술이나 부상기술을 지원하기 위한 국가적 차원의 정책을 반영하므로 시의성(timeliness) 및 정시성(punctuality) 속성이 두드러지는 데이터인 반면, 출판된 연구논문은 각 분야의 전문가 조직 내에서 논의되고 있는 연구주제들을 고르게 다루고 있으며, 연구주제의 우수성을 기반으로 과학적 품질이 검증된 객관적인 연구결과가 축적된 다양한 학문분야의 학제 및 과학기술 연구분야를 구성하고 있는 신뢰성(reliability) 및 명확성(clarity) 속성이 뚜렷한 데이터이다. 본 연구에서는 국가 감염병 연구분야의 체계적이고 과학적인 분류체계 기반의 학습 모델을 구축하고자, 논문데이터의 과학분류체계인 ASJC (All Science Journal Classification)를 활용하여 단일화된 분류체계 기준으로 과학적 품질이 검증된 객관적인 데이터를 수집하였으며, 이를 분류모델을 학습시키는데 활용하였다. 해당 모델을 개발하기 위해 순환 신경망 (Recurrent Neural Network, 이하 RNN)을 사용했다. RNN은 순차적인 정보 형태를 나타내는 자연어나 시계열 데이터를 처리하는데 강력한 성능을 보이는 모델이다 (Mikolov et al., 2010). 이러한 특징으로 RNN은 문서 분류와 관련된 연구들에서도 많이 활용되고 있다. 예를 들어 RNN을 이용해 연예 기사에서 악성 댓글을 분류하거나(Kim et al., 2019), 소셜 네트워크 서비스(Social Network Service) 공간에서 빠르게 확산되고 있는 소문을 감지하기도 한다(Ma et al., 2016). 이상의 선행 연구를 바탕으로 본 연구는 국가간 다른 내용으로 수행되고 있는 연구개발 과제를 비교하기 위해 필요한 공통된 분류체계를 개발하기 위해 RNN을 활용한다. 모델을 훈련시키는 구체적인 프로세스는 다음과 같다. 먼저 구분하고자 하는 분류체계의 핵심적인 내용이 잘 포함될 수 있는 검색식을 이용해 Scopus의 DB에서 관련 논문을 수집한다. 이렇게 검색식으로 수집된 Scopus의 논문 데이터에 이에 상응하는 분류체계 인덱스를 라벨로 설정한다. 최종적으로 각각의 분류체계 별로 논문의 초록과 함께 분류체계가 라벨링된 텍스트 데이터를 모델에 학습시킨다. 이상의 프로세스는 준지도 학습(Semi-Supervised Learning)으로 반응 값이 없는 데이터를 지도학습(Supervised Learning)에 사용하는 기법으로 이해할 수 있다. 실제로 준지도 학습은 분류 문제를 푸는데 주로 활용된다(Seok, 2016). 이후, 이렇게 학습한 모델에 미국, 유럽, 일본의 비정형화된 과제정보의 초록을 입력시킨 뒤 앞서 설정한 분류체계에 맞춰 데이터를 분류한다. 최종적으로 감염병 R&D 과제의 분류체계 별로 국가간 비교분석을 통해 정책 입안자에게 국가 감염병 R&D를 위한 의사결정에 유용한 시사점을 도출하였다. 이상의 모든 연구 프로세스



〈Figure 1〉 Research Model Design

를 도식화한 자료가 <Figure 1>과 같다.

2.4. 분류체계 설정 및 연구 데이터 수집

본 연구는 국가과학기술지식정보사이트 (National Science & Technology Information System, NTIS)에서 감염병과 관련된 R&D 과제로 검색된 과제데이터를 대상으로, 다음과 같은 프로세스를 통해 총 7개의 카테고리로 분류체계를 설정하였다. 감염병 관련 R&D 과제의 기술분류를 유형화하는 과정에서 감염병 전문가 집단(대학 및 민간기업의 바이오신약개발 전문가, 한국생명공학연구원 감염병연구센터, 바이오나노연구센터 소속의 전문가)을 구성하였다. 1차적으로는 질병관리본부가 국내 감염병분야 전문가 위원회(질병관리본부, 출연연, 대학, 민간 소속의 관련 전문가 등 총 47인 구성)를 통해 수립한 국가 감염병 위기대응 추진전략을 위한 범부처 R&D 시행계획을 참고하였으며(Ministry of Health and Welfare, 2019), 전문영역별로 특정되는 전문가의 지식을 반영하여 정성적 검토를 수행하기 위해 국가전략기술, 6T 기술분류, 과학기술표준분류체계, 연구과제명, 요약문 및 연구내용 등을 정성적 판단기준으로 세워 검토하였다. 최종적으로 전문가들의 의견을 종합하여, 범부처 R&D 시행계획의 감염병 연구분야별 중점기술 중에서 비R&D 성격의 정책과 인프라를 제외한 총 7개의 카테고리로 분류체계를 설정한 후, 각 카테고리의 연구과제명 중에서 대표성이 높은 키워드로 명명하였다. 각각의 분류체계는 다음과 같다: Diagnosis_biomarker, Drug_discovery, Epidemiology, Evaluation_validation, Mechanism_signaling pathway, Prediction, Vaccine_therapeutic antibody. 본 연구에서 구분하고자 하는 분류체계 별로 Scopus의 연구 데이터를 수집하기 위해 <Table 1>과 같은 검색식을 만들었다. 이상의 검색식에 Scopus에서 분류해 놓은 All Science Journal Classification (이하 ASJC) 코드를 설정해 바이오 분야의 논문들만 수집할 수 있었다. 바이오 분야의 ASJC는 Cancer

Research (1306), Cell Biology (1307), Molecular Biology (1312), Molecular Medicine (1313), Infectious Diseases (2725), Oncology (2730), Pharmacology_medical (2736), Pulmonary and Respiratory Medicine (2740), Pharmacology, Toxicology and Pharmaceutics_all (3000), Pharmacology, Toxicology and Pharmaceutics_miscellaneous (3001), Drug Discovery (3002), Pharmaceutical Science (3003), Pharmacology (3004), Toxicology (3005)이다.

