기계학습을 이용한 기록 텍스트 자동분류 사례 연구*

A Study on Automatic Classification of Record Text Using Machine Learning

김해찬솔 (Hae Chan Sol Kim)**, 안대진 (Dae Jin An)***
임진희 (Jin Hee Yim)****, 이해영 (Hae-Young Rieh)*****

초 록

기록이나 문헌의 자동분류에 관한 연구는 오래 전부터 시작되었다. 최근에는 인공지능 기술이 발전하면서 기계학습이나 딥러닝을 접목한 연구로 발전되고 있다. 이 연구에서는 우선 문헌의 자동분류와 인공지능의 학습방식이 발전해 온 과정을 살펴보았다. 또 기계학습 중 특히 지도학습 방식의 특징과 다양한 사례를 통해 기록관리 분야에 인공지능 기술을 적용해야 할 필요성에 대해 알아보았다. 그리고 실제로 지도학습 방식으로 서울시의 결재문서를 ETRI의 엑소브레인을 통해 정부기능분류체계로 자동분류해 보았다. 이를 통해 기록을 다양한 방식의 분류체계로 자동분류하기 위한 각 과정의 고려사항을 도출하였다.

ABSTRACT

Research on automatic classification of records and documents has been conducted for a long time. Recently, artificial intelligence technology has been developed to combine machine learning and deep learning. In this study, we first looked at the process of automatic classification of documents and learning method of artificial intelligence. We also discussed the necessity of applying artificial intelligence technology to records management using various cases of machine learning, especially supervised methods. And we conducted a test to automatically classify the public records of the Seoul metropolitan government into BRM using ETRI's Exobrain, based on supervised machine learning method. Through this, we have drawn up issues to be considered in each step in records management agencies to automatically classify the records into various classification schemes.

키워드: 자동분류, 인공지능, 지도학습, 분류체계, 한국전자통신연구원 엑소브레인, 기계학습
automatic classification, artificial intelligence, supervised learning, classification scheme, ETRI Exobrain, machine learning

---

* 본 연구는 2017년 국가기록원 R&D사업 '차세대 기록관리 모델 재설계 연구'의 일환으로 수행된 연구임. 본 연구는 미래창조과학부 및 정보통신기술진흥센터의 정보통신·방송 연구개발 사업의 일환으로 하였음. [2013-0-00131, (엑소브레인-1세부) 휴먼 지식증강 서비스를 위한 지능진화형 WiseQA 플랫폼 기술 개발]
** 아카이브랩 연구원(abook123@naver.com) (제1저자)
*** 아카이브랩 대표, 명지대학교 기록정보과학전문대학원 박사과정(daejin@archivelab.co.kr) (공동저자)
**** 정보인권연구소 연구위원(yimjhkr@empas.com) (교신저자)
***** 명지대학교 기록정보과학전문대학원 교수(hyrieh@gmail.com) (공동저자)
■ 논문접수일자 : 2017년 12월 8일 ■ 최초심사일자 : 2017년 12월 13일 ■ 게재확정일자 : 2017년 12월 27일
■ 정보관리학회지, 34(4), 321-344, 2017. [http://dx.doi.org/10.3743/KOSIM.2017.34.4.321]

1. 서 론

현재 4차 산업혁명은 전 세계가 주목하는 화두이다. 국내에서도 이세돌과 알파고의 대국을 기점으로 인공지능 기술 등 4차 산업혁명에서 언급되는 신기술에 관한 관심이 커지고 있다. 2016년 1월 클라우스 슈바프(Klasu Schwab)는 세계경제포럼(World Economic Forum, WEF)에서 물리적 공간, 디지털 공간 및 생물학적 공간의 경계가 희석되는 기술융합의 시대를 주창하며 이를 4차 산업혁명이라 정의하였다. 현재 사물인터넷(IoT), 빅데이터, 인공지능, 블록체인 등 다양한 기술들이 4차 산업혁명의 대표적인 기술로 언급되고 있다(WEF, 2016).

4차 산업 혁명을 주도하는 기술은 인공지능이라 할 수 있다. 인공지능(Artificial Intelligence, A.I) 기술의 발전이 심화됨에 따라 이에 대한 정의도 점차 변화하고 있다. 김인택, 안대진, 이해영(2017)은 인공지능에 대한 다양한 정의가 존재함을 지적하며 "지능이 요구될 것이라고 생각되는 행위를 보여주는 컴퓨터화된 시스템 또는 실제 환경에서 일어나는 문제를 해결하기 위해 복잡한 문제를 풀 수 있거나 적절한 행동을 취할 수 있는 시스템"이라고, White House(2016)의 정의를 인용하고 있다. 즉, 일상생활이나 일상 업무의 다양한 문제 해결에 인공지능이 사용될 것이라는 관점으로 이해할 수 있다. 실제로 IBM, Google 등은 일상생활에 적용할 수 있는 다양한 기술을 상용 클라우드 서비스로 제공하고 있다. 최근에는 인공지능을 이용한 자연어 이해, 이미지 인식, 음성인식 등의 기술이 급속도로 발전하고 있다. 스탠포드 비전랩의 안드레 카파시(Andrej Karpathy)는 2016년도 딥러닝 스쿨에서 현재 인공지능은 이미지 인식 대회인 이미지넷(ImageNet)의 연도별 정확도 수치를 보면 2012년 80%에서 2015년에는 96%로 인간의 능력은 인간과 동등한 수준이라 언급하였다(Fridman, 2016). 이처럼 사람의 눈이나 귀를 대신하는 제한적인 인공지능(narrow AI)은 상당한 수준에 도달했다. 사람처럼 생각하고 학습하는 일반적인 인공지능(Artificial General Intelligence, AGI)은 인간의 신경망을 모방한 뉴럴 네트워크(Neural Network) 기반의 딥러닝 알고리즘을 통해 그 가능성이 연구되고 있다. IBM, 구글, 아마존(Amazon), 마이크로소프트(Microsoft) 등 글로벌 IT기업들은 이러한 인공지능 기술과 서비스를 지속적으로 발전시키고 있다. 국내에서도 2013년부터 한국전자통신연구원(ETRI)의 주도로 엑소브레인(Exobrain)이라는 인공지능 플랫폼을 개발하고 있다. 엑소브레인은 3단계(Phase)에 걸쳐 10년간 수행하는 과제이며 올해 4년 차, 2단계 사업이 시작되었다. 엑소브레인은 IBM Watson을 넘어서는 것을 목표로 전문가 수준의 질의응답(Question Answering)을 통한 지식 서비스를 제공할 수 있는 SW 개발을 목표로 하고 있다(김현기, 허정, 임수종, 이형직, 이충희, 2017).

기록관리 영역에서는 기관별 한 명의 기록관리 담당자만 채용하는 1인 기록관 체제에 대한 문제점이 지속적으로 제기되어 왔다(한국기록전문가협회, 2017). 전문 인력은 부족한 실정이고, 기록관리 업무는 관련 정보를 저장하고 관리하는 수준에 그치고 있다(한성산, 2010). 이러한 문제점을 해결하기 위해 기록학계에서도 인공지능 기술을 기록관리 업무에 접목할 수 있는지에 대한 관심이 커지고 있다. 국가기록

원은 2017년 '차세대 기록관리 재설계 연구'를 통해 빅데이터, 인공지능 등의 신기술을 이용한 지능형 기록관리 방안에 대해 연구했다. 이 사업의 3세부과제에서는 실제로 인공지능 플랫폼을 이용한 기록물 자동분류 테스트베드를 수행했다. 이 연구에서는 인공지능 기술이 방대한 기록의 일괄 처리를 통한 효율성 제고, 업무의 정확도 향상, 기록 내용 기반의 지능화된 이용자 서비스 제공 등을 가능하게 할 것이라고 제시하고 있다.

이 연구의 목적은 기계학습을 이용한 기록 자동분류를 실제 수행해보고 실무에의 적용 가능성을 확인한 후, 이를 위해 기록관리기관이 준비해야 할 사항과 실제 수행 과정의 고려사항을 파악하는 것이다. 따라서 본 연구는 기록 자동분류를 직접 이행하고, 이를 위한 각 과정의 고려사항을 도출하였다. 이를 위해 실험연구를 중심으로 사례분석 등의 문헌연구를 병행하였다. 실험연구에서는 ETRI의 엑소브레인을 이용하여 서울시 결재문서의 자동분류 테스트를 수행하고 결과를 분석하였다. 엑소브레인을 선택한 이유는 툴 선정 시점 기준으로 한글 처리가 가능한 유일한 툴이었기 때문이다. IBM, 구글 등 국외 인공지능 플랫폼은 기술력은 뛰어나지만, 당시에 한글 처리를 지원하지 않았고, 국내 민간기업의 플랫폼은 테스트를 위한 기술지원에 많은 비용이 소요되었다. 연구팀은 기록물 자동분류를 위한 개념 증명(Proof of Concept, PoC)의 툴로 엑소브레인을 사용했다. 실제 학습과 테스트는 엑소브레인을 개발한 ETRI SW콘텐츠연구소의 언어지능연구그룹이 수행했다.

본 연구팀은 연구 초기에 테스트 시나리오를 작성하여 ETRI와 함께 엑소브레인을 이용한 기록 자동분류 방법에 대해 논의한 후 구체적인 테스트 방법을 결정하였다. 파일럿 테스트와 본격적인 본 테스트에 쓰일 기록 분류체계의 클래스 수, 클래스별 학습 데이터의 수, 학습 데이터의 레이블링 항목, 학습셋과 평가셋의 비중, 학습에 쓰일 알고리즘 등을 합의하였고 이에 따라 테스트가 진행되었다.

문헌연구에서는 자동분류 관련 선행연구와 기계학습, 지도학습 알고리즘의 특징 분석, 인공지능 기술을 업무에 적용한 사례 등을 정리하였다. 또한, 기계학습 기술이 본격화되기 전에 존재했던 자동분류 기술과 기계학습 기반의 자동분류의 차이점을 파악하기 위해 이에 대한 선행연구와 ECM 등 엔터프라이즈 기록관리 솔루션들의 자동분류 기능을 살펴보았다.

