한국기록관리학회지
24(2), 65-88, 2024.5
https://doi.org/10.14404/JKSARM.2024.24.2.065
pISSN 1598-1487 eISSN 2671-7247

JKSARM
Journal of Korean Society of Archives and Records Management

자연어 처리의 개체명 인식을 통한 기록집합체의 메타데이터 추출 방안

A method for metadata extraction from a collection of records using Named Entity Recognition in Natural Language Processing

송치호(Chiho Song)

E-mail: chihosong@gmail.com

(사)한국국가기록연구원 원장

논문접수 2024.4.16
최초심사 2024.4.20
게재확정 2024.5.10

ORCID

Chiho Song
https://orcid.org/0009-0004-8028-1988

초 록

본 연구는 인공지능의 하위분야인 자연어 처리(NLP)의 개체명 인식(NER)을 통하여 기록에 내재된 메타데이터 값과 기술 정보를 추출하는 방안에 대한 시험적 연구이다. 연구 대상은 1960~1970년대에 생산된 구로공단 수기 기록물(약 1,200 쪽, 8만여 단어)을 대상으로 하였다.
디지털화를 포함하는 전처리 과정과 함께 기록 텍스트에 대해서 구글의 BERT 언어 모델에 기반하여 구현되어 공개된 언어 API를 사용하여 개체명을 인식하였다. 그 결과로 구로공단의 과거 기록에 포함된 173개의 인명과 314개의 조직 및 기관 개체명을 추출할 수 있었고, 이는 기록의 내용에 대한 직접적인 검색어로 사용될 수 있다고 기대된다. 그리고 자연어 처리의 이론적 방법론을 반·비정형의 텍스트로 이루어진 실제 기록물에 적용할 때 발생하는 문제점을 파악하여 해결 방안과 고려해야 할 시사점을 제시했다.

ABSTRACT

This pilot study explores a method of extracting metadata values and descriptions from records using named entity recognition (NER), a technique in natural language processing (NLP), a subfield of artificial intelligence. The study focuses on handwritten records from the Guro Industrial Complex, produced during the 1960s and 1970s, comprising approximately 1,200 pages and 80,000 words.
After the preprocessing process of the records, which included digitization, the study employed a publicly available language API based on Google's Bidirectional Encoder Representations from Transformers (BERT) language model to recognize entity names within the text. As a result, 173 names of people and 314 of organizations and institutions were extracted from the Guro Industrial Complex's past records. These extracted entities are expected to serve as direct search terms for accessing the contents of the records. Furthermore, the study identified challenges that arose when applying the theoretical methodology of NLP to real-world records consisting of semistructured text. It also presents potential solutions and implications to consider when addressing these issues.

Keywords: 인공지능, 자연어 처리, 메타데이터, 언어 모델, 개체명인식
AI, NLP, Metadata, LLM, NER

▪ 본 논문은 송치호의 박사 학위논문 「인공지능을 적용한 기록관리의 지적 통제방안 - 자연어 처리를 중심으로 -」(2024)를 수정·보완한 것임.

1. 서론

1.1 연구 배경 및 목적

기록관리는 논리적으로 구조화된 관례와 방법 및 절차 규칙에 따라 만들어진 범주 속에서 업무 활동과 기록을 체계적으로 관리하는 것이다(ISO 15489-1 : 2016, 3.5, 8.3, 9.4). 이는 기능별로 단위화한 업무 과정과 기록의 생산을 결합하는 것이며, 계층화된 업무체계의 구조에 대응하여 기록물은 생산과정에서부터 해당 업무맥락과 계층의 정보를 메타데이터 형태의 기록물 기술정보(Description)로 부여받는다(Bak, 2012). 이를 통해 기록물은 출처 확인 및 분류, 기술 과정을 거쳐 기록의 내용과 생산 맥락을 파악할 수 있도록 하며, 이용자가 원하는 기록을 찾고 이해하는 것이 가능해진다. 또한 기록물은 단순히 증빙의 근거로 보존될 뿐만 아니라, 물화(物化)된 기억으로, 데이터와 정보 자산으로도 활용되어야 한다. 그렇지만 2024년 현재에도 영구 기록 관리기관, 일부 대학기록관을 제외한 기초자치단체의 기록관은 온라인에서의 기록검색은 도서 등에 대한 '소장자료' 검색만을 지원하거나 기초자치단체 홈페이지의 통합 검색과 연계되어 실질적인 기록물의 내용에 대한 검색은 지원하지 않는 경우가 다수이다.

검색의 과정에서 내용을 직접 반영하는 키워드, 혹은 주제나 토픽, 기술정보가 있다면 검색의 정확성이 높아지고 이후 선별 과정이 훨씬 쉬워지지만, 현재의 기록관리 체계에서는 적절한 메타데이터와 기술 작성이 제대로 이루어지지 않고 있다. 즉, 현재의 기록 메타데이터는 기록의 4대 속성인 진본성, 신뢰성, 무결성을 기술하는 것에 초점을 맞추어 생산자, 생산 부서, 보존 기간 등을 기술하고 있지만, 기록이 실제로 담고 있는 내용에 직접 접근하여 그 이용성을 보장하는 검색과 활용에는 상대적으로 부족하다. 내용을 반영하는 메타데이터를 작성하는 것은 업무 담당자, 기록관리자 등이 직접 기록물의 내용을 독해해서 작성해야 하는데, 현재의 기록관리 체계에서는 기록관 인력 부족과 기록관리자의 업무 부담이 장애로 작용한다.

지금까지 단어 빈도 분석을 응용한 토픽 추출이나 시맨틱웹 등 정보기술을 기록관리에 적용하고자 하는 관심과 시도는 꾸준히 있었다(한미경, 2020). 그렇지만 이런 관심과 시도는 기록과 기록 집합체가 이미 존재한다는 가정하에, 다중 분류와 같은 개선된 검색 도구의 일종으로서 접근하거나, 내부의 기록물을 자원 식별 체계(URI)를 부여하여 기록 집합체 간의 연결을 확장하는 것이었다. 이러한 시도가 성공하기 위해서는 기록 자체가 가지고 있는 정보의 양적 확보와 함께 질적 수준의 향상이 동시에 필요하다. 그러나, 기록관의 정보량은 꾸준히 양적으로 증가해 왔지만, 기록의 색인과 기술로서 메타데이터의 질적 수준이 현시대의 새로운 정보기술이 요구하는 수준에는 다소 부족한 것이 현실이다.

본 연구에서는 기록의 내용 중 기록검색에 직접 활용할 수 있는 방법으로서, 인공지능과 자연어 처리 기술 중 개체명 인식(Named Entity Recognition)에 초점을 맞추어 그 가능성을 검토한다. 대규모 언어 모델(Large Language Model)의 발전은 이제 언어 텍스트를 직접 독해하고 그 특질을 스스로 찾아내는 수준까지 진화했다. 개체명 인식은 대규모 언어 모델의 학습 내용을 기반으로 텍스트가 포함하고 있는 특질 중 인명, 기관명, 지명 등의 고유명사를 인식하여 텍스트가 가지고 있는 고유한 특질을 파악할 수 있는 기술인데, 이를 기록 건에 적용하여 이용자가 검색할 수 있는 색인으로, 기록의 내용을 유추할 수 있는 일종의 키워드로 기능할 수 있다. 이를 통해 결과적으로 메타데이터 기술정보의 수준을 이용성 측면에서 향상할 수 있을 것으로 판단된다.

1.2 연구 방법론

본 연구는 자연어 처리의 개념과 함께 언어 모델을 개괄하고, 자연어 처리의 절차와 기록관리에 적용할 수 있는

https://doi.org/10.14404/JKSARM.2024.24.2.065

구체적인 응용 영역들을 살펴보았다. 이런 이론적인 개념과 방안이 실제 기록물에 어떻게 적용될 수 있는지를 실증적으로 검토하기 위하여, 1960년대에서 1970년대에 생산된 구로공단의 기록물을 대상으로 개체명 인식을 통한 색인정보 추출에 초점을 맞추어 시험 연구를 수행하였다. 시험 연구 대상이 된 구로공단 기록물은 다음과 같은 특징을 가지고 있다.

첫째, 수기로 작성된 국·한문 혼용의 종이 문서를 쪽 단위로 스캔한 tiff 포맷의 이미지 파일을 건별로 묶은 PDF 포맷으로 구성되어 있다.

둘째, 디지털화한 기록물의 기계 가독성과 표준 호환성이 보장되지 않았다.

셋째, 5, 60년 전에 생산된 과거의 기록물로서, 철-건 구조 등 계층별 구조 정보가 부재하거나 신뢰성을 확신하지 못하며, 메타데이터의 기술 수준과 분류가 제목과 과거 보존 장소 등만 간략하게 기재된 상태이다.

본 연구는 과거에 생산된 종이 문서를 대량으로 디지털화했지만, 기계 가독성과 철·건 구조의 부재로 이용자의 지적 접근이 어려운 기록물에 대하여 기록물의 내용에서 고유명사에 해당하는 이름과 직위, 기관명, 지명 등에 대한 정보를 색인화하고 기술정보에 대한 메타데이터를 보완하여 접근성과 사용성을 높이는 것을 목표로 진행되었다. 이를 위하여 데이터 전처리에서부터 인식 언어 모델을 적용했을 때의 구체적인 인식 결과까지 개체명 색인의 전체 과정을 절차별로 진행하였다. 자연어 처리에서 공통으로 수행해야 하는 데이터 전처리의 부분은 특히 한국어로 작성된 기록물로서 가지고 있는 특성과 어려운 점을 제시하였다. 그리고 개체명 인식 언어 모델의 적용은 보안과 사용성을 고려하여 해외에서 개발된 대규모 언어 모델을 API 형태로 재가공한 국내의 공공 연구기관의 언어 모델을 사용하였다. 이러한 개별적인 세부 절차에서 실제로 발생하는 문제점과 해결 방안을 파악하고, 이를 통해 기록관리에서의 시사점을 제시했다.

1.3 선행연구

본 연구는 자연어 처리의 응용 영역과 활용 가능성, 자연어 처리의 사례연구, 기계 가독성과 사용성이라는 관점에서 선행연구를 검토하였다. Rolan et al.(2019)는 전문가 시스템에서 딥러닝으로의 인공지능 변화라는 관점에서 규칙 기반 시스템, 통계 모형, 딥러닝 모형을 소개하고, 인공지능이 기록관리 업무에서 자동화된 분류와 처분에 적용될 수 있는 잠재적 가능성이 있다고 밝히고 있다. 특히 TNA의 eDiscovery 도구를 소개하며 이관 과정에서 디지털 기록으로 생산된 기록(Born-Digital Records)은 특히 법무 관련 영역에서 평가, 선별에 도움을 줄 수 있으며, 이외의 영역은 상대적으로 적은 성공 사례를 보여준다고 언급했다.

Colavizza, Ehrmann, Bortoluzzi(2019)는 디지털화할 자료의 선택 문제에 주목한다. 이는 어디서 시작해야 하며, 모든 것을 디지털화해야 하는지에 대한 질문에 대한 답변의 문제이며 결국 명확한 디지털화 전략이 필요하다는 것이다. 디지털화를 통해 기록에 접근하게 되면 이는 거의 전적으로 맥락정보로서의 메타데이터에 기반을 두게 되는데, 검색 수단으로서의 메타데이터뿐만 아니라 내용 자체에 대한 처리가 있어야 디지털 역사 기록에 대한 접근성 및 이용률을 향상할 수 있다고 주장했다. 그러므로 유럽의 문화유산 컬렉션의 디지털화 가능성은 내용 수준에서의 색인화에 있다고 전제하고, 역사 콘텐츠의 색인화와 구조 정보 추출이 어휘적 지평과 참조 지평에서 이루어져야 한다고 설명했다.