이렇게 수집한 논문의 개수는 각각의 라벨의 순서대로 Diagnosis_biomarker에 4,355건, Drug_discovery에 3,398건, Epidemiology에 10,000건,

〈Table 1〉 Classification system definition and search terms

Classification	Count	Description	Search Terms
Diagnosis_biomarker	4,355	Diagnostic biomarker R&D refers to the research and development of biological parameters that aid in the diagnosis of diseases and help determine disease progression or treatment success.	TS=((diagnostic OR diagnosis) AND detection AND biomarker) ASJC=(3002 OR 3001 OR 3000 OR 3003 OR 3004 OR 3005 OR 1312 OR 1313 OR 1307 OR 1306 OR 2736 OR 2730 OR 2740 OR 2725)
Drug_discovery	3,398	Drug candidate R&D refers to research and development conducted with the aim of discovering new candidate drugs in the fields of medicine, biotechnology and pathology.	TS=(drug* AND candidate* AND discover* AND novel) ASJC=(3002 OR 3001 OR 3000 OR 3003 OR 3004 OR 3005 OR 1312 OR 1313 OR 1307 OR 1306 OR 2736 OR 2730 OR 2740 OR 2725)
Epidemiology	10,000	Epidemiological R&D refers to research and development that deals with the distribution, patterns and determinants of health and disease states in a defined group.	TI=(epidemiolog*) ASJC=(3002 OR 3001 OR 3000 OR 3003 OR 3004 OR 3005 OR 1312 OR 1313 OR 1307 OR 1306 OR 2736 OR 2730 OR 2740 OR 2725) PY=(2010-2020)
Evaluation_validation	3,744	Effectiveness evaluation R&D refers to research and development conducted for the purpose of verifying the efficacy of a new drug candidate or a developed drug.	TS=((evaluation OR evaluating) AND validation AND (safety OR efficacy OR marker)) ASJC=(3002 OR 3001 OR 3000 OR 3003 OR 3004 OR 3005 OR 1312 OR 1313 OR 1307 OR 1306 OR 2736 OR 2730 OR 2740 OR 2725)
Mechanism_signaling pathway	1,810	Signal transduction regulation mechanism R&D refers to research and development that deals with how a signal transducer binds to a receptor to generate a secondary signal transducer to regulate cell activity.	TI=(mechanism AND (signal* OR pathway OR transduction) ASJC=(3002 OR 3001 OR 3000 OR 3003 OR 3004 OR 3005 OR 1312 OR 1313 OR 1307 OR 1306 OR 2736 OR 2730 OR 2740 OR 2725)
Prediction	7,856	Predictive R&D refers to research and development that deals with predicting the success of clinical trials at each stage of new drug development or predicting toxicity that causes side effects.	TI=(prediction) ASJC=(3002 OR 3001 OR 3000 OR 3003 OR 3004 OR 3005 OR 1312 OR 1313 OR 1307 OR 1306 OR 2736 OR 2730 OR 2740 OR 2725) PY=(2014-2020)
Vaccine_therapeutic antibody	3,820	Vaccine and therapeutic antibody R&D refers to research and development for the purpose of developing vaccines and therapeutic agents.	TS=(vaccine* AND immuno* AND antibod* AND therapeutic) ASJC=(3002 OR 3001 OR 3000 OR 3003 OR 3004 OR 3005 OR 1312 OR 1313 OR 1307 OR 1306 OR 2736 OR 2730 OR 2740 OR 2725)

Evaluation_validation에 3,744건, Mechanism_signaling pathway에 1,810건, Prediction에 7,856건, Vaccine_ therapeutic antibody에 3,820건이다. 이상의 논문 데이터는 연구방법론에서 전술한 것과 같이 초록의 비정형화된 텍스트 데이터와 라벨로 이뤄진 튜플 데이터셋이다.

3. 연구결과

3.1. 전처리 및 워드 임베딩

수집한 데이터는 비정형화된 텍스트 데이터와 라벨로 구성되어 있기 때문에 분류 모델을 학습하기 위해서는 이상의 데이터를 워드 임베딩 (Word Embedding)해야 한다. 워드 임베딩이란 단어를 벡터로 표현하기 위한 방법으로 희소 표현된 벡터들을 밀집 표현된 벡터로 변환하는 것을 목표로 한다. 이를 위해서는 먼저 긴 문장의 텍스트를 단어 단위로 토큰화 (Tokenization)시켜야 한다. 이때 단어 토큰화를 수행하기 위해서 각 단어가 어떤 유형의 품사 정보인지를 나타내는 포스태깅(Part of speech) 분석을 통해 이를 파싱하는 태깅 작업 (Tagging Task)을 수행하게 된다. 본 연구에서는 이상의 과정을 통해 학습 데이터에서 추출된 단어 중 상위 빈도 15,000번까지만 사용했다. 그리고 서로 다른 초록의 길이를 고려해 추출된 단어의 차원을 512로 설정하였다. 예컨대 상대적으로 초록의 길이가 짧은 논문에서 추출된 단어의 개수는 상대적으로 초록의 길이가 긴 논문에서 추출된 단어의 개수 보다 작을 것이다. 본 연구에서는 이러한 문제를 개선하기 위해 추출된 단어의 차원을 모두 512로 고정시키고 최종적으로 34,983의 행과 512의 열을 가진 학습용 데이터를 정제할 수 있었다.

3.2. 모델학습 및 평가 환경설정

정제된 데이터로 모델을 학습하기 위해서는 같은 각각의 라벨 별로 불균형한 데이터의 분포를 맞춰주는 작업이 필요하다. 불균형한 라벨의 분포를 갖고 있는 데이터를 학습한 모델은 왜곡된 분류 결과를 낳기 때문에 모델의 심각한 성능 저하를 야기한다. 따라서 본 연구에서는 이상의 문제를 개선하기 위해 데이터가 적은 표본을 더 많이 추출하도록 하는 업 샘플링 (Up sampling)을 수행했다. 이때 주의해야할 점은 샘플링된 데이터는 학습에만 사용되야 한다는 점이다. 만약 모델의 성능을 확인하는 검증용 데이터나 시험용 데이터에 샘플링된 데이터가 사용되면 특정 라벨을 더 많이 추출하도록 하는 모델이 생성되기 때문이다.