2. 지도학습 기술의 기록관리 적용 필요성

문헌의 자동분류에 대한 연구는 1960년대부터 시작되었다. 초기의 연구는 주로 텍스트 범주화에 관한 연구였고 1990년대에 기계학습 이론이 도입되면서 활성화되었다. 이에 따라 텍스트를 대상으로 하는 분류기의 성능이 크게 향상되었다(Sebastiani, 2002). 국내외의 관련 연구는 대부분 로이터(Reuters), 20뉴스그룹(20Newsgroup) 등 공개된 텍스트 데이터 세트를 이용하여 분류기의 성능을 측정하고 정확도를 향상시키는 방안을 모색하는 것이었다. 이러한 연구들은 다양한 응용분야에서 문헌 분류 프로그램의 성능을 개선하기 위해 특정 또

는 일부 영향 요소들에 중점을 두어 실험을 수행한 결과를 보고하였다(박찬정, 성동수, 이건배, 2012; 송성전, 정영미, 2012; 이용구, 2009, 2013; 이재윤, 2005a, 2005b; Foulds & Frank, 2010; Khan, Baharudin, & Lee, 2010).

기록관리 분야에서는 기록의 맥락정보를 이용하여 자동분류하는 시스템 등이 연구되었다. 장지숙, 이해영(2009)은 기록 자동분류가 기록담당자의 업무의 편의성을 가져올 수 있으며, 분류 시간을 단축할 수 있다고 했다. 이 연구에서는 계층구조를 갖는 기록의 특수성을 고려하여 개별 문서가 아닌 기록의 집합적 맥락정보를 이용한 자동분류시스템을 설계하였다. 연구결과 맥락정보의 품질이 시스템의 성능에 큰 영향을 미친 것을 확인하였다. 남은경, 안혜림, 송민(2013)은 경기도 홈페이지의 민원 게시판 게시물을 자동분류하는 시스템을 구현하였다. 이 연구에서 자동분류 시스템의 성능에 영향을 미치는 요소는 분류체계의 전처리, 한글 형태소 분석기의 성능, BRM의 단계별 가중치 등이었다. 정확도를 높이기 위해서는 맥락에 따라 달라지는 단어의 의미를 변별해내고, 유사어를 연결해주는 시스템과 서포트 벡터 머신(Support Vector Machine, SVM) 등의 알고리즘의 도입이 필요하다고 제시하였다. 이러한 선행연구들을 통하여 자동분류 시스템의 동작 원리와 개념을 확인하였고 자동분류 시 언어 학습 모델의 성능에 따라 자동분류 시스템의 성능이 영향을 받는다는 것을 확인하였다.

최근 언어 학습 분야의 연구에도 기계학습이 도입되고 있다. 나승훈, 민진우(2016)는 개체명 인식을 위한 단어 표상을 보완하기 위해 LSTM(Long Short Term Memory network)과 합성곱신경망(Convolutional Neural Network, CNN)을 기반으로 문자 단위 표상을 결합하는 방식을 제안하였고, 이를 통해 기존의 영어 및 한국어 개체명 인식 성능을 향상시킨 결과를 제시하였다. 또한, 최윤수, 차정원(2016)은 워드임베딩(word embedding)방법의 CBOW(Continuous Bag-of-Words)모델과 K평균(K-means) 알고리즘을 이용하여 개체명 인식 및 분류 성능을 향상시켰다. 김다혜, 이지형(2016)은 기존 LSTM 모델을 변형하여 문서 주제와 해당 주제에서 단어가 가지는 문맥적인 의미를 단어 벡터 표현에 반영할 수 있는 새로운 언어 학습 모델을 제안하였다. 이처럼 인공지능을 활용한 자동분류의 성능에는 언어 학습모델이 영향을 미친다.

하지만 인공지능 기술이 아니더라도 기록의 자동분류는 가능하다. 스팸 메일인지 아닌지를 이용자가 판단하여 처리 방식을 결정하는 규칙 기반(rule based)의 알고리즘만으로도 간단한 자동분류 기능을 구현할 수 있다. HP RM(Record Manager), IBM Datacap, Alfresco 등 상용 ECM(Enterprise Contents Management) 솔루션들은 인공지능 기술이 확산되기 이전인 2000년대 초반부터 이미 이메일이나 문서의 자동분류 기능을 제공하고 있었다. 최근 이러한 솔루션들에서도 기계학습이나 딥러닝 기반으로 자동분류 기능이 발전되고 있지만, 여기서 확인해야 할 사항은 규칙 기반의 자동분류와 지도 학습 기반의 자동분류 차이점이 무엇이냐 하는 것이다. 여러 사례를 통해 그 차이를 살펴볼 필요가 있다.

HP는 자사의 검색엔진인 아이돌(IDOL)을 이용하여 HP RM, HP ControlPoint 솔루션에

서 문서를 자동으로 분류하는 기능을 제공하고 있다(HP, 2014). 'ControlPoint'라는 툴이 기업의 문서 생산시스템이나 저장소로부터 기록을 식별(identification)하고 획득(declaration)하면 HP RM에서 미리 정의된 상위 분류 카테고리에 문서를 자동으로 분류하고 이어서 하위 분류 카테고리로 재분류하게 된다. 만약 서울시 공무원이 '서울시청 광장 사용신청' 민원 문서를 HP RM에 등록하면 '일반행정지원>민원관리' 등으로 분류되는 셈이다. HP RM의 자동분류 기능은 스팸메일 규칙처럼 분류된 이후 생산 날짜별로 묶거나 새로운 폴더를 만들어 이동시키는 등의 규칙을 미리 설정하도록 하고 있다. 또한, 자동분류 알고리즘 정확도에 임계치(threshold)를 두어 예상되는 정확도 수치가 특정 수치 이상일 때에만 자동분류 되도록 하거나 기존에 분류가 완료된 문서들로부터 학습(training)시키는 기능도 제공하고 있다(Whitefield, 2016). 최근 HP의 ECM 제품군들은 텍스트뿐만 아니라 비디오나 이미지 분석, 음성인식, 감성분석, 트렌드 분석 등으로 기능을 확장시키고 있다. 이러한 기능들은 자연어 처리(Natural Language Processing, NLP)나 빅데이터 분석, 기계학습, 딥 뉴럴 네트워크 등의 기술이 적용되었기 때문에 가능해졌다(Microfocus, 2017).

IBM 파일넷(FileNet)은 'Datacap' 응용프로그램을 통해 문서 이미지에서 정보를 추출한 후 IBM 왓슨(Watson)의 자연어 처리 등을 통해 텍스트의 자동분류를 수행한다(Lau, 2015). 오픈소스(Opensource) ECM 솔루션인 알프레스코(Alfresco)는 문서 자동분류의 규칙을 설정하기 위한 관리도구(Admin tool)와 분류 가이드(Classification Guides)를 제공한다. 추가로 입수되는 문서나 파일은 이미 작성된 분류 가이드의 규칙에 따라 자동으로 분류된다(Alfresco, 2017).

이러한 ECM 솔루션들은 규칙 기반의 자동분류를 수행하는 것이다. 규칙 기반 자동분류는 이용자가 미리 설정해 놓은 규칙에 따라 자동분류를 수행한다. 즉, 이용자가 얼마나 규칙을 세부적으로 설정해주는가에 따라 분류기의 성능이 결정된다. 예를 들어 문서의 출처나 특정 키워드 포함 여부에 따라 정해진 분류체계로 배정하는 식이다. 하지만 기계학습을 통한 자동분류의 경우 기계가 학습 데이터의 특징(feature)을 통해 현행 분류체계의 레이블을 학습한 후 분류체계를 결정한다. 따라서 규칙기반 자동분류는 점차 기계학습 기반의 지도학습 모델로 대체되고 있다. 그리고 ECM 솔루션에서 쓰이던 다양한 규칙들은 지도학습을 위한 학습데이터로 유용하게 활용되고 있다. 즉, 텍스트 분석 등 기존의 자동분류를 위한 기술들이 기계학습 기술로 이어지며 확장되고 있는 것으로 이해할 수 있다.

최근 기록관리 분야에서도 지도학습 기반의 연구들이 시작되었다. ㈜스토리안트는 공공기록물 목록 수백만 건을 텐서플로우 기반의 딥러닝 알고리즘에 학습시키고 이를 이용한 기록물 보유목록 관리, 공개관리 등의 기능을 개발했다. 이 솔루션은 기록 목록을 입력받아 기계학습을 수행한 후 새로운 데이터가 입력되었을 때 기존에 학습되어있는 데이터를 통해 기록물의 보존기간 검수 및 제안, 기록물 공개 비공개 여부 검사, 민감정보 검사 등의 기능을 수행한다(방재현, 2014, 2015).

국립문화재연구소는 텐서플로우를 이용하여 문화재 이미지 인식 프로토타입을 개발하였다.

학습을 위하여 남대문 등 각 클래스별로 최소 100건 이상의 이미지를 수집하였다. 이미지는 해당 클래스의 특징이 잘 나타나는 것을 선별하여 폴더로 구분하였고 폴더명과 클래스명을 일치시켜 레이블로 이용하였다. 국립문화재연구소는 이 프로토타입을 통하여 학습에 최적화된 이미지 수집과 특징 추출 알고리즘 연구가 선행되어야 함을 강조하고 있다(국립문화재연구소, 2017).

한국중부발전은 '신기술을 적용한 기록 관리 시스템 개발' 사업을 준비하고 있다. 이 사업은 기록관리시스템에도 점차 인공지능 기술이 도입될 것임을 시사하고 있다. 사업의 제안요청서에는 기록물의 자동분류, 접근권한 및 공개여부 추천, 챗봇 개발, 개인 맞춤형 검색과 알림, 분류체계 추천, 민감정보 탐지 및 기계학습 기반의 오류 관리, 폐기업무 시의 보류항목 자동제안 등 자동분류에 필요한 학습 기반의 기술이 대거 포함되어 있다(국가기록원, 2017). 이처럼 자동분류는 업무의 효율성 향상을 위해 지속적으로 연구됐다. 기록의 자동분류를 위해서는 학습에 최적화된 기록을 선별하고 이를 통해 학습해 가는 전략이 필요하다.