2000년대 중후반부터 본격화한 종이 기록의 디지털화는 앞으로 다가올 활용에 대한 고려보다는 종이 문서의 멸실과 훼손을 막기 위하여 일단 디지털 파일 형태로의 전환에 초점을 맞추었다. 임진희(2021)는 이런 문제의식에서 기록 텍스트, 특히 공문서를 기계가 읽을 수 있는 상태로 전환해야 할 필요성과 요건을 제시하고 있다. 구체적으로는 공개 문서표준(ODF)으로 가독성뿐만 아니라 구조 정보를 인식하여 전거 정보, 서식 태깅의 과정을 거치면

https://doi.org/10.14404/JKSARM.2024.24.2.065

식별, 무결성, 문서 서식을 포함하는 자기 기술 메타데이터화 할 수 있다는 점을 제시하고 있다. 그리고 이를 통해 빅데이터 분석, 조직 및 인사 관리의 정량적 지표, 지식 관리에서의 편이성, 단위과제 오 분류와 공개설정의 오류 해결, 스토리지 이용 효율성 향상의 성과를 달성할 수 있다고 언급했다.

임진희(2021)와 유사한 문제의식에서 출발한 안세진, 황현호, 임진희(2022)의 연구는 기존의 OCR 사업에서 문제가 되었던 OCR 엔진의 낮은 인식률 문제를 해결하기 위해 인공지능 기술을 도입하여 OCR 엔진 스스로 학습하고 교정하는 방안을 제시했다. 이 연구는 기초자치 단체 단위에서 최초로 AI-OCR 기록물 통합검색시스템을 성공적으로 구축한 사례연구로서 현업에서 맞닥뜨릴 수 있는 현실적인 문제와 기록관의 역할과 준비 과정부터 검수 과정까지의 절차를 구체적으로 소개했다는 점에서 의의가 크다.

김인후, 김성희(2022)는 한국어 데이터로 학습된 구글의 BERT 모델을 기반으로 문헌정보학 분야의 문서에 대한 자동 분류를 수행하고 성능을 분석하였다. 분류 클래스는 13개였으며, 문장의 토큰 수에 따라 64개의 Short Model, 128개의 Middle Model, 256개의 Long Model로 구분하고 70%의 학습데이터와 30%의 평가데이터로 5, 457개의 학술 논문 분류를 수행하였다. 평가 결과는 주제의 성격이 명확한 경우, 데이터양이 많을 경우, 데이터양이 적어도 주제와 관련된 명확한 키워드를 사용하여 데이터 품질의 좋은 경우 성능 향상이 나타났다. 데이터의 품질과 양을 축으로 정교하게 분석한 결과 데이터의 양보다는 품질이 분류 성능에 영향을 미친다는 것을 실증적으로 증명했다.

실제 자연어를 처리하는 과정에서 시간이 많이 소요되고 자원이 투자되는 영역은 알고리즘 적용이나 언어 모델 생성보다는 데이터의 기계 가독성을 확보하고 데이터의 구조를 정형화하는 데이터 전처리와 정규화 과정이다. 그러므로 자연어 처리 전반에 대해서 기술 도입과 적용뿐만 아니라, 기록관리라는 전체 환경에서 실천적인 효능을 조망하여 자연어 처리를 고민해야 한다. 김학래(2022)는 기록관리 분야에서 한국어 자연어 처리에 대하여 기록물 디지털화 과정에서 메타데이터 자동 추출과 자동화의 가능성을 제시했다. 이때 고려해야 할 점으로 공동 활용을 위한 기반 확보, 생산 등록과 활용을 고려한 디지털화의 개선, 기록관리 환경의 변화에 따른 기술적 적용을 들었다. 이 연구에서 주목할 점은 자연어 처리 기술이 기록 생애주기 전반에 반영되어야 한다는 점이다. 즉, 자연어 처리가 기술 그 자체로, 혹은 부분적인 요소기술로써 사용되는 것이 아니라, 전체적인 기록관리 영역이라는 조망을 가지고 세부적인 영역에 적용해야 한다는 점이다. 이는 기록 생애주기 전반에 자연어 처리를 위한 준비와 반영이 이루어져야 한다는 문제의식을 본격적으로 제기했다는 점에서 시사하는 바가 매우 크다.

2. 기록관리와 자연어 처리

2.1 자연어 처리 방법론의 전제: 계산할 수 있는 언어 표현

자연어 처리는 의미의 최소 단위를 단어라고 가정한다. 단어는 규칙(문법)으로 구성된 문장으로 생성되며, 문장은 순차적으로 배열되어 화제와 의미의 일관성을 가진 문단으로 구성되고, 문단이 모여 최종적으로 문서가 된다(강범모, 2014). 자연어 처리는 일상 언어를 대상으로 하되, 일상어를 말뭉치(Corpus)라는 자연어 문서들의 집합으로 변형시켜 처리를 수행한다. 말뭉치는 자연어 연구를 위해 특정한 목적을 가지고 언어의 표본을 추출한 집합이며, 목적에 따라 적절히 특질과 주석을 첨부하여 말뭉치를 데이터로 사용할 수 있게 한다. 언어 모델은 언어와 관련된 사용자와의 상호작용을 수행하는 자연어 처리의 핵심인데, 구체적인 실체는 말뭉치의 상태와 구조를 표현하고 있는 메타데이터로서의 특질과 알고리듬으로 구성되어 있으며, 알고리듬은 텍스트를 구성하고 있는 단어 집합에서 그 순서의 확률 분포를 계산하는 함수의 집합이다.

https://doi.org/10.14404/JKSARM.2024.24.2.065

대규모 언어 모델은 언어 모델의 학습 파라미터를 극적으로 늘려 성능을 향상한 언어 모델이다. 최근에 주목받고 있는 대규모 언어 모델은 확률 계산에 필요한 데이터 세트와 변수(차원)를 10조 개 이상으로 산정하여 확률 예측의 정확도를 높이고 있다. 대규모 언어 모델은 문맥을 사전 학습하여 입력 문장에 포함된 단어의 문법과 의미를 이해하고 분석하여 텍스트 분류, 키워드 추출 등에 활용할 수 있는 언어이해 모델, 대용량 데이터를 미리 학습하여 주어진 단어 열에 가장 적합한 단어를 예측하여 문장을 생성할 수 있는 실시간 번역, 질의응답, 챗봇 등에 사용할 수 있는 언어 생성 모델, 양자를 혼합한 언어이해 및 생성 모델로 분류할 수 있다(임수종, 2021).

<표 1> 자연어 처리 언어 모델의 유형별 특성

구분	언어이해 모델	언어 생성 모델	언어이해 및 생성 모델
개념	자연어 문장에서 단어의 주변 문맥을 사전 학습하여 입력 문장에 포함된 단어의 문법과 의미를 이해	자연어 문장을 사전 학습하여 순서대로 주어진 단어 열에 가장 적합한 다음 단어를 예측하여 생성	언어이해와 생성을 같이 사용하는 모델로 입력 문장을 이해한 결과를 바탕으로 출력 문장을 생성하는 모델
특징	활용 사례 및 후속 연구가 가장 활발한 모델	자동번역, 요약 같은 언어 생성 태스크에 가장 적합한 모델	언어이해 및 언어 생성 모델을 모두 포함
학습 방법	주변 단어를 이용하여 목표 단어를 예측	이전 단어(들)를 기반으로 다음에 나올 단어를 예측	입력된 문장에 해당하는 문장을 출력
모델	구글 BERT 페이스북 RoBERTa, Allen 인공지능 연구소의 SpanBERT, 스탠퍼드대학교 ELECTRA	OpenAI GPT, 카네기 맬런 대학 & 구글 브레인 XLNet, 페이스북 BART	구글 T5

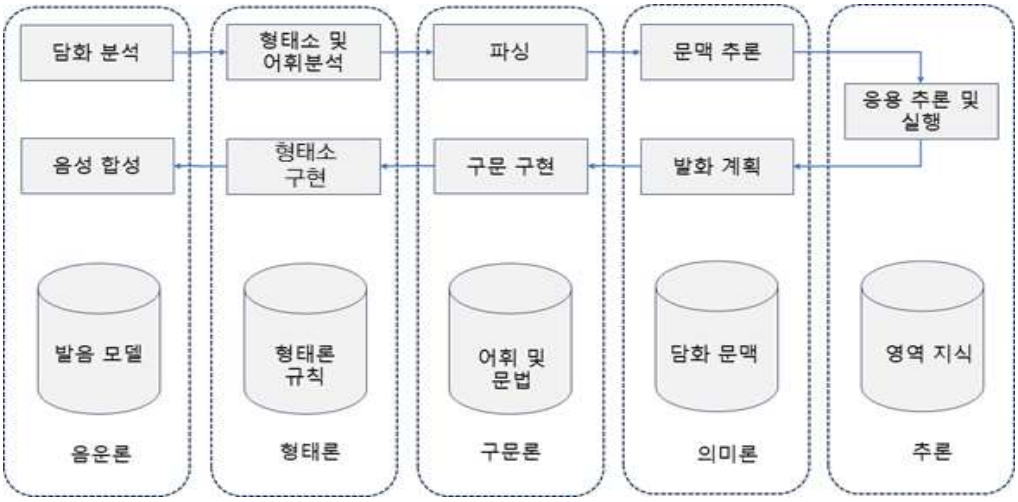
2.2 자연어 처리의 절차

자연어 처리는 언어라는 특수한 대상을 다룬다. 일반적인 자연어 처리 절차의 전형은 담화 분석에서 시작하여 단어 수준에서 형태소 및 어휘를 분석하는 과정 등을 거쳐 전체 말뭉치에 대한 응용 추론을 실행한다. <그림 1>은 정제된 언어로서의 데이터 소스에 해당하는 말뭉치가 이미 준비되었다는 과정을 전제하고 음성과 텍스트를 입력하여 처리 후 발화하는 자연어 처리의 전반적인 절차이다. 언어 데이터 역시 반/비정형이라는 성격을 가지는 데이터이므로 전통적인 데이터 분석 과정과 유사한 절차를 거친다(<표 2> 참조).

<표 2> 데이터 분석 절차와 자연어 처리 절차 비교

영역	데이터 처리 절차	자연어 처리 절차	
상호작용(Interaction)	입수(Ingest)	말뭉치 변환	
데이터(Data)	정제(Wrangling)	토큰화를 통한 형태소 및 어휘 분석	데이터 전처리
저장소(Storage)	정규화(Normalization)	정규화	
연산(Computation)	특질 분석/모델 생성(Feature Analysis/Model Build)	문맥 추론 응용 추론	
교차 검증(Cross Validation)	교차 검증(Cross Validation)		
저장소(Storage)	모델 선택 & 모니터링(Model Selection & Monitoring)		
데이터(Data)	응용프로그램 인터페이스(API)	서비스 구현, 배포	
상호작용(Interaction)	피드백(Feedback)	피드백	

https://doi.org/10.14404/JKSARM.2024.24.2.065



<그림 1> 자연어 처리 절차
(출처: https://ratsgo.github.io/natural%20language%20processing/2017/03/22/lexicon/)

데이터 분석의 첫 번째 단계인 Ingestion은 데이터 처리 모델이 데이터 소스의 위치 지정, 주석 작성 등을 하는 작업 전반을 의미한다. 자연어 처리 단계에서는 데이터 소스에 해당하는 말뭉치를 만들고, 말뭉치 내부에 주석을 달거나 라벨링 하는 작업에 해당한다. 다음 단계인 Wrangling과 Normalization은 데이터 처리 모델이 데이터를 사용할 수 있도록 데이터를 변환하고 응용할 수 있는 준비 단계이다. Wrangling은 토큰화와 어휘 분석으로, Normalization은 불용어 처리, 대소문자 통합 등을 포괄하는 정규화를 의미하며 자연어 처리의 전처리 단계에 해당한다. 전처리 단계는 말뭉치 텍스트를 단어 단위에서 사전에 조작하여 단어의 원형에 해당하는 어근을 뽑아내어 데이터로써 사용할 수 있게 하는 일련의 과정 전반을 의미하는데, 굴절어 계열의 서유럽 언어는 구두점 등 문장 기호와 격 변화 등을 고려해야 하고, 교착어 계열의 한국어, 일본어 등은 접사 제거와 띄어쓰기까지 고려해야 한다. 이 과정에서 언어의 특성을 반영한 토크나이저, 형태소 분석과 품사 태깅의 기법이 사용되어 결과적으로 말뭉치는 최소의 의미가 있는 단어의 집합으로 재가공된다. Computation 단계는 교차 검증을 통해 언어 모델을 만들고 문맥과 응용 추론을 수행하는 단계에 해당한다. Computation 이후 단계는 출력에 해당하며, 입력의 반대 방향으로 Storage에 저장 Data로서 API 제공, 피드백의 절차를 밟는다.