이에 본 연구는 이상의 전처리 과정을 거친 최종적인 데이터를 훈련용 70%, 검증용 20%, 시험용 10%로 구분한 뒤 홀드아웃 검증을 통해 학습된 모델의 성능을 평가했다. 전술한 것과 같이 훈련에 사용된 데이터는 각각의 라벨 별로 10,000건씩 업샘플링을 수행하였다.

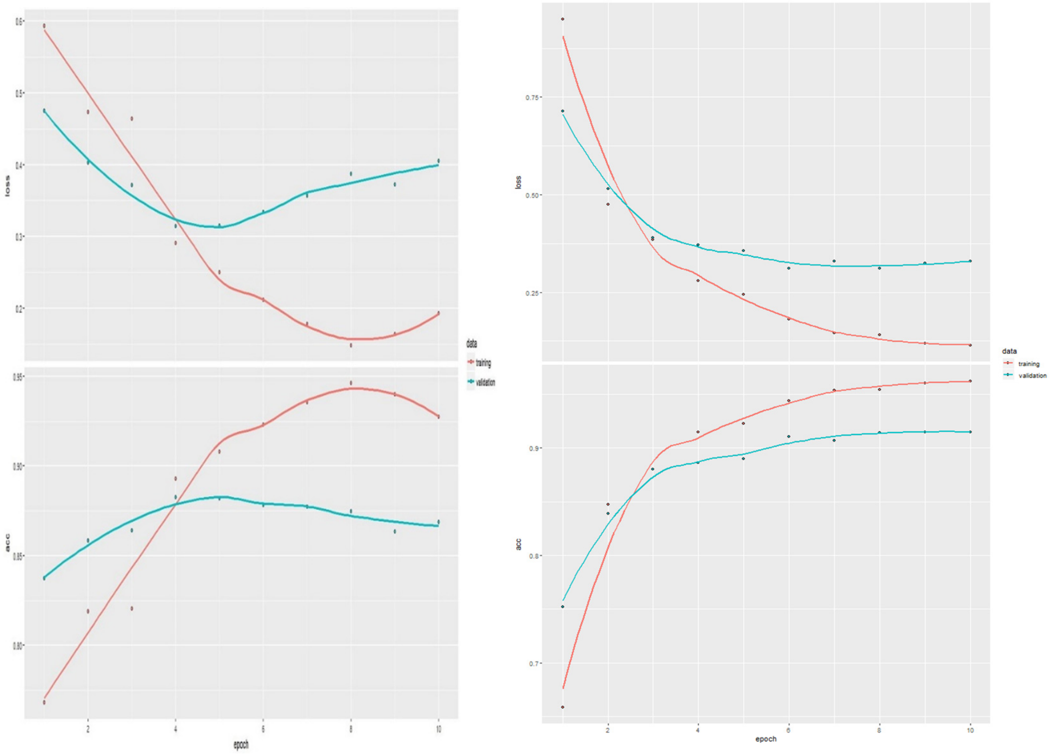
3.3. 모델학습 및 평가 환경설정

본 연구에서 사용된 모델은 양방향 RNN (Bidirectional RNN)이다. 앞서 2.3 연구방법론에서 RNN이 순차적인 정보 형태를 나타내는 데이터를 처리하는데 강력한 분석 기법이라고 설명

하였다. 만약 처리하고자 하는 데이터가 텍스트라고 할 때 일반적인 RNN은 순차적인 형태로 텍스트의 왼쪽에서 오른쪽으로 배열된 단어의 순서를 고려하며 모델이 학습된다. 한편 양방향 RNN은 역방향으로도 텍스트에 배열된 단어의 순서를 고려하는 모델로 이해할 수 있다. 즉 다시 말해, 양방향 RNN은 순방향의 RNN 모델과 역방향의 RNN 모델을 각각 만든 뒤 이 둘의 학습결과를 합치는 방법으로 신경망 모델을 구축한다. 따라서 양방향 RNN은 기존의 RNN 보다 텍스트의 의미와 맥락을 더 잘 보존해 상대적으로 높은 성능을 나타내는 것으로 알려져 있다. 최종적으로 본 연구에서 사용된 모델의 구조는 512차원의 입력 데이터를 양방향 RNN 층으로 훈련을 시키고 7종의 라벨로 분류되게끔 소프트함수(softmax)를 활성화 함수로 사용했다. 이상의 모델에서 파라미터를 최적화는 아담(adam)을 사용했다. 이때 파라미터의 업데이트는 categorical crossentropy 방법에 따른 손실 값을 기반으로 수행하였다.

3.4. 모델 학습 결과

최종적인 모델의 분류 성능은 Figure 2에 나타낸 바와 같다. 홀드아웃 검증을 통해 평가된 모델의 성능은 훈련 데이터를 통해 파라미터를 학습하는 훈련 단계에서 96.20%로 나타났다. 그리



〈Figure 2〉 Visualization of model training results (Left: LSTM, Right: Bi-LSTM)

