한국기록관리학회지
22(4), 129-149, 2022.11
https://doi.org/10.14404/JKSARM.2022.22.4.129
pISSN 1598-1487 eISSN 2671-7247

JKSARM
Journal of Korean Society of Archives and Records Management

기록관리 분야에서 한국어 자연어 처리 기술을 적용하기 위한 고려사항

Considerations for Applying Korean Natural Language Processing Technology in Records Management

김학래(Haklae Kim)

E-mail: haklaekim@cau.ac.kr

중앙대학교 사회과학대학 문헌정보학과 교수

논문접수 2022.10.21
최초심사 2022.10.25
게재확정 2022.11.25

ORCID

Haklae Kim
https://orcid.org/0000-0002-2616-421X

초 록

기록물은 과거와 현재를 포함하는 시간적 특성, 특정 언어에 제한되지 않는 언어적 특성, 기록물이 갖고 있는 다양한 유형을 복합적으로 갖고 있다. 기록물의 생성, 보존, 활용에 이르는 생애주기에서 텍스트, 영상, 음성으로 구성된 데이터의 처리는 많은 노력과 비용을 수반한다. 기계번역, 문서요약, 개체명 인식, 이미지 인식 등 자연어 처리 분야의 주요 기술은 전자기록과 아날로그 형태의 디지털화에 광범위하게 적용할 수 있다. 특히, 딥러닝 기술이 적용된 한국어 자연어 처리 분야는 다양한 형식의 기록물을 인식하고, 기록관리 메타데이터를 생성하는데 효과적이다. 본 논문은 한국어 자연어 처리를 기술을 소개하고, 기록 관리 분야에서 자연어 처리 기술을 적용하기 위한 고려사항을 논의한다. 기계번역, 광학문자인식과 같은 자연어 처리 기술이 기록물의 디지털 변환에 적용되는 과정은 파이썬 환경에서 구현한 사례로 소개한다. 한편, 자연어 처리 기술의 활용을 위해 기록관리 분야에서 자연어 처리 기술을 적용하기 위한 환경적 요소와 기록물의 디지털화 지침을 개선하기 위한 방안을 제안한다.

ABSTRACT

Records have temporal characteristics, including the past and present; linguistic characteristics not limited to a specific language; and various types categorized in a complex way. Processing records such as text, video, and audio in the life cycle of records' creation, preservation, and utilization entails exhaustive effort and cost. Primary natural language processing (NLP) technologies, such as machine translation, document summarization, named-entity recognition, and image recognition, can be widely applied to electronic records and analog digitization. In particular, Korean deep learning-based NLP technologies effectively recognize various record types and generate record management metadata. This paper provides an overview of Korean NLP technologies and discusses considerations for applying NLP technology in records management. The process of using NLP technologies, such as machine translation and optical character recognition for digital conversion of records, is introduced as an example implemented in the Python environment. In contrast, a plan to improve environmental factors and record digitization guidelines for applying NLP technology in the records management field is proposed for utilizing NLP technology.

Keywords: 기록관리, 자연어 처리, 인공지능, 머신러닝, 딥러닝
Records management, Natural language processing, Artificial intelligence, Machine learning, Deep learning

1. 서론

데이터, 인공지능, 클라우드 컴퓨팅으로 대표되는 새로운 디지털 시대는 인류의 사회문화적 변화를 이끌고 있다. 전 세계적으로 데이터 규모가 급속히 증가하고, 인공지능 기술이 비약적으로 발전함에 따라 다양한 영역에서 파괴적 혁신(disruptive innovation)이 촉진되고 있다. 예컨대, 최근 유행하는 메타버스(metaverse)는 가상과 현실이 상호작용하는 혼합현실인데, 인공지능, 가상현실(virtual reality), 증강현실(augmented reality), 블록체인(blockchain)과 같은 기술을 통해 서로 다른 산업 영역을 결합되거나 확장시키고 있다(Dwivedi et al., 2022). 인공지능은 인지능력, 학습능력, 추론능력, 이해능력 등과 같이 인간의 고차원적인 정보처리 능력을 인공적으로 구현한 것이며, 인간 사고와 지능적인 행위의 메커니즘을 이해하고 인공 기계에 구현하는 목표를 갖고 있다(McCarthy et al., 2006). 존 매카시(John McCarthy)는 인공지능을 '고도의 지능을 가진 컴퓨터 디바이스를 만드는 과학과 공학' 또는 '사람이 그렇게 행동했다면 지능적이라고 말할 수 있는 행동을 하는 기계'로 정의하고 있다(카플란, 2017). 인공지능 기술은 머신러닝, 음성인식, 이미지 인식, 로봇, 자연어 처리 등 다양한 분야로 세분화되고 있고, 응용 분야가 광범위하게 확산되고 있다. 특히 머신러닝(machine learning)과 딥러닝(deep learning)은 입력된 데이터로부터 규칙을 학습하고 결과를 예측하는 기술로 인공지능 분야의 진화를 이끌고 있다(Bengio, LeCun, & Hinton, 2021).

기록관리는 종이기록에서 전자기록으로, 보존위주에서 기록물의 지식정보화로 패러다임이 전환되고 있다. 이에 따라 기록물은 종전에 경험하지 못한 수준으로 양적 확대가 이루어지고, 기록물 관리를 위한 정보통신기술의 융합에 대한 요구가 확산되고 있다. 국가기록원은 차세대 전자기록관리체계 구축을 목표로 인공지능과 빅데이터의 기술을 적용하고 있다(강민균, 2021; 류한조, 2021; 오효정, 2019; 2021). 기록관리 분야에서 인공지능 기술과 관련된 선행연구는 지능형 기록관리 방안을 위한 인공지능 기술의 적용(김인택 외, 2017), 기록 텍스트 자동분류(김해찬솔 외, 2017), 지능형 전자기록서비스 모델(김태영 외, 2018)이 있다. 그러나 기록물의 유형과 매체가 다양하고 복잡해지고 있지만, 기록관리 분야에서 인공지능 기술을 적용 또는 활용하기 위한 연구가 미흡한 것이 현실이다.

인공지능 기술은 기록관리의 업무 전반에 광범위하게 적용될 수 있다. 예컨대, 기록 메타데이터의 생성과 활용에 있어 자연어 처리 기술은 현장의 문제를 효과적으로 개선할 수 있다. 기록 메타데이터는 시간에 관계없이 "영역(domain)의 안과 영역 간의 기록의 생산, 관리, 이용을 가능하게 하는, 구조화되거나 반구조화된 정보"로 정의하고 있다(국가기술표준원, 2021). 기록 메타데이터는 기록의 생산, 보존, 활용에 관련된 사람, 시스템, 지침, 정책을 식별하고 승인하는데 매우 중요하다. 기록관리에서 메타데이터는 소수의 도메인 전문가 또는 비전문가가 소프트웨어의 효과적 지원 없이 노동집약적으로 처리되는 것이 현실이다. 카네기멜론대학에서 개발한 'ReCAPTCHA'는 고전을 디지털화하는 과정에서 스캔한 글자를 사람이 직접 입력하도록 만든 것으로 기록의 차원을 넘어 웹에서 사람을 인식하는 도구로 응용되고 있다(Von Ahn et al., 2008). 집단지성을 활용하는 방법은 대규모 참여 환경을 전제하기 때문에 일반적인 기록관리 현장에서 적용하는데 한계가 있다. 반면, 인공지능 기술은 이미지나 동영상 기록물에서 특정한 인물, 사물의 인식, 다국어 텍스트의 번역, 영상 형식의 구술 기록물에서 채록하는 과정에 폭넓게 적용할 수 있다. 더불어 기록 메타데이터의 생산, 식별, 보존, 활용의 전 단계에 걸쳐 업무 담당자의 전문성을 보완할 수 있다.

본 논문은 인공지능의 대표적인 기술인 자연어 처리에 대해 소개하고, 기록 관리 분야에서 자연어 처리(NLP: Natural Language Processing) 기술을 적용할 수 있는 방안을 살펴본다. 자연어 처리 기술은 최근, 음성, 영상 등 구조화되지 않은 데이터에서 텍스트를 추출하고 처리하는데 활발하게 적용되고 있다. 본 논문은 기록관리 분야에서 실용적으로 적용할 수 있는 기술을 검토하고, 실제 사례에 적용하여 설명한다. 특히, 아날로그 형태의 기록물을 디지털화하는 단계에서 자연어 처리 기술의 적용과 고려사항을 논의한다. 본 논문의 구성은 다음과 같다. 2장은 자연어 처리 기술의 개념, 딥러닝 기술과 자연어 처리 기술의 특징, 한국어를 위한 자연어 처리 기술과 개발 동향을

 https://doi.org/10.14404/JKSARM.2022.22.4.129

소개한다. 3장은 기록관리 분야에 적용할 수 있는 기계번역, 개체명인식, 문서요약 기법을 한국어 자연어 처리 라이브러리를 적용하여 설명한다. 4장은 기록관리 분야에서 자연어 처리 기술을 적용하기 위한 고려사항을 논의하고, 5장은 논문의 결과와 향후 방향을 요약한다.

2. 자연어 처리 기술의 개요

2.1 개념

자연어 처리는 인간의 언어를 기계가 이해할 수 있도록 하는 인공 지능의 한 분야다. 자연어 처리는 기계가 인간의 음성과 텍스트를 이해하고 해석하고 조작할 수 있도록 처리하는 기술이며, 비정형 언어 데이터를 구조화된 데이터 형식으로 변환한다(Nadkarni, Ohno-Machado, & Chapman, 2011). 자연어 처리는 구문론, 의미론, 화용론, 형태론을 다루는 언어학, 텍스트와 음성을 처리하기 위한 컴퓨터 과학이 융합되어 있다. 자연어 처리는 토큰화, 어간추출, 표제어 추출 과 같은 전처리(pre-processing) 작업을 수행하고, 목적에 따라 문서 분류, 문서 요약과 같은 응용 영역에 적용할 수 있다(김학래, 2022). <표 1>은 자연어 처리와 관련된 개념을 요약하고 있다.