2.3 자연어 처리의 응용 영역

자연어 처리는 여러 분야에서 다양한 기법으로 사용된다. 특정 영역에서 사용된 기법이 다른 분야에서 해당 도메인 영역의 특징에 맞추어 변형되어 사용되며, 도메인 영역의 특징에 따라 다양한 기법이 조합되어 사용되기도 한다. 자연어 처리의 종류 역시 영역과 사례, 기법에 따라 분류하는 방식이 다르며, 상황에 따라 적절히 혼용된다. 기록관리의 영역에서 자연어 처리는 전통적인 기록관리 업무에서 기록관리자가 기록 집합체에 대해 지적 통제를 직접 수행하는 지점에 적용될 수 있다. 이관받거나 수집한 기록에 대한 정리·기술, 검색과 추천 서비스 등이 그것인데, 업무의 성격에 따라 기록관리시스템에서 수행하는 업무 지원과 이용자에게 제공하는 정보서비스 영역으로 분류할 수 있다. <표 3>은 '지능형 기록정보 서비스'를 위하여 기록관리 업무 영역별로 적용할 수 있는, 자연어 처리를 포함한 인공지능 세부 기술이다.

https://doi.org/10.14404/JKSARM.2024.24.2.065

<표 3> 기록관리 업무 영역별 지능화 기술 적용(출처: 김태영 외(2018))

구분	성격	세부 기술 내용
업무지원 영역	정리	클러스터링, 기계 학습/딥러닝
	기술	문서 요약, 키워드/메타데이터 추출
	이관 및 보존	클라우드 컴퓨팅, 블록체인
정보서비스 영역	검색 고도화	택소노미/폭소노미: 자동 태깅
		시청각 검색 지원: 이미지/영상/음성/안면 인식 기술
	추천 서비스	빅데이터 분석: 데이터 마이닝, 텍스트 마이닝
		챗봇
	콘텐츠 제공	LOD, 온톨로지
		스마트 디바이스: 사물인터넷, 웨어러블, VR/AR, 로봇 디바이스

기록물은 종이 문서, 문서파일, 사진/필름, 음성/소리, 동영상, 행정 박물 등 다양한 형태로 구성되어 있다. 그러므로 기록물의 물성에 따라 그에 맞는 자연어 처리의 적용 영역이 필요하다. 예를 들어 <표 4>는 기록물의 유형별로 분류한 자연어 처리의 응용 영역이다. 예를 들어 종이 문서의 경우 스캐닝을 통해 전자 문서로 변형한 후 문자인식을 기반으로 한 광학 인식 기술을 통해 전자 문서로 전환할 수 있으며, 사진 등 이미지는 컴퓨터 비전(CV) 기술을 기반으로 한 피사체 인식을 통한 분류가 가능하다.

<표 4> 기록 유형별 자연어 처리의 응용 영역

세부 구분		세부 기술 내용
문서	비전자	문자인식
	전자	개체명 인식, 토픽 모델링/문서 분류, 클러스터링, 문서 요약
시청각류	사진·필름·도면	이미지 인식을 이용한 분류, 이미지 메타데이터(EXIF)를 통한 태깅
	음성	음성인식을 이용한 녹취 전사, 내용 요약, 주제 분류, 정서 분석
	동영상	음성인식을 이용한 녹취 전사, 내용 요약, 주제 분류, 자막 자동 생성, 신(Scene)별 태깅(시작/종료 등)
도서 및 간행물		특징 추출, 개체명 인식, 문서 분류, 클러스터링, 문서 요약

본 연구에서는 자연어 처리의 응용 영역 중 개체명 인식을 사용하여 기록에 대한 색인 추출을 통한 메타데이터의 이용성의 향상 가능성을 제시하였다. 개체명 인식은 “미리 정의해 둔 사람, 회사, 장소, 시간, 단위 등에 해당하는 단어(개체명 Entity)를 문서에서 인식하여 추출 분류하는 기법. 추출된 개체명은 인명(Person), 지명(Location), 기관명(Organization), 시간(Time) 등으로 분류된다. 개체명 인식은 정보 추출을 목적으로 시작되어 자연어 처리, 정보 검색 등에 사용된다.”라고 정의된다(한국정보통신기술협회 정보통신 용어 사전). 좀 더 상세하게는 문장 내 어절을 기본 단위로 분석하여, 인명, 지명, 기관명 등과 같은 고유명사나 명사구를 의미하는 모든 개체명의 경계를 인식하고, 해당 개체명이 어떤 태그로 분류될 수 있는지 자동으로 인식하는 것을 말한다(고명현 외, 2019). 그러므로 개체명 인식을 통해 기록이 포함하고 있는 내용에 대한 직접 접근을 통해 기록물에 대한 이용성을 확보할 수 있다고 판단하였다.

https://doi.org/10.14404/JKSARM.2024.24.2.065

3. 자연어 처리의 적용: 구로공단 기록물을 중심으로

본 장에서는 2장에서 소개한 자연어 처리의 개념과 절차, 응용 영역과 방법론을 실제 기록물에 대해서 적용하였다. 본 연구의 목적인 메타데이터와 기술정보의 추출을 위하여 실제로 1960년대에서 70년대까지 한국수출산업공단에서 생산한 구로공단 기록물에 대한 사람과 조직에 대한 개체명 인식을 수행하였다. 이를 통해 이론에서 놓칠 수 있는 현실적인 문제를 실증적 시험을 통해 파악하고 해결책을 제시했다.

3.1 구로공단 기록물 개요

2000년대 이후 구로구의 공단지역 일대가 디지털 산업 위주의 지식산업센터로 재개발되며, 서울시에서는 지역의 명칭을 'G밸리'로 변경하였다. 그리고 G밸리의 역사와 기억을 저장하는 공간 계획을 2018년에 수립하고 3년여의 준비 기간을 거쳐 2022년에 'G밸리 산업박물관'을 개관하였다. 개관 준비 과정에서 다양한 기록물과 유물들이 수집되었는데, 그중에는 과거 한국수출산업공단에서 생산한 공공기록물이 포함되었다. 그리고 G밸리 산업박물관은 한국산업단지공단에서 보존하고 있던 기록물 중 362권의 기록물을 2020년에 대여받아 2021년 상반기까지 디지털화하였다.

<표 5> 구로공단 기록물의 구성

	문서		사진 필름	도면		지도		해제
	철	건	철	철	건	철	건	
1960년대	70	1,138	1	5	13	4	4	22권
1970년대	157	3,163	4	12	121	9	9	26권
1980년대	9	108	9	3	69	3	3	1권
1990년대	37	538	3	3	32	0	0	1권
2000년대 이후	1	12	0	0	0	0	0	0권
연대 미상	7	34	4	10	59	11	11	0권
총계	281	4,993	21	33	294	27	27	50권

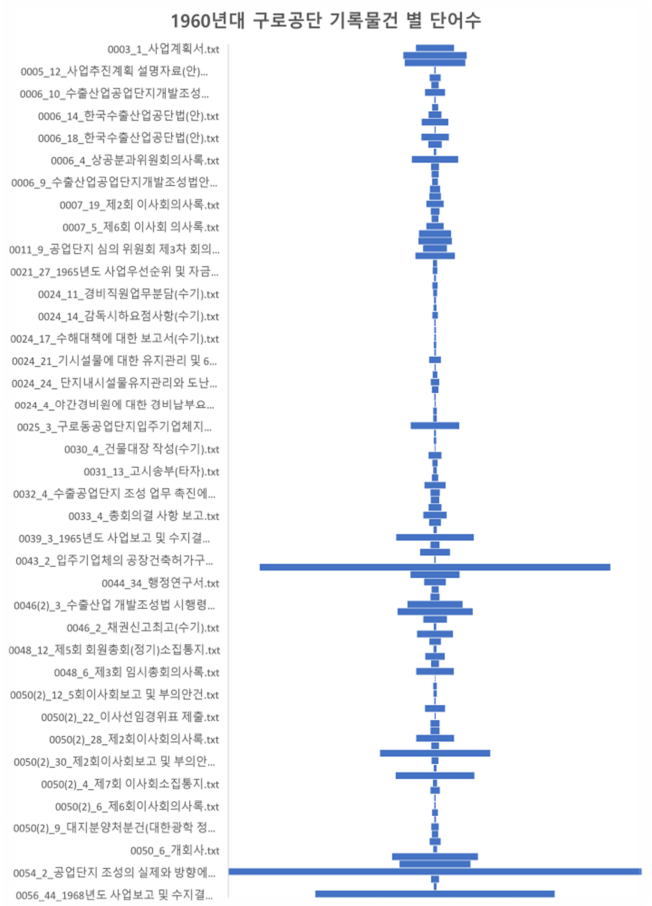
구로공단 기록물은 개발도상국으로서 대한민국의 산업사 초기 정책 집행을 파악할 수 있는 귀중한 기록으로서, 향후 현대사 연구와 대국민 서비스로 활용 가능성이 크다. 그리고 수기와 타자로 작성된 종이 기록물을 원본으로 디지털화하였으므로 자연어 처리를 포함하는 인공지능의 실제적인 기여 가능성을 확인하는 좋은 대상이 될 것으로 판단했다. 그리고 필자가 2021년과 2023년, 2차례에 걸쳐 G밸리 산업박물관의 구로공단 기록물에 대한 기초적인 정리·기술과 해제를 수행하고 전문가와 전시를 위한 활용 콘텐츠를 만드는 연구 용역에 참여한 경험에 비추어 볼 때, 가공되지 않은 원(原)자료로서의 기록을 수집한 후 기록관리자가 어떻게 이 기록에 대한 개요를 파악하고, 데이터로서 정제하여, 자연어 처리를 통하여 지적 통제를 강화할 수 있는가에 대한 사고 실험과 시험적 실증의 현실적인 사례로 충분히 의미가 있다고 판단하였다.

https://doi.org/10.14404/JKSARM.2024.24.2.065

3.2 구로공단 기록물의 특징

디지털화한 구로공단 기록물 281권의 철-건 편철 구조를 분석한 결과, 기록물 건의 약 25%는 명확한 업무 단위나 성격 별로 편철되기보다는, 유사한 성격으로 보이는 건들을 '관계', '참고' 등으로 묶어 한꺼번에 편철한 것으로 파악된다. 이는 G밸리 산업박물관이 디지털화 사업의 과정에서 기존의 철 상태를 그대로 보존하고, 일부 편철이 되지 않은 건은 충분한 검토 없이 유사한 제목을 기준으로 철로 묶었을 것으로 판단된다.

그리고 개별 기록 건과 그 단어 분포를 분석해 보았을 때, 개별 건 각각의 정보량에 편차가 큰 것이 확인되었다. '0054_2_공업단지 조성의 실제와 방향에 관한 소고' (12,848 단어), '0043_9_약정서(수기)' (10,901 단어), '0056_44_1968년도 사업 보고 및 수지결산서' (7,437 단어) 3건이 10, 000건 이상으로 단어 수 평균인 766개를 크게 웃돌았다.