고 검증용 데이터를 통해 확인한 모델의 성능은 91.49%로 나타났다. 마지막으로 훈련이나 검증 단계에서는 전혀 사용하지 않은 시험용 데이터를 통해 살펴본 최종적인 모델의 분류 정확도는 90.80%로 도출됐다. 이상의 과정에서 모델 훈련 성능과 모델 검증 성능이 각각 96.20%와 91.49%로 어느 정도 과적합이 존재하는 것을 알 수 있다. 실제로 모델을 적합시키기 위해 훈련 모델의 성능과 검증 모델의 성능을 비교 평가한 내용이 <Figure 2>의 오른쪽과 같은데, 해당 자료를 보면 최적의 파라미터 세팅이 4번째 에포크(epoch)에서 이뤄진 것을 알 수 있다. 하지만 모델 검증 성능과 모델 시험 성능을 비교해본 결과 10번째 에포크에서 각각 91.49%와 90.80%로 그 차이가 가장 적어서 본 모델을 최종적으로 채택할 수 있었다. 왜냐하면 이상의 결과는 본 모델이 새로운 데이터임에도 불구하고 90% 이상의 정확도를 안정적으로 기록하여 상당히 강건한 것을 알 수 있기 때문이다. 즉 다시 말해, 비록 본 모델을 학습하기 위해 관련 분류체계 별로 상응하는 논문 데이터를 수집한 뒤 모델을 학습했지만, 궁극적인 본 연구의 목적은 우리나라를 포함하는 여러 국가들의 연구개발 과제 데이터에 원하는 분류체계를 성공적으로 할당하는데 있다. 따라서 전혀 새로운 연구개발 과제 데이터라고 해도 강건하게 분류체계를 구분해내는 모델의 구축이 본 연구의 목적에 더 부합한다고 판단했다. 이상의 모델은 <Figure 2>의 왼쪽에 위치한 일반적인 LSTM과 비교해도 훨씬 안정적으로 모델링이 되는 것을 알 수 있다.

3.5. 최종 분류 결과

코로나 바이러스 관련 국내외 연구과제를 분석하기 위하여 미국, 일본, 유럽 등 각 국가별 기초연구와 관련된 과제 투자 정보 데이터베이스(data base, DB)로부터 국가 R&D 과제 정보를 수집하였다. 각 국가별 데이터의 수집원은 <Table 2>에 나타내었다.

해외 주요 국가(미국, 유럽, 일본)의 R&D를 비교분석하기 위해 한국과학기술정보연구원(KISTI)에서 해당 시스템에서 제공하는 국가 R&D 과제 정보를 표준화하여 단일 데이터베이스(database)로 구축하였다(Lee et al., 2020). 국가 R&D 과제 정보는 2012년~2018년에 수행된 모든 과제를 대상으로 하였고, 수집된 총 과제정보는 1,172,174건이다. 이중 주요 국가의 코로나 바이러스관련 R&D 동향을 분석하기 위해 해당 기술 내용을 포함하는 검색식을 작성하였고, 검색건수는 미

〈Table 2〉 The sources of national scientific research funding data of the major countries

Countries	Sources
US	• NSF(NationalScienceFoundation)-https://www.nsf.gov/ • NIH–STARMetrics–https://www.starmetrics.nih.gov/
EU	• CORDIS(CommunityResearch&DevelopmentInformationService) - http://cordis.europa.eu/
JP	• KAKEN (Database of Grants-in-Aid for Scientific Research, KAKEN) - https://kaken.nii.ac.jp/ja/

〈Table 3〉 Search terms of coronavirus-related research and the number of selected patents in the three countries.

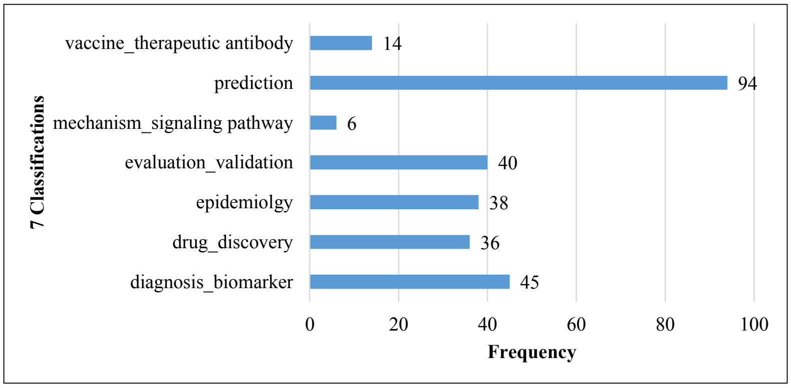
Search terms	Search count			
	US	EU	JP	Total
(corona* and virus*) or coronavirus* or mers-cov or sars-cov or covid or "severe acute respiratory syndrome" or "middle east respiratory syndrome"	515	10	35	560

국 515건, 유럽 10건, 일본 32건으로 총 557건의 R&D 과제 데이터를 추출하였다. 이상의 내용이 <Table 3>과 같다.

이렇게 구축한 모델에 최종적으로 분류하고자 하는 데이터는 유럽 (CORDIS), 미국 (NSF, FEDRIP), 일본 (KAKEN)으로 구성된 총 4종의 연구개발데이터 셋을 대상으로 한다. 이 중에서 특히 국가 전염병 R&D와 관련이 높다고 여겨지는 273개의 연구개발과제를 정성적인 기준을 통해서 추출했는데, 유럽의 연구개발 데이터 셋 (CORDIS)에서 10건 미국에서 228건 (NSF 2건, FEDRIP 226건). 일본에서 35건 (KAKEN)으로 구분됐다.