3. 엑소브레인을 활용한 기록 텍스트 자동분류

3.1 기록 텍스트 자동분류의 목적 및 필요성

기록 텍스트를 자동으로 분류하는 목적은 크게 두 가지이다.

첫째, 기록관리 업무의 효율성과 정확성을 제고하는 것이다. 국가기록원뿐만 아니라 많은 기록소장기관들은 소수 인력이 대량의 기록물을 관리하고 있다. 공공기록은 단위과제별로, 보존기간, 공개 여부, 비공개사유 등의 메타데이터가 정확하게 관리되어야 하는데 생산 과정의 실수로 누락되거나 오분류되는 경우가 많다. 자동분류 기능은 기록 생산시스템에 적용되어 단위과제 등의 관리 값이나 기타 메타데이터를 추천하거나 기록시스템에 보관하고 있는 방대한 기록으로부터의 오분류나 개인정보 항목을 검출하여 기록관리 담당자의 업무 부담을 줄이고 정확한 업무결과를 유지하도록 해 줄 것이다. 또한 정부기능분류체계(Business Reference Model, 이하 BRM)가 적용 이전의 기록을 2007년부터 사용된 현재의 BRM으로 분류하여 일원화된 관리의 틀로 활용할 수도 있을 것이다.

둘째, 기록 내용 기반의 향상된 검색을 제공하는 것이다. 국가기록원은 생산기관별 검색, 기술계층별 검색, 관련용어 검색 등 주로 출처 중심의 온라인 검색도구를 제공한다. 하지만 이러한 도구들만을 이용하여 원하는 기록이 어디에 있는지 탐색하기란 쉽지 않다. 이용자들은 대부분 찾으려는 기록이 어떤 기관에서 생산되었는지, 어느 컬렉션의 하위에 있는지를 알지 못하기 때문이다. 또한, 기록의 특성상 군(records group), 계열(series) 등 상위계층의 기술(description)은 상세하지만, 철(file)이나 건(item) 단위의 기술은 제공하지 않는다. 만약 국가기록원이 기록 건 단위로 텍스트를 추출하여 일반적인 주제 분류나 생산 당시의 기능분류체계로 제공할 수 있다면 보조적인 검색도구로 유용하게 활용될 것이다.

이를 기반으로, 본 연구에서는 ETRI의 인공지능 플랫폼인 엑소브레인을 이용하여 자동분류 테스트베드를 구축하고 기록물의 텍스트를 자동으로 분류해 보았다. 이 테스트의 목적은 우선 현재 활용 가능한 인공지능 툴로 기록을 자동분류할 수 있는지 기술 수준을 파악하는 것이다. 또한, 각 과정에서 얻은 시사점을 바탕으로 실제 기록관리기관에서 기록을 자동분류하고자 할 때 준비해야 할 사항과 고려할 사항을 제시하는 것이다.

3.2 테스트 수행 과정

자동분류 테스트는 플랫폼 선정, 기록 선별, 분류체계 선별, 텍스트 추출, 태깅, 정제, 학습, 평가, 분석(시사점 도출)의 순서로 진행되었다.

플랫폼 선정을 위해 연구팀은 활용 가능한 국내외 후보를 조사한 후 활용 가능성을 타진하였고 한글 처리가 가능한 유일한 인공지능 플랫폼인 엑소브레인을 활용하여 자동분류 테스트를 수행하게 되었다. 엑소브레인은 질의응답 기반의 지식 서비스를 제공하는 인공지능 플랫폼으로 전체 10년 개발단계 중 올해 4년차가 진행되고 있다. 본 테스트베드에서는 엑소브레인의 기술 중 형태소분석, 개체명인식, 분류(classification) 등 3종의 기술을 활용하여 기록 자동분류를 수행하였다. 엑소브레인의 형태소 분석 기술은 45개의 세종 태그셋[1]을 기반으로 하며, 기계학습방법론으로는 시퀀스 레이블링(Sequence labeling) 기반 음절단위 품사 태깅 방법을 사용한다. 분류 알고리즘으로는 'Structural SVMs"[2]을 사용하고 전처리 및 후처리 단계에 대용량 형태소 사전을 결합하여 성능을 개선하였다(ETRI, 2017).

연구팀은 테스트 시나리오를 설계한 후 ETRI의 SW컨텐츠 연구소의 언어지능 연구 그룹과 자세한 테스트 방법에 대해 논의하였다. 연구팀은 기록의 분류(classification)뿐만 아니라 군집화(clustering)를 이용한 분류체계 자동생성 테스트를 수행하고자 했으나, 엑소브레인은 군집화 기술이 지원되지 않아 지도학습 기반의 자동분류 테스트만을 진행하였다. 본 연구는 파일럿 테스트(Pilot test)와 본 테스트로 구분하여 수행하였다. 파일럿 테스트는 당초 예상보다 긴 3개월 동안 수행되었다. 소량의 학습데이터를 만드는 데에도 상당한 시간이 소요되었기 때문이다. 실제 기록 분류 테스트에 의미 있는 결과를 얻기 위해 먼저 파일럿 테스트(Pilot Test)를 통해 개선방향을 도출하고 이를 바탕으로 본격적인 테스트를 수행하였다.

테스트용 기록은 "서울시 정보소통광장"에서 제공되고 있는 결재문서를 이용하였다. 국가기록원의 소장기록은 디지털화 비중이 매우 낮았고, 스캔된 기록의 경우에도 OCR 기술이 미적용 되어 있거나, 한자 및 수기로 작성되었거나, 문서의 상하가 바뀌어 있는 경우도 많아, 텍스트 추출 작업에 상당한 시간이 소요되리라 예

1) 세종 태그셋이란 세종계획의 산출물로 한글 형태소의 품사를 '체언, 용언, 관형사, 부사, 감탄사, 조사, 어미, 접사, 어근, 부호, 한글 이외'와 같이 나누고 각 세부 품사를 태그로 구분한 것이다.
2) Structural SVM이란 기존의 SVM을 확장한 기계학습 알고리즘으로, 기존의 SVM이 바이너리 분류, 멀티클래스 분류 등을 지원하는 반면에, structural SVM은 더욱 일반적인 구조의 문제를 지원한다.

상되었기 때문이다. 서울시의 결재문서들은 이미 텍스트 추출 작업을 거친 후 제목, 생산자 등의 각 항목을 DB화하여 제공하고 있어 학습데이터 구축이 용이하였다. 테스트용 기록을 선별하는 기준은 제목이나 본문에 해당 분류체계에 적합한 키워드를 얼마나 가지고 있느냐 하는 대표성이었다. 개별 대기능에 포함된 기록 건을 모두 확인할 수 없었기 때문에 단위과제 설명문을 통해 적합한 키워드를 추출하기 쉬운 단위과제를 선정하였다.

분류체계는 서울시의 기능분류체계인 BRM을 활용하였는데, 분류체계를 선정하는 기준은 일종의 주제분류체계로 활용이 가능한 일반적인 명칭이거나, 클래스명으로 활용할 경우 변별력이 있는 대상으로 하되, 가급적 다양한 주제영역의 분류명을 선택하였다. 이러한 방식으로 파일럿 테스트에서는 50개의 대기능을 선별하고, 각 대기능별로 100건 이상의 기록을 선정하여 총 5,982건을 활용하였다. 본 테스트에서는 108개의 대기능을 먼저 선별하였는데, 선정

〈표 1〉 파일럿 테스트용으로 선정된 대기능 및 기록 수량

	대기능명	건수		대기능명	건수
1	가족기능강화	127	26	생활폐기물	97
2	건설건축	147	27	아동보호	118
3	고용정책	136	28	야생동식물	111
4	공원녹화	141	29	여성복지	119
5	관광산업	85	30	의료지원	96
6	관광정책	114	31	일반수질	135
7	근로자복지	117	32	자원봉사관리	105
8	노동정책	116	33	자활서비스	112
9	노인생활안정	126	34	장묘	116
10	대기보전	143	35	장애인생활안정지원	110
11	도시환경	115	36	재활용	108
12	문화사업관리	130	37	정보자원관리	134
13	문화산업정책	123	38	정보통신망	119
14	문화예술정책	122	39	정보통신정책	136
15	문화재보존정책	132	40	정보화지원	113
16	방송통신	107	41	주택건설	142
17	보육복지	116	42	중하수도	140
18	훈관리	78	43	지역및도시계획	140
19	부랑인의사상지원	107	44	청소	109
20	사업장폐기물	110	45	청소년정책	109
21	사회복지기반조성	123	46	체육정책	138
22	산업단지개발	109	47	토양	105
23	상수도	146	48	토지	110
24	생태계	119	49	하천	131
25	생활보호대상지원	112	50	환경보전	128

〈표 2〉 본 테스트용으로 선정된 대기능 및 기록 수량

	대기능명		건수
	모호한 대기능명	통합 대기능명	
1	관광산업	관광산업	344
	관광정책		
2	사회복지기반조성	사회복지기반조성	244
	사회복지행정일반		
3	산림자원관리	산림자원관리	263
	산림정책		
4	여성권익증진	여성복지	244
	여성복지		
	요보호여성보호		
5	장애인생활안정지원	장애인생활안정지원	263
	장애인직업재활지원		
6	정보통신정책	정보통신정책	546
	정보화지원		
7	지방세외기획	지방세외기획운용	252
	지방세외운용		
8	지방세운용	지방세기획운용	254
	지방세정책기획		
9	청소년정책	청소년정책	252
	청소년주변 환경관련정책		
10	축산물	축산물	213
	축산업진흥		
11	가족기능강화		270
12	건설건축		231
13	고용정책		228
14	공원녹화		243
15	공중위생		278
16	교통관리		248
17	구조구급		259
18	국제교류		248
19	군 정책		245
20	근로자복지		258
21	기업지원		290
22	노동정책		259
23	노인생활안정		251
24	농림 진흥		256
25	농산물		242
26	농업기술연구개발		242
27	대기보전		241
28	대중교통		259