일반적으로 자연어 처리 기술은 인간의 언어를 기계가 읽고, 이해하도록 하는 기술인 자연어 이해(NLU: Natural Language Understanding)와 데이터를 인간이 읽을 수 있는 언어로 변환하는 기술인 자연어 생성(NLG: Natural Language Generation)으로 구분한다(Locke et al., 2021). 자연어 이해는 기계가 인간 언어의 의미를 인식하도록 하는 기술이다. 자연어 이해는 텍스트, 음성, 이미지 등 비정형 데이터를 구조화된 데이터로 변환하고, 맥락에 맞게 정확한 정보를 추출하여 언어의 정확한 의미를 이해할 수 있도록 지원한다. 자연어 이해는 규칙을 정의하여 처리하였지만, 워드 임베딩(word embedding), 언어 모델(language model) 등 딥러닝 기반의 다양한 알고리즘이 개발되어 인간의 개입이 최소화된 방향으로 기술이 발전되고 있다(Bengio, LeCun, & Hinton, 2021). 자연어

〈표 1〉 자연어 처리 기술의 주요 용어

용어	설명
문서 (Document)	하나의 텍스트 집합으로 문서의 집합을 말뭉치(corpus)라고 함 문서의 수준에 따라 말뭉치의 수준이 문장, 문단, 페이지, 댓글과 같이 다를 수 있음
어휘사전 (Lexicon)	어휘(lexical)의 집합 또는 어휘에 대한 정의를 갖고 있는 사전 인물사전, 영어사전과 같이 도메인의 특징에 따라 어휘사전이 존재함
불용어 (Stopword)	텍스트 분석에 있어 의미 없는 단어의 집합. 정보전달이 아닌 문법적·기능적 역할을 하는 단어 예: 은, 는, 이, 가, 이제, 여기, 저기 등
빈출어 (Common word)	특정한 문서에서 빈도가 매우 높은 단어로 의미 또는 중요도가 떨어지는 단어의 집합 예: 영화에 대한 리뷰 문서: 제목, 배우, 평점, 공감, 출연진
형태소 (Morpheme)	뜻을 갖고 있는 가장 작은 말의 단위 동사, 명사, 조사, 문장부호 등 품사(Part of Speech: POS) 단위를 의미함
단어주머니 (Bag of words)	문서에 함께 사용된 단어의 집합 중복된 단어는 하나로 취급하고, 순서의 의미를 고려하지 않음
토큰화 (Tokenisation)	의미가 있는 단위로 함께 존재하는 일련의 문자열(token) 문헌 단위의 문자열을 토큰 단위로 문자열을 분리하는 작업
어간추출 (stemming)	단어에서 어형이 변화된 접두사 (prefixes) 또는 접미사(affixes)를 제거하고 기본형으로 알려진 어간(stem)으로 줄이는 과정
표제어추출 (Lemmatisation)	표제어는 사전에 등재된 단어이며 단어의 기본형인 표제어(lemma)를 추출하는 작업
품사 태깅 (Part-of-Speech)	문장을 형태소 단위로 분리 한 후, 해당 형태소의 품사를 태깅하는 작업

https://doi.org/10.14404/JKSARM.2022.22.4.129

생성은 입력된 데이터를 기반으로 텍스트를 생성한다. 2020년에 등장한 GPT-3(Generative Pretrained Transformer-3)는 비교적 적은 학습 데이터를 기반으로 다양한 자연어 텍스트를 생성할 수 있다(Floridi & Chiriatti, 2020). <표 2>에서 보듯이, 자연어 처리 기술은 정보검색, 질의응답, 문서 요약, 문서 분류, 이미지 인식, 음성 인식, 기계번역, 지식그래프에 적용되고 있다(박상언, 2021; 박진호, 2019).

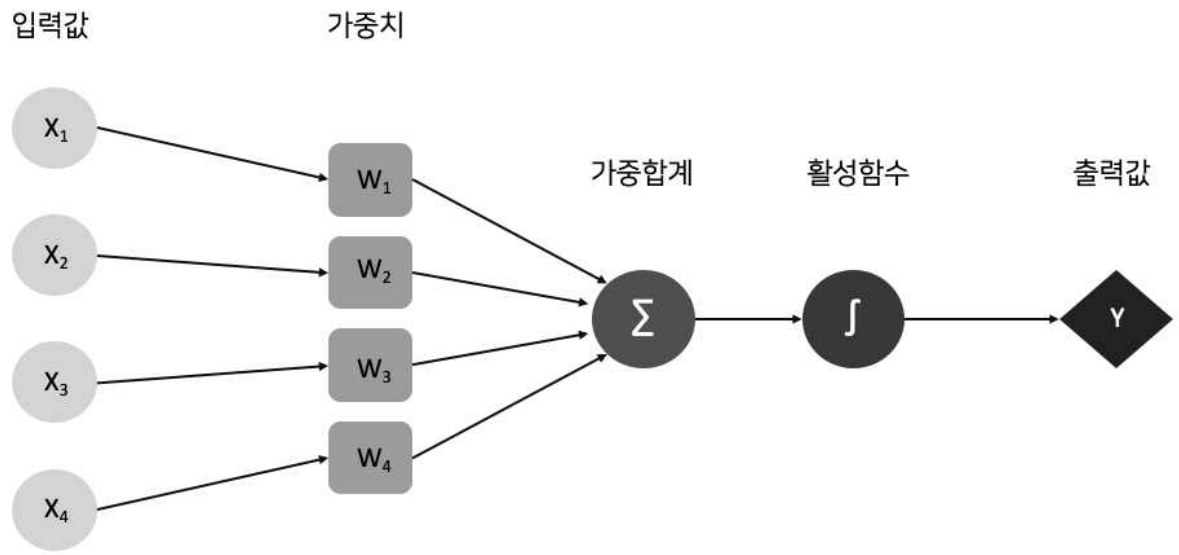
〈표 2〉 자연어 처리 기술의 응용 서비스

응용서비스 분야	설명
정보검색 (Information Retrieval)	웹 문서, 이미지, 동영상 데이터로부터 입력된 질의와 관련된 정보를 찾는 기술
질의응답 (Question Answering)	컴퓨터가 자연어 질문의 맥락을 파악하여 자동으로 응답을 제공하는 기술
문서 요약 (Document Summarisation)	주어진 문서에서 중요한 내용을 추출하거나, 새로운 문장을 생성해 요약문을 만드는 기술
문서 분류 (Document Classification)	텍스트의 내용을 기준으로 주어진 범주(Category)에 따라 나누어 할당하는 기술
문장 생성 (Document Generation)	입력 데이터 없이 기계가 자동으로 문장을 생성하는 기술 이미지를 인식해서 자동으로 캡션을 생성하거나 농담이나 이야기를 생성할 수 있음
이미지 인식 (Image Recognition)	이미지로부터 인물, 객체, 텍스트 등을 인식하는 기술 자연어 처리 분야에서는 디지털화한 문서에서 텍스트 데이터를 인식하는 데 주로 사용됨
음성 인식 (Voice Recognition)	사람이 말하는 음성 언어를 컴퓨터가 해석하여 텍스트 데이터로 전환하는 기술
기계 번역 (Machine Translation)	컴퓨터를 이용하여 특정 언어의 단어, 문장, 문서를 다른 언어로 번역하는 기술
지식그래프 (Knowledge Graph)	객체 사이의 관계를 의미적으로 표현한 그래프 형태의 데이터

2.2 인공신경망과 딥러닝 기술

인공신경망은 1956년에 등장한 퍼셉트론(perceptron)으로 시작되었다. 퍼셉트론은 가장 간단한 신경망으로 생물학적 뉴런(neuron)의 개념을 바탕으로 만들어졌으며 신호에 따라 입력과 출력을 전달하는 구조다(Rosenblatt, 1958). 퍼셉트론에서 입력값은 가중치를 곱해서 합산하고 결과값이 0보다 크면 1을 출력하고 나머지는 0을 출력한다. 활성 함수(activation function)는 퍼셉트론의 활성여부를 결정한다. <그림 1>의 퍼셉트론 구조는 현대의 신경망에서 그대로 활용되고 있는데, 최근 신경망은 은닉층(hidden layer) 구조가 복잡해지고 활성 함수가 다양해지고 있다. 딥러닝(deep learning)이란 용어는 2006년에 처음 등장한다(Bengio, LeCun, & Hinton, 2021). 딥러닝은 여러 층(layer)을 가진 인공신경망을 사용해 머신러닝을 수행하는 기술로, 기계가 자동으로 학습데이터에서 특징을 추출하여 모델을 학습한다(Wang & Raj, 2017). 딥러닝 기술은 인공 신경망의 이론을 계승하고 있지만, 컴퓨터 성능과 속도 문제로 한동안 일반화되는데 한계가 있었다. 2000년대 이후 GPU(Graphics processing unit)를 갖춘 고성능 컴퓨터의 등장으로 이런 문제가 해결되면서 딥러닝 기술에 대한 연구와 산업 현장의 적용이 가속화되고 있다. 딥러닝 기술은 사람의 개입 없이 대규모 언어 데이터를 처리하는데 효과적이며 텍스트, 오디오, 이미지, 비디오 등 비정형 데이터를 포함하는 컴퓨터 비전, 음성인식, 자연어 처리 분야에서 광범위하게 사용되고 있다.

https://doi.org/10.14404/JKSARM.2022.22.4.129



〈그림 1〉 퍼셉트론의 구조와 구성요소

2.3 자연어 처리와 딥러닝 모델

자연어 처리 모델은 언어의 문법적 규칙과 패턴을 사전에 정의하는 방식과 조건부 확률의 수학적 개념을 통계적으로 계산하는 방식을 중심으로 발전되어 왔다. 20세기 초, 마르코프(Markov)는 사슬(chain, 다른 의미로 시퀀스)의 마지막 요소를 사용하여 다음 요소를 예측하는 마르코프 체인(Markov Chains)을 소개한다(Ranjan et al., 2016). 마르코프 체인은 특정 상태에서 다른 상태로 이전할 때 과거의 상태에 의존하는 확률적인 특성을 따르는 것을 말한다. 클로이드 쉐논(Claude Shannon)은 마르코프 체인을 응용하여 확률 기반의 정보이론을 정립한다(Weikum, 2002). 1956년에 인공지능의 개념이 처음 사용된 이후, 언어 구조를 미리 정의한 규칙 목록을 실행하는 규칙시스템에 대한 연구가 활발했다. 그러나 다양한 형식의 자연어 처리를 위해 미리 규칙을 정의하는 것이 현실적으로 어렵고, 복잡한 규칙을 처리하는데 정확도가 낮은 한계가 있었다.