이상의 데이터에 대해 본 연구가 구축한 모델을 통해 본 연구에서 설정하고 있는 분류체계로 분류한 최종적인 결과가 다음과 같다. 본 연구에서 설정하고 있는 분류체계는 각각 diagnosis_biomarker, drug_discovery, epidemiology, evaluation_validation, mechanism_signaling pathway, prediction, vaccine_therapeutic antibody였는데, 글로벌 국가 감염병 R&D는 각각 5건, 36건, 38건, 40건, 6건, 94건, 14건으로 분류됐다. 이상의 결과를 통해 국가 감염병과 관련된 글로벌 연구개발 추세가 대부분 신약개발 단계별 임상에 대한 성공 예측이나, 부작용을 유발하는 독성을 예측 등을 다루는 prediction 연구분야에 집중되어 있는 것을 알 수 있었다. 그리고 새로운 후보 약물의 발견을 목적으로 수행되는 drug discovery 분야와 건강 및 질병 상태의 분포, 패턴 및 결정 요인 등을 다루는 epidemiology 분야, 그리고 신약 후보 물질 또는 개발된 약물의 효능을 검증하는 것을 목적으로 수행하는 연구개발 분야인 evaluation & validation 분야가 비슷하게 진행되고 있음을 알 수 있었다. 한편 질병의 진단을 돕고 질병 진행이나 치료 성공을 결정하는데 도움이 되는 생물학적 파라미터를 다루는 diagnosis_biomarker 연구분야와 신호전달물질이 수용체에 결합해 2차 신호전달 물질을 발생시켜 세포의 활동을 조절하는 방법 등을 다루는 mechanism_signaling pathway 분야는 상대적으로 연구가 부족한 편인 것을 알 수 있었다. 주목할만한 부분은 다른 나라에서도 백신과 치료제의 개발을 목적으로 하는 연구개발 영역인 vaccine_therapeutic antibody 연구분야에 투자가 상당히 적었다는 점인데, 이상의 결과는 세계적으로 유행하는 바이러스의 예상치 못한 창궐이 확산되는 속도에 비해 백신과 치료제 개발이 제대로 이뤄지지 않는 원인의 배경과 어느정도 부합한다고 보인다. 이상의 내용이 <Figure 3>과 같다.

이상의 결과를 국가별로 세분하여 자세하게



〈Figure 3〉 Classification results of infectious diseases in major countries by classification system

〈Table 4〉 Classification result of R&D information by classification system

Classification system	US	EU	JP	Total cost per classification
Diagnosis_biomarker	$ 921,992.14	$ 5,125,340.50	$ 67,60.00	$ 1,219,677.13
	18%	29%	16%	23%
Drug_discovery	$ 362,247.30	$ 3,505,285.00	$ 46,800.00	$ 432,029.06
	7%	20%	11%	8%
Epidemiology	$ 436,128.93	$ 244,569.00	$ 80,807.88	$ 356,283.45
	9%	1%	20%	7%
Evaluation_validation	$ 718,692.03	-	$ 61,700.00	$ 636,568.03
	14%	0%	15%	12%
Mechanism_signaling pathway	$ 668,993.50	$ 1,498,514.00	$ 21,700.00	$ 699,364.67
	13%	9%	5%	13%
Prediction	$ 1,460,341.99	$ 7,075,419.00	$ 55,250.00	$ 1,462,136.96
	29%	41%	13%	28%
Vaccine_therapeutic antibody	$ 474,323.00	-	$ 80,166.67	$ 389,860.93
	9%	0%	19%	8%
Total cost per country	$ 5,042,718.89	$ 17,449,127.50	$ 414,024.54	$ 5,195,920.21
	100%	100%	100%	100%

살펴본 결과가 <Table 4>와 같다. 하기 결과는 과제 투자비용을 과제 건수로 나눈 과제당 투자비용을 의미한다. 그리고 과제당 투자비용과 함께 제시되어 있는 비율은 국가를 기준으로 했을 때 전체 과제당 연구투자 비용 대비 비중을 의미한다. 미국의 경우 prediction (29%), diagnosis_biomarker (18%), evaluation_valdiation (14%), mechanism_signaling pathway (13%), epidemiology (9%), vaccine_therapeutic antibody (9%), drug_discovery (7%) 순으로 연구개발 투자를 많이 하는 것으로 나타났다. 유럽의 경우는 prediction (41%), diagnosis_biomarker (29%), drug_discovery (20%), 〈Table 5〉 Major research organizations by country and classification system

Classification	Major Institutions or Organizations		
	US	EU	JP
Diagnosis_biomarker	- NIVERSITY OF MINNESOTA TWIN CITIES - UNIVERSITY OF NORTH CAROLINA CHAPEL HILL - UNIVERSITY OF MARYLAND BALTIMORE, - UNIVERSITY OF IOWA - NEW YORK BLOOD CENTER - MEDICAL COLLEGE OF WISCONSIN - ARIZONA STATE UNIVERSITY-TEMPE CAMPUS	- EIDGENOESSISCHES DEPARTEMENT DES INNERN - ERASMUS UNIVERSITAIR MEDISCH CENTRUM ROTTERDAM - UNIVERSIDAD POLITECNICA DE MADRID - UNIVERSITAT DE VALENCIA	- Juntendo University - National Agriculture and Food Research Organization - National Institute of Infectious Diseases - University of Yamanashi
Drug_discovery	- UNIVERSITY OF PENNSYLVANIA - UNIVERSITY OF NORTH CAROLINA CHAPEL HILL	- KATHOLIEKE UNIVERSITEIT LEUVEN	- Kyoto Pharmaceutical University - Nippon Veterinary and Life Science University
Epidemiology	- THE UNIVERSITY OF PITTSBURGH - CORNELL UNIVERSITY ITHACA - NEW YORK BLOOD CENTER - DYNPORT VACCINE COMPANY, LLC - VANDERBILT UNIVERSITY - UNIVERSITY OF UTAH - DREXEL UNIVERSITY - NORTHWESTERN UNIVERSITY AT CHICAGO - CLEVELAND CLINIC LERNER COL/MED-CWRU	- UNIVERSITY OF BRISTOL	- National Institute of Infectious Diseases
Evaluation_validation	- UNIVERSITY OF TEXAS MEDICAL BR GALVESTON - UNIVERSITY OF NORTH CAROLINA CHAPEL HILL - UNIVERSITY OF IOWA		- Tokyo University of Pharmacy and Life Science - Yamaguchi University - Nippon Veterinary and Life Science University - Kitasato University - Jikei University School of Medicine
Mechanism_signaling pathway	- UNIVERSITY OF IOWA - UNIVERSITY OF VIRGINIA CHARLOTTESVILLE - UNIV OF MASSACHUSETTS MED SCH WORCESTER - UNIVERSITY OF CALIFORNIA SAN DIEGO	- Victoriano Martinez Riomansa SL	- Yamaguchi University
Prediction	- UNIVERSITY OF NORTH CAROLINA CHAPEL HILL - UNIVERSITY OF PENNSYLVANIA - UNIVERSITY OF MARYLAND BALTIMORE - LOYOLA UNIVERSITY CHICAGO - UNIVERSITY OF WISCONSIN MADISON	- ERASMUS UNIVERSITAIR MEDISCH CENTRUM ROTTERDAM	- The University of Tokyo
Vaccine_therapeutic antibody	- VANDERBILT UNIVERSITY		- Osaka University - Tokyo University of Pharmacy and Life Science - Kitasato University