<그림 2> 1960년대 구로공단 기록물 건 단어 분포

3.3 구로공단 기록물의 데이터 전처리

3.3.1 기계 가독성 확보

G벨리가 수집한 전체 4,993건의 기록물 중 1960년대에서 70년대까지 생산된 기록 중에서 숫자 등이 중심이 된 회계 증빙자료, 도면 등을 제외하고 사업계획서, 총회 자료 등 중요하다고 판단된 기록물 115건을 일차로 선별

https://doi.org/10.14404/JKSARM.2024.24.2.065

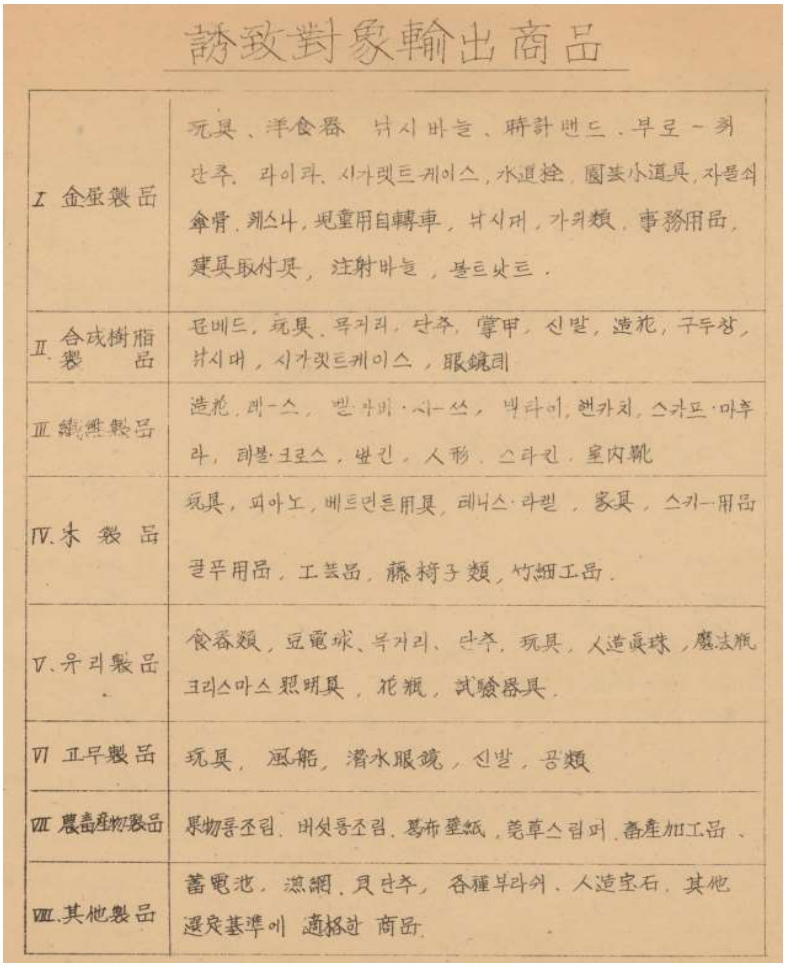
하였다. 그리고 업무 참고 자료로 확인된 외부 학위 논문 1건을 제외한 114건을 대상으로 자연어 처리를 진행하였다. 114건의 형태는 주로 수기로 작성된 종이 문서의 개별 페이지를 TIFF 이미지로 스캔하여 PDF 문서로 합본한 디지털 파일이다. 기계 가독성을 확보하기 위하여 PDF 문서에 기재된 한문을 독음(讀音)하고, 한글로 기재된 내용과 병합하여 최종적으로 114개의 텍스트 파일이 생성되었다.

3.3.2 오·탈자 교정과 사어(死語)·신조어 처리

114건의 문서에서 발견된 오·탈자의 대부분은 띄어쓰기 오류로 확인되었고, 일부 독음 과정에서 잘못 입력한 내용이 다시 수정되었다. 이 과정에서 당시에 통용되었던 사어가 다수 발견되었는데. '써-비스(서비스)', '데자인(디자인)' 등 당시의 외래어 단어는 빈도-역문서(TF-IDF, Text Frequency-Inverse Document Frequency) 분석 과정에서 중요한 단어로 판별되지 않아 그대로 두었다. <표 6>은 1964년에 생산된 '사업계획서' 기록물건의 '유치 대상 수출 상품 목록'에서 식별된 사어의 사례이다.

<표 6> 사어 식별 사례

'유치 대상 수출 상품 목록' 원문



誘致對象輸出商品

I 金屬製品	玩具、洋食器 남시바늘、時計밴드、부로-취 단추、라이타、시가렛트케이스、水道栓、園藝小道具、자물쇠 傘骨、체스나、兒童用自轉車、남시레、가위類、事務用品、建具取付具、注射바늘、볼트낫트.
II 合成樹脂製品	핸베드、玩具、목거리、단추、掌甲、신발、造花、구두창、남시대、시가렛트케이스、眼鏡테
III 纖維製品	造花、레-스、벨·카바·시-쓰、넥타이、핸카치、스카프·마후라、테불·크로스、뱁킨、人形、스타킨、室內靴
IV 木製品	玩具、피아노、베드멘트用具、레닉스·라켈、家具、스키-用品 골푸用品、工藝品、藤椅子類、竹細工品.
V 유리製品	食器類、豆電球、목거리、단추、玩具、人造眞珠、魔法瓶 크리스마스 照明具、花瓶、試驗器具.
VI 고무製品	玩具、風船、潛水服鏡、신발、공類
VII 農畜産物製品	果物통조림、버섯통조림、葛布蛋紙、乾草스립퍼、畜産加工品
VIII 其他製品	蓄電池、漁網、貝단추、各種부라쉬、人造宝石、其他 選定基準에 適格한 商品.

식별 가능 단어: 부로-취(브로치), 시가렛트 케이스(담배 케이스), 볼트낫트(볼트, 너트), , 핸카치(손수건), 마후라(머플러), 테불·크로스(테이블보), 냅킨(냅킨), 스타킨(스타킹), 배트민튼(배드민턴), 라켈(라켓), 골푸용품(골프용품), 스립퍼(슬리퍼), 부라쉬(브러시)
식별 불가 단어: 레-스(식별 불가), 벨·카바·시-쓰

https://doi.org/10.14404/JKSARM.2024.24.2.065

3.3.3 불요 문자 제거 및 불용어 처리

불요 문자 제거와 불용어 처리는 자연어 처리에서 부딪히게 되는 희박성과 차원 축소 문제의 상당 부분 해결할 수 있지만, 거꾸로 분석자의 선입관이 작용하여 텍스트의 의미를 왜곡할 수 있다. 본 연구에서는 한국어 자연어 처리에서 관행적으로 불요 문자로 통용되는 한 개 음절 문자, 숫자, 영문 단어만을 제거했다. 불용어는 공문서라는 특성상 문서 공통 서식 형태, 결재문 등 특정 형태 문서의 서식에 고정적으로 사용된 단어는 불용어로 처리하였다. '계', '계장', '과장', '부장', '상무이사', '전무이사', '사장' 등의 직위도 결재 서식에 반복되어 나오지만, 본문에서 성명과 직위가 병기되는 경우도 많으므로 불용어로 처리하지 않았다.

<표 7> 불용어 목록

문서 공통 서식	'문서번호', '수신', '참조', '제목', '쪽 번호', '한국수출산업공단', '본문', '첨부', '경유', '경유 수신 참조', '유첨', '품의', '별표', '상태 명'
결재 서식	'결재', '기안자', '기안 년 월일', '결제 년 월일', '시행 연월일', '정서', '기장', '발송'
사후 독음 과정에서 삽입된 용어	'철 번호', '건 목차', '건 번호', '건 제목', '원본 쪽 번호', 0(독음 불가 문자 처리)

수기로 작성된 종이 문서를 광학 인식으로 인식시켰을 때, 종이에 사전 인쇄된 서식과 수기 작성 후 추가된 관인, 날인 역시 문자로 인식할 수 있어 불용어 처리 단계에서 처리되어야 한다. 구로공단 기록물에서는 문서 오른쪽 아래에 인쇄된 '한국수출산업공단'이 불용어 처리되었고, 관인의 경우 인식이 불가한 경우가 대부분이었다.

3.3.4 개인정보 처리

1960년대 구로공단 기록물은 주민등록번호 제도가 시행된 1968년 이전에 생성된 문서가 대부분이므로 주민등록번호, 여권 번호 등 개인식별정보에 해당하는 내용은 발견되지 않았다. 1968년 이후 기록물 역시 개인정보에 해당하는 내용은 정규표현식을 통해 검색하였으나 발견되지 않았다.

3.3.5 말뭉치 생성

전처리 과정에서 생성된 114개의 텍스트 파일이 포함된 기록물 전체 원본에 대한 정보는 지역(보존 장소), 생산 연도, 권 번호, 면수(쪽수), 문서 유형, 권 제목, 건 제목, 건 쪽 번호, 업무적 가치, 역사적 가치, 독음 여부, 해제 여부를 메타데이터로 기술하였다. 앞서 언급한 것처럼, 철-건 구조에서 편철 정보는 단순한 물리적 정리 수준의 정보로 파악되었으며, 개별 건의 구조를 분석한 결과 개별 건은 개조식으로 작성된 소제목-내용, 소제목-문자/숫자 번호-내용 구성, 문자/숫자 번호-내용으로 구성될 수는 있으나, 그 규칙이 통일되어 있지 않아 일정한 규칙 구조를 찾아낼 수는 없었다. 최종적으로 텍스트 파일을 합치고 원본에 대한 구조 정보를 작성하여 하나의 텍스트 파일(1,200쪽, 8만여 단어)로 말뭉치를 생성했다.

3.4 개체명 인식을 통한 색인 및 기술정보 추출

개체명 인식은 데이터 정제 이후 자연어 처리에서 핵심 절차인 형태소 분석과 품사 태깅의 결과와 문맥에 추론을 결합하여 개체의 성격을 인식하여 최종적인 개체명으로 돌려주게 된다. 이는 기록물에 포함된 인물, 직위/직급, 조직(기관, 회사, 위원회 등), 사건에 대한 일종의 다차원 색인으로서, 아카이브가 보유한 기록물에 대한 검색과

https://doi.org/10.14404/JKSARM.2024.24.2.065

활용 차원에서 구체적인 결과를 보일 수 있을 것으로 기대된다. 이렇게 추출된 개체명은 일종의 패싯으로서, 영구기록물 기술 규칙의 '색인어'(국가기록원, 2022a), 메타데이터와 기록관리 메타데이터 표준의 '주제-유형 & 주제명' 메타데이터(국가기록원, 2022b)로 입력될 수 있다.

<표 8> 기록물 철/건 별 색인 및 기술정보

구분	기록물 철		기록물 건
영구 기록물 메타 데이터	색인어		기능어: 업무 과정, 활동 등을 나타내는 명사나 명사구 인명: 인물 이름 지명: 장소 이름 단체명: 기관, 회사, 법인, 학교 이름 주제명: 주제어 사건명: 행사명, 회의명, 기념일, 사건/사고 명
한시 기록물 메타 데이터	주제	유형	일반주제명: 인명, 단체명, 지명을 제외한 일반주제 인명: 기록물의 내용에 있거나 관련된 주요한 인명 단체명: 기록물의 내용에 있거나 관련된 주요한 단체명 지명: 기록물의 내용에 있거나, 관련된 주요 지역의 지명 중요한 내용을 간결하게 표현해 주는 단어
	주제명		

3.4.1 모델 선택

개체명 인식을 수행하기 위하여 한국전자통신연구원(이하 ETRI) SW-SoC 융합 연구개발센터가 운영하는 OpenAI API·DATA의 언어 분석 기술 API를 사용하였다(https://aiopen.etri.re.kr/serviceList). 이를 선정한 이유는 ETRI는 공공기관으로서의 보안에 대한 신뢰성과 접근성을 가지고 있으며, 제공되는 API가 영역별로 잘 구분되어 있고, 해당 영역의 설명과 문서화, 고객지원이 충실하다고 판단했기 때문이다. 그리고 API 내에서 형태소 분석 및 품사 태깅까지 모두 지원하여 별도의 한글 형태소 분석이 필요 없이 End-To-End로 모델을 사용할 수 있다는 장점도 고려하였다. ETRI API의 언어 모델은 구글의 BERT(Bidirectional Encoder Representations from Transformers)를 기반으로 한국어 데이터로 학습한 언어 모델이다. 다중 언어를 처리할 수 있는 멀티모달 모델로서, CJK(Chinese, Japan, Korea)에 대한 별도의 문자 처리 모듈을 가지고 있으며, 12개의 은닉층과 약 30,000개의 형태소/어근의 한국어 사전을 내장하고 있다.