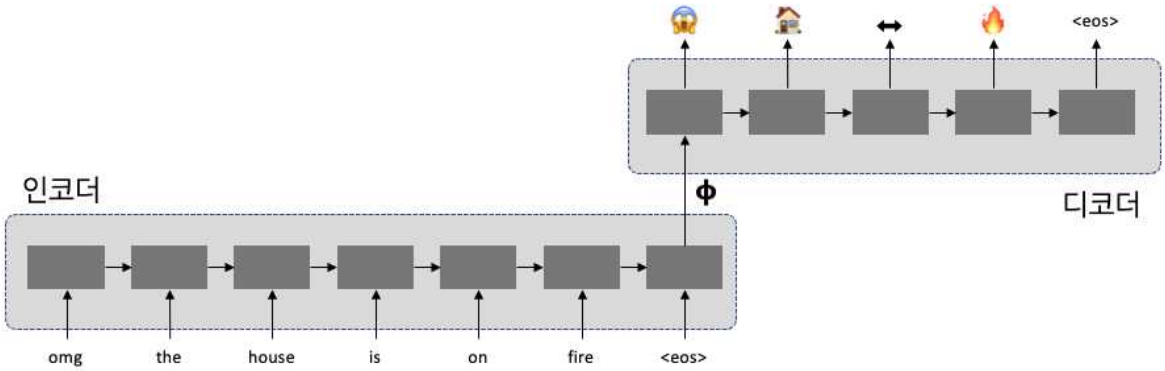
딥러닝 모델 기반의 자연어 처리는 2010년 중반부터 본격화되었지만, 인공지능의 발전과 함께 지속적으로 연구되었다. 존 홉필드(John Hopfield)는 홉필드 네트워크로 알려진 RNN(Recurrent Neural Networks) 을 1982년에 소개했는데(Sutskever, Martens, & Hinton, 2011), 이는 LSTM(Long Short-Term Memory)으로 발전했다(Gonzalez-Dominguez et al., 2014). 얀 르쿤(Yann LeCun)과 요슈아 벤지오(Yoshua Bengio)는 텍스트 시퀀스에 적용할 수 있는 CNN(Convolutional Neural Networks)을 1995년에 발표하였다(LeCun & Bengio, 1995). CNN은 합성곱(Convolution) 전처리 작업이 들어가는 신경망(Neural Network) 모델이며, 데이터에서 특징을 파악 후 추출하는 작업으로 정사각형 모양의 필터로 추출하여 위치마다 포개지는 숫자를 곱해서 모두 더하는 방법으로 계산한다.

신경망에서 자연어 처리를 하기 위해 데이터 구조가 변경되어야 한다. 신경망은 '오늘'과 '날씨'와 같은 단어 자체를 처리할 수 없기 때문에 단어를 고정 길이의 벡터(vector)로 변환해야 한다. 원핫 벡터(one-hot vector)는 벡터의 원소에서 단 하나의 원소만 1이고 나머지는 모두 0으로 표현된 벡터를 말한다(Camacho-Collados & & Pilehvar, 2021). 임베딩(embedding)은 단어와 문장과 같은 자연어를 벡터로 변환하여 벡터 공간(vector space)으로 표현하는 것을 말한다(Tang, et al., 2014). 임베딩은 차원이 큰 원핫 벡터를 연속값 형식의 축소된 벡터로 변환하고, 말뭉치의 의미 정보를 벡터로 표현하여 단어나 문장 수준의 관계성을 계산할 수 있다. 자연어 처리를

https://doi.org/10.14404/JKSARM.2022.22.4.129

위한 임베딩 기법은 단어와 문장 수준으로 구분한다. Word2Vec(Mikolov et al., 2013), GloVe(Global Vectors for Word Representation)(Pennington, Socher, & Manning, 2014), FastText(Bojanowski et al., 2017)와 같은 기법은 단어 수준의 임베딩 기법이고, BERT(Bidirectional Encoder Representations from Transformers)와 GPT는 문장 수준의 기법으로 분류된다. 문장 수준의 임베딩은 2018년에 소개된 ELMo(Embeddings from Language Models)(Peters et al., 2018)를 통해 본격화되었는데, 개별 단어가 아닌 문맥적 의미를 함축하기 때문에 단어 임베딩 기법에 비해 성능이 좋다. 예컨대, 먹는 '배'와 운송수단인 '배'는 동음이의어(homonym)로 단어 임베딩 기법은 의미를 식별하지 못하지만, 문장 수준 임베딩 기법은 두 단어의 의미를 구분할 수 있다.

딥러닝 모델은 매우 빠르게 발전하고 있어 해결하려는 문제에 따라 적합한 모델을 선택해야 한다. 인코더-디코더(encoder-decoder) 모델은 입력값을 요약하는 인코더와 출력 데이터를 생성하는 디코더로 구성된다(Badrinarayanan, Kendall, & Cipolla, 2017). 이때, 입력과 출력의 길이는 서로 다를 수 있다. 이 모델은 순차적으로 구성된 값을 다른 순차열로 변환하기 때문에 시퀀스-투-시퀀스(sequence-to-sequence)라고 한다. 시퀀스(sequence)란 연관된 연속의 데이터로, 문장의 단어 나열, 음성, 시계열 데이터를 표현할 수 있다. 이런 의미에서 시퀀스-투-시퀀스는 단어와 같은 특정한 속성을 갖고 있는 시퀀스를 다른 속성의 시퀀스로 변환하는 작업이다(Wolf et al., 2020). 예를 들어, <그림 2>는 영어로 입력된 문장을 이모지(emoji)로 변환하는 모델이다.



〈그림 2〉 시퀀스-투-시퀀스 모델: 6개의 입력값과 4개의 출력값 처리 결과 (재인용: https://tinyurl.com/3zprnn9d)

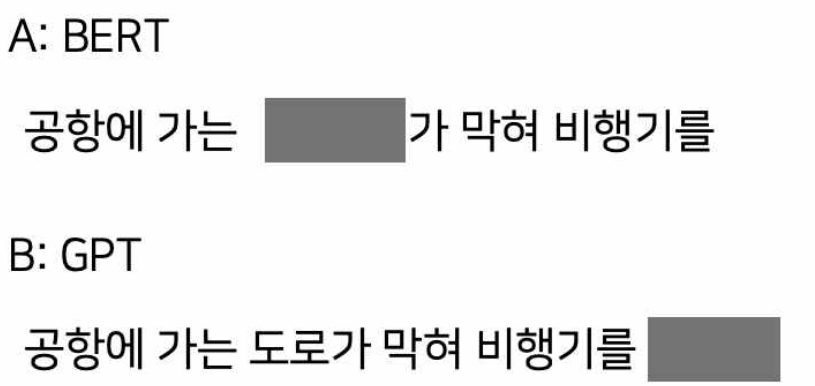
어텐션(attention)은 신경과학 등 여러 분야에서 사용되는 개념이다. 예컨대, 사람이 책을 읽을 때, 모든 문장을 한 번에 기억하지 않고, 개별 문장과 단어에 집중하고 이후 전반적인 내용을 이해한다. 이 때, 맥락에 맞게 개별적인 단어와 문장에 집중하는 것을 어텐션으로 이해할 수 있다. 신경망에서 어텐션은 문장의 특정 단어를 분석하기 위해 문장 내의 다른 단어의 연관성을 고려한다. 어텐션은 시퀀스의 마지막 요소가 아닌 다른 토큰을 연산하는 방법으로 모든 단어의 관계를 단어 대 단어(word-to-word)로 연산한다(Vaswani et al., 2017). 트랜스포머(transformer)는 순환 신경망을 사용하지 않고 어텐션 메커니즘으로 구현된 인코더-디코더 모델이다(Wolf et al., 2020). 순환신경망은 단어를 순차적으로 처리하며 문장 패턴을 인식하지만, 트랜스포머는 디코더에서 출력값을 예측하는 매 시점마다 입력값을 다시 참고하여 단어 사이의 상관관계를 해석한다. 즉, 트랜스포머는 시퀀스-투-시퀀스(sequence-to-sequence)를 수행하는 모델이며, 디코더에서 결과 값을 예측할 때 인코더의 입력 값을 다시 참고하는 어텐션(attention) 기법을 적용한다.

자연어 처리 분야에서 딥러닝 모델의 적용을 위해 전이 학습(transfer learning)의 개념은 매우 중요하다. 전이 학습은 특정 문제를 해결하기 위해 학습한 모델을 다른 딥러닝 모델의 입력값으로 사용하는 기법이다(Weiss, Khoshgoftaar, & Wang, 2016). 전이 학습은 사전 학습(pre-training)과 파인 튜닝(fine-tuning)으로 구성된다. 사전 학습은 임의값으로 초기화된 모델의 가중치를 다른 태스크에 학습시킨 가중치로 초기화하는 방법이다. 예컨대, 감성 분석(sentiment analysis)을 위해 학습한 모델의 가중치를 텍스트 분류 모델의 가중치로 활용할 수 있다. 이 때,

사전학습 모델을 업스트림 태스크(upstream task), 사전 학습한 가중치를 활용하여 해결하는 대상은 다운스트림 태스크(downstream task)라고 한다. 업스트림 태스크는 다운스트림 태스크에 앞서 해결되어야 하는 과제로 감성 분석이 대상이고, 다운스트림 태스크는 텍스트 분류가 해당된다. 파인 튜닝은 사전 학습한 가중치와 다운스트림 태스크를 위한 최소한의 가중치를 모델에 추가하고 미세하게 조정하는 방법이다 (김학래, 2022).

2.4 자연어 처리를 위한 언어모델

언어 모델(Language Model)은 단어 또는 문장에 확률을 할당하여 기계가 처리할 수 있도록 만든 모델이며, 문장의 특정한 위치에 적합한 단어가 출현하는 것을 예측한다(Radford et al., 2018). 언어 모델은 통계 기반 방법과 인공 신경망을 이용한 방법으로 구분하며, 최근 인공 신경망을 기반의 BERT, GPT가 활발하게 적용되고 있다 (Singh, Rutten, & Lefever, 2021). BERT는 모든 출력 요소가 모든 입력 요소에 연결되고 이들 사이의 가중치가 연결에 따라 동적으로 계산되는 트랜스포머에 기반하고 있다(Rogers, Kovaleva, & Rumshisky, 2020). BERT는 주변 텍스트를 사용하여 컨텍스트를 설정하고 컴퓨터가 텍스트에서 모호한 언어의 의미를 처리할 수 있다. 대부분의 언어 모델은 왼쪽에서 오른쪽으로 또는 오른쪽에서 왼쪽으로 순차적으로 텍스트 입력을 읽을 수 있지만, BERT는 동시에 양방향으로 읽고 처리하는 장점이 있다(Ethayarajh, 2019). GPT는 OpenAI에서 개발한 언어모델로, 트랜스포머의 디코더 부분을 사용하여 이전 단어를 기반으로 다음 단어를 단방향으로 계산한다(Ding & Chen, 2020). 2018년에 발표한 GPT-1은 트랜스포머의 디코더 부분을 사용한 구조로, 원시 텍스트의 장기 의존성(Long-term dependencies)에 높은 성능을 보여준다. 2020년 OpenAI는 초대규모 매개변수를 가진 GPT-3를 공개했고, 최신 언어모델 중 가장 주목받는 언어모델로 평가되고 있다. <그림 3>은 BERT와 GPT의 동작 방식을 보여준다. BERT는 문장의 중간에 있는 단어를 앞뒤 단어의 시퀀스를 통해 예측할 수 있다. 반면, 단어 시퀀스가 있는 문장에서 GPT는 문장의 마지막에 있는 빈칸에 들어갈 단어의 확률을 계산한다. <그림 3-A>의 빈칸은 문맥으로 볼 때 '도로', <그림 3-B>의 빈칸은 '놓쳤다'라고 예측할 수 있다.