	대기능명	건수
44	보훈관리	254
45	복구지원관리	251
46	부랑인의사상자지원	248
47	사업장폐기물	390
48	산업기술개발육성	251
49	하천	131
50	환경보전	128
51	산업진흥	251
52	상수도	266
53	생태계	249
55	생활폐기물	243
56	선거관리	246
57	소방시설및장비관리	300
58	소방인력관리	242
59	소방일반관리	210
60	소비자보호	312
61	수사	357
62	식품안전	244
63	아동보호	250
64	야생동식물	248
65	에너지관리	298
66	육상및광역교통정책	250
67	의료지원	275
68	일반수질	260
69	일반행정지원	251
70	자원봉사관리	339
71	자활서비스	280
72	장묘	244
73	재난상황관리	242
74	재난예방관리	230
75	재난재해시설및장비관리	374
76	재활용	234
77	정보자원관리	277
78	정보통신망	238
79	주소정책	244
80	주택건설	229
81	중하수도	242
82	지방공기업	243
83	지역경제	221
84	지역및도시계획	232

	대기능명		건수
	모호한 대기능명	통합 대기능명	
29	도로건설		239
30	도로관리		335
31	도로시설물관리		239
32	도로정책		372
33	도시철도건설		247
34	도시환경		263
35	문화사업관리		256
36	문화산업정책		258
37	문화예술정책		323
38	문화재보존정책		297
39	물류정책		262
40	민방위		362
41	방송통신		205
42	병력동원		244
43	보육복지		228

	대기능명	건수
44	보훈관리	254
85	지역보건	265
86	지역산업	217
87	청소	249
88	체육정책	217
89	초중등교육	245
90	축산재해질병관리	250
91	토양	270
92	토지	217
93	통상협력	212
94	평생직업교육기반	240
95	하천	257
96	항만운영및해상운송정책	240
97	환경보전	240

된 대기능 중 21개는 클래스 간의 모호성과 데이터 수량을 고려하여 10개로 통합하였다. 예를 들어 '관광산업'과 '관광정책'을 '관광산업'으로 '사회복지기반조성'과 '시화복지행정일반'을 '사회복지기반조성'으로 통합하여, 본 테스트에서는 97개의 대기능을 활용하고자 하였다. 각 대기능 별로는 200건 이상의 기록물을 선정하여, 총 25,553건의 기록이 실제 활용되었다.

실제 텍스트 추출 작업은 웹 크롤러를 개발하여 자동화된 방식으로 수행하였다. 서울시 정보소통광장의 시스템 관리자와 합의된 크롤링 방법(1건 크롤링 후 쉬는 시간 설정, 크롤링 건 수)에 따라 약 5~6일이 소요되었다. 실제 크롤링 작업에 소요된 시간은 짧았으나 크롤러의 소스 코드를 작성하고 테스트를 반복하는 과정에 약 1개월이 소요되었다.

그런데, 나중에 파악한 결과 실제 크롤링된 대기능의 수는 연구팀이 선별한 대기능의 수와 달랐다. 파일럿 테스트에서는 대기능 수가 선정되었던 50개에서 69개로 늘었고, 본 테스트에서는 대기능 수가 선정되었던 97개에서 120개로 증가하였다. 이는 서울시 정보소통광장에서 이용자 편의를 위해 일부 BRM의 명칭을 변경하거나 통합하여 제공하고 있었기 때문이다.

태깅(tagging)은 엑소브레인이 학습할 수 있는 형태로 각 기록의 제목(title), 본문(content), 분류체계(class), 생산부서(origin), 보존기간(retention), 관련문서(relation), 태그(tag) 등의 레이블을 입력하는 작업이다. 즉, 태깅 말뭉치를 구축하는 과정으로, 이는 지도학습 과정에서 각 태그(레이블, label)를 학습하여 분류가 가능하도록 하는 핵심적인 학습데이터이다. 이 작업 또한 웹크롤러를 이용하여 수행하였다. 정보소통광장 홈페이지에서는 각 기록의 메타데이터를 구조화된 형태로 제공하여 웹크롤러가 텍스트를 추출한 후 정제 작업을 통해 자동으로 태깅 작업을 수행할 수 있었다.

〈표 3〉 학습데이터 태깅 사례

〈class〉문화관광/문화재보존정책/박물관운영/박물관교육및문화행사운영/박물관문화행사운영〈/class〉
〈origin〉한성백제박물관/교육홍보과〈/origin〉
〈title〉2017 한성백제박물관 사계콘서트 봄 행사 〈하모니 뒤 스와〉 운영〈/title〉
〈contents〉 1. 교육홍보과-87(2017.2.7.) 관련입니다.
2. 2017 한성백제박물관 사계콘서트 봄 행사 〈하모니 뒤 스와〉를 아래와 같이 운영하고자 합니다.
가. 행 사 명 : 아트 콘서트 〈하모니 뒤 스와〉
나. 운영일시 : 2017년 4월 29일(토) 오후 5시 ~ 6시 30분
다. 운영장소 : 한성백제박물관 한성백제홀
라. 공연단체 : 파리뮤직포럼
마. 공연내용 : 피아노 연주
바. 참가대상 : 일반시민
사. 참가인원(예상) : 200명 내외
아. 참가방법 : 당일 선착순 무료 입장
붙임: 1. 프로그램(안) 1부. 끝.〈/contents〉

다음으로, 정제 작업은 엑셀, 텍스트 편집기, 서브라임 텍스트(Sublime text)[3] 등의 툴을 이용하여 연구팀이 수동으로 수행하였다. 주로 불필요한 공백 제거, 본문이 없는 콘텐츠 삭제, 특수 문자 삭제, 표 서식 수정 등 엑소브레인의 학습에 방해가 되는 요소들을 제거하는 과정이었다. 또한 크롤링 대상 문서의 공개여부(공개, 비공개, 부분공개), 텍스트 추출 가능 여부에 따라 해당 기록 건을 학습데이터로 포함하거나 대상 데이터에서 삭제하는 작업에 많은 반복 작업이 필요했다. 이처럼 본 테스트뿐만 아니라 파일럿 테스트에서도 수천 건의 텍스트를 육안으로 검수하며 정제하는 작업에 상당한 시간이 소요되었다. 태깅 및 정제가 완료된 결과물을 1개의 텍스트 파일로 만든 후 학습과 평가를 위해 ETRI에 전달하였다.

학습과 평가 작업은 엑소브레인을 개발한 ETRI의 SW연구소의 언어지능연구그룹이 수행하였다. 학습에 사용된 알고리즘은 지도학습 방식의 분류(classification)이며, 태깅 말뭉치의 90%는 학습에, 10%는 평가에 사용하는 'Ten-fold Cross Validation' 방법을 적용하였다. 학습에 사용된 태그(레이블)는 파일럿 테스트의 경우 분류체계(class), 제목(title), 본문(contents) 등 3종, 본 테스트에서는 여기에 다른 관련 문서 5-10개 정도가 제시되는 관련문서(relation) 태그를 추가하여 4종을 활용하였다. 생산부서(origin), 보존기한(retention) 등의 태그는 기록 자동분류를 위한 학습에 도움이 요소가 없다고 판단하여 학습용 레이블로 활용하지 않았다. 태그(tag)의 경우 기록물의 분류체계명과 동일한 경우가 대부분이어서 학습할 경우 항상 정답을 맞추게 되므로 제외하였다.

파일럿 테스트에서는 90%의 학습데이터(5,362건)와 10%의 테스트데이터(597건)로 학습 및 평가를 진행하였다. 다음 진행된 본 테스

3) 서브라임 텍스트란 코드 에디터(code editor)로 프로그래밍 언어를 작성 할 때 사용하는 프로그램으로 빠르고 가볍고 확정성이 좋은 에디터이다.

트에서는 가설을 검증하기 위해 네 가지의 방법으로 학습 및 평가를 진행하였다.

- 방법 1. 태깅 말뭉치의 90%(22,998건)를 학습하고 10%(2,555건)를 평가하는 방법이다. 이 방법은 파일럿 테스트와 동일한 방식이며, 인공지능의 성능을 평가하는 가장 일반적인 방식이다.
- 방법 2. 풍부한 키워드를 학습하게 되어 자동분류의 정확도가 향상될 것이라는 가설을 증명하기 위해 관련문서(relation)태그를 추가하여 학습하고 평가하는 방법을 사용하였다. 태깅 말뭉치의 90%(22,998건)를 학습하고 10%(2,555건)를 평가하였다.
- 방법 3. 태깅 말뭉치의 10%(2,555건)를 학습하고 90%(22,998건)를 평가하는 방법이다. 적은 수의 태깅 말뭉치만으로 얼마나 분류 정확도가 도출되는지를 확인하기 위한 방법이다. 만약 분류 정확도가 80% 이상이라면 학습데이터 구축 과정에 들이는 노력을 줄일 수 있기 때문이다.
- 방법 4. 관련문서(relation)태그를 추가한 10%(2,555건) 학습, 90%(22,998건)평가 방법이다. 이 방법 역시 적은 수의 태깅 말뭉치로 얼마만큼 정확한 분류 결과가 도출되는지 확인하기 위한 방법이다.

학습과 테스트에는 총 2주가 소요되었다. ETRI의 연구원들은 연구팀과 합의한 방식에 따라 엑소브레인을 학습시키고 클래스별 성능 평가 결과의 점수표를 연구팀에게 제공하였다.

