國土計劃
pISSN: 1226-7147, eISSN: 2383-9171
https://doi.org/10.17208/jkpa.2024.12.59.7.169

Journal of Korea Planning Association Vol.59, No.7 (2024)
대한국토·도시계획학회지 「국토계획」 제59권 제7호 pp.169-180

개체명 인식 모델링을 통한 120 다산콜 민원텍스트의 공간정보 추출 및 기초분석

Spatial Information Extraction and Basic Analysis from 120 Dasan Call Civil Complaint Texts through Named Entity Recognition Modeling

박진홍* · 강민규**
Park, Jin-Hong · Kang, Min-Gyu

Abstract

This study aimed to fine-tune an advanced Named Entity Recognition (NER) model optimized for spatial information extraction to enhance the usability of urban complaint text data, such as Dasan Call. We trained an existing NER model—which previously could not differentiate various spatial hierarchies—with a fine-grained, tagged named entity corpus. As a result, the updated NER model demonstrated excellent performance in recognizing address information and distinguishing between different layers of spatial information: districts, roads, and building numbers. By integrating this model with geocoding and GIS techniques, we successfully extracted address information from Dasan Call civil complaint texts and converted it into point data for basic analysis and visualization based on Seoul district boundaries. Prior to this research, opportunities to analyze civil complaint big data were limited, despite its value in reflecting urgent urban issues and providing active, real voices of citizens. This study can foster local authorities' ability to identify citizens' urban policy demands previously unexplored due to a lack of AI and big data technologies. Future studies will be conducted with higher-level statistical analyses based on micro spatial units, using the new NER model.

주제어 자연어 처리, 개체명 인식, 미세 조정, 120 다산콜, 모두의 말뭉치
Keywords Natural Language Processing (NLP), Named Entity Recognition (NER), Fine-tuning, 120 Dasan Call, Modu Corpus

Ⅰ. 서 론

최근 ChatGPT를 비롯한 거대언어모형(Large Language Model, LLM)의 대중화를 앞당긴 자연어 처리는 텍스트를 수치적으로 변환하거나, 이에 대해 통계적인 방법론을 적용하여 컴퓨터에게 텍스트를 이해시키거나 활용할 수 있도록 하는 기술이다(델립 라오·브라이언 맥머핸, 2021). 인공지능(Artificial Intelligence, AI)에 대한 관심과 중요성이 지속적으로 증가하는 가운데, AI 관련 연구 동향은 모델링 중심에서 데이터 중심으로 전환되고 있다. 즉, 모델 성능 개선을 위한 코딩 및 알고리즘 개발에 중점을 두기보다, 데이터의 품질 향상이 더 중요하고 효과적이라는 주장에 힘이 실리고 있다(Yang, 2021).

이처럼 자연어 처리 기법과 데이터 중심의 AI 연구가 각광을 받는 가운데, 서울시정에 대한 민원상담 서비스를 제공하는 120다산콜재단(이하 다산콜) 데이터는 국토·도시 연구 분야의 새로운 가능성을 열어주고 있다. 2017년 출범한 다산콜은 약 420명의 전문 상담 인력을 바탕으로 서울시정에 대한 종합적인 민원상담 서비스를 제공하는 공공 콜센터다. 전화, 문자, 챗봇, 수어, 외국어 등

* Ph.D Student, Department of Urban Administration, University of Seoul (First Author: park21@uos.ac.kr)
** Associate Professor, Department of Urban Administration, University of Seoul (Corresponding Author: mgkang23@uos.ac.kr)

여러 채널을 통한 다양한 유형의 민원을 처리하고 있으며, 그 양은 하루 평균 2만여 건에 달하고, 연간 약 700만 건에 달하는 민원텍스트 빅데이터를 실시간 생산하고 있다(120다산콜재단, 2022). 또한 전문상담 인력이 직접 민원 정보를 처리하는 데이터 레이블링(labeling) 작업을 수행하고 있어 다산콜 자료는 텍스트 빅데이터 기반 AI 도시 연구에 적합한 특성을 갖추고 있다.