순서로 많은 투자를 했으며, mechanism_signaling pathway (9%), epidemiology (1%) 분야를 상대적으로 적게 투자하고 evaluation_valdiation (0%) 분야와 vaccine_therapeutic antibody (0%) 분야는 투자가 없는 것으로 나타났다. 마지막으로 일본의 경우는 epidemiology (20%), vaccine_therapeutic antibody (19%), diagnosis_biomarker (16%), evaluation_valdiation (15%), prediction (13%), drug_discovery (11%), mechanism_signaling pathway (5%) 순서로 연구개발 투자를 많이 하는 것으로 나타났다. 이상의 결과를 살펴보면 미국과 일본은 상대적으로 모든 영역에 골고루 연구개발 투자를 하고 있는 것으로 나타났다. 한편 유럽은 diagnosis_biomarker와 같이 다른 나라에 비해 상대적으로 특정 연구분야에 많은 투자를 하는 집중화 전략을 취하는 것으로 나타났다. 본 연구는 이상의 결과에 기초해 국가별 주요 연구기관에 대한 내용도 함께 제시하고 있는데, 이에 대한 내용이 <Table 5>와 같다.

4. 결론 및 시사점

본 연구에서는 감염병 질환을 대응하기 위한 국가 방역과 관련된 R&D 전략분야에 대한 투자 전략 및 정책수립의 기초적인 토대로 삼을 수 있는 맞춤형 분류체계에 대한 필요성에 기반하여 기계학습 기반의 탐색적 연구를 수행하고자 하였다. 이를 위해 국가 방역 R&D와 관련된 맞춤형 분류모델을 고안하였다. 먼저 학습모델을 구축하기 위해서 Scopus 데이터를 수집하여 데이터셋을 설계하였으며, 워드 임베딩 단계를 거쳐 학습용 데이터를 정제하고, 양방향 RNN 모델을 사용하여 학습시킨 후, 최종적으로 구축된 모델의 성능을 검증하는 과정을 수행하였다. 모델 성능의 검증 결과, 본 분류모델의 단계별 정확도는 훈련단계에서 96.2%, 검증단계에서 91.49%, 시험단계에서 90.8%로 매우 유의미한 분류 성능을 나타냈다. 이를 통해 보다 정량적이고 객관적으로 R&D에 대한 맞춤형 분류 작업이 가능하다는 점을 확인할 수 있었다.

본 연구 결과를 통해 세 가지 정책적 의미를 도출할 수 있다. 첫째, 데이터기반 과학기술정책 관점에서 수요자 관심분야에 대한 국가 R&D사업의 정보를 글로벌 기준으로 분류하는 방안을 제시하였다. 국가적 차원에서 최대 현안인 감염병을 대상으로 국가 R&D 사업을 분류하는 본 연구에서 제안하고 실증하였지만, 고령화, 기후변화 등 다양한 국가현안에 대한 글로벌 기준으로 R&D 사업의 정보를 일관된 분류체계를 기반으로 도출하여 이해관계자들간 정책 수립 및 공동연구 전략의 협의와 토론의 기초적 자료로 제공할 수 있는 가능성이 있다. 예를 들어, Lee et al (2020)이 감염병의 기초연구영역인 바이러스학(Virology)의 국가 R&D과제 데이터 기반의 클러스터링 분석을 통해 도출 분야를 대상으로 공통 분류체계를 수립하고 국가별 R&D현황을 분석한다면, 이해당사자들이 국제 공동 R&D 전략과 관련한 의사 결정 시 유용한 다양한 정보를 제공할 수 있을 것이다. 둘째, 감염병관련 국가 R&D사업 영역에 대한 정보분석 서비스 기획의 기반을 마련하였다. 본 연구에서는 감염병관련 분류체계를 기반으로 국가간 비교 및 조직정보를 제공하는 실증을 통해 분석서비스의 구체적

인 활용안을 제시하였다. 기존에 국내 R&D 과제 중심의 정보제공에서 해외 주요국가의 R&D과제와의 동일한 기준으로 분석서비스 제공하는 방안을 제시하였다. 즉, 이질적인 국가별 R&D과제를 대상으로 단일 체계로 분석하여 정책을 수립할 때 기초 자료를 제공하는 틀을 제공하였다. 마지막으로 국가 감염병 R&D 분류체계 수립을 통해 분류 체계의 궁극적 목표인 산업, 기업, 정책 정보를 제공할 수 있는 기반을 마련한 것이다. 특히 생명과학 및 보건의료분야의 R&D 사업은 학제 간, 산업 간, 기술 분야간 융합 및 연계 기반의 속성이 매우 두드러지는 분야로서, 전략적인 예산 배분을 수립하고 추진하기 위해서는 R&D 목적과 각 사업 내용에 맞는 적절한 분류체계가 필요하다. 이를 위해서는 기존의 과학기술표준분류체계, 한국표준산업분류체계, 바이오산업분류체계, 생명공학기술분류체계 등을 대상으로 각 분류체계의 속성과 특징을 분석하고, 데이터 기반의 근거를 제공할 수 있는 다양한 분류체계 연계 방법론에 대한 연구가 지속적으로 필요한데, 본 연구는 이러한 이종 분류체계의 연계구조를 분석하기 위한 초석 역할을 할 수 있다.