〈그림 3〉 BERT와 GPT의 작동 방식

2.5 한국어를 위한 자연어 처리 기술

한국어는 어근과 접사에 의해 단어의 기능이 결정되는 교착어로 문맥 정보의 이해를 위한 자연어 처리의 난이도가 높다. 한국어 처리를 위해 KoNLPy, Soynlp, Khaiii 등 한국어 형태소 분석기와 KrBERT, KoGPT-2, HyperCLOVA와 같은 대규모 언어모델을 적용한 도구가 개발되고 있다(Lee et al., 2020). 한국어 형태소 분석기는 한국어의

https://doi.org/10.14404/JKSARM.2022.22.4.129

언어적 특징을 반영하여 품사 태그를 분석하는 기능을 제공한다(최용석, 이공주, 2020). KoNLPy는 한국어 자연어 처리를 위한 형태소 분석 패키지이며, Hannanum(한나눔), Kkma(꼬꼬마), Komoran(코모란), Mecab(메캅), Open Korea Text 형태소 분석기를 통합하여 제공한다(박은정, 조성준, 2014). Soynlp는 학습데이터를 이용하지 않고 텍스트에 존재하는 단어를 찾거나 단어열로 분해할 수 있는 비지도 학습 방식을 사용하고 있다. 카카오에서 개발한 Khaiii는 딥러닝 기반의 형태소 분석기로 신경망 알고리즘 중 CNN을 사용하였고, 약 85만 문장, 1000만 코퍼스를 포함하고 있다.

한국어 처리를 위한 언어 모델은 IT 업계를 중심으로 활발하게 진행되고 있고, 일부 프로젝트는 오픈소스 소프트웨어 형식으로 제공되고 있다. SKTBrain에서 개발한 KoBERT는 BERT의 한국어 성능 한계를 극복하기 위한 모델이다(김인후, 김성희, 2022). 한국전자통신연구원(ETRI)이 개발한 KorBERT는 신문기사와 백과사전 등 23GB의 대용량 텍스트를 대상으로 약 47억 개의 형태소를 사용하여 학습한 2가지 모델을 공개하고 있다. 네이버에서 개발한 HyperCLOVA는 한국어 기반의 대규모 언어모델이며, 네이버에서 서비스하는 블로그, 카페, 뉴스 기사와 댓글, 지식IN 데이터와 국립국어원 모두의 말뭉치, 위키백과 데이터를 전처리한 대규모 한국어 데이터를 포함하고 있다. 한국어의 자연어 처리는 임베딩(강형석, 양장훈, 2019), 지식그래프(김상운, 신원철, 2021), 질의응답(김영민 외, 2020), 챗봇(유소엽, 정옥란, 2019), 분류 모델 (황상흠, 김도현, 2020; 조희련 외, 2021) 등 다양한 기법이 적용되고 있고, 일반 상식에서 전문적인 도메인으로 확장되고 있다.

3. 한국어 자연어 처리 기술의 활용

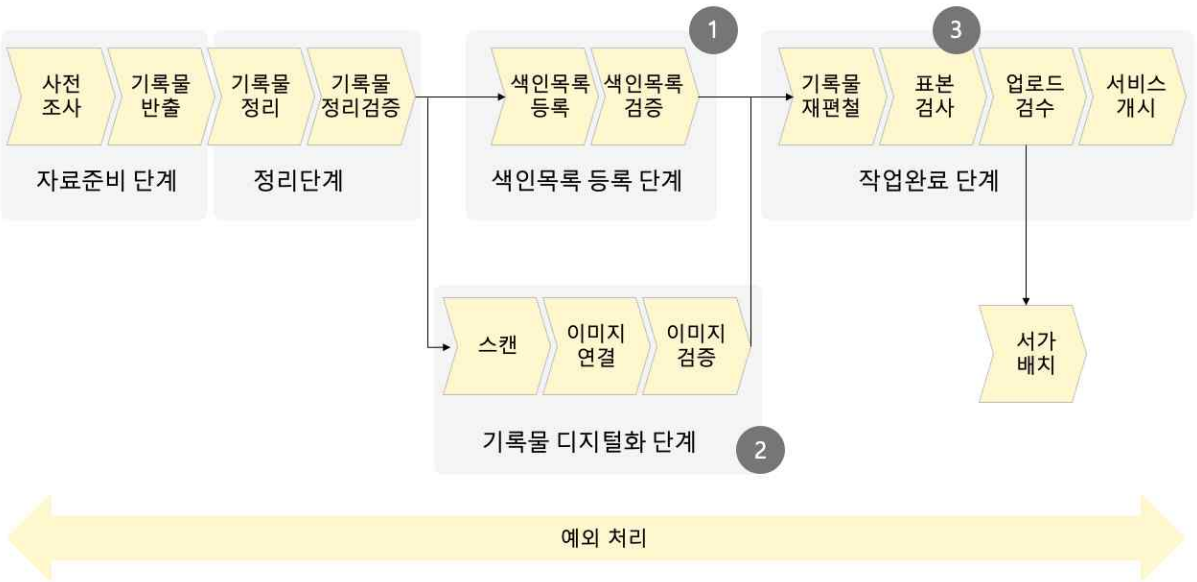
3.1 기록물 디지털화를 위한 자연어 처리 기술

공공기록물법 시행령(행정안전부, 2022)에 따르면, 전자기록물은 “정보처리능력을 가진 장치에 의하여 전자적인 형태로 작성하여 송신·수신 또는 저장되는 전자문서, 웹기록물 및 행정정보 데이터세트 등의 기록정보자료”로 정의하고 있다. 전자기록물은 정보시스템에서 전자형태로 생산된 기록(born digital)과 아날로그 형태로 생산된 기록이 디지털로 변환된 기록을 모두 포함한다. 비전자기록물은 스캐닝 또는 인코딩 장비를 통해 디지털 형태로 변화하는데 이 과정을 디지털화(digitalisation)라고 한다(국가기록원, 2018). 모든 기록은 내용과 메타데이터로 구성되는데, 메타데이터는 기록의 고유한 내용, 구조, 맥락과 기록물 관리사항을 기술하기 위한 요소를 포함한다. ISO 15489-1에 따르면, 기록 메타데이터는 기록의 내용에 대한 기술, 기록의 구조(예: 형태, 포맷 등), 기록이 생산, 접수, 이용되었던 업무맥락, 다른 기록과 다른 메타데이터와의 관계, 기록을 검색하고 재현하는데 필요한 식별자나 기타 정보, 기록과 관련되었던 업무행위와 사건(행위의 일자와 시간, 행위를 수행한 행위자와 메타데이터에 대한 변경사항을 포함)을 포함한다(국가기술표준원,2021). 기록 메타데이터는 그 자체가 기록으로 관리되어야 하며, 기록의 생산과 활용에 이르는 전 주기의 정보를 연계하는데 핵심적인 역할을 한다.

자연어 처리 기술은 다양한 전자기록물에서 메타데이터를 추출하고 서로 다른 기록물 관리 단계에 일관성 있는 데이터를 제공할 수 있다. 디지털화 과정에서 자연어 처리 기술은 스캐닝 또는 인코딩한 기록물과 관련된 메타데이터를 자동으로 추출하고, 수작업으로 처리하는 등록 작업을 신속하고 정확하게 개선할 수 있다. <그림 4>에서 보듯이, 기록물 목록 작성과 디지털화 작업은 자료준비 단계, 정리단계, 색인목록 등록 단계, 기록물 디지털화 단계, 작업 완료단계로 구성된다(국가기록원, 2017). 각 단계의 업무는 지침을 통해 구체적으로 기술하고 있지만, 대부분의 작업은 업무 담당자가 해당 업무를 숙지해서 처리해야 한다: ① 색인목록 등록은 기록물 정리가 완료된 기록물에서 철/건에 대한 색인항목을 발췌해서 시스템에 입력해야 한다. 등록된 색인 정보는 기록물철 레이블번호, 처리과와 기록물 철 제목으로 조회할 수 있다. 색인항목의 등록은 디지털화가 완료되고 기록관리시스템에서 재사용되는 정보이기 때문에 정확한 정보를 등록하는 것이 중요하다. 이 단계에서 통제어휘 기반의 자동완성 기능은 입력 정보의

 https://doi.org/10.14404/JKSARM.2022.22.4.129

오류를 감소시켜 색인목록의 등록에 대한 전수검사를 효율적으로 수행할 수 있다. ②디지털화 단계는 아날로그 형식의 기록을 디지털 정보로 변환하고, 이 과정에서 생산되는 다양한 메타데이터를 기록한다. 디지털화 단계에서 변환되는 기록물의 물리적 특성은 자동으로 메타데이터로 입력될 수 있다. 그러나, 아날로그 기록물에 대한 설명, 배경, 관련 정보는 일반적으로 충실하게 등록되지 않는 경향이 있다. 자동요약, 이미지 또는 비디오의 정보에서 자동캡션 생성과 같은 자연어 처리 기술은 기록 메타데이터를 충실하게 표현하는데 활용될 수 있다. ③ 작업완료 단계에서 표본검사는 색인목록과 디지털 기록물을 검토하는 과정으로 전체 구축 물량에서 최소 3% 이상을 권장하고 있다(국가기록원, 2018). 자연어 처리 기술은 색인목록과 디지털화 단계의 메타데이터를 자동으로 비교하고 검증하는데 적용할 수 있고, 검사 환경을 자동화해서 디지털화된 모든 기록물의 검증을 자동화할 수 있다.