3.3 결과 확인 및 시사점 도출

학습 및 평가 결과 엑소브레인은 파일럿 테스트에서는 83.08%, 본 테스트에서는 97.81%(방법2 기준)의 정확도로 자동분류를 수행했다. 본 테스트에서 정확도가 획기적으로 향상된 이유는 파일럿 테스트를 통해 발견된 문제점들을 개선했기 때문이다. 즉 선정된 분류체계 클래스간의 모호성 제거, 관련문서(relation)태그의 추가, 데이터 정제 품질의 향상 등이다. 그 중 관련문서(relation) 태그의 경우 학습에 포함한 경우(97.81%)와 포함하지 않은 경우(91.26%)의 정확도 차이가 6.55%나 발생하며 큰 차이를 보였다. 파일럿 테스트와 동일한 조건인 방법1의 경우 91.26%의 정확도가 도출되었는데, 이는 학습용 기록의 선별, 클래스간 모호성 제거 등 학습

〈표 4〉 학습 및 평가 방법

구분	파일럿 테스트	본 테스트
학습 데이터	서울시 기록물 5,982건	서울시 기록물 25,553건
분류 체계	서울시 BRM (50개 대기능 선정) (69개 대기능 크롤링)	서울시 BRM (97개 대기능 선정) (120개 대기능 크롤링)
학습용 태그	〈class〉, 〈title〉, 〈content〉	〈class〉, 〈title〉, 〈content〉, 〈relation〉
학습/평가 방법론	학습셋 90%(5,362건), 평가셋 10%(597건)	방법1: 학습셋 90%(22,998건), 평가셋 10%(2,555건) 방법2: 학습셋 90%(22,998건), 평가셋 10%(2,555건)(relation 포함) 방법3: 학습셋 10%(2,555건), 평가셋 90%(22,998건) 방법4: 학습셋 10%(2,555건), 평가셋 90%(22,998건)(relation 포함)

〈표 5〉 테스트 결과

	파일럿 테스트	본 테스트			
분류체계(BRM)	학습90%, 평가10%	방법1	방법2	방법3	방법4
정책영역	86.61%	93.42%	98.32%	85.74%	92.13%
대기능	83.08%	91.26%	97.81%	77.68%	88.11%

데이터의 품질이 자동분류의 정확도 향상에 최대 8.18%만큼의 효과가 있었음을 말해준다. 반면 방법3(77.68%)과 방법4(88.11%)에서는 소량의 학습데이터만으로는 높은 정확도를 얻기가 불가능함을 확인하였다.

본 테스트의 방법2를 활용하였던 자동분류 평가 결과에서, 오답 항목 56건(2.19%)을 분석한 결과는 다음과 같다.

첫째, 분류 클래스의 포괄성으로 인한 오분류가 14건(25%)이었다. 예를 들어 〈표 6〉과 같이, '행정/일반행정지원'에 해당하는 '공공안전관 대체근무 명령'을 인공지능은 '환경/공원녹화'로 분류했다. 서울시의 '행정/일반행정지원' 클래스에는 행정 처리와 관련된 아주 다양한 주제의 키워드들이 포함되는데, 이는 인공지능이 '행정/일반행정' 클래스의 특징(feature)을 제대로 뽑아내거나 학습하지 못한 것으로 볼 수 있다.

둘째, 학습데이터 품질에 따른 오분류가 13건(23.21%)이었다. 이 경우는 본문의 내용과 분류체계의 클래스 간에 관련성이 없는 경우와 학습데이터의 클래스 오류의 경우로 나뉜다. 여기서 관련성이 없는 경우는 기록물의 내용과 제목을 분석하였을 때 서울시와 인공지능의 분류 결과 모두 실제 문서를 반영하지 않았다. 예를 들어 〈표 7〉의 기록은 '행사개최에 따른 청소지원 요청'인데, 서울시에서는 '경제/농림진흥'으로 분류하였지만 인공지능은 '건강/지역보건'으로 분류하였다. 추출된 데이터 본문의 내용은 농림진흥이나 지역보건과는 상관이 없는 것이었다.

또 학습데이터 클래스의 오류는 서울시 정보소통광장에서 이용자 편의를 위해 일부 BRM의 명칭을 변경하거나 통합해서 제공하여 나타난 결과였다. 예를 들어 〈표 8〉에서 보이듯이, '서울김장문화제 행사 시 도로명주소 홍보 계획'이라는 기록에 대해 서울시의 분류는 '기타/주소정책'으로 되어있고 인공지능은 '행정/주소정책'으로 분류

〈표 6〉 분류 클래스의 포괄성으로 인한 오분류 예시

공공안전관 대체근무 명령

투명하고 신뢰받는 청렴서울, 천만시민의 자랑입니다.
서부공원녹지사업소 수신 내부결재(경유) 제목 공공안전관 대체근무 명령
월드컵공원에서 근무중인 공공안전관의 연가 실시에 따른 대체근무를 아래와 같이 명령하여 공원관리에 만전을 기하고자 합니다.
□ 대체근무 명령 내역 근 무 일 당초 근무자 대체 근무자
근 무 지 사 유2017.07. 12.(수) 주간 ***(C조) ***(A조) 노을정문 안내소(116) 연가(1일)

〈표 7〉 학습데이터 품질(본문의 내용과 클래스의 관련성이 없는 경우)에 따른 오분류 예시

광화문광장 행사개최에 따른 청소지원 협조요청

1. 서울시에 협조하여 주시는 귀 기관에 감사드립니다.
2. 우리시에서는 도농교류 활성화를 위하여 시민들이 참여하는 농부시장 및 영화시사회(파밍보이즈)를 아래와 같이 개최할 예정이므로 행사장 청결유지를 다음과 같이 요청하니 적극 협조바랍니다.
   가. 행사개요
      - 행사명: 농부시장 및 영화시사회 개최
      - 일 시: 2017.6.25.(일) 11:00~22:00
      - 장 소: 광화문광장
   나. 협조요청: 행사종료 후 정리된 쓰레기 수거 요청

※ 행사문의: 쌈지농부(총괄책임자 **********). 끝

〈표 8〉 학습데이터 품질(학습데이터 클래스의 오류)에 따른 오분류 예시

서울김장문화제 행사 시 도로명주소 홍보 계획

문서번호자치행정과-21542
결재일자2016.10.12.
공개여부 대시민공개방침번호 시민주무관행정관리팀장자치행정과장박지수정문철전결 10/12임동국협조
서울김장문화제 행사 시 도로명주소 홍보 계획 2016.10. 행정국(자치행정과)

하였다. 이는 같은 대기능이라도 정책영역이 다르기 때문에 자동분류가 오분류로 판단되었다.

셋째, 원인 파악이 어려운 오분류가 11건(19.64%)이었다. 인공지능이 기록의 제목이나 내용과 관련 없는 분류를 수행한 경우이다. 예를 들어 〈표 9〉의 지역축제 안전관리실적을 제출하라는 공문에서 서울시 분류는 '안전/재난예방관리'였는데, 인공지능의 분류는 '환경/중하수도'였다. 제목과 본문 어디에도 중하수도와 관련된 내용은 없었다.

〈표 9〉 원인 파악이 어려운 오분류 예시

2017년 2분기 지역축제 안전관리 추진실적 제출

1. 국민안전처 안전점검과-2310(2017.7.3.)호와 관련입니다.
2. 2017년 2분기 지역축제 안전관리 추진실적을 붙임서식에 따라 7.7(금)까지 제출하여 주시기 바랍니다.
   가. 작성대상
   나. 제출방법
   - 市 부서 및 기관: 안전총괄과 제출(자치구 및 시에서 안전관리계획 심의를 받은 축제 제외)
   - 자치구: 문화체육부서의 자료 포함하여 재난안전부서 수합제출(안전관리 실적이 없는 자치구 '해당없음' 제출)

   붙 임 1. 2분기 지역축제 안전관리 추진실적 양식 1부.
   2. 관련 공문 및 조사기준 참고자료 각1부.

넷째, 클래스 간의 모호성으로 인한 오분류가 9건(16.07%)이었다. 본 테스트에서는 직관적으로 보았을 때 클래스 간의 모호성 이 확인된 대기능을 통합하여 테스트를 수행하였으나, 결과를 확인해보니 기록물 내 의 주요 키워드(사업, 정책)가 유사하여 모 호성이 생기는 경우가 대부분이었다. 예를 들어 〈표 10〉의 'ISBN 및 간행물 발간등록 번호 부여 신청'에 대한 기록의 경우, 서울 시에서는 '주택도시계획/지역 및 도시계획'으로 분류하였지만 인공지능은 '건설/건설 건축'으로 분류하였다.

다섯째, 학습데이터 내의 자주 언급되는 키워드에 지나치게 의존하는 분류가 5건(8.93%)이었다. 이는 지도학습 과정에서 자주 발생하는 오버피팅(Overfitting)을 원인으로 추측해볼 수 있다. 즉, 인공지능이 학습데이터의 키워드를 지나치게 학습하여 특정 키워드가 포함된 경우 편향적으로 특정 클래스로 분류하는 경우이다. 예를 들어 〈표 11〉에서 보면, 지역아동복지센터의 지원사업 설명회 문서는 서울시 분류는 '행정/초중등교육'이었는데 인공지능은 '여성가족/아동보호'로 분류하였다. 문서 내의 '아동복지센터'라는 키워드가 '여성가족/아동보호' 클래스에서 자주 언급되었기 때문에 인공지능은 '행정/초중등교육'보다 '여성가족/아동보호' 클래스로 분류한 것이다. 만약 클래스별 학습데이터가 200건보다 훨씬 많았다면 이렇게 특정 키워드를 과학습하는 사례는 줄어들 것으로 예상된다.