본 연구의 한계점으로 우리나라 감염병 R&D 과제를 대상으로 제안한 분류체계로 동시에 분류하여 국가간 비교를 통해 우리나라 감염병관련 연구 역량을 비교 분석을 한다면, 우리나라 중심의 국가 감염병 R&D 전략수립에 필요한 보다 세부적인 정보를 제공할 것이다. 감염병 관련 국가 R&D 과제의 영문화를 통해 향후 분석 서비스의 내용을 더욱 우리나라 중심으로 제공한다면 수요자 중심의 정보제공으로 서비스의 품질이 높아질 것으로 판단된다.

참고문헌(References)

Chen, T.-K., H.-H. Liao, and H.-J. Kuo, “Internal liquidity risk, financial bullwhip effects, and corporate bond yield spreads: Supply chain perspectives”, *Journal of Banking & Finance*, Vol.37, No.7(2013), 2434-2456.

Foltz, P.-W., and S.-T. Dumais, “Personalized information delivery: An analysis of information filtering methods”, *Communications of the ACM*, Vol.35, No.12(1992), 51-60.

Hankookilbo, Live Issue, 2020. Available at https://www.hankookilbo.com/News/Read/A2020081406480005122 (Access 15 August, 2020).

Hong, S-K, “Research on classification criteria for national R&D projects for systematic information provision”, *Research Report*, Korea Institute of Science and Technology Evaluation and Planning, 2016.

Houseman, E.-M., and D.-E. Kaskela, “State of the art of selective dissemination of information”, *IEEE Trans. Eng. Writing Speech III,* (1970), 78-83.

Jeong, H., Y.-K. Kim, and J. Kim, “An evaluation-committee recommendation system for national R&D projects using social network analysis”, *Cluster Computing*, Vol.19, No.2(2016), 921-930.

Joseilbo, Economic news, 2020. Available at http://www.joseilbo.com/news/htmls/2020/07/20200723402630.html (Access 15 August, 2020).

Kwak S.-J., H.-Y. Lee, “A Study on Classification and Metadata for R&D Reports in the Field of Energy”, *Annals of Social Science*, Vol.24, No.2(2013), 361-378.

Kim J., C. Yeo, “New infectious disease crisis response technology (diagnosis, treatment, vaccine)", *Research Report*, Korea Advanced Institute of Science and Technology Evaluation, 2020.

Kim J.-W., H.-I. Jo, B.-G. Lee, "A Comparison Study on Performance of Malicious Comment Classification Models Applied with Artificial Neural Network", *Journal of Digital Contents Society*, Vol. 20, No. 7(2019), 1429-1437.

Kim S., C. Kim, H. Jang, S. Ye, M. Song, "A Classification for Research Projects in Oriental Medicine Field", *Journal of the Korean society for information management*, Vol.25,No.4(2008), 309-326.

Korea Centers for Disease Control and Prevention, "Implementing national emergency response strategy: A study on R&D execution plan and new project planning". 2020.

Korea Health Industry Development Institute, "New Convergence Industry Discovery Research (Convergence Technology): Classification System Study of HT Convergence Technology". 2015.

Lee, D, J. Kang, K. Kim, "Global collaboration research strategies for sustainability in the post COVID-19 era: analyzing virology-related national-funded projects", *Sustainability*, Vol.12(2020), 6561.

Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K. F., and M. Cha, "Detecting rumors from microblogs with recurrent neural networks", *Proceedings of the 25th International Joint Conference on Artificial Intelligence,* (2016), 3818-3824.

Mikolov, T., M. Karafiat, L. Burget, J. Cernocky, and S. Khudanpur, "Recurrent neural network based language model", *INTERSPEECH 2010 11th Annual Conference of the International Speech Communication Association*, (2010), 1045-1048.

Ministry of Economy and Finance, "Held the 3rd Emergency Economy Central Countermeasure Headquarters Meeting (Infectious Disease Response Industry Promotion Plan)", 2020. Available at http://www.moef.go.kr/nw/nes/detailNesDtaView.do?searchBbsId=MOSFBBS_000000000028&searchNttId=MOSF_000000000036711&menuNo=4010100 (Access 15 August, 2020).

Ministry of Health and Welfare, "A Study on the Research Plan for the Drive Plan of Multi-Agency R&D Program to Implement the National Responding strategy of Infectious Diseases", 2019.

Ministry of Science and ICT, "2021 Government R&D investment direction and standard revision", 2020.

Moon, S.-Y, "Classification scheme of biotechnology R&D for strategic budget allocation", *Research Report*, Korea Institute of Science and Technology Evaluation and Planning, 2015.

Sebastiani, F., "Machine learning in automated text categorization", *ACM Computing Surveys*, Vol.34, No.1(2002), 1–47.

Seo, T. and B. Kim, "Understanding classification, thesaurus, and ontology for information services", *Research Report*, Korea Institute of Science and Technology Information, 2015.

Seok, K., "Smoothing parameter selection in semi-supervised learning", *Journal of the Korean Data & Information Science Society*, Vol.27, No.4(2016), 993-1000.

Sohn, S.-H., and B.-K. Yoo, "New drug classification system in accordance with global harmonization", *Korean Journal of Clinical Pharmacy*, Vol.22, No.3(2012), 260-267.

Tang, J., S. Wu, J. Sun, and H. Su, "Cross-domain collaboration recommendation", *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining,* (2012), 1285-1293.

Tian, Q., J. Ma, J. Liang, R.-C. Kwok, and O. Liu, "An organizational decision support system for effective R&D project selection", *Decision support systems*, Vol.39, No.3(2005), 403-413.

Trappey, A.-J., C.-V. Trappey, C.-Y. Wu, C.-Y. Fan, and Y.-L. Lin, "Intelligent patent recommendation system for innovative design collaboration", *Journal of Network and Computer Applications*, Vol.36, No.6(2013), 1441-1450.

Xu, W., J. Sun, J. Ma, and W. Du, "A personalized information recommendation system for R&D project opportunity finding in big data contexts", *Journal of Network and Computer Applications*, Vol.59 (2016), 362-369.

Yang, Y, "An evaluation of statistical approaches to text categorization", *Inform. Retr.* Vol.1 (1999), 69–90.