특히 민원자료는 시민들의 자발적 신고로 취합되며, 도시문제 및 정책 수요를 진단할 수 있는 특수한 도시 데이터다(Hong et al., 2016). 시민에게 지역 현안과 관련된 도시공공서비스를 효율적이고 신속하게 제공해야 하는 지방정부에게 있어 민원 데이터는 도시정책 수요를 발굴하고 서비스 시설의 입지선정을 돕는 핵심적인 근거자료라고 볼 수 있다(Lee et al., 2019; Olivos et al., 2022).

다산콜과 유사한 민원서비스 플랫폼으로 뉴욕 311 콜센터(이하 NYC 311)가 있다. 이는 뉴욕시의 주요 정보를 구득하거나 응급서비스를 제외한 모든 정부 서비스를 제공받을 수 있는 공공 채널이다. 주목할 점은 NYC 311의 경우 'Look up Service Requests'라는 민원 맵핑 서비스에서 주소 정보 기반의 해상도 높은 민원 분포 현황을 실시간으로 공개하고, 이를 지속적으로 업데이트하고 있다는 점이다. 이 맵핑 서비스는 지도의 공간적 범위에 따라 개별 민원 또는 민원 다발지역(cluster)의 주소 정보뿐만 아니라 민원의 유형 및 처리 현황 등도 함께 제공한다. 이러한 공공 서비스는 시민들의 커뮤니티에 대한 관심을 도모하고 민원 처리 현황을 실시간으로 파악할 수 있어 정책 효능감을 높일 수 있다(서울연구원, 2014). 또한, NYC OpenData 플랫폼에서 민원 맵핑 서비스의 원시 자료(311 Service Requests from 2010 to Present)를 별도로 제공한다. 이 원시 자료는 민원 접수 일자, 관할 부서, 민원 유형뿐만 아니라 민원 발생 지점의 세부 주소명과 좌표와 같은 중요 정보를 비롯한 총 41개 변수를 포함하고 있어 이를 활용한 민원분석연구의 가능성이 높다고 볼 수 있다.

한편, 다산콜은 스마트서울맵의 시민말씀지도 플랫폼과의 연계하에 민원 맵핑 서비스를 제공 중이다. 그러나 민원 정보는 자치구, 행정동 등 행정구역 단위에서 집계되며, 그 분류 항목이 도로, 환경, 주택 등 대분류 체계로 제공된다는 한계가 있다. 물론 기간 설정에 따른 데이터의 집계가 매우 편리하다는 장점이 있으나 민원과 관련된 기타 세부 정보는 공개되지 않고 있다. 이처럼 다산콜의 민원 맵핑 서비스는 공간정보를 비롯하여 데이터의 해상도가 높지 않아 상대적으로 도시정책의 수요 파악에 어려움이 따른다.

따라서 본 연구의 목적은 다산콜 데이터와 같이 시민의 정책 수요 및 효능감을 높일 수 있는 민원 데이터의 활용 가능성을 제고할 수 있는 방법으로, 텍스트 데이터 기반의 세부 공간정보 추출이 가능한 새로운 개체명 인식 모델을 개발하는 것이다. 이 모델의 개발 과정은 다산콜뿐만 아니라 민원을 비롯한 여러 지자체와 기관의 텍스트 데이터의 활용도를 높이기 위해 상세히 서술되었다. 이를 활용한다면 NYC OpenData 플랫폼에서 제공하는 311 콜센터 데이터와 같이 좌표 단위의 민원위치정보를 추출할 수 있다. 이러한 고해상도의 민원위치정보 식별 및 추출 성과는 민원 맵핑 서비스의 품질을 개선하고 다양한 민원활용연구의 가능성이 커진다는 것을 시사한다.

본 논문의 구성은 다음과 같다. 2장에서 개체명 인식 모델의 학문적 배경인 자연어 처리의 개념과 분석기법을 검토하고, 민원데이터를 활용한 텍스트마이닝 선행연구를 정리하였다. 그리고 이론 및 선행연구 검토 내용을 바탕으로 연구의 차별성을 도출하였다. 3장에서 세부 공간정보 추출이 가능한 새로운 개체명 인식 모델의 개발 및 데이터 전처리 과정을 상세히 공개하였다. 4장과 5장에 걸쳐 새로운 모델의 성능 평가와 해당 모델링으로 추출된 민원데이터를 활용한 시범적 연구의 기초 분석 및 시각화 결과를 보고하고 결론을 작성하였다.