<표 9> ETRI AI API·DATA의 API 목록 (출처: https://aiopen.etri.re.kr/serviceList)

기술명	API
언어 분석(문어)	형태소 분석, 개체명 인식, 동음이의어 분석, 다의어 분석, 의존 구문분석, 의미역 인식)
언어 분석(구어)	형태소 분석, 개체명 인식
어휘 관계 분석	문장 인식, 어휘 정보, 동음이의어 정보, 다의어 정보, 어휘 간 유사도 분석, 개체 연결, 상호참조 해결
질의응답	질문분석, 기계독해, 위키백과 QA, 법률 QA
음성인식	중국어, 일본어, 독일어, 불어, 스페인어, 러시아어, 베트남어, 아랍어, 태국어, 이탈리아어, 포르투칼어 , 말레이어, 인도네시아어, 광둥어
발음평가	한국어, 영어
이미지 인식	객체 검출, 사람 속성 검출, 얼굴 비식별화, 사람 상태 이해
동영상 인식	장면 분할

https://doi.org/10.14404/JKSARM.2024.24.2.065

ETRI가 제공하는 API 중에서 개체명 인식은 언어 분석 기술 영역에서 개체명 인식 API로 제공된다. 개체명 인식 API의 기본 구조는 API 서비스 요청을 포함하는 HTTP Request 헤더와 API 서비스에서 처리 후 JSON(객체 표현 문서규격) 형태의 텍스트 데이터로 결과를 전송하는 HTTP Response 헤더로 이루어져 있다.

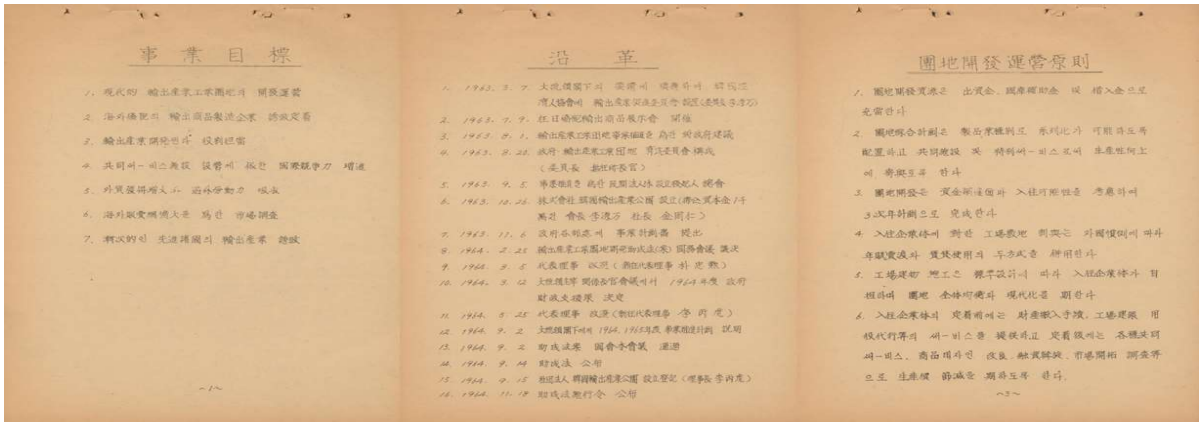
<표 10>은 API 사용자가 직접 입력하는 파라미터 정보이며, 인증을 위한 접근키와 분석 코드, 분석 문장으로 이루어진 argument이다. argument의 개별 인자 중에서 분석 코드 필드에서 형태소 분석, 어휘 의미 분석, 개체명 인식, 의존 구문 인식, 의미역 인식을 지원하는 것을 확인할 수 있다. 이를 통해 개체명 인식 API를 통해 텍스트 전처리에서 필요한 형태소 분석 등 다양한 텍스트 처리를 포괄적으로 지원할 수 있다는 것을 알 수 있다.

<표 10> 개체명 인식 API의 파라미터 (출처: https://aiopen.etri.re.kr/guide/WiseNLU)

Field 명	타입	필수 여부	설명
access_key	String	○	API 사용을 위해 ETRI에서 발급한 사용자 API Key
argument	Object	○	API 사용 요청 시 분석을 위해 전달할 내용
analysis_code	String	○	요청할 분석 코드로서 요청할 수 있는 분석 요청은 아래와 같음 형태소 분석 (문어/구어) : "morp", 어휘 의미 분석 (동음이의어 분석)(문어) : "wsd" 어휘 의미 분석 (다의어 분석)(문어) : "wsd_poly" 개체명 인식 (문어/구어) : "ner" 의존 구문 분석 (문어) : "dparse" 의미역 인식 (문어) : "srl"
text	String	○	분석할 자연어 문장으로서 UTF-8으로 인코딩된 텍스트만 지원

결과는 전체 텍스트를 문장 단위로 나누고 분석 내용을 계층화해서 구성된 JSON 파일 형태로 제공된다. 주요 분석 항목은 'text(개체명 어휘)', 'type(개체명 형태)', 'begin(개체명을 구성하는 첫 형태소의 ID)', 'end(개체명을 구성하는 끝 형태소의 ID)', 'weight(개체명 인식 가중치)', 'common_noun(일반명사 여부)'이다.

3.4.2 모델 평가


事業目標

沿革

團地開發運營原則

<그림 3> 사업계획서 총론 원문

기록물의 내용에 포함된 개체명의 색인화가 개체명 인식을 통해 가능한지를 확인하기 위해 우선 기록물 1건에 대하여 개체명 인식을 수행했다. 시험 방법은 특정 기록물 건의 일부에 대해 개체명 인식을 진행하여 그 결과를

https://doi.org/10.14404/JKSARM.2024.24.2.065

분석하고, 결과가 유효하면 점차 범위를 넓혀 적용하는 순서로 진행했다.

1차 시험 대상은 구로공단 기록물 중 1964년에 생산된 '사업계획서'의 총론 일부를 선정했다(<그림 3 참조>). 해당 건은 공단 조성 사업계획을 전체적으로 조망한 문서이며, 공단 조성 초기의 주요 인명, 세부 사업 명칭 등이 다수 출현하여 개체명 인식의 유효성 검증에 적절할 것으로 판단했기 때문이다. 그리고 자연어 처리의 특성상 긴 텍스트 처리에 드는 시간을 고려할 때, 사업계획서의 총론 부문만을 먼저 시험 적용하는 것이 좋을 것으로 판단했다. 총론은 사업 목표, 연혁, 단지개발 운영 원칙으로 구성되어 있으며 3쪽의 페이지와 1, 127개의 문자(구두점 포함)로 구성되어 있다. 먼저 기록물에서 개체명 인식을 수행할 수 있는지 확인하기 위하여 API를 호출하는 모듈을 구현하고, ETRI로부터 받은 인증키와 함께 분석 코드 argument로 'ner', 텍스트 argument에 총론의 원문 텍스트(<그림 3> 참조) 중 '연혁'과 '단지개발 운영 원칙'의 특정 문단 하나씩을 대상으로 개체명 인식을 수행했다.

<표 11> 서비스 API 호출 코드

```
# -*- coding:utf-8 -*-

import urllib3
import json

# 언어 분석 기술(문어) API 호출 URL
openApiURL = "http://aiopen.etri.re.kr:8000/WiseNLU"

# # 언어 분석 기술(구어) API 호출 URL
# openApiURL = "http://aiopen.etri.re.kr:8000/WiseNLU_spoken"

# API 접근키. 뒷 자리는 익명 처리
accessKey = "d5c346f2-48bb-45d1-b087-OOOOOO"

# 개체명 인식 분석 코드
analysisCode = "ner"

# 분석 대상 텍스트 파일 생성(gVally_ner.py)
with open("/gVally_ner.py", "r", encoding='utf8') as text:
    while True:
        line = text.readline()
        if not line:
            break
# 요청 파일 생성
requestJson = {
    "argument": {
        "text": text,
        "analysis_code": analysisCode
    }
}
# http 전송
http = urllib3.PoolManager()
response = http.request(
    "POST",
    openApiURL,
    headers={"Content-Type": "application/json; charset=UTF-8",
```


https://doi.org/10.14404/JKSARM.2024.24.2.065

```
        "Authorization":  accessKey},
    body = json.dumps(requestJson)
)
#  결과 출력
print("[responseCode] " + str(response.status))
print("[responBody]")
print(str(response.data, "utf-8"))
```

다음은 개체명 분석 결과의 성공과 오류 사례이다. 개체명 분석의 내용은 텍스트를 개별 문장으로 분리하고 문장의 단어를 정규화된 형태소로 분류하여 개체를 식별하게 된다. JSON 포맷으로 해당 문장별 id, 원문(text), 개체 형태(type), 해당 개체의 문장에서의 토큰 위치(begin, end) 등을 키(key): 값(value)으로 결과를 제시한다.

<표 12> 사업계획서 총론의 개체명 인식 결과 사례

분석 문장	분석 결과	개체

“5. 25 대표이사 개선(신임 대표이사 이병호)”	{
	"id":0,
	"text":"5 ", "type":"QT_ORDER ",
	"begin":0, "end":0, "weight":.297218, "common_noun":0
	},
	{
	"id":1,
	"text":"25 ", "type":"QT_AGE ",
	"begin":2, "end":2, "weight":.151487, "common_noun":0
	},
	{
	"id":2,
	"text":"대표이사 ", "type":"CV_POSITION ",
	"begin":3, "end":4, "weight":.393578, "common_noun":0
	},
	{
	"id":3,
	"text":"신임대표이사 ", "type":"CV_POSITION ",
	"begin":7, "end":9, "weight":.646705, "common_noun":0
	},
	{
	"id":4, "text":"이병호 ", "type":"PS_NAME ",
	"begin":10, "end":10, "weight":.607636, "common_noun":0
	}
		서수 “5”
		나이 “25”(오류)
		직급 “대표이사”, “신임 대표이사”
		이름 “이병호”

“입주기업체의 정착 전에는 재산 반입 수속, 공장건축 용역 대행 등의 써-비스를 제공하고 정착 후에는 각종 공동 써-비스, 상품 데자인 개량, 융자 간선, 시장개척 조사 등으로 생산가 절감을 기하도록 한다.”	"id":0 , "text":"비스 ", "type":"PS_NAME ", "begin":33, "end":33, "weight":.0722384, "common_noun":0	이름: “비스”(오류)

잘못 인식된된 ‘써-비스’라는 단어는 형태소 분석 과정에서 ‘써’, ‘-’, ‘비스’로 어근을 분리하고 ‘써’는 조사, ‘-’는 기호, ‘비스’를 이름으로 인식했다. 이는 과거에 외래어를 표기할 때 장모음을 하이픈으로 표시하는 관행을 반영하지 않은 것으로, 데이터 전처리 과정에서의 사어·신조어 처리에 오류가 있었던 것이 원인으로 판명되었다. 날짜를 의미하는 ‘25’가 나이로 인식된 것은 아직 충분히 데이터를 학습하지 않은 것으로 판단된다. 이에 따라

https://doi.org/10.14404/JKSARM.2024.24.2.065

사어를 다시 정리하고, 이번에는 메타데이터로서 가치가 없다고 판단되는 숫자 개체(QT_ORDER), 형태소 타입(morp, morp_eval)을 제외하고 총론 전체에 대해 개체명 인식을 다시 진행했다. <표 13>은 그 결과를 요약한 내용이다.