마지막으로 서울시의 분류보다 인공지능이 더 적합한 분류를 제시한 경우가 4건(7.14%)이었다. 예를 들어 〈표 12〉의 '소방장비 관리운용 계획' 공지문을 서울시는 '안전/소방인력관리'에 분류했는데 인공지능은 '안전/소방시설및장비관리'로 분류한 경우이다. 이는 업무 담당자가 오분류했다기보다는 일반적인 주제 분류의 측면에서 더 적합한 분류로 판단했을 것인

〈표 10〉 클래스 간의 모호성으로 인한 오분류 예시

ISBN 및 간행물 발간등록번호 부여 신청

* 필수입력 사항 저자명: 간행물 생산부서명(제1저자) / 용역보고서의 경우 용역수행기관명(혹은 개인)(제2저자)둘을 세미콜론(;) 구두점을 두고 기입해준다
※ 대외적으로 배포되는 자료(단행본, 용역보고서 등)가 아닐 경우 ISBN 부여 제외 대상입니다.
※ 발간자료에 대해서 과업지시서 및 저작권 이용허락양식을 통해 자료의 저작권이 서울시에 귀속될 수 있도록 조치바랍니다.
※ 수신처: 서울도서관 정보서비스과(ISBN/ISSN 담당) / 정보공개정책과(간행물 발간등록번호 담당) 간행물 발간등록번호 부여 신청
※ 발간자료에 대해서 과업지시서 및 저작권 이용허락양식을 통해 자료의 저작권이 서울시에 귀속될 수 있도록 조치바랍니다.
※ 수신처: 서울도서관 정보서비스과(ISBN/ISSN 담당) / 정보공개정책과(간행물 발간등록번호 담당)
추진실적을 붙임서식에 따라 7.7(금)까지 제출하여 주시기 바랍니다.
가. 작성대상
나. 제출방법
- 市 부서 및 기관: 안전총괄과 제출(자치구 및 시에서 안전관리계획 심의를 받은 축제 제외)
- 자치구: 문화체육부서의 자료 포함하여 재난안전부서 수합제출(안전관리 실적이 없는 자치구 '해당없음' 제출)
붙 임 1. 2분기 지역축제 안전관리 추진실적 양식 1부.
2. 관련 공문 및 조사기준 참고자료 각1부.

〈표 11〉 지역아동복지센터 자기주도학습 지원 사업 설명회 참석 안내

1. 서울시정에 협조하여 주시는 귀 기관의 무궁한 발전을 기원합니다.
2. 서울시는 2016년부터 사회적 배려계층 초·중·고 학생의 기초학습력 배양을 위하여 자기주도학습을 지원하는 사업을 진행하고 있습니다.
3. 2018년에는 위 사업을 서울시지역아동복지센터 학생들에게도 도움이 되도록 지원하고자 합니다.
4. 이를 위하여 사업 내용 및 방향에 대한 설명을 드리고, 지역아동복지센터의 의견 청취 기회를 마련하고자 하오니 바쁘시더라고 꼭 회의에 참석하여 주시기 바랍니다.
   가. 회의일시: 2017. 7.13.(목) 15:00(약 1시간 내외 소요 예상)
   나. 회의장소: 서울시립창동청소년수련관 3층 창의융합과학공방
   다. 참석대상: 각 지역아동복지센터 팀장(과장) 1인
   라. 회의내용
      - 창동자기주도학습관 견학 및 설명
      - 2018년 자기주도학습 사업 안내 및 의견 청취
   마. 지역아동복지센터 자기주도학습 지원 사업 추진 형태(안)
   바. 기타사항: 7월 중 참여희망기관 조사 예정
   붙 임 창동청소년수련관 약도 1부. 끝.

〈표 12〉 서울시 분류보다 더 적합한 경우 예시

2017년 7월중 소방장비 관리운용 계획 알림

"119와 함께 만드는 안전, 시민과 함께 누리는 행복"중부소방서 수신자 수신자참조(경유) 제 목 2017년 7월중 소방장비 관리운용 계획 알림
1. 소방행정과-228(2017.1.16.)호『2017년 소방장비 관리운용 기본계획 알림』와 관련 입니다.
2. 2017년도 7월중 소방장비 관리운용계획을 다음과 같이 알리니, 각 부서에서는 시행에 차질이 없도록 만전을 기하기 바랍니다.

데, 인공지능은 데이터의 제목 및 본문에서 자주 언급된 '소방장비'란 키워드 등을 통해 가장 주제와 가까운 클래스로 분류한 것이다.

이 테스트들은 '서울시 정보소통광장'에 공개된 결재문서와 서울시의 업무기능분류체계를 활용한 것이며, 자동분류에 사용된 툴 또한 업무협약에 의해 제한적으로 제공받은 것이다. 따라서 타 기관에서 이 연구 결과를 바로 적용하기는 쉽지 않을 것이다. 하지만 이 테스트를 통하여 기록 텍스트 자동분류의 도입 가능성을 확인하고 많은 시사점을 얻었다. 이후 타 기관에서도 활용하기 위한 고려사항을 제안하고자 한다.

4. 기록 텍스트 자동분류를 위한 기록관리 기관의 고려사항

8개월 동안 엑소브레인을 이용하여 테스트한 결과를 바탕으로 기록 텍스트 자동분류를 위해 각 과정에서의 고려사항을 정리해보면 다음과 같다.

첫째, 인공지능 기술의 활용을 위한 텍스트 추출과 디지털화 전략이 필요하다. 이번 자동분류 테스트 과정에서는 학습데이터를 만드는 데 가장 많은 시간이 소요되었다. 이미 텍스트화된 서울시 정보소통광장의 데이터를 이용했음에도 웹 크롤러를 개발하고 원하는 텍스트를 추출해 내는 작업이 별도로 필요했다. 국가기록원의 경우처럼 수기 종이기록이 많을 경우 손글씨 인식 등의 기술 또한 요구된다. 이처럼 소장기록의 유형별로 효과적으로 텍스트를 추출하기 위한 단계적 전략이 수립되어야 한다.

대상 기록은 크게 종이기록과 전자문서로 나누어서 생각해 볼 수 있다. 우선 종이기록은 디지털화와 텍스트 추출 작업을 동시에 수행하는 것을 고려해 보아야 한다. 최근 딥러닝 기술을 적용한 OCR도 등장했고 순환신경망(Recurrent Neural Networks, RNN)이나 LSTM(Long Short Term Memory networks) 구조를 사용하여 손글씨 인식의 성능을 획기적으로 향상시킨 사례도 발표되었다(Graves, 2012). 따라서 한자나 일본어 등 외래어 인식, 한글 필기체 인식 등을 위한 딥러닝 기반의 OCR이나 추출 도구를 선별하고 그 성능에 따라 단계별 전략을 수립할 필요가 있다. 전자기록의 경우에도 HWP, PDF/A 등 문서 포맷별로 적합한 텍스트 추출 도구를 선별해야 한다.

둘째, 자동분류를 위한 학습 모델이 필요하다. 자동분류 하고자 하는 목적에 따라 지도학습(Supervised learning), 비지도학습(Unsupervised learning) 등 적용할 학습 모델이 달라지고 학습시킬 데이터도 달라진다. 예를 들어 2007년 이전의 기록을 현재의 BRM으로 분류하고자 한다면 기존 BRM의 레이블을 학습시키는 지도학습 모델, 그 중에서도 분류(Classification) 모델이 적용되어야 한다. 이번 연구에서처럼 BRM의 각 클래스 간 모호성을 제거하고 학습시킬 클래스별로 가장 적합한 기록을 대량으로 선별하고 정제해야 한다. BRM이 아닌 일반적인 주제분류를 하고자 할 때도 이와 유사한 방식의 학습 모델이 필요하다. 기록관의 경우 이미 정해진 분류의 틀이 있고 그에 따라 분류된 학습데이터를 이미 보유하고 있으므로 지도학습 방식을 운용하는 것이 효율적일 것이다. 만약 기존의 분류체계나 학습데이터가 부족하여 새로운 틀로 분류하고자 한다면 처음부터 비지도학습 방식을 적용해야 한다. 분류(Classification)와 군집(Clustering) 모델이 모두 활용될 수 있을 것이다. 이번 자동분류 테스트에서는 분류 모델만을 활용했다. 군집 모델을 자동분류에 적용하기 위해서는 더 고도화된 기술력이 요구되므로 추후 기술 발전에 따라 단계적으로 적용하는 것이 바람직하다. 비지도학습은 레이블링된 기록으로부터 학습하는 것이 아니라 방대한 데이터 안에서 인공지능 에이전트가 특징이나 패턴을 추출하여 스스로 학습하는 방식이다. 비지도학습을 적용하려면 대량의 기록 텍스트를 제공해야 한다.

빅데이터 환경에서는 딥러닝 기반의 비지도 학습 모델의 효율성이 점차 증가할 것이다. 비지도 학습을 통한 자동분류를 위해서는 소장기록뿐만 아니라 외부의 LOD(Linked Open Data)[4] 나 소셜 미디어에 이용자가 태깅한 분류 등을

4) 기술적인 개념으로, 웹페이지가 서로 연결된 것처럼 데이터들끼리 다양한 관계에 의해 연결되어 있는 형태를 말함

통해 스스로 데이터를 학습하는 전략으로 점차 확장시켜 나가는 것이 바람직하다. 일부 학습 데이터에는 지도학습을 적용하고 나머지 트레이닝셋에 대해서는 비지도 학습을 병행할 수도 있을 것이다. 추가적인 조치로 LOD를 통해 딥러닝 알고리즘만으로 해결되지 않는 특화된 도메인의 내용을 인식하고 처리할 수 있도록 하는 것 또한 고려해볼 수 있을 것이다.

학습모델을 수립하고 발전시켜 가는 과정에서 반드시 고려할 사항 중 하나로 오버피팅(Overfitting)이 있다. 오버피팅은 샘플 데이터의 수가 부족하여 알고리즘이 학습 데이터에 필요 이상으로 최적화되는 현상이다. 다시 말해, 분류 기준이 학습 데이터에 너무 치우쳐서 학습 데이터에 없는 다른 값에 대해서는 제대로 분류하지 못하게 되는 것이다. 예를 들어, 얼굴인식 모델을 만들 때 동양인의 얼굴로만 구성되면, 서양인 얼굴을 제대로 분류할 수 없게 된다. 이는 기계학습 과정에서 굉장히 자주 겪는 문제 중 하나이므로, 모델을 만들 때 오버피팅의 발생 가능성을 항상 염두에 두어야 한다. 또한, 소량의 데이터만을 학습한 후 정확도가 높다고 해서 알고리즘의 성능을 과신하면 안 된다.