II. 자연어 처리 및 관련 선행연구 검토

1. 자연어 처리의 개념과 분석 기법

자연어 처리(Natural Language Processing, NLP)란 텍스트를 수치적으로 변환하거나 이에 대해 통계적인 방법론을 적용하여 컴퓨터에게 텍스트를 이해시키거나 활용할 수 있도록 하는 학문 분야다(델립 라오·브라이언 맥머핸, 2021). 이는 크게 컴퓨터에게 사람의 언어를 이해시키는 과정인 자연어 이해(Natural Language Understanding, NLU)와 컴퓨터가 사람이 이해할 수 있는 언어로 생각을 표현하는 과정인 자연어 생성(Natural Language Generation)으로 분류된다(쇼홈 고시·드와이트 거닝, 2020). 이때의 자연어란 구체적으로 인간이 상호 의사소통하기 위해 사용하는 단어, 소리 등이 포함된 언어를 뜻한다(임희석·고려대학교 자연어 처리연구실, 2020). 컴퓨터가 이러한 자연어를 이해하거나 분석하기 위해서는 컴퓨터에게 수치화된 텍스트 데이터를 입력시키는 작업이 선행되어야 하며, 이를 단어 표현(word representation)이라고 한다. 또는 벡터로 텍스트 데이터를 수치화하기에 단어 벡터(vector) 또는 단어 임베딩(embedding)으로도 불린다(전창욱 외, 2022).

자연어 처리 분야에서는 수치화된 텍스트 데이터를 다양하게 전처리 또는 모델링할 수 있는 방법이 지속적으로 개발되어 왔다. 본격적인 처리 작업에 앞서 기본 분석 단위인 '토큰(token)'의 개념을 이해하는 것이 필요하다. 한글의 경우 글자 형태로 본다면 음절 또는 자모 수준까지 최소 분석 단위로 설정할 수도 있으나, 기본적으로 "더 이상 나누면 안 되는 최소의 의미 있는 단위"를 뜻하는 토큰(token)이라는 용어를 분석 단위로 사용한다(박건숙, 2022).[1)]

이러한 토큰 기준으로 문장 또는 말뭉치(corpus)를 분리해주는 토큰화 이외에도 토큰의 어법상 품사를 할당해주는 PoS 태그

(Parts-of-Speech tagging), 고유 명사를 식별 및 사전 정의된 범주에 매핑하는 개체명 인식(Named Entity Recogniton, NER), 조사·관형사 등 문장의 의미에 실질적 영향을 주지 않는 어법적 요소에 속하는 불용어(stop word) 제거, 여러 단어 형태를 하나의 표준 형식으로 변환하는 텍스트 정규화(text normalization), 동음이의어를 다르게 매핑하는 단어 중의성 해결(word sense disambiguation), 그리고 마침표를 비롯하여 다양한 방식으로 문장 경계를 인식하는 작업(sentence boundary detection) 등이 있다(쇼홈 고시·드와이트 거닝, 2020). 그 밖에도 텍스트 데이터를 대상으로 정보 추출, 질의 응답, 번역, 요약, 생성, 추론, 분류 등의 다양한 응용 작업이 가능한 모델링도 수행할 수 있다(임희석·고려대학교 자연어 처리연구실, 2020).

자연어 처리 분야는 어텐션(attention) 메커니즘에 기반한 트랜스포머(transformer) 모델이 등장하며(Vaswani et al., 2017) 정보 손실 또는 기울기 소실(vanishing gradient) 문제 등 기존 순환 신경망(Recurrent Neural Network, 이하 RNN)[2] 구조의 한계를 크게 보완할 수 있었다. 순차적으로 나열되는 시퀀스 데이터의 특성상 문장이 길어질수록 참고하는 전체 입력 문장이 많아지며 정보가 손실되거나 출력 문장의 정확도가 떨어지는 문제가 발생하는데, 특정 시점에 예측 단어와 관련된 단어만 집중적으로 보는 어텐션으로 해결할 수 있었기 때문이다(유원준·안상준, 2022). 이처럼 트랜스포머 모델은 입력 및 출력 문장 간의 거리가 멀어질수록 성능이 저하되는 장기 의존성(long-term dependency) 문제를 크게 개선하며 이후 언어모형의 대중화에 기여한 BERT(Bidirectional Encoder Representations from Transformers) 또는 GPT(Generative Pre-trained Transformer) 모델 개발의 토대를 제공하며 자연어 처리 분야의 발전을 견인하였다(쇼홈 고시·드와이트 거닝, 2020; 루이스 턴스톨 외, 2022).