Abstract

The Classification System and Information Service for Establishing a National Collaborative R&D Strategy in Infectious Diseases: Focusing on the Classification Model for Overseas Coronavirus R&D Projects

Doyeon Lee* · Jae-Seong Lee** · Seung-pyo Jun*** · Keun-Hwan Kim***

The world is suffering from numerous human and economic losses due to the novel coronavirus infection (COVID-19). The Korean government established a strategy to overcome the national infectious disease crisis through research and development.

It is difficult to find distinctive features and changes in a specific R&D field when using the existing technical classification or science and technology standard classification. Recently, a few studies have been conducted to establish a classification system to provide information about the investment research areas of infectious diseases in Korea through a comparative analysis of Korea government-funded research projects. However, these studies did not provide the necessary information for establishing cooperative research strategies among countries in the infectious diseases, which is required as an execution plan to achieve the goals of national health security and fostering new growth industries. Therefore, it is inevitable to study information services based on the classification system and classification model for establishing a national collaborative R&D strategy.

Seven classification - Diagnosis_biomarker, Drug_discovery, Epidemiology, Evaluation_validation, Mechanism_signaling pathway, Prediction, and Vaccine_therapeutic antibody - systems were derived through reviewing infectious diseases-related national-funded research projects of South Korea. A classification system model was trained by combining Scopus data with a bidirectional RNN model. The classification performance of the final model secured robustness with an accuracy of over 90%. In order

---

* Division of Data Analysis, Korea Institute of Science and Technology Information
** Department of Science and Technology Management Policy, University of Science and Technology
*** Data Analysis Platform Center, Korea Institute of Science and Technology Information
**** Corresponding author: Keun-Hwan Kim
Division of Data Analysis, Korea Institute of Science and Technology Information
66 Hoegi-ro, Dongdaemun-gu, Seoul 130-741, Korea
Tel: +82-2-3299-6072, Fax: +82-2-3299-6041, E-mail: khkim75@kisti.re.kr

to conduct the empirical study, an infectious disease classification system was applied to the coronavirus-related research and development projects of major countries such as the STAR Metrics (National Institutes of Health) and NSF (National Science Foundation) of the United States(US), the CORDIS (Community Research & Development Information Service)of the European Union(EU), and the KAKEN (Database of Grants-in-Aid for Scientific Research) of Japan.

It can be seen that the research and development trends of infectious diseases (coronavirus) in major countries are mostly concentrated in the prediction that deals with predicting success for clinical trials at the new drug development stage or predicting toxicity that causes side effects. The intriguing result is that for all of these nations, the portion of national investment in the vaccine_therapeutic antibody, which is recognized as an area of research and development aimed at the development of vaccines and treatments, was also very small (5.1%). It indirectly explained the reason of the poor development of vaccines and treatments. Based on the result of examining the investment status of coronavirus-related research projects through comparative analysis by country, it was found that the US and Japan are relatively evenly investing in all infectious diseases-related research areas, while Europe has relatively large investments in specific research areas such as diagnosis_biomarker. Moreover, the information on major coronavirus-related research organizations in major countries was provided by the classification system, thereby allowing establishing an international collaborative R&D projects.

Key Words : National-Funded Project, Infectious diseases, Classification, Bidirectional RNN, Coronavirus, Collaboration

Received : August 18, 2020 Revised : September 11, 2020 Accepted : September 18, 2020
Corresponding Author : Keun-Hwan Kim

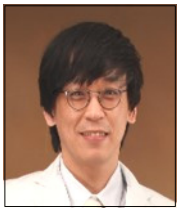
저자소개

이도연

중앙대학교에서 생명과학 학사, 의학부 석사 및 박사학위를 취득하고 현재 한국과학기술정보연구원 데이터분석본부 연구원으로 재직 중이다. Nature Nanotechnology, PLOS ONE, European Journal of Cell Biology, Healthcare, 한국융합학회 등 국내외학술지에 다수의 논문을 게재하였다. 관심 연구분야는 생명과학, 신경과학 등 바이오 헬스케어 분야의 데이터 분석, 데이터기반 융합R&D 영역 탐색, 데이터 기반 의사결정지원, 중소기업 R&BD 혁신 역량 제고 등이다.

이재성

과학기술연합대학원대학교 과학기술경영정책학과에서 석사 및 박사과정을 수료하고 현재 한국과학기술정보연구원 데이터분석본부 학생연구원으로 재직 중이다. 주요 관심 연구분야로는 기계학습 또는 인공지능을 활용한 맞춤형 정부지원 또는 국가 연구개발과제 관리 등으로, 국가 혁신성과 제고를 위한 데이터 기반의 다양한 정책 연구를 수행했다.



전승표

KAIST에서 경영학으로 석사학위를 취득하고, 고려대학교에서 과학관리학 전공으로 이학박사를 취득했다. 현재 한국과학기술정보연구원 데이터분석플랫폼센터에 책임연구원으로 재직 중이며, 과학기술연합대학원대학교 과학기술정책학과 부교수로 재직중이다. Technological forecasting and social change, Scientometrics, Energy policy, Internet research 등 해외학술지와 한국기술혁신학회지, 지능정보연구 등 국내학술지에 주저자로 다수의 논문을 게재했다. 주요 관심분야는 빅데이터를 활용한 수요 예측, 유망 기술 탐색, 기술가치평가, 산업시장분석 등을 위한 지능형 정보 시스템 연구이다.

김근환

University of Wisconsin (Milwaukee)에서 MBA를 취득하고, UST에서 응용정보과학으로 공학박사를 취득했다. 현재 한국과학기술정보연구원 데이터분석플랫폼센터에 선임연구원으로 재직 중이다. PLOS ONE, Healthcare, 한국기술혁신학회 등 국내외학술지에 다수의 논문을 게재하였다. 관심분야는 데이터기반 융합R&D 영역 탐색, 데이터기반 기술역량 지표 개발, 데이터 기반 의사결정지원 등이 있다.