<표 13> 사업계획서 총론의 개체명 인식 최종 결과

대분류	세분류	개체명	개체 수	오류 수	인식 오류 개체	정확도 (%)
PERSON	PS_NAME (사람 이름)	'이병호'(4), 이원만'(1), '김주인''(1), '박충훈''(1)	8	0	-	100
CIVILIZATION	CV_POSITION (직위/직책 명칭, 스포츠 포지션)	'대통령'(2),재일교포'(1)', '위원장'(1), '무임소장관'(1), '회장(1)', '사장'(1), '대표이사'(2), 신임대표이사'(1), '이사장'(1)	13	2	재일교포, 신임대표이사	84.61
	CV_LAW (법/법률 명칭)	'수출산업공업단지개발조성법'(1), '조성법'(1), '조성법시행령'(1)	3	0	-	100
ORGANIZATION	OGG_POLITICS (정부/행정기관, 공공기관, 정치기관)	'수출산업촉진위원회'(1), '정부'(2), '국무회의'(1), '국회'(1)	6	1	정부각	83.33
	OGG_ECONOMY (경제 관련 기관/단체, 기업)	'한국 경제인협회'(1), '한국수출산업공단'(2)	3	0	-	100
LOCATION	LC_OTHERS (LC계열의 세부 유형이 아닌 기타 장소)	'수출산업공업단지'(3)	3	0	-	100
EVENT	EV_OTHERS (기타 사건/사고 명칭, ~사태)	'대통령 주재 관계 장관 회의'(1)	1	0	-	100

시험 결과, 전체 37개 중 34개를 정확하게 인식하여 91.98%의 정확도를 보여주었다. 7개 영역 중 <이름>, <법/법률 명칭>, <경제 관련 기관/단체, 기업>, <기타 장소>, <기타 사건/사고> 등 5개 영역은 100%의 정확하게 인식하였고, <직위/직책>과 <정부/행정기관, 공공기관, 정치기관>의 2개 영역에서 각각 84.61%, 83.33%의 정확도를 보여주었다.

'재일교포'와 '신임대표이사' 개체는 <직위/직책> 영역으로 잘못 인식했다. '재일교포'는 민족/종족 명칭에 해당하는 <CV_POSITION>으로 인식해야 하는데 <직위/직책>으로 인식된 것은 학습데이터의 불충분이 원인으로 생각되며, '신임대표이사'는 '신임'과 '대표이사'를 형태소 단위에서 분리해야 하는데, 단일 개체로 인식한 것이 원인으로 생각된다. '정부각'은 '정부각부처에 사업계획서 제출'이라는 원문에서 '정부각부처에'라는 음절에서 '정부', '각', '부처'로 형태소를 분리해야 하는데, 정부각'을 <경제 관련 기관/단체, 기업>으로 인식하고'부처'를 일반명사(NNG)로 인식한 것이 원인으로 판단된다.

학습데이터의 불충분은 앞으로 데이터가 축적되거나, 데이터 세트에 반영하여 해결될 수 있는 문제지만, 형태소 단위의 분리 과정에서의 오류는 한국어의 특성인 띄어쓰기 문제에서 발생했다. '신임대표이사', '정부각부처'를 '신임 대표이사', 정부 각 부처'로 띄어쓰기 오류를 교정하면 '대표이사'와 '정부'를 정확하게 개체명으로 인식했다. 이는 개체명 인식 이전 단계인 데이터 전처리 과정에서 철저하게 오·탈자 교정과 정규화가 필요한 것을 시사한다.

https://doi.org/10.14404/JKSARM.2024.24.2.065

3.4.3 개체명 인식 결과

1960년대~1970년대 구로공단 기록물 중 연도별로 사업계획, 이사회 및 총회, 감사, 토지, 타 기관 협력, 계약, 기타 등 주요 주제를 담은 대표적인 기록물 건 114건(약 1,200쪽 8만여 단어)에 등장하는 사람(PERSON)과 조직(ORGANIZATION)에 대한 개체명 인식을 진행했다. 두 번에 걸친 시험에서 개선점으로 파악되었던 사어 처리와 오·탈자, 띄어쓰기 등 데이터 전처리 과정에서의 정규화를 반복하여 최종적으로 314개의 조직/기관과 173개의 사람 개체를 인식하여 추출할 수 있었다(<표 14> 참조). 이렇게 추출된 개체명은 영구기록물의 색인어 영역의 인명과 단체명, 한시 기록물의 주제 유형의 인명, 단체 유형의 주제명으로 색인화하여 기록물의 검색과 활용에 즉시 사용될 수 있을 것으로 판단된다.

<표 14> 구로공단 기록물의 개체명 인식 최종 결과

연도	기록물 철	개체명	
		조직/기관	사람
1963	기공식 관계철	한국수출산업공단	이병호
1964	사업 보고 및 수지 결산서	중촌전기공업주식회사, 평화안경공업주식회사, 고등금속공업주식회사, 옥강라이트 공업소, 동화화성주식회사, 삼화제관주식회사, 백양산업주식회사, 동성산업주식회사주식회사, 조광공업주식회사, 조일공업주식회사, 삼화합성공업, 고속제지주식회사, 일흥스텐레스주식회사, 광양정밀보석제작소, 유광비니루공업소주식회사, 길천아동승물제작소, 연공업협동조합, 동화산업 주식회사, 고려석면고무공업, 국제산업공사, 동양미싱제조주식회사, 일동봉공주식회사, 대원수출전구제작소, 진해축전지공업소, 대판상공진흥회, 모국산업기술시찰단, 삼협프라스틱 주식회사, 일지출자동차 상회, 삼흥화학공업소, 동양광금공업소, 안전공업소	나종렬, 배찬두, 고시종, 장봉호, 김성선, 강병준, 이상범, 여상배, 오병수, 윤병원, 오복채, 하문상, 박칠동, 곽태석, 유용갑, 유한수, 신명호, 최태섭, 박을철, 이태원, 최한조, 이동일, 김홍조, 김익성, 송종호, 오인규 , 김창해, 유용갑, 탁시갑, 김려옥, 안인권, 강두화, 송종호, 오인규, 김창해, 유용갑, 탁시갑, 김려옥, 안인권, 강두화
	이사회 관계철	한국수출산업공단, 대일건설, 동화	이병호, 홍재선
	국방부 관계철	국방부, 한국수출산업공단, 서울신문	김성은, 이병호, 박충훈
1965	감사 관계철	중촌전기공업주식회사, 옥강라이트 공업소, 한국수출산업공단, 동성산업 주식회사, 한국식품공사, 남진양행 주식회사, 평화공업주식회사, 고등금속공업주식회사, 삼화합성공업주식회사, 조일공업주식회사	나종렬, 장봉호, 오상원, 윤희중, 임종수, 배찬두, 고시종, 오복심, 윤병원
	사업계획 관계철	한국수출산업공단	이원만, 김주인, 박충훈, 이병호
	대지조성 관계철	대일건설주식회사, 한국수출산업공단	
	총회 관계철	한국수출산업공단, 중촌전기공업주식회사, 평화안경공업주식회사, 고등금속공업주식회사, 옥강라이트공업소, 동화화성주식회사, 삼화제철주식회사, 백양산업주식회사, 동성산업주식회사, 조광공업주식회사, 조일공업주식회사, 고속제지주식회사, 일흥스텐레스주식회사, 광양정밀실석제작소, 유광비니루공업주식회사, 길천아동승물제작소, 연공업협동조합, 동화산업주식회사, 고려석면고무공업, 국제산업공사, 동양미싱제조주식회사, 일동봉공주식회사, 대원수출전구제작소, 진해축전지공업소	나종열, 배찬두. 고시종, 장봉호, 김성선, 강병준, 이근범, 여상배, 오병수, 윤병원, 하문상, 박칠동, 곽태석, 유용갑, 정한수, 신명호, 최태섭, 박응철, 이태윤, 최한희, 이동일, 전홍조, 김익성
1966	단지관리 관계철	한국수출산업공단, 대일건설주식회사, 심산산업주식회사, 조일공업주식회사	박남식, 명계복
	심의회 관계철	동남미네론화학공업㈜, 한국수출산업공단, 삼화제조주식회사, 인천공업단지 ,코리아크리스탈 공업사, 고등금속공업주식회사, 중촌전기공업주식회사, 동흥전기주식회사, 평화공업주식회사,	이규태, 정환무, 임호, 고시종, 나종열, 유일룡, 배찬두, 윤병원, 장봉호, 김선풍, 곽태석, 박성진,

https://doi.org/10.14404/JKSARM.2024.24.2.065

		조일공업주식회사, 옥강라이트 공업소, 대판화섬주식회사, 광양정밀보석주식회사, 대경물산주식회사, 협동신탁주식회사, 삼화합성공업주식회사, 협동신탁주식회사, 한국이기, 남진양행, 다옥편물, 한국공업은행, 한국산업은행	김용태
	소심의 관계철	구로동수출공업단지, 풍전공업주식회사, 안양모방주식회사, 주식회사 아동	장중균, 박용완
1967	서울도시계획사업(일단의공업단지조성) 조성도서	국방부, 한국수출산업공단	김성은, 이병호
1968	총회 서류철	한국수출산업공단	연일수
	이사회 서류철	경국산업주식회사, 한미산업사, 오륙실업, 한국수출산업공단, 한국모수공예주식회사, 우신산업, 진양화성, 유풍섬유공업주식회사, 대한전자공업주식회사, 인데어리어 디자이너스, 범한물산주식회사, 삼립화학주식회사, 협우산업주식회사, 국제냉동공업주식회사, 동국무역주식회사, 삼리염직주식회사, 유풍상사주식회사, 조일광학공업주식회사, 평화안경주식회사, 삼양금속주식회사, 아이맥주식회사, 효성물산주식회사, 일신산업주식회사, 유풍섬유공업주식회사, 삼흥화학, 삼주무역 주식회사, 대한조화공업사, 시대복장, 중앙공업, 홍주산업, 코스모산업, 원림산업 주식회사, 중앙공예물사, 홍우산업 주식회사, 마산방직, 대한통운, 상영산업, 동진기업 주식회사, 중앙계량국, 금광업연구소, 다옥편물, 대도섬유, 아동, 대양물산, 우미산업, 한국이기	방한규, 권오선, 김동석, 박용학, 안정희, 허민, 장기열, 염래문, 윤응노, 오광열, 박기주, 박종영, 주성일, 임광상, 지광열, 주동준, 배찬두, 박병주, 아서 타일러, 백영기, 박중영, 김원종, 김인수, 황신하, 이위형, 만연옥, 홍정호, 김영규, 공근초, 김명선, 오상근, 이종영, 도환, 박용완, 최용운, 배효갑, 김인득
1970	1~5단지 토지 관계철(용지과)	서울시, 한국수출산업공단, 한국산업은행, 감정원	김두남
	제1~4단지 소유권 이전 관계철	아이맥전자주식회사, 영등포구청, 오리온전자공업주식회사, 외환은행, 삼화완구, 기업은행	김성택
1971	제 3~4단지 계약 관계철	한국수출산업공단	김명화, 유지현
1972	기본 운영 관계철	한국수출산업공단	김재길
	정기총회 회의록	한국수출산업공단, KOTRA, 건설부, 인천시, 외환은행	
1973	공업단지 현황철	한국수출산업공단, 중소기업은행, KOTRA, 건설부, 인천시, 외환은행, 삼화합성주식회사, 상영금속주식회사, 동진기업주식회사, 대월 브레키 주식회사, 이시다베전기공업주식회사,삼브벨브 주식회사, 삼브벨브 주식회사 ,뉴코리아전자주식회사, Veriton West 사, 아남산업㈜, 도리우미제작소, 뉴코리아전자주식회사, 한일은행	우규호, 오복침, 오상은,김희범, 홍구표, 박순석 ,김향수
1977	도시산업 선교회 참고철	한국 교회사회 선교협의회, 한국 도시산업 선교 연합회, 영등포 도시산업 연합회, 감리교계 영등포 도시산업 선교회, 장로교계 도시산업 선교회, 경수도시산업 선교회, 영락교회, 세검교회, 중앙교회, 동부서울 도시산업선교회, 괴산동교회, 양평동교회, 홍릉교회, ㈜대협, ㈜한국마벨	조화순, 정진동, 안광수, 인명진, 요한페실 메일리, 조지송, 김경락, 인명진, 김봉태, 안경수, 황효남, 이정학, 김정현, 김석재, 김정현, 한승대, 이명만, 김용태, 김성희, 안광주

4. 시사점

3장의 실증적 시험 연구를 통하여 파악할 수 있었던 문제점과 원인, 고려 사항들은 반·비정형 데이터로서의 자연어 텍스트에서 발생하거나, 한국어로서의 특징에 의해서 발생한 것, 그리고 기록으로서의 고유한 특징에 의해 발생한 것들이 혼재되어 있다. 본 장에서는 개체명 인식 등 특정한 하위 영역으로서의 자연어 처리뿐만 아니라

https://doi.org/10.14404/JKSARM.2024.24.2.065

문서 분류, 정서 분석, 문서 요약 등 일반적인 자연어 처리 절차에서 공통으로 고려해야 할 상황과 시사점을 제시하고자 한다.