〈그림 4〉기록물 색인목록과 기록물 디지털화 절차에서 NLP 기술의 적용

3.2 자연어 처리 기술의 응용

기록 관리 분야에서 활용할 수 있는 자연어 처리 기술은 검색, 문서요약, 분류, 이미지 인식 등 매우 다양하다. 한국어 처리를 위한 환경은 자바, 파이썬, R 등 대부분의 프로그래밍 언어에서 지원한다. 기록관리에서 자연어 처리 기술의 적용 방안을 소개하기 위해 구글 코랩(Google Colab) 환경을 구성하고 대표적인 응용 사례를 적용한다. 본 논문에서 활용한 모든 코드, 텍스트, 이미지는 깃허브(https://github.com/haklaekim/rm-nlp)에 공개되어 있다. 한국어 처리를 위한 라이브러리는 Pororo(Platform Of neuRal mOdels for natuRal language prOcessing)와 Spark NLP를 적용한다. Pororo는 한국어, 영어, 중국어 등 다국어를 처리하기 위한 30개 이상의 자연어 처리 모델이 구현되어 있고, Spark NLP는 아파치 스파크(Apace Spark) 기반의 텍스트 처리를 지원하고 있다. 두 라이브러리는 딥러닝 모델을 적용하여 자연어 처리가 가능하고, 비교적 적은 코드로 기능을 테스트할 수 있는 장점이 있다. 한국어 처리를 위한 텍스트와 이미지는 구글에서 '기록', '아카이브', '디지털 아카이브', '구술채록'으로 검색하고, 임의적으로 결과를 선택했다. 한국어 처리를 위해 사용한 텍스트는 <표 3>에 요약되어 있다.

https://doi.org/10.14404/JKSARM.2022.22.4.129

〈표 3〉 자연어 처리에 사용된 텍스트 예제

번호	기관	유형	내용
1	서울기록원 (https://archives.seoul.go.kr/aggregation/45)	텍스트	서울대 정진성연구팀은 서울시 여성가족정책실의 '일본군 위안부 기록물 관리사업(2016~2019 계속)'을 수행하면서 미국 국립문서기록관리청(NARA, National Archives and Record Administration)과 영국 국립문서보관소(TNA, The National Archives of UK), 태국 현지를 방문하여 연합군 자료를 중심으로 일본군 '위안부' 관련 자료를 발굴?조사하였고, 그 결과 연합군이 생산한 문서, 사진, 영상등 가치 있는 자료를 다수 수집함. 이 중 해제작업이 완료된 138건의 수집기록(영상 2건, 사진 32건, 문서 104건)을 2018년 서울기록원에 기증함
2	한국예술디지털아카이브 (https://www.daarts.or.kr/handle/11080/77382)	텍스트	2015 여름, 다시 찾아온 히즈피아노 두 대의 피아노가 선사하는 뮤지컬의 밤. 두 대의 피아노가 선사하는 뮤지컬의 밤. 다채로운 음색과 폭넓은 표현력은 물론, 오케스트라 악기 전체가 연주할 수 있는 음역을 모두 소화할 수 있기에'작은 오케스트라'라는 별명을 가지고 있는 단 하나의 악기, 피아노. 지난해 10월, 가을밤을 촉촉이 적셨던 투 피아노 뮤지컬 콘서트 '히즈피아노 온 브로드웨이'가 최고의 피아니스트들과 함께 한층 더 업그레이드 된 모습으로 올여름 다시 찾아옵니다.

한국어 처리를 위한 코드는 Pororo에서 제공하는 기본 기능으로 구현하고, 전처리나 코드를 통한 성능을 개선하기 위한 옵션은 별도로 조정하지 않는다. 기능별로 사용한 코드는 <표-4>와 같다.

〈표 4〉 자연어 처리를 위해 사용한 주요 파이썬 코드

기능	코드
기계번역	mt = Pororo(task="translation", lang="multi") mt(text, src="ko", tgt="en")
문서요약	summarizer = Pororo(task="text_summarization", lang="ko", model="abstractive") summarizer(text, beam=5, len_penalty=0.6, no_repeat_ngram_size=3, top_k=50, top_p=0.7)
NER	ner = Pororo(task="ner", lang="ko") ner(text)
OCR	ocr = Pororo(task='ocr') ocr('arko.png')

3.2.1 기계 번역

기계번역은 특정 언어를 다른 언어로 자동으로 번역하는 기술이며, 자동번역과 번역 지원 두 가지 유형으로 구분된다. 자동번역은 인간의 개입이 최소화된 기계적인 번역기로 구현되고, 번역 지원은 전문 번역가의 번역 작업을 보조하는 전문 소프트웨어가 대표적이다. <표 5>는 서울기록원과 한국예술디지털아카이브에서 추출한 텍스트를 기계 번역으로 처리한 결과다. 구글 번역, 파파고와 같이 기계 번역은 일반 상식에 대한 번역 성능을 갖추고 있다. 그러나, 특정한 맥락을 갖고 있는 기록물, 다국어 형식의 기록물은 한국어로 번역이 쉽지 않기 때문에 기계 번역을 통해 한국어로 변환하는 것이 효과적일 수 있다.

https://doi.org/10.14404/JKSARM.2022.22.4.129

〈표 5〉 주어진 텍스트의 기계번역 결과

번호	결과
1	The Seoul National University\'s Chung Jin-sung research team conducted the ""Japanese Military comfort women\'s Records Management Project(2016~2019 Continued)"" of the Seoul Women\'s Family Policy Office and conducted research on the U.S. National Architecture Agency(NARA), the National Archives of the United Kingdom, and Thailand, focusing on the data of the Allied Army, and found data related to the Japanese military\'s ""comfort women"" at the center.
2	In the summer of 2015, the night of the musical presented by the two pianos, presented by two different Hezpianos. Just one instrument with the nickname ""Small Orchestra,"" as well as colorful sound and wide expression, as well as all the music that the orchestra instruments can play. In October last year, the two piano musical concerts ""Hizpiano On Broadway,"" which had a moisturizing autumn night, will come up with the best pianos.

3.2.2 문서 요약

문서 요약은 원본 문서의 핵심적인 내용을 자동으로 요약하여 제공하는 기술이다. 문서 요약 기술은 텍스트에서 핵심 문장과 단어구를 추출해 요약문을 완성하는 추출 요약(extractive summarisation)과 텍스트의 관계를 계산해 문맥을 파악하고 새로운 문장을 직접 생성하고 요약하는 생성 요약(abstractive summarisation) 방식으로 구분하고 있다. <표 6>은 주어진 문장을 요약한 결과를 보여준다. 1번 결과는 비교적 문장을 압축하여 요약하고 있지만, 2번 항목은 요약된 결과가 주어진 문장과 의미가 다르고, 요약이 제대로 수행되지 않은 것을 볼 수 있다.

〈표 6〉 기록물 문서의 일부를 자동요약한 결과

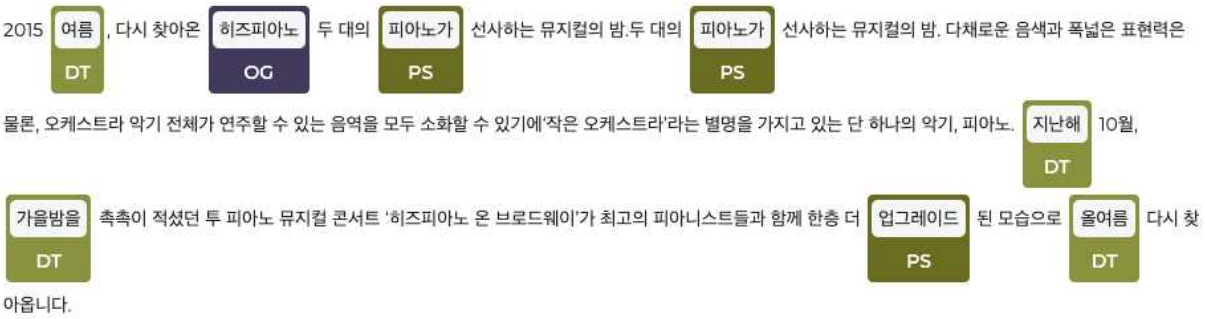
번호	결과
1	용인서울대 정진성연구팀은 2018년 서울기록원에 138건의 수집기록을 기증하면서 '일본군 위안부 기록물 관리사업' 수행 경험이 있다.
2	맛있 음색과 폭넓은 표현력은 물론, 오케스트라 악기 전체가 연주할 수 있는 음역을 모두 소화할 수 있고작은 오케스트라는 별명을 가지고 있는 단 하나의 악기, 히즈피아노 두 대의 피아노가 찾아온 뮤지컬의 밤이었다.('('휜 브로드웨이'), 스키 스키 야회 등 작년 10월, 겨울밤을 촉촉이 적셨던 투 피아노 뮤지컬 콘서트 '히즈피아 노 온 브로드웨웨이'가 최고의 피아니스트들과 함께 한층 더 업그레이드 된 모습으로 올여름 다시 찾아옵다보이는 일본가야 피아니스트와 함께 한층 성과 성과 남긴('('한겨레'), 별밤밤 우주(('('), 스키 야스카스카야'), 유대 유대감이 끝난 후에는('('('다'다', 로켓 로켓웨이웨이웨이)' 등이 다시 찾아왔다.

3.2.3 개체명 인식

개체명 인식은 이름을 가진 개체를 인식하는 것으로 미리 정의된 인명(person), 지명(location), 기관명(organization), 시간(time) 등에 해당하는 개체명을 인식하여 추출·분류하는 기법이다(민진우 외, 2021). 일반적으로 개체명은 인물이나 장소와 같은 일반적인 개체명(generic NE), 전문 분야의 용어에 해당되는 특정 분야의 개체명(domain-specific NE)으로 구분한다. <그림 5>의 개체명은 주로 기관(OG), 장소(LC)에 관련된 개체가 추출되었다. 그러나, '영국'과 'UK'가 의미적으로 동일하지만, 추출한 개체명은 'LC'와 'OG'로 서로 다른 개체명으로 인식되었다. 반면, '서울시', '여성가족정책실', '연합군'과 같은 개체는 인식되지 않았다. <그림 6>은 음악과 악기와 관련된 공연정보를 포함하고 있지만, 대부분의 개체들이 인식되지 않는다.

https://doi.org/10.14404/JKSARM.2022.22.4.129

〈그림 5〉 개체명 인식 결과 - 텍스트 항목 1번(OG: Organization, LC: Location)

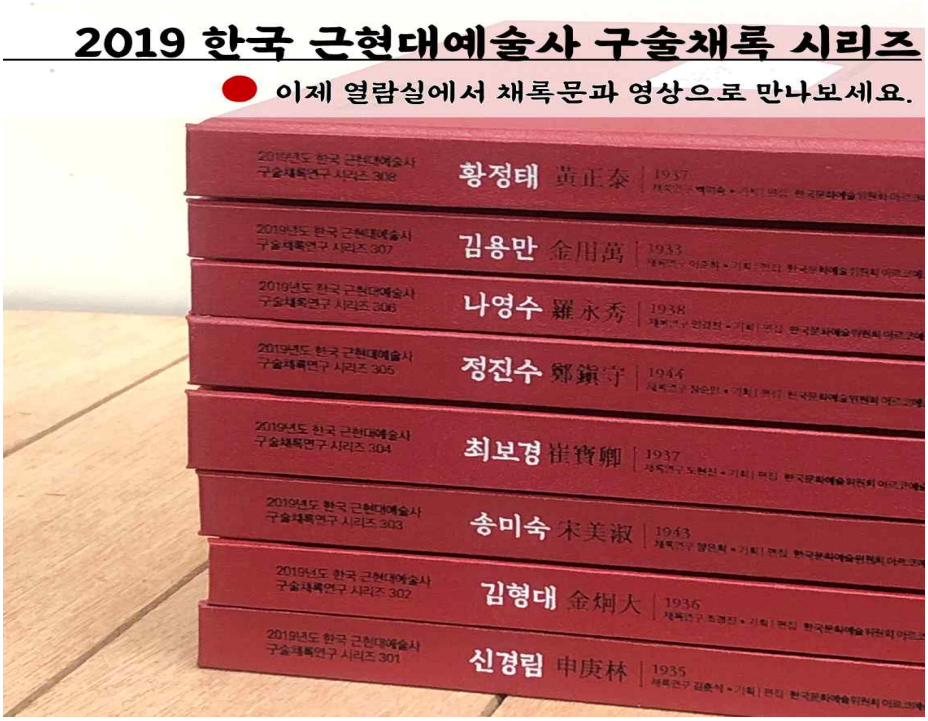


〈그림 6〉 개체명 인식 결과 - 텍스트 항목 2번(DT: Determiner, OG: Organization)