셋째, 솔루션 선별 및 적용방안이 필요하다. 우선 선택 가능한 자동분류 솔루션은 이번 연구에서 검증해 본 ETRI의 엑소브레인 뿐만 아니라 IBM, Google, Microsoft 등에서 제공하는 클라우드 방식의 상용 서비스까지 다양하다. 실제로 IBM 등의 클라우드 서비스들은 기록물 자동분류에 필요한 문서 변환, 자연어 처리 등의 서비스를 유료로 제공하고 있다(김인택, 안대진, 이해영, 2017). ETRI의 엑소브레인 등 국내 솔루션들은 한글 처리에 대한 강점이 있다. 이를 도입하는 가장 현실적인 방안은 해당 기술을 이전받은 개발업체들과 수년간의 자동분류 프로젝트를 진행하는 것이다. 학습모델을 수립하고 학습데이터를 구축하여 성능을 향상시키는 과정이 단기간에 이루어지지 않기 때문이다. 첫 해에는 파일럿 테스트 등을 통해 PoC를 수행하고 점차 활용 가능한 수준으로 발전시켜야 한다. 또 다른 방안은 기록물관리기관이 직접 ETRI의 기술이전을 받거나, 국회도서관이나 특허청의 사례처럼 국가 인공지능 프로젝트에 참여하여 공동으로 연구를 수행하며 기술을 획득하는 것이다. 국가기록원 등 큰 규모의 기관이 먼저 독자적으로, 또는 업체와 공동으로 기술을 확보한 이후에 각급 기록물관리기관에 솔루션 이전이나 기술교육 등을 제공하는 방식이 가장 현실적이며 효과적일 것이다.

넷째, 인공지능의 활용 관점에서 문서의 생산 및 보존포맷은 ODF(Open Document Format)가 유리하다. ODF는 제목, 생산자, 본문 등 문서의 구조가 XML 형태로 분리되어 표현되므로 별도의 처리를 하지 않고도 각 항목을 메타데이터로 취득하거나 인공지능이 처리하기에 유용한 태깅 말뭉치 형태로 추출하는 것이 가능하다. 현재 대부분의 공공기관에서 사용하고 있는 HWP 포맷은 txt나 ODF 포맷과 달리 입력된 텍스트를 독자적인 비트스트림으로 변형하여 저장하므로 이를 위한 별도의 텍스트 추출도구가 필요하다. PDF/A 또한 텍스트 추출작업을 별도로 수행해야 하고 정확도 또한 완벽하지 않으므로 인공지능의 처리 관점에서는 ODF가 유리하다. 또한, ODF로 생산된 문서는 PDF/A로 변환하지 않고 문서보존포맷으로 활용될 수 있다. 이미 HWP 등의 포맷으로 생산되어 PDF/A로

변환된 문서는 그대로 유지하되 ODF로 생산하는 문서의 비중을 늘려 인공지능의 처리 가능성 또한 도모할 필요가 있다.

다섯째, 한글 자연어 처리를 위한 말뭉치와 전거데이터 구축이 필요하다. 엑소브레인 프로젝트에서는 한글의 문법과 내용을 이해하고 처리하기 위한 수십 종의 학습자원 즉 말뭉치 개발을 병행했다. 기록관리 도메인에 특화된 지식을 처리하기 위해서는 그에 상응하는 말뭉치나 전거데이터를 보조 수단으로 활용하는 것이 좋다. 행정용어나 행정연혁, 인물사전, 시소러스 등 기존에 만들어진 결과물을 활용하고 추가로 개발하여 해당 기록이 담고 있는 내용을 인공지능이 제대로 분류하도록 해야 한다.

여섯째, 기록관리 담당자의 디지털 리터러시 확보 방안이 필요하다. 기록관리 현장에 적용할 만한 자동분류기를 만들기 위해서는 인공지능 개발업체와 기록 전문가의 협업이 필수적이다. 기록관리 담당자는 소장 기록에 대한 전문성을 바탕으로 분류의 목적별 학습 모델을 설계하고 적합한 학습 데이터를 선별하며 테스트 과정을 주도해야 한다. 그러기 위해서는 인공지능의 학습 원리를 이해하고 다양한 학습 알고리즘의 차이를 구분할 수 있어야 한다. 엑소브레인 테스트를 예로 들면, 테스트 결과 보고서에 표기된 정확도(precision), 재현율(recall), 가중치(weight), 바이어스(bias) 등의 용어를 이해하지 못하면 문제점이 무엇인지 판단이 불가능했다. 이와 같이 인공지능을 도입하고자 하는 기록관리 담당자는 전문업체의 도움을 받아 학습 데이터와 알고리즘의 변수를 조정해 가며 스스로 분석 결과를 해석할 수준의 디지털 리터러시 능력을 갖춰야 한다. 적어도 올해 수행된 국가기록원의 '차세대 기록관리 모델 재설계' 연구에서 제시된 새로운 기록관리 개념이나 신기술에 대해 교육을 제공하는 것을 적극 검토해야 할 것이다. 희망적인 것은 인공지능 분야에 양질의 교육자원과 오픈소스 알고리즘이 풍부해졌다는 점이다. 기록관리 담당자 스스로 무크(Mooc) 공개강의나 Youtube, 깃허브(Github) 등에 공개된 인공지능 분야의 학습자료를 익히거나, 국가기록원 등에 인공지능 관련 교육과정을 마련하여, 기록관리 담당자들이 인공지능 리터러시가 되도록 하는 것이 필요하다.

5. 결 론

인공지능은 점차 일상생활과 업무의 영역에 영향을 미치고 있다. 공공 영역에서 인공지능 기술이 각광받는 이유는 기존에 처리하기 어려웠던 방대한 데이터를 한꺼번에 처리하거나, 보다 정확하게 처리하거나, 풀기 어려운 아주 어려운 문제를 해결해 줄 것이라는 기대 때문일 것이다. 기록 텍스트를 자동으로 분류하는 것의 목적 또한 이 모두에 해당된다. 소수의 기록연구사가 방대한 양의 기록을 처리하는 것은 그 동안 아주 풀기 어려운 문제였다. 하지만 최근의 자연어 처리와 지도학습 기술의 진화로 인해 가능한 수준이 되었다.

이 연구에서는 현재 활용 가능한 인공지능 툴을 이용하여 기록을 자동분류해 보았다. 테스트용 플랫폼으로 국내 인공지능 플랫폼인 엑소브레인을 이용하였고 ETRI의 연구팀과 협의하여 테스트를 진행하였다. 테스트용 기록 및 분류체계는 서울시 정보소통광장의 결재문서들와 서

울시 BRM을 이용하였다. 테스트 과정에서 얻은 시사점을 바탕으로 실제 기록관리기관에서 기록을 자동분류하고자 할 때 준비해야 할 사항과 고려사항을 제시하였다.

학습 및 평가 결과 엑소브레인은 파일럿 테스트에서 83.08%, 본 테스트에서는 97.81%(방법2, 관련문서를 학습시키고, 90%의 데이터를 학습시킨 후 10%의 데이터에 적용하는 방법, 기준)의 정확도로 자동분류를 수행했다. 본 테스트에서 정확도가 향상된 이유는 클래스간의 모호성 제거, 관련문서(relation)태그의 추가, 데이터 정제 품질의 향상 등이다. 그 중 관련문서(relation) 태그는 학습에 포함한 경우에는 97.81%(방법2)와 포함하지 않은 경우 91.26%(방법1, 관련문서를 학습하지 않고, 90%의 데이터를 학습시킨 후 10%의 데이터에 적용하는 방법)의 정확도 차이가 6.55%나 되는 것으로 나타났다. 파일럿 테스트(83.08%)와 동일한 조건인 방법1의 경우는 91.26% 정확도가 도출되었는데, 이는 학습용 기록의 선별, 클래스간 모호성 제거 등 학습데이터의 품질이 자동분류의 정확도 향상에 최대 8.18%만큼의 효과가 있었음을 말해준다. 반면 방법3(관련문서를 학습하지 않고, 10%의 데이터를 학습시킨 후 90%의 데이터에 적용하는 방법, 77.68%의 정확도)과 방법4(관련문서를 학습시키고, 10%의 데이터를 학습시킨 후 90%의 데이터에 적용하는 방법, 88.11%의 정확도)에서 보듯이 소량의 학습데이터만으로는 높은 정확도를 얻기가 불가능했다. 적어도 명확한 학습 전략과 학습데이터를 통한 지도학습 방식을 적용하면 업무 현장에 적용할 만한 높은 정확도로 기록 텍스트의 자동분류가 가능하다는 것을 확인하였다.

이 연구에서 제시한 기록 텍스트의 자동분류를 위한 준비 및 고려사항은 다음과 같다. 첫째, 인공지능 기술을 활용한 텍스트 추출과 디지털화 전략이 필요하다. 소장하고 있는 종이기록과 전자문서의 세부 유형별로 효과적인 텍스트 추출 도구를 선별해야 한다. 종이기록의 경우 손글씨 인식 등을 고려하여 디지털화 단계부터 기계학습 기반의 인식 툴을 활용할 필요가 있다. 둘째, 기관에 맞는 자동분류 학습 모델이 필요하다. 이번 연구에서는 BRM을 이용하여 지도학습 방식으로 자동분류했지만 국가기록원이나 대통령기록관 등에서는 주제분류를 위한 딥러닝 기반의 비지도학습이나 분류체계 생성을 위한 군집 모델이 유용할 수도 있다. 테스트 결과에서도 보았듯이 양질의 대량 학습데이터를 구축하지 않으면 높은 정확도를 도출하기 어려우며, 학습데이터에 지나치게 의존하는 오버피팅 현상 또한 발생할 수 있다. 셋째, 솔루션 선별 및 적용방안이 필요하다. 엑소브레인뿐만 아니라 국내외의 다양한 인공지능 솔루션들을 검토하고 개발업체들과 PoC(Proof of Concept) 등을 수행하며 기술을 확보해야 한다. 넷째, ODF 포맷을 문서의 생산 및 보존포맷으로 채택하는 것이 좋겠다. ODF는 문서의 구조가 XML 형태로 분리되어 인공지능이 처리하기에 적합하며 별도의 변환 없이 문서보존포맷으로 활용될 수 있다. 다섯째, 한글 자연어 처리를 위한 말뭉치와 전거데이터 구축이 필요하다. 세종말뭉치 2차 사업 등 국가적 차원에서의 노력과 함께 각 영역에 특화된 학습자원으로서의 전거데이터가 개발되고 공유되어야 한다. 기록 분야에서는 행정용어나 행정연혁, 인물사전, 시소러스 등이 이에 해당한다. 여섯째, 기록관리 담당자의 디

지털 역량을 강화해야 한다. 인공지능 개발업체와 협업을 수행할 때 기록관리 담당자는 소장 기록에 대한 전문성을 기반으로 자동분류의 목적에 따른 학습모델을 설계하고 테스트를 주도해야 한다. 그러기 위해서는 인공지능의 학습 원리를 이해하고 다양한 알고리즘의 차이와 특성을 구분할 수 있어야 한다.