첫째, 목표 설정 및 모델 선정, 둘째, 데이터 전처리와 구조화, 셋째, 실제 언어 모델 적용으로 나누어 제시하였다. 이는 자연어 텍스트, 한국어의 특성, 기록의 특징에 따른 문제점들은 실제 기록관리 업무에 자연어 처리를 적용할 때의 구체적인 절차에서 그 역순으로 발생하는 것이 확인되었기 때문이다. 즉, 자연어 텍스트의 문제는 언어 모델의 적용에서 구체적으로 발현되며, 한국어로서의 특징과 기록으로서의 특징은 주로 데이터 전처리 과정에서 발현되는 문제인 것이 확인되었다. 그러므로 실제 기록 업무에서 자연어 처리를 적용하는 절차에 따른 고유한 문제로 환원하여 그 틀 안에서 해결책을 제시하는 것이 실천적으로 유의미하다고 판단했기 때문이다.

4.1 목표 설정 및 특징 파악과 모델 선정

자연어 처리를 실제로 적용할 때는 명확한 목적을 가져야 한다. 개체명 인식을 통한 색인어 추출로 메타데이터 기술 이용성 향상, 평가와 폐기를 위한 문서 요약, 주제 토픽을 통한 분류 등 뚜렷한 목표를 설정해야 하고, 목표를 달성했다는 것을 사후에 판단할 수 있는 지표를 설정해야 한다. 목표가 결정되었으면, 그에 해당하는 기록물의 구체적인 특징을 파악해야 한다. 대상의 구조(철-건-디지털 컴포넌트, 분류체계 등)와 단어 수, 쪽 수, 대상 전체와 주요 단어, 단어-문서 관계에서의 주요 출현 빈도, 상대적 중요도 등은 색인어 추출을 위한 개체명 인식과 기록 집합체 차원에서 시리즈 차원까지를 포괄하는 문서 분류 등에서는 해당 기록물의 계층별 색인과 기록의 층위별(시리즈별 등) 문서 분류에서는 중요하지만, 기록물 건 단위의 문서 요약, 정서 분석에서는 상대적으로 불필요하므로 과감히 생략할 수 있다.

<표 15> 특징 파악의 지표

구분	지표	세부 내용
질적 특징 파악	편철 여부 및 철 정보의 유효성	생산자가 단위 업무 분류체계 등 명확한 근거에 따라 편철을 진행하였는지에 대한 여부 과거의 종이 기록물의 경우 적절한 편철, 분철 작업이 이루어지지 않을 가능성이 상존
양적 특징 파악	기계 가독성 정도	종이 기록물은 광학 인식을 사용하여 텍스트 추출 시 추정 오류율 산정 디지털 기록물은 텍스트로 추출할 수 있는지, 별도의 광학 인식이 필요한 포맷인지 파악 종합적으로 기계 가독성이 확보된 기록물 건수와 전체 비율을 파악하여 대상 조정이나 기계 가독성 추가 확보 필요
	철-건 분포	전체 철별 건수의 분포 파악 전체 분포에서 특이한 철(과소 혹은 과대)은 편철을 무성의하게 수행했거나 분류가 어려운 건을 한꺼번에 편철하는 등 철 정보의 유효성이 낮을 가능성이 큼
	건 별 쪽수 및 단어 분포	건 별 단어 수의 분포 파악 특히 건 별 단어가 전체 분포와 비교해서 과소할 경우 언어 모델링 과정에서 특징 증가만 가져올 수 있으므로 필요에 따라서는 과감하게 대상에서 배제 필요
	TF-IDF	문서에 자주 등장하는 단어의 빈도수를 분석함으로써 문서별 토픽 파악 다수의 문서가 대상일 경우 TF-IDF 분석을 병행 목적에 부합하지 않는 토픽의 문서는 대상에서 제외할 수 있음

목적과 대상, 범위가 결정되고 대상 기록의 특징이 파악되었다면 그에 가장 적합한 모델을 선택한다. 이때 고려해야 할 점은 최신 기술, 뛰어난 성능보다는 검증된 신뢰성과 접근성이다. 예를 들어 텍스트 전처리 과정에서의 개인정보 처리에서는 사례를 통해 성능과 신뢰성이 검증된 정규표현식 같은 기법이 충분히 유효하며, 텍스트에

https://doi.org/10.14404/JKSARM.2024.24.2.065

대한 기본적인 분석은 최신의 트랜스포머 모델보다는 기존의 텍스트 처리에서 적용 사례가 많은 순환신경망 분석 기법이 자원의 배분과 그 결과의 품질을 비교했을 때 더 좋은 방법일 수 있다. 문서 분류 역시 문어와 구어를 포괄하는 대규모 언어 데이터를 학습한 언어 모델보다는 언론 기사, 도서 등 특정 영역의 데이터를 집중적으로 학습한 언어 모델의 성능이 높을 수 있다. 그리고 많은 시간과 컴퓨팅 자원을 소모하는 문제에 대한 현실적인 대안으로, 오픈 소스로 공개되거나 충분한 문서가 제공된 솔루션/모델을 사용하여 처리 과정의 과정과 결과를 최대한 교차 검증하여 기록뿐만 아니라 기록을 정보화하는 과정의 설명 책임을 확보할 필요가 있다.

4.2 데이터 전처리와 구조화

4.2.1 기계 가독성 확보와 정규화

기록물이 종이 문서를 스캔한 파일이라면 텍스트로 읽을 수 있는 포맷으로 변형해야 한다. 데이터베이스 구축 사업을 통해 과거의 종이 기록물들이 다수 디지털화되었지만, 결과물이 텍스트로 판독할 수 없는 이미지 기반의 pdf일 경우가 많다. 이런 경우는 광학 인식을 통해 텍스트로 변환해야 한다. 기록물이 디지털 파일이지만 서식을 사용한 상업용 문서작성 소프트웨어로 작성되었을 경우, 서식을 제거한 텍스트 파일이나 공개 데이터 포맷으로 변형해 주어야 한다. 실제로 데이터 분석이나 자연어 처리를 위하여 가장 시간과 자원을 소모하는 부분은 언어 모델의 적용이 아니라 기계 가독성 확보이다.

한글은 로만어 계열의 알파벳처럼 개별문자들이 나열되어 단어를 구성하지 않고, 초성과 중성, 종성을 조합하여 하나의 글자를 구성하는 방식이다. 그래서 UTF-8 이전 초기 표현 방식인 조합형과 완성형으로 인코딩된 텍스트를 UTF-8을 지원하는 텍스트 뷰어로 읽을 경우, 글자 깨짐 현상이 발생한다. 그러므로 1990년대 중후반까지 작성된 구형의 문서작성 소프트웨어의 인코딩 형식을 사용한 텍스트는 UTF-8 방식으로 변환이 필요하다.

4.2.2 오·탈자 교정과 사어·신조어 처리, 불요 문자와 불용어 제거

회의 녹취록 등 음성 기록을 텍스트로 전환한 경우, 수기로 작성한 종이 기록을 디지털화하였을 때 오·탈자에 대한 교정이 특히 필요하다. 음성 기록에서 사용되는 사투리 등 비표준어는 표준어로 통일시키거나 시소러스 등으로 비표준어와 표준어의 관계를 사상하여 표기된 기호 이면에 있는 의미 기반의 자연어 처리를 수행할 수 있어야 한다. 그리고 특정 단어가 오기되거나 연음 등으로 음절이 잘못 표기될 때 자연어 처리 과정에서 해당 단어는 별도의 단어로 파악되므로 오·탈자에 대한 기본적인 교정이 필요하다.

사전에 등재되어 있지 않은 사어와 신조어 역시 오·탈자로 인식된다. 과거의 기록 텍스트에서 출현하는 사어 처리는 기록관리 관점에서 중요한 작업이다. 관점에서 과거의 행정 문서의 경우 국어사전에 등재되지 않은 일본식 한자 조어(예. '유첨')가 다수 존재하며, 특정 영역에서 사용하는 특수한 단어(예. '추레라') 등이 존재한다. 이렇게 과거 기록물에 사용된 사어와 현대 수집기록에서 사용된 신조어는 사전 분석 단계의 빈도 분석(TF, Text Frequency) 분석 등을 통하여 중요성을 판단해야 한다. 중요하지 않은 단어일 경우 제거하며. 중요한 단어일 경우, 해당 단어의 대체 단어로 변경해야 한다. 특히 개체명 인식은 기록 사용성 확보를 위한 색인어 추출을 목표로 하므로 사어·신조어는 현재 사용되는 단어로 변경하는 것이 중요하다.

느낌표, 마침표 등 문장부호와 등호, 부등호 등 수학 표기 기호, 괄호, 기타 기호(#, & 등)는 자연어 처리에 불필요하므로 제거한다. 텍스트에서 무의미하거나 너무 반복적으로 사용되는 단어는 텍스트 말뭉치의 차원만 늘리거나 잡음으로 처리되므로 이런 단어는 불용어 처리 단계를 통해 원문 텍스트에서 사전에 제거해야 한다. 단, 감탄사가 텍스트의 정서를 반영하고 있고 향후 정서 분석이 필요할 수 있으면 이를 고려해서 처리해야 할 필요가

https://doi.org/10.14404/JKSARM.2024.24.2.065

있다. 불용어 처리는 통상적으로 불용어 사전 등을 사용하여 수행하지만, 해당 영역에 따라서 기록관리자의 질적인 판단이 필요할 수 있으므로 충분한 확인이 필요하다.

4.2.3 띄어쓰기 처리와 토큰화 및 품사 태깅

교착어 계열인 한국어는 어근에 조사가 붙어 격, 시제 등이 달라지므로 띄어쓰기를 기준으로 단어를 구분할 수 없다. 그리고 두 단어가 조합된 합성어의 경우 하나의 단어로 보고 붙여 쓰는 것이 원칙이나, 지켜지지 않는 경우가 유의미하게 많이 발견된다. 그러므로 띄어쓰기 규정에 최대한 맞추어 정리하는 과정이 필요하다. 그리고 과거 타자기로 작성된 기록의 경우 글자와 글자 사이의 공백 때문에 한 음절 단위로 띄어쓰기가 된 것으로 인식되는 사례가 있다. 이럴 때는 모든 띄어쓰기를 제거한 다음 형태소 분석기 등을 사용하여 단어를 추출하는 방법도 사용할 수 있다.

한국어의 체언(명사, 대명사, 수사)과 용언(동사, 형용사), 수식사와 조사 중 자연어 처리의 대상으로 포함될 품사를 선택한다. 앞서 언급한 자연어 처리의 목표와 대상, 범위를 설정할 때 그 목적이 개체명 인식이라면 체언 중 고유명사만을 대상으로 선택할 수 있고, 온톨로지 구성이라고 한다면 체언과 용언을 선택할 수 있다. 텍스트 분류, 특히 정서 분석이 목적이라고 한다면 용언만을 선택할 수 있을 것이다.

4.2.4 개인정보 처리

오래된 기록의 경우 개인정보가 익명 처리되지 않고 표기된다. 주민등록번호, 여권 번호 등 일정한 패턴이 있는 개인정보는 정규표현식이나 패턴 인식 등을 통해 찾아내어 블랙 마킹, 익명/가명 처리하거나 분석 대상에서 불용어 처리를 해야 한다. 문자로 이루어져 있고 특정한 패턴이 없는 출생지, 이름 등은 개체명 인식 등을 사전에 적용하여 처리할 수 있을 것이다.

4.3 개체명 인식 언어 모델의 확장 가능성

자연어 처리 중 개체명 인식은 개별 기록 건에 대한 개체명 인식 결과에서 더 나아가 기록 건과의 정보를 연결하여 다음과 같은 활용 방안을 생각할 수 있다.