3.2.4 광학문자인식

광학문자인식(OCR: Optical character recognition)은 이미지 데이터로부터 인물, 객체, 텍스트를 인식하는 기술이며, 이미지 내의 특정 사물을 분류, 추출, 식별할 수 있다. <표 7>은 <그림 7>의 이미지를 OCR로 추출한 텍스트를 포함하고 있다(https://artsarchive.arko.or.kr/upload2/userfiles/images/2021/06/07/204819669640831330.png). 인물의 한글명은 비교적 정확하게 인식되었으나, 한자로 표현된 이름은 대부분 인식되지 않았다. '한국 근현대예술사 구술채록연구 시리즈'는 '근현대', '채록'에 대한 단어는 부정확한 결과를 포함하고 있다. OCR 결과를 보면, 단어가 짧은 한글은 비교적 인식율이 높지만, 한자와 짧은 숫자는 대부분 인식하지 못하고 있다. OCR은 이미지의 해상도, 글씨체, 폰트 등 인식에 영향을 미치는 요소가 많기 때문에, 실제 활용하는 단계에서 이에 대한 검토가 필요하다.

https://doi.org/10.14404/JKSARM.2022.22.4.129



〈그림 7〉 이미지 인식을 위한 이미지: 한국 근현대 예술사 구술채록사업 2019년도 채록문 알림
(이미지 출처:
https://artsarchive.arko.or.kr/upload2/userfiles/images/2021/06/07/204819669640831330.png)

〈표 7〉 OCR 처리 결과

번호	결과
1	['2019 한국 근현대예술사 구술채록 시리즈 이제 열람실에서 채록문과 영상으로 만나보세요.', '209년도 한국 근헌대예술사 구승체육연구 시리즈 308', '황정태국포소 1 j :', '~*', '2019년도 한국 근현대예술사 구&대록연구 시리즈 307', '김용만 中오 :() 3 3 이준히', '한국된 : 이예술 :트씨 아곳.', '2019년도 한국 근훤대예술사 구술체록연구 시리즈 306', '나영수 돼 가주 1938 가*." , 헌경철', '판 국문 화예술위현씨 마근', '2013년도 한국 근현대예술사 구술체축연구 시리즈 305', '정진수 헷: 삶\'j\' 1044 질(^: N소?! * ": : 국무 싸예효 위펠써 어르', '2119년도 한국 근헌대예술사 구술채록민구 시리즈 304', "최보경남뉴럼 137 경속*- 노면인-* *, 한국무파하술5씨비 아고코 2019년도 한국 근현대예술사 구술차록연구 시리즈 303 중미주**| 193 가록*, 양문& -)(괜심 '극*크 2019년도 한국 근현대예술사 구술챠록연구 시리즈 302 김형대 中치〈 1936 '록* 구 조겹진 ' '일	편집 2019년도 한국문 화예술 위원피 아르 구술체복연구 한국 근천대예술사 시리즈' 301 신경립 타1차 #* 1935 채록연구 * 김출식 ~ 기획	편진 한국문바예술위원회 영료코"]

4. 기록관리를 위한 한국어 자연어 처리 기술의 고려사항

자연어 처리 기술은 비디오, 오디오, 문서와 같은 다양한 형식의 기록물을 디지털화하고 탐색하는데 효과적이다. 인공지능에 대한 사회적 관심이 광범위하게 확산되는 상황에서 기록관리에 적용할 수 있는 방안을 검토하는 것은 중요한 의미가 있다. 기록관리 분야는 사회·문화·역사·환경 등 주제 분야가 매우 광범위하고 다양한 매체를 포함하는 특수성이 있다. 따라서, "딥러닝 기반" 또는 "인공지능 기반"과 같은 기술 중심의 과제 추진을 지양하고, 기록의 생성과 활용에 이르는 생애주기 관점에서 효율적인 방안을 모색해야 한다.

https://doi.org/10.14404/JKSARM.2022.22.4.129

4.1 공동 활용을 위한 기반 확보

자연어 처리 기술은 데이터, 소프트웨어, 하드웨어의 균형 있는 조합을 통해 구현된다. 3장에서 도출한 결과는 학습 과정이 제외되어 자연어 처리 성능을 객관적으로 평가하는데 제약이 있다. 딥러닝 모델의 성능은 훈련 세트를 학습시키고 테스트 세트에 적용한 결과를 의미하며, 학습을 위한 적합한 양질의 데이터 확보는 자연어 처리의 성능에 매우 큰 영향을 미친다. 과학기술정보통신부에서 운영하는 AI 허브(AI Hub)는 인공지능 분석에 활용할 수 있는 한국어, 영상이미지, 헬스케어, 재난안전환경, 교통물류 분야의 데이터세트를 제공하고 있다(https://aihub.or.kr). 일부 데이터세트는 기록관리 분야의 자연어 처리를 위해 활용할 수 있다. 예컨대, 한국어 글씨체 데이터세트는 OCR 분석에 적용할 수 있고, 한국어 학습용 어휘 목록 데이터세트는 한국어 처리를 위해 범용적으로 사용할 수 있다. 그러나, 기록관리 분야의 데이터는 현재 미흡한 수준이기 때문에, 기록 분야에 특화된 학습용 데이터를 지속적으로 발굴하는 것이 필요하다.

소프트웨어와 하드웨어는 개방형 환경을 적극적으로 검토해 볼 필요가 있다. 자연어 처리 기술은 방대한 규모의 입력과 출력 데이터를 사용하기 때문에, 고성능의 컴퓨팅 인프라가 필요하다. 국가기록원은 약 4천여권의 타자기록물을 인식하는 OCR 과제를 수행하고, 디지털 컬렉션에서 OCR로 인식한 결과를 함께 제공하고 있다. 연구보고서에 따르면, 기록물 한 면의 처리시간은 2.6초, 텍스트의 탐지율과 인식율은 각각 98.9%, 95.3%로 비교적 높은 수준이다(강지홍, 2021). 그러나, 예산과 인력의 제약으로 대부분의 기록관리 기관은 유사한 과제를 수행하기 어렵다. 따라서, 기록관리 분야에서 공동으로 활용할 수 있는 소프트웨어를 개발하고, 클라우드 컴퓨팅 기반의 하드웨어를 적용 방안을 검토할 필요가 있다. 소프트웨어는 오픈소스 형식으로 개발하고 관련 기관이 함께 사용할 수 있도록 공유하는 방식을 검토해야 한다. 오픈소스 소프트웨어는 기록 관리 분야의 특징을 반영하여 지속적으로 업데이트시킬 수 있고, 관련 연구자와 커뮤니티와 협력할 수 있는 기회가 될 수 있다. 예컨대, FinBERT(Yang, Uy, Huang, 2020), Clinical BERT(Alsentzer et al., 2019), BioBRT(Lee et al., 2020), SciBERT(Beltagy, Lo, & Cohan, 2019), TourBERT(Arefieva & Egger, 2022)와 같이 주제 영역의 특성에 맞는 딥러닝 모델을 개발하고 공동으로 활용할 수 있다. 한편, 개별 기관이 딥러닝 기술을 적용하기 위해 하드웨어를 구매하지 않고 클라우드 컴퓨팅을 활용하는 방안을 고려할 수 있다.

4.2 기록물 디지털화의 개선

전자 기록은 전자 형태로 생산된 기록(born digital)과 아날로그 형태를 디지털화시킨(digitalised) 기록을 포함한다. 미국·영국·호주 등 각국의 국가기록원은 전자기록관리를 위한 전략을 수립하고 적극적으로 추진하고 있다. 디지털화는 전자기록관리의 핵심이며, 인공지능을 포함한 다양한 정보통신기술이 적용될 수 있다. 다만, 단편적인 기술의 적용이 아닌 기록관리 전반에 걸친 변화가 함께 수반될 필요가 있다. 예컨대, 기록물 목록 작성 및 디지털화 작업 지침은 목록 작성과 디지털화 단계를 세분화하고, 기록물관리자, 품질관리자, 작업장관리자의 역할을 명확하게 구분하고 있다(국가기록원, 2013; 2017; 2018). 그러나 현재 기록물의 디지털화는 맥락의 유기적 연결이 아닌 개별 단계의 완료에 초점이 맞춰질 수 있다. 아날로그 기록물은 색인 목록 등록단계, 디지털화 단계를 진행하고 기록물 메타데이터를 작성하게 된다. 각 단계에서 색인, 디지털 파일과 같은 정보가 생성되지만, 기록물의 활용 단계에서 업무 담당자의 경험과 맥락을 연결할 방법은 존재하지 않는다. 이와 같은 문제를 해결하기 위한 디지털화의 각 단계는 자연어 처리 단계와 연결하고, 디지털 기록의 생산·등록과 활용이 함께 고려되는 것이 바람직하다. 자연어 처리 단계는 색인어의 추출, 문장의 자동요약 등 자연어 처리를 위한 공통활용기술, 한국어 처리를 위한 공통언어자원, 기록관리 분야에 특성화된 기록관리언어자원을 포함한다. 기록물 활용 단계는 수요자의 목적에 맞게 전자기록을 활용할 수 있는 기능을 포함한다. 예를 들어, 시청각 기록물은 기록의 활용 단계에서 텍스트 추출과 변환이 진행되는

 https://doi.org/10.14404/JKSARM.2022.22.4.129

〈그림 8〉 기록물 디지털화의 개선 방안: 자연어 처리와 기록물 활용 단계의 추가

것이 일반적이다. 종전의 방식으로 보면, 디지털화와 기록물 활용은 업무적·시기적으로 구분된 작업이고, 디지털화 단계에서 명시적으로 기술하지 않은 정보를 활용하기 어렵다. 수정 제안된 디지털화 단계는 단계별 업무에 따라 자연어 처리 기술을 적용하고, 언어자원을 공유함으로써 공통적이고 일관된 정보를 활용할 수 있다. 이런 관점에서 디지털화 작업 지침과 기록물 메타데이터 표준이 개선되어야 하고, 기록관리 분야에서 적극적으로 논의해야 한다.