자동분류는 오분류 교정 등 기록의 관리 측면에서도 유용할 수 있지만, 궁극적으로는 기록물 디스크립션에 의존했던 전통적인 검색에서 기록 내용 기반의 검색과 활용으로 크게 한 걸음 나아가는 계기를 제공할 것이다. 이 연구가 기록물관리기관에 산재해 있는 방대한 기록 처리 문제를 해결하는 데 도움이 되길 바란다.

참 고 문 헌

국가기록원 (2017). 2017년도 국가기록원 주요업무 참고집.

국립문화재연구소 (2017). 국립문화재연구소 ISP 보고서.

김다해, 이지형 (2016). 문서 주제에 따른 문장 생성을 위한 LSTM 기반 언어 학습 모델. 한국컴퓨터정보학회 학술발표논문집, 24(2), 17-20.

김인택, 안대진, 이해영 (2017). 인공지능을 활용한 지능형 기록관리 방안. 한국기록관리학회지, 17(4), 225-250. http://dx.doi.org/10.14404/JKSARM.2017.17.4.225

김현기, 허정, 임수종, 이형직, 이충희 (2017). 엑소브레인 한국어 분석 및 질의응답 기술의 개발 현황 및 고도화 계획. 정보과학회지, 35(8), 51-56.

나승훈, 민진우 (2016). 문자 기반 LSTM CRF를 이용한 개체명 인식. 한국정보과학회 학술발표논문집, 729-731.

남은경, 안혜림, 송민 (2013). 공공사이트 게시판 자료의 기록관리를 위한 자동분류 시스템. 제20회 한국정보관리학회 학술대회 논문집, 175-178.

박찬정, 성동수, 이건배 (2012). 기계 학습을 이용한 특허 문서의 자동 IPC 분류. 한국정보기술학회논문지, 10(4), 119-128.

방재현 (2014). 기록 보유목록 관리 방법 및 장치. KR101672522B1. Retrieved from https://patents.google.com/patent/KR101672522B1/ko?assignee=%EC%A3%BC%EC%8B%9D%ED%9A%8C%EC%82%AC+%EC%8A%A4%ED%86%A0%EB%A6%AC%EC%95%88%ED%8A%B8

방재현 (2015). 기계학습 기반 지능형 기록물 공개관리 시스템. KR101627750B1. Retrieved form https://patents.google.com/patent/KR101627550B1/ko

송성전, 정영미 (2012). 용어의 문맥활용을 통한 문헌 자동분류의 성능 향상에 관한 연구. 정보관리학회지, 29(2), 205-224. http://doi.org/10.3743/KOSIM.2012.29.2.205

이용구 (2009). 기계번역을 이용한 교차언어 문서 범주화의 분류 성능 분석. 한국문헌정보학회지, 43(1), 313-332. http://doi.org/10.4275/KSLIS.2009.43.1.313

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

이재윤 (2005a). 문헌간 유사도를 이용한 SVM 분류기의 문헌분류성능 향상에 관한 연구. 정보관리학회지, 22(3), 261-287.

이재윤 (2005b). 자질 선정 기준과 가중치 할당 방식간의 관계를 고려한 문서 자동분류의 개선에 대한 연구. 한국문헌정보학회지, 39(2), 123-146.

장지숙, 이해영. (2009). 맥락정보를 이용한 기록 자동분류시스템 설계. 한국기록관리학회지, 9(1), 151-173.

최윤수, 차정원 (2016). Word Embedding 자질을 이용한 한국어 개체명 인식 및 분류. 정보과학회논문지, 43(6), 678-685. http://dx.doi.org/10.5626/JOK.2016.43.6.678

한국기록전문가협회 (2017). 기록정책포럼 2017 3호.

한성산 (2010). 경기도 기록관의 기록관리 체계의 현황과 개선방안. 석사학위논문, 중부대학교, 기록관리 전공.

Alfresco (2017). Alfresco governance services 2.6 - Create classification guides. Retrieved from https://youtu.be/OiVRg0hgWMw

ETRI (2017). 엑소브레인 한국어 언어분석 툴킷 v2.0. Retrieved form https://itec.etri.re.kr/itec/sub02/sub02_01_1.do?t_id=1210-2017-00440

Foulds, J., & Frank, E. (2010). A review of multi-instance learning assumptions. The Knowledge Engineering Review, 25(1), 1-25.

Fridman, L. (2016). Deep learning for computer vision (Andrej Karpathy, OpenAI)(video) Retrieved from https://youtu.be/u6aEYuemt0M

Graves, A. (2012). Supervised sequence labelling with recurrent neural networks. Studies in Computational Intelligence, 385. Heidelberg: Springer.

Khan, A., Baharudin, B., Lee, L. H., & Khan, K. (2010). A review of machine learning algorithms for text-documents classification. Journal of Advances in Information Technology, 1(1), 4-20.

Lau, R. (2015). The total economic impact™ of IBM Datacap. IBM.

Microfocus (2017). IDOL: Text, video, image and speech data analytics. Retrieved from https://software.microfocus.com/en-us/products/information-data-analytics-idol/overview

Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys (CSUR), 34(1), 1-47.

WEF (2016). The global competitiveness report 2016-2017. Retrieved from

http://www3.weforum.org/docs/GCR2016-2017/05FullReport/TheGlobalCompetitivenessReport2016-2017_FINAL.pdf

White House (2016). Preparing for the future of artificial intelligence. Executive Office of the President National Science and Technology Council. Committee on Technology. Retrieved from https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf

Whitefield, B. (2016). HP records manager auto classification module (demo video). Retrieved from https://www.youtube.com/watch?v=MqAPqYF9GkM

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

Bang, Jae Hyun (2014). Method and apparatus for managing record retention lists. KR101672522B1. Retrieved from https://patents.google.com/patent/KR101672522B1/ko?assignee=%EC%A3%BC%EC%8B%9D%ED%9A%8C%EC%82%AC+%EC%8A%A4%ED%86%A0%EB%A6%AC%EC%95%88%ED%8A%B8

Bang, Jae Hyun (2015). Intelligent documentary disclosure management system based on machine learning. KR101627750B1. Retrieved form https://patents.google.com/patent/KR101627550B1/ko

Choi, Yunsu, & Cha, Jeongwon (2016). Korean named entity recognition and classification using Word Embedding features. Journal of KIISE, 43(6), 678-685. http://dx.doi.org/10.5626/JOK.2016.43.6.678

Han, Sung San (2010). The present condition of records management system and improvement plan in GyeongGi-Do archives. Master's thesis, Chungbu University, Records Management Major.

Jang, Ji-Sook, & Rieh, Hae-Young (2009). Design of automatic records classification system using contextual information. Journal of Korean Society of Archives and Records Management, 9(1), 151-173.

Kim, Dahae, & Lee, Jee-Hyong (2016). LSTM based language model for topic-focused sentence generation. Proceedings of the Korean Society of Computer Information Conference, 24(2), 17-20.

Kim, Hyun-Ki, Hur, Jeong, Lim, Soo-Jong, Lee, Hyung-Jik, & Lee, Chung-Hee (2017). Development

status and upgrade plan of Korean analysis and question answering technology in Exobrain software. Communications of the Korean Institute of Information Scientists and Engineers, 35(8), 51-56.

Kim, Intaek, An, Dae-Jin, & Rieh, Hae-young (2017). Intelligent records and archives management that applies artificial intelligence. Journal of Korean Society of Archives and Records Management, 17(4), 225-250. http://doi.org/10.14404/JKSARM.2017.17.4.225

Korea Records Expert Association (2017). Record Policy Forum 2017 Issue 3.

Lee, Jae-Yun (2005a). Improving the performance of SVM text categorization with inter-document similarities. Journal of the Korean Society for Information Management, 22(3), 261-287.

Lee, Jae-Yun (2005b). An empirical study on improving the performance of text categorization considering the relationships between feature selection criteria and weighting methods. Journal of the Korean Society for Library and Information Science, 39(2), 123-146.

Lee, Yong-Gu (2009). Classification performance analysis of cross-language text categorization using machine translation. Journal of the Korean Society for Library and Information Science, 43(1), 313-332. http://doi.org/10.4275/KSLIS.2009.43.1.313

Lee, Yong-Gu (2013). A study on feature selection for kNN classifier using document frequency and collection frequency. Journal of Korean Library and Information Science Society, 44(1), 27-47. http://doi.org/10.16981/kliss.44.1.201303.27

Na, Seung-Hoon, & Min, Jinwoo (2016). Character-based LSTM CRFs for named entity recognition. Proceedings of Korean Institute of Information Scientists and Engineers, 729-731.

Nam, Eunkyung, Ahn, Hye-Rim, & Song, Min (2013). Automatic classification system for record management of bulletin board on public website. Proceedings of 20th Conference of Korean Society for Information Management, 175-178.

National Archives of Korea (2017). 2017 National Archives of Korea' manual for major duties.

National Research Institute of Cultural Heritage (2017). National Research Institute of Cultural Heritage ISP report.

Park, Chanjeong, Seong, Dongsu, & Lee, Keonbae (2012). Automatic IPC classification for patent documents using machine learning. The Journal of Korean Institute of Information Technology, 10(4), 119-128.

Song, Sung-Jeon, & Chung, Young-Mee (2012). A study on improving the performance of document classification using the context of terms. Journal of the Korean Society for Information Management, 29(2), 205-224. http://doi.org/10.3743/KOSIM.2012.29.2.205