첫째, 진화된 검색 방법이다. 기록 군의 개별 기록물 건 각각에서 개체명을 인식하여 서수, 관형 조사 등을 제외하고 일반명사까지 포함된 통합색인목록을 만들고, 개별 기록물 건 색인목록과 동일 관계를 연결하여 해당 색인을 포함하는 기록 건을 검색할 수 있다.

둘째, 기록 건의 주요 토픽을 파악할 수 있다. 첫째와 반대로 특정 기록 건에 포함된 색인을 검색할 수 있는데, 기록 건의 색인은 일반명사를 제외한 인명, 기관명 등 고유명사 개체명을 별도로 색인화하고 해당 색인의 TF 분석과 결합하면 특정 기록 건의 주요 토픽으로 생각할 수 있다.

셋째, 다차원 분류체계를 구성할 수 있다. 인명/직위/조직/지명 등 개체명의 태그 분류 시대 분류 등 다른 분류체계와 결합하는 아카이브의 분류체계를 고도화할 수 있다.

https://doi.org/10.14404/JKSARM.2024.24.2.065

5. 결론

기록관/아카이브의 수집 과정에서 발생했던 실제 기록물과 데이터를 대상으로 사전에 분석과 기획 단계에서의 유의점과 데이터 전처리 과정에서의 고려 사항, 그리고 가능한 최신의 언어 모델로 실제 텍스트 데이터를 처리하면서 파악할 수 있었던 교훈과 시사점을 통해 다음과 같은 결론을 얻을 수 있었다.

첫째, 데이터 전처리 과정에서의 기계 가독성과 정보 보전의 교환 관계였다. 이는 기계 가독성을 위해 전처리 과정에서 텍스트로 변환하는 과정에서 진행의 편의와 언어 모델의 처리 부담을 줄일 목적으로 말뭉치 텍스트를 순수 텍스트 파일이나 CSV 파일로 정규화하는 과정에서 수치 데이터, 임베딩 된 개체, 표 등의 정보가 제거되는 문제였다. 그러므로 이 문제는 문서의 구조와 표현, 내용 정보를 내부 패키지로 보전하는 ODF 도입을 통해 해결될 수 있을 것으로 생각된다. ODT(ODF에서 워드프로세서 응용프로그램의 문서 포맷) 내에 기술되어 있는 테이블 형태의 표 정보와 임베딩 된 개체 정보는 정규화 과정에서 별도의 데이터로 추출하여, 일종의 목차, 혹은 색인을 통해, 빅데이터 분석 등의 자료로 가공할 수 있을 것이다.

둘째, 적절한 학습데이터와 정보의 불균형 문제였다. 기록과 언어 모델의 말뭉치와 훈련 데이터 성격의 조화 여부는 모델의 성능과 결과 신뢰도와 직결되므로, 언어 모델의 선택과 방법론뿐만 아니라 학습데이터와 정보 균형에 대한 충분한 검토가 필요하다. 이를 해결하기 위하여 기록관리라는 특정한 업무 영역에서 전문 말뭉치와 균형 있는 데이터 세트를 만들고 기록관리 영역에 특화된 작은 거대 언어 모델을 만드는 점을 고려할 수 있을 것이다. 여기에서 발생하는 보안 문제는 기관 내부의 시스템에서 독자적으로 운영하는 형태로 해결하거나 공공 클라우드의 사례처럼, 공동의 인프라와 서비스를 구축하는 방안을 생각할 수 있다.

셋째, 자연어 처리를 특정 업무에 특정 요소기술을 반영하는 것에 그칠 것이 아니라, 다른 관련 기술과 조합하여 새로운 가치를 만들고, 기록 집합체 전체에 대해 지속적인 갱신 활동이 필요하다(김학래, 2022). 이는 원자료로서의 구체적인 기록의 내용에 접근하는 자연어 요소기술의 결과물이 상위 계층의 원자료로서 다시 발전하는 상승 구조를 생각할 수 있다. 예를 든다면 개체명 인식으로 기록의 식별과 메타데이터를 추출하고, 가까운 의미 거리에 있는 개체와의 관계성과 결합하여 사회 연결망, 혹은 토픽 맵으로 발전시킬 수 있으며, 식별된 조직 정보의 식별자를 키로 하여 LOD와 연계한 기록 네트워크를 생성할 수 있을 것이다.

더 나아가 생각해 볼 수 있는 문제는 인간 전문가와 인공지능의 협업 문제이다. 이는 인공지능의 성과를 인간 전문가가 어디까지 수용할 것이냐, 그렇다면 범위에 대한 기준이 있는가의 문제로 귀결된다. 기록관리의 인공지능 적용에서 공통으로 언급되는 것이 기록 보유(Retention)와 처분(Disposal)에 대해서 인공지능이 보유와 처분 여부를 기계적이고 객관적으로 판단할 수 있을 것이라는 가능성이다. 이는 인공지능이 가진 형식적 중립성에 대한 신뢰를 빌어, 전문가의 주관성에 좌우되는 기록의 가치평가를 수행한다는 것이다.

여기에 대해서 어떤 기록은 무엇이 중요하고 주요한 것인지 해독(Deciphering)이 필요하고, 평가에는 상식이 중요한 역할을 한다는 반론이 제기된다(Rolan, 2019). 이 문제는 인공지능에 의한 정량적, 기계적 판단과 기록 전문가의 정성적, 인간적 판단에 관한 문제로서, 인공지능의 설명 책임(XAI)과 기록관리자의 윤리 문제와 연결된다. 그러므로 기록 공동체와 사회 전반의 토론과 합의가 필요하며 과학 철학과 연구 윤리, 규범 윤리학 등의 학제간 연구가 이루어질 수 있을 것이다.

이제 인공지능은 경탄의 대상이 아니라 일상생활 속까지 깊이 침투했으며, 기록관리 영역 역시 이런 흐름에서 예외가 아닐 것이다. 기록관리 영역은 기술의 사용뿐만 아니라 서비스, 평가와 같은 질적 가치를 판단할 수 있는 공인된 기록 전문가에 의해 수행된다. 이런 흐름에서 기록관리자들이 전문가로서의 적극적인 대응과 실천에 본 연구가 작은 기여가 될 수 있기를 기대한다.

https://doi.org/10.14404/JKSARM.2024.24.2.065

참고문헌

강범모 (2014). 텍스트 맥락과 단어 의미: 잠재 의미 분석. 언어학, 68, 3-34. https://doi.org/10.17290/jlsk.2014..68.3

고명현, 김학동, 임헌영, 이유림, 지민규, 김원일 (2019). 효율적 대화 정보 예측을 위한 개체명 인식 연구. 방송공학회 논문지, 24(1), 58-66. https://doi.org/10.5909/JBE.2019.24.1.58

기록관리 메타데이터 표준. NAK 8:2022(v2.3).

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

김태영, 강주연, 김건, 오효정 (2018). 지능형 기록정보 서비스를 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). 기록관리 분야에서 한국어 자연어 처리 기술을 적용하기 위한 고려사항. 한국기록관리학회지, 22(4), 129-149. https://doi.org/10.14404/JKSARM.2022.22.4.129

문헌정보 - 기록관리 - 제1부: 개념과 원칙. ISO 15489-1 : 2016, 3.5, 8.3, 9.4.

안세진, 황현호, 임진희 (2022). 종이 기록 데이터화를 위한 AI-OCR 적용 사례연구. 정보관리학회지, 39(3), 165-193. https://doi.org/10.3743/KOSIM.2022.39.3.165

영구기록물 기술 규칙. NAK 13 : 2022(v2.1).

임수종 (2021). 초거대 인공지능 언어 모델 동향 분석. KOSTAT 통계플러스, 16, 70-85.

임진희 (2021). 공문서의 기계 가독형(Machine Readable) 전환 방법 제언. 기록학연구, 67, 99-138. https://doi.org/10.20923/kjas.2021.67.099

한국전자통신연구원 SW-SoC 융합 R&BD 센터 [발행년불명]. 언어 분석 기술. 공공 인공지능 오픈 API·Data 서비스 포털. 출처: https://aiopen.etri.re.kr/guide/WiseNLU

한국전자통신연구원 SW-SoC 융합 R&BD 센터 [발행년불명]. 제공 API. 공공 인공지능 오픈 API·Data 서비스 포털. 출처: https://aiopen.etri.re.kr/serviceList

한미경 (2020). 내한 선교사 편지(1884-1942)와 디지털 아카이브. 파주: 보고사.

ratsgo (2017). NLP의 기본 절차와 Lexical Analysis. Ratsgo's blog. 출처: https://ratsgo.github.io/natural%20language%20processing/2017/03/22/lexicon/

Bak, G. (2012). Continuous classification: Capturing dyanamic relationships among information resources. Archival Science, 12(3), 287-318. https://doi.org/10.1007/s10502-012-9171-8

Colavizza, G., Blanke, T., Jeurgens, C., & Noordegraaf, J. (2021). Archives and AI: An Overview of Current Debates and Future Perspectives. Journal on Computing and Cultural Heritage, 15(1), 1-15. https://doi.org/10.1145/3479010

Colavizza, G., Ehrmann, M., & Bortoluzzi, F. (2019) Index-Driven Digitization and Indexation of Historical Archives. Front. Digit. Humanities, 6, 1-16.

Rolan, G., Humphries, G., Jeffrey, L., Samaras, E., Antsoupova, T., & Stuart, K. (2019). More human than human? Artificial intelligence in the archive. Archives and Manuscripts, 47(2), 1-25. https://doi.org/10.1080/01576895.2018.1502088

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

Ahn, Sejin, Hwang, Hyunho, & Yim, Junhee (2022). A Case Study on the Application of AI-OCR for Data Transformation of Paper Records. Journal of the Korean Society for Information Management, 39(3), 165-193. https://doi.org/10.3743/KOSIM.2022.39.3.165

Archival Description Rules. NAK 13 : 2022(v2.1).

ETRI SW-SoC Convergence R&BD Center [n.d.]. Language analysis techniques. Public AI Open API·Data Service Portal. Available: https://aiopen.etri.re.kr/guide/WiseNLU

ETRI SW-SoC Convergence R&BD Center [n.d.]. Provision API. Public AI Open API·Data Service Portal. Available: https://aiopen.etri.re.kr/serviceList

Go, Myunghyun, Kim, Hakdong, Lim, Heonyeong, Lee, Yurim, Jee, Minkyu, & Kim, Wonil (2019). A Study on Named Entity Recognition for Effective Dialogue Information Prediction. Journal of Broadcast Engineering, 24(1), 58-66. https://doi.org/10.5909/JBE.2019.24.1.58

Han, Mi-Kyoung (2020). Letters from protestant missionaries in Korea (1884-1942) & digital archive. Paju: bogosa.

Information and documentation - Records management - Part 1: Concepts and principles. ISO 15489-1 : 2016, 3.5, 8.3, 9.4.

Kang, Beom-mo (2014). Text Context and Word Meaning: Latent Semantic Analysis. EONEOHAG, 68, 3-34. https://doi.org/10.17290/jlsk.2014..68.3

Kim, Haklae (2022). Considerations for Applying Korean Natural Language Processing Technology in Records Management. Journal of Korean Society of Archives and Records Management, 22(4), 129-149. https://doi.org/10.14404/JKSARM.2022.22.4.129

Kim, In hu & Kim, Seong hee (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. https://doi.org/10.3743/KOSIM.2022.39.3.293

Kim, Tae-Young, Gang, Ju-Yeon, Kim, Geon, & Oh, Hyo-Jung (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. https://doi.org/10.14404/JKSARM.2018.18.4.149

Lim, Soojong (2021). An Analysis of Trends in the Super-Gigantic AI Language Model. Datascience. KOSTAT Statistics Plus, 16, 70-85.

Metadata Standard for Records and Archives Management. NAK 8:2022(v2.3).

ratsgo (2017). Basic Procedure for NLP & Lexical Analysis. Ratsgo's blog. Available: https://ratsgo.github.io/natural%20language%20processing/2017/03/22/lexicon/

Yim, Jin Hee (2021). Suggestions on how to convert official documents to Machine Readable. The Korean Journal of Archival Studies, 67, 99-138. https://doi.org/10.20923/kjas.2021.67.099