4.3 기록관리 환경의 변화

디지털 시대라는 환경 변화에 기록관리가 능동적으로 대응하기 위해 산학연이 함께 협력할 수 있는 방안을 검토할 필요가 있다. 먼저, 기록학과 인접 학문의 협력 체계를 구축해야 한다. 기록물은 보관하는 장소가 특정되어 접근이 수월하지 않거나, 저작권의 제한으로 활용이 쉽지 않다. 기록물이 데이터 과학자, 인공지능 전문가가 활용할 수 있도록 접근성을 높이고, 데이터세트로 활용할 수 있는 기회를 제공해야 한다. 국가기록원에서 수행한 공통 학습데이터 세트 구축은 기록관리 분야에서 데이터를 공통으로 활용하기 위한 방안을 논의하고 있다(오효정, 2021). 이와 같은 접근은 기록물과 관련된 업무담당자와 대중의 이해를 높이는 기회가 될 수 있고, 동시에 다양한 분야의 전문가가 참여하고 논의할 수 있는 출발점이 될 수 있다.

기록은 역사적 관점에서 실시간성을 포함하는 방향으로, 인쇄물 중심의 기록은 전자기록과 클라우드 컴퓨팅 환경으로 전환되고 있다. 정보통신기술은 기록학 전반에 영향을 줄 수 있기 때문에, 학계와 현장의 업무 담당자가 디지털 리터러시에 대해 적극적으로 대응해야 한다. 기록학의 정규 커리큘럼에서 인공지능, 자연어 처리와 같은 기술을 접목하고 실무적으로 적용하는 방안을 검토하는 것이 필요하다. 기록관리의 현장 업무에서 기록물의 관리와 시스템 관리는 분리되어 있기 때문에, 새로운 정보통신기술의 검토와 도입이 효과적으로 진행되지 않는 것이 현실이다.

https://doi.org/10.14404/JKSARM.2022.22.4.129

정보기술과 관련된 업무가 용역과제 중심으로 진행되는 것은 이와 같은 현실을 반영하고 있다고 해석할 수 있다. 이런 문제는 기록관리 분야에 광범위하게 나타나고 있기 때문에, 기록관리 분야의 산·학·연·관 전문가가 함께 논의하고 대안을 모색해야 한다.

5. 결론

본 논문은 기록관리 분야에서 응용할 수 있는 자연어 처리 기술을 소개하고, 실제 사례를 통해 적용 방안을 소개했다. 자연어 처리는 인공지능 기술의 주요한 주제이며, 최근 머신러닝과 딥러닝 기술의 비약적 발전과 함께 다양한 분야에서 주목받고 있다. 기록의 생애주기 관점에서 보면, 기록의 생성과 활용에 이르는 전 단계에서 자연어 처리 기술의 적용이 효과적일 수 있다. 예컨대, 기계번역은 고문서에 포함된 한자 또는 영어 중심의 기록물을 한글로 번역하는데 적용할 수 있고, 기록물에 포함된 중요한 인물, 장소와 같은 정보는 개체명 인식을 적용할 수 있다.

본 논문에서 예제로 소개한 기계번역, 문서요약, 개체명인식, 이미지 인식 기술은 기록물과 관련된 실제 텍스트, 이미지를 대상으로 적용하였다. 적용 결과는 텍스트에 따라 성능이 다르고, 기술 분야에 따라 정확성에 차이가 있다. 기록관리 분야에서 자연어 처리 기술을 적용하려면 기술적 측면에서 고려사항이 있다. 첫째, 기록 관리에 특화된 자연어 처리를 위한 말뭉치의 구축, 기록 관리 분야에서 활용할 수 있는 개체명 등 관련 데이터의 구축과 보급이 필요하다. 둘째, 머신러닝과 딥러닝은 대규모 데이터를 학습하고 이를 통해 성능이 좋은 모델을 개발할 수 있다. 대부분의 기록물은 관리기관에서 보존하고 있고 라이선스 또는 이용약관의 제약을 받는 것이 현실이다. 특히, 자연어 처리 기술은 대규모 데이터 기반의 학습과 테스트가 필수적이기 때문에, 공동으로 활용할 수 있는 오픈 데이터의 발굴과 보급이 시급하다. 마지막으로, 새로운 기술의 적용의 목적과 목표에 대해 구체적으로 정의하고, 객관적으로 기술 도입을 검토해야 한다. 인공지능을 포함한 대부분의 기술은 명확하게 문제를 정의하고, 해결하는 목표가 구체적이어야 한다. 막연하게 신기술을 도입하는 것은 기술을 수단이 아닌 목표로 오해하게 만들고, 기대하는 결과를 도출하지 못하게 만드는 가장 큰 요인이다.

자연어 처리 기술은 기록관리 분야에서 폭넓게 활용될 가능성이 있다. 사람이 직접 수작업으로 수행하던 업무가 자동화될 수 있고, 국가의 기록물이 서로 연결되고 탐색 가능한 체계로 전환하는데 큰 역할을 할 수 있다. 본 논문은 한국어 자연어 처리를 위한 기본적인 적용 방안을 소개했다. 향후 연구에서는 다양한 기록물에서 활용할 수 있는 개체명을 구축하고 활용할 수 있는 방안, 대규모 기록물에 있는 텍스트를 분석하고 기계가 이해할 수 있는 구조로 표현하는 방안을 연구하는 것이 필요하다.

참고문헌

강민균 (2021). 빅데이터·인공지능 기술을 활용한 대학 기록관리 표준모델 연구(11-1741050-000079-01), 국가기록원.
강지홍 (2021). 기록물 OCR 데이터셋 구축 및 한글 필기체 인식 성능 개선 연구(11-1741050-000078-01), 국가기록원
강형석, 양장훈 (2019). 한국어 단어 임베딩을 위한 Word2vec 모델의 최적화. 디지털콘텐츠학회논문지, 20(4), 825-833. http://doi.org/10.9728/dcs.2019.20.4.825
공공기록물 관리에 관한 법률 시행령. 대통령령 제32772호.
기록관리 메타데이터 표준(v2.1). NAK/S 8:2016(v2.1).
기록물 디지털화 지침-제1부: 종이 및 사진·필름(v1.0). NAK/G 8-1:2013(v1.0).
기록물 디지털화 지침. NAK 26:2018(v2.0).
기록물 목록 작성 및 디지털화 작업 지침(v1.2). NAK 23:2017(v1.2).

김상운, 신원철 (2021). GPT-2를 이용한 지식 그래프 완성. 정보과학회논문지, 48(12), 1281-1288. http://doi.org/10.5626/JOK.2021.48.12.1281

김영민, 임승영, 이현정, 박소윤, 김명지 (2020). KorQuAD 2.0: 웹문서 기계독해를 위한 한국어 질의응답 데이터셋. 정보과학회논문지, 47(6), 577-586. http://doi.org/10.5626/JOK.2020.47.6.577

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

김인후, 김성희 (2022). 딥러닝 기반의 BERT 모델을 활용한 학술 문헌 자동분류. 정보관리학회지, 39(3), 293-310. http://doi.org/10.3743/KOSIM.2022.39.3.293

김태영, 강주연, 김건, 오효정 (2018). 지능형 기록정보서비스를 위한 선진 기술 현황 분석 및 적용 방안. 한국기록관리학회지, 18(4), 149-182. https://doi.org/10.14404/JKSARM.2018.18.4.149

김학래 (2022). 머신러닝 기술을 활용한 국회전자도서관 지식검색 서비스 모델 연구, 국회도서관

김해찬솔, 안대진, 임진희, 이해영 (2017). 기계학습을 이용한 기록 텍스트 자동분류 사례 연구. 정보관리학회지, 34(4), 321-344. https://doi.org/10.3743/KOSIM.2017.34.4.321

류한조 (2021). 빅데이터 분석기술을 활용한 자치단체 기록관리기준표 현황분석 및 개선방안 연구(11-1741050-000075-01), 국가기록원.

문헌정보 — 기록관리 — 제1부: 개념과 원칙. KS X ISO15489-1.

민진우, 나승훈, 김현호, 김선훈, 강인호 (2022). LUKE를 이용한 한국어 자연어 처리: 개체명 인식, 개체 연결. 정보과학회 컴퓨팅의 실제 논문지, 28(3), 175-183. http://doi.org/10.5626/KTCP.2022.28.3.175

박상언 (2021). 딥러닝 중심의 자연어 처리 기술 현황 분석. 한국빅데이터학회 학회지, 6(1), 63-81. http://doi.org/10.36498/kbigdt.2021.6.1.63

박은정, 조성준 (2014). KoNLPy: 쉽고 간결한 한국어 정보처리 파이썬 패키지. 제 26회 한글 및 한국어 정보처리 학술대회 논문집, 2014, 133-166.

박진호 (2019). 딥러닝 기반 자연어 처리에서 도메인 지식의 역할. 인공지능인문학연구, 4, 135-166. http://doi.org/10.46397/JAIH.4.7

오효정 (2019). 지능형 전자기록관리 기술연구 개발 기획연구(11-1741050-000033-01), 국가기록원

오효정 (2021). 기록관리 AI 기술적용을 위한 공통 학습데이터 세트 구축 연구(11-1741050-000073-01), 국가기록원

유소엽, 정옥란 (2019). BERT 모델과 지식 그래프를 활용한 지능형 챗봇. 한국전자거래학회지, 24(3), 87-98. http://doi.org/10.7838/jsebs.2019.24.3.087

조희련, 이유미, 임현열, 차준우, 이찬규 (2021). 딥러닝 기반 언어모델을 이용한 한국어 학습자 쓰기 평가의 자동 점수 구간 분류 -KoBERT와 KoGPT2를 중심으로-. 한국언어문화학, 18(1), 217-241. http://dx.doi.org/10.15652/ink.2021.18.1.217

최용석, 이공주 (2020). 트랜스포머와 BERT로 구현한 한국어 형태소 분석기의 성능 분석. 정보과학회논문지, 47(8), 730-741. http://doi.org/10.5626/JOK.2020.47.8.730

황상흠, 김도현 (2020). 한국어 기술문서 분석을 위한 BERT 기반의 분류모델. 한국전자거래학회지, 25(1), 203-214. http://doi.org/10.7838/jsebs.2020.25.1.203

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W. H., Jin, D., Naumann, T. & McDermott, M. (2019). Publicly available clinical BERT embeddings. arXiv preprint arXiv:1904.03323.

Arefieva, V. & Egger, R. (2022). TourBERT: A pretrained language model for the tourism industry. arXiv preprint arXiv:2201.07449.

Badrinarayanan, V., Kendall, A. & Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence, 39(12), 2481-2495.

Beltagy, I., Lo, K. & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. arXiv preprint

arXiv:1903.10676.

Bengio, Y., Lecun, Y. & Hinton, G. (2021). Deep learning for AI. Communications of the ACM, 64(7), 58-65. https://doi.org/10.1145/3448250

Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the association for computational linguistics, 5, 135-146. https://doi.org/10.1162/tacl_a_00051

Camacho-Collados, J. & Pilehvar, M. T. (2020). Embeddings in natural language processing. Proceedings of the 28th international conference on computational linguistics: tutorial abstracts, 10-15.

Ding, J. & Chen, J. (2020). Assessment of Empirical Troposphere Model GPT3 Based on NGL's Global Troposphere Products. Sensors, 20(13), 3631. https://doi.org/10.3390/s20133631

Dwivedi, Y. K., Hughes, L., Baabdullah, A. M., Ribeiro-Navarrete, S., Giannakis, M., Al-Debei, M. M. & Wamba, S. F. (2022). Metaverse beyond the hype: Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. International Journal of Information Management, 66.

Ethayarajh, K. (2019). How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. arXiv preprint arXiv:1909.00512. https://doi.org/10.1016/j.ijinfomgt.2022.102542

Floridi, L. & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. Minds and Machines, 30(4), 681-694. https://doi.org/10.1007/s11023-020-09548-1

Gonzalez-Dominguez, J., Lopez-Moreno, I., Sak, H. & Gonzalez-Rodriguez, J. (2014). Automatic language identification using long short-term memory recurrent neural networks. Proceedings of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH), 2155-2159.

Kaplan, J. (2016). Artificial Intelligence: What Everyone Needs to Know. 신동숙 옮김 (2017). 제리 카플란 인공지능의 미래 : 상생과 공존을 위한 통찰과 해법들. 한즈미디어.

LeCun, Y. & Bengio, Y. (1995). Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks, 3361(10),

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H. & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234-1240.

Lee, S., Jang, H., Baik, Y., Park, S. & Shin, H. (2020). KR-BERT: A Small-Scale Korean-Specific Language Model. CoRR, arXiv:2008.03979. https://doi.org/10.48550/arXiv.2008.03979

Locke, S., Bashall, A., Al-Adely, S., Moore, J., Wilson, A., & Kitchen, G. B. (2021). Natural language processing in medicine: a review. Trends in Anaesthesia and Critical Care, 38, 4-9. https://doi.org/10.1016/j.tacc.2021.02.007

McCarthy, J., Minsky, M. L., Rochester, N. & Shannon, C. E. (2006). A proposal for the dartmouth summer research project on artificial intelligence. AI magazine, 27(4), 12-12. https://doi.org/10.1609/aimag.v27i4.1904

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient estimation of word representations in vector space. In 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings, 2013.

Nadkarni, P. M., Ohno-Machado, L. & Chapman, W. W. (2011). Natural language processing: an introduction. Journal of the American Medical Informatics Association, 18(5), 544-551. https://doi.org/10.1136/amiajnl-2011-000464

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

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. & Zettlemoyer, L. (2018). Deep contextualized word representations. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1, 2227-2237

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. & Sutskever, I. (2018). Language Models are Unsupervised Multitask Learners. Technical report, OpenAi.

Ranjan, N., Mundada, K., Phaltane, K., & Ahmad, S. (2016). A Survey on Techniques in NLP. International Journal of Computer Applications, 134(8), 6-9.

Rogers, A., Kovaleva, O. & Rumshisky, A. (2020). A Primer in BERTology: What we know about how BERT works. Transactions of the Association for Computational Linguistics, 8, 842-866. https://doi.org/10.1162/tacl_a_00349

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review, 65(6), 386-408. https://doi.org/10.1037/h0042519

Singh, P., Rutten, G. & Lefever, E. (2021). A pilot study for BERT language modelling and morphological analysis for ancient and medieval Greek. Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2021), 128-137.

Sutskever, I., Martens, J. & Hinton, G. E. (2011). Generating text with recurrent neural networks. Proceedings of the 28th International Conference on International Conference on Machine Learning (ICML'11), 1017-1024.

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T. & Qin, B. (2014). Learning sentiment-specific word embedding for twitter sentiment classification. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 1, 1555-1565. https://doi.org/10.3115/v1/P14-1146

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), 6000-6010.

Von Ahn, L., Maurer, B., McMillen, C., Abraham, D. & Blum, M. (2008). recaptcha: Human-based character recognition via web security measures. Science, 321(5895), 1465-1468. https://doi.org/10.1126/science.1160379

Wang, H. & Raj, B. (2017). On the origin of deep learning. arXiv preprint arXiv:1702.07800.

Weikum, G. (2002). Foundations of statistical natural language processing. ACM SIGMOD Record, 31(3), 37-38. https://doi.org/10.1145/601858.601867

Weiss, K., Khoshgoftaar, T. M. & Wang, D. (2016). A survey of transfer learning. Journal of Big data, 3(1), 1-40. https://doi.org/10.1186/s40537-016-0043-6

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A. & Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, 38-45.

Yang, Y., Uy, M. C. S. & Huang, A. (2020). Finbert: A pretrained language model for financial communications. arXiv preprint arXiv:2006.08097.

● 국문 참고자료의 영어 표기
(English translation / romanization of references originally written in Korean)

Cho, Heeryon, Yi, Yumi, Im, Hyeonyeol, Cha, Junwoo & Lee, Chankyu (2021). Automatic Score Range Classification of Korean Essays Using Deep Learning-based Korean Language Models -The Case of KoBERT & KoGPT2-. International Network For Korean Language And Culture, 18(1), 217-241. http://dx.doi.org/10.15652/ink.2021.18.1.217

Choi, Yongseok & Lee, Kongjoo (2020). Performance Analysis of Korean Morphological Analyzer based on Transformer and BERT. Journal of KIISE(JOK), 47(8), 730-741. http://doi.org/10.5626/JOK.2020.47.8.730

Digitization Guidelines for Records. NAK 26:2018(v2.0)

Enforcement Decree of The Public Records Management Act. Presidential Decree No. 32772.

Guideline for Digitalization of Records - Part 1: Papers, Photographs and Films Version 1.0. NAK/G 8-1:2013(v1.0).

Guideline for Records Cataloging and Digitalization Project Version 1.2. NAK 23:2017(v1.2).

Hwang, Sangheum & Kim, Dohyun (2020). BERT-based classification model for Korean documents. The Journal of Society for e-Business Studies, 25(1), 203-214. http://doi.org/10.7838/jsebs.2020.25.1.203

Information and documentation — Records management — Part 1: Concepts and principles. KS X ISO15489-1.

Kang, Hyungsuc & Yang, Janghoon (2019). Optimization of Word2vec Models for Korean Word Embeddings. Journal of Digital Contents Society, 20(4), 825-833. http://doi.org/10.9728/dcs.2019.20.4.825

Kang, Jihong (2021). Establishment of OCR Dataset for Records and Research on Improvement of Korean Handwriting Recognition (11-1741050-000078-01). National Archives of Korea.

Kang, Mingyun (2021). A Study on the Standard Model for University Records Management Using Big Data and Artificial Intelligence Technology (11-1741050-000079-01). National Archives of Korea.

Kim, Haklae (2022). A Study on Recycled Electronic Library Search Service Model Using Machine Learning Technology, National Assembly Library.

Kim, Inhoo & Kim, Sunghee (2022). Automatic Classification of Academic Articles Using BERT Model Based on Deep Learning, Journal of the Korean Society for Information Management, 39(3), 293-310. http://doi.org/10.3743/KOSIM.2022.39.3.293

Kim, Intaek, An, Daejin & Rieh, Haeyoung (2017). Intelligent Records and Archives Management That Applies Artificial Intelligence. Journal of Korean Society of Archives and Records Management, 17(4), 225-250. https://doi.org/10.14404/JKSARM.2017.17.4.225

Kim, Sangwoon & Shin, Wonchul (2021). GPT-2 for Knowledge Graph Completion. Journal of KIISE(JOK), 48(12), 1281-1288. http://doi.org/10.5626/JOK.2021.48.12.1281

Kim, Taeyoung, Gang, Juyeon, Kim, Geon & Oh, Hyojung (2018). A Study on the Current Status and Application Strategies for Intelligent Archival Information Services. Journal of Korean Society of Archives and Records Management, 18(4), 149-182. http://doi.org/10.14404/JKSARM.2018.18.4.149

Kim, Youngmin, Lim, Seungyoung, Lee, Hyunjeong, Park, Soyoon & Kim, Myungji (2020). KorQuAD 2.0: Korean QA Dataset for Web Document Machine Comprehension. Journal of KIISE(JOK), 47(6), 577-586. http://doi.org/10.5626/JOK.2020.47.6.577

Kim, HaeChansol, An, Daejin, Yim, Jinhee & Rieh, Haeyoung (2017). A Study on Automatic Classification of Record Text Using Machine Learning. Journal of the Korean Society for Information Management

34(4), 321-344. https://doi.org/10.3743/KOSIM.2017.34.4.321

Metadata Standard for Records and Archives Management Version 2.1. NAK/S 8:2016(v2.1)

Min, Jinwoo, Na, Seunghoon, Kim, Hyunho, Kim, Seonhoon & Kang, Inho (2022). LUKE for Korean Natural Language Processing: Named Entity Recognition and Entity Linking. KIISE Transactions on Computing Practices (KTCP), 28(3), 175-183. http://doi.org/10.5626/KTCP.2022.28.3.175

Oh, Hyojung (2019). Intelligent Electronic Records Management Development Planning (11-1741050-000033-01). National Archives of Korea.

Oh, Hyojung (2021). Study on Common Training Dataset Construction for Applying AI Technology for Records Management (11-1741050-000073-01). National Archives of Korea.

Park, Eunjeong & Cho, Sungjun (2014). KoNLPy: Korean natural language processing in Python. Proceedings of the 26th Annual Conference on Human & Cognitive Language Technology, 2014, 133-136.

Park, Jinho (2019). The Role of Domain Knowledge in Deep Learning-Based Natural Language Processing. Journal of AI Humanities(JAIH), 4, 135-166. http://doi.org/10.46397/JAIH.4.7

Park, Sangun (2021). Analysis of the Status of Natural Language Processing Technology Based on Deep Learning. The Korea Journal of BigData, 6(1), 63-81. http://doi.org/10.36498/kbigdt.2021.6.1.63

Ryu, Hanjo (2021). A Study on the Status Analysis and Improvement of Local Government Record Management Reference Table Using Big Data Analysis Technology (11-1741050-000075-01), National Archives of Korea.

Yoo, Soyeop & Jeong, Okran (2019). An Intelligent Chatbot Utilizing BERT Model and Knowledge Graph. The Journal of Society for e-Business Studies, 24(3), 87-98. http://doi.org/10.7838/jsebs.2019.24.3.087