2. 민원데이터의 텍스트 마이닝 연구

자연어 처리의 넓은 개념에서 텍스트는 사람의 언어뿐만 아니라 통화, 로그 기록 등 일련의 숫자와 문자 전반을 총칭한다(쇼홈 고시·드와이트 거닝, 2020). 다만, 도시계획 및 정책 분야에서는 주로 텍스트 마이닝이라는 키워드 아래 민원을 비롯한 실생활의 언어 텍스트에 관한 분석적 연구가 수행되었다(김선재·이수기, 2020; 박건철, 2020; 이재혁 외, 2018). 이중 민원분석연구는 다양한 지자체에서 생산된 민원에 대하여 텍스트 마이닝을 수행한 사례와 서울디지털재단에서 편찬한 민원데이터 분석 및 활용에 대한 연구보고서로 크게 나뉜다.

텍스트 마이닝 연구는 대체로 빈도 정보를 바탕으로 핵심어를 추출하고 이를 활용한 네트워크 분석, 연관성 분석, 단어 구름, 토픽 모델링 등의 절차로 이루어진다. 이재혁 외(2018)는 시흥시 생태계서비스 정책수립을 위한 근거자료를 마련하기 위해 환경 관련 민원과 도시계획 문건에 대한 텍스트 마이닝을 수행하였다. 박영빈 외(2022)는 다양한 장애인 정책수요를 반영하기 위해 장애인 키워드로 추출한 국민권익위원회 민원데이터에 대한 텍스트 분석을 통해 기존의 일상생활에서 간과되었던 점자 표기와 보도 턱에 대한 실질적인 장애인 관련 민원수요를 확인할 수 있었다.

한편, 김현종 외(2018)가 지적한 바와 같이, 일반적인 빈도 분석은 연구 주제와 무관하지만 빈번히 등장하는 일상어를 핵심어로 추출하는 문제를 발생시킨다. 이 연구는 이러한 한계를 보완하기 위해 핵심어 분석과 연관성 분석을 연계하는 동시에 이를 두 차례에 걸쳐 수행하는 계층적 연관성 분석을 적용하여 부산시 민원을 심층적으로 분석하였다. 다만, 이처럼 추출된 토픽들이 현실의 도시문제와 동떨어져 있을 가능성이 있다. 이를 보완하기 위해 하재현 외(2019)의 연구와 같이 전문가 델파이 조사를 추가하는 연구 설계도 가능하다. 이로써 분석 결과로 도출된 서울시의 주요 도시문제 및 이슈에 대한 토픽을 재검토함으로써 실질적인 도시 정책수요를 파악할 수 있다. 그러나 이러한 연구들은 정형적인 텍스트 분석 패턴을 답습하므로 방법론적 참신성이 부족하다는 한계가 있다고 볼 수 있다. 또한, 정형의 키워드를 활용한 응용 연구이므로 현재 다산콜 데이터와 같이 복잡다단한 전처리가 필요한 민원자료의 활용 연구에 대한 직접적인 시사점을 찾기 어렵다.

서울디지털재단에서 수행한 세 편의 연구보고서(박건철 외, 2019; 박건철, 2020)는 본 연구와 같이 도시민원에 대한 실질적인 전처리 및 활용 가능성 제고를 목표하였다는 점에서 일맥상통하다. 박건철 외(2019)는 구체적으로 데이터 분석 환경 및 코드 사용법을 공개하고 이를 적용한 서울시 민원데이터 활용 사례를 소개하였다. 박건철·백수진(2018)과 박건철(2020)은 구체적으로 각각 구로구와 강남구 민원 데이터 사례를 집중적으로 분석하여 상세한 분석 결과와 정책적 시사점을 도출하였다. 그러나 이 역시 앞선 정형적인 텍스트 분석 패턴에 대한 적용에 머무르고 있어 자연어 처리 기법에 대해 상세히 참고하기 어렵다.

3. 연구의 차별성

본 연구의 차별성은 기존의 정형적인 텍스트 분석 패턴에서 벗어나 자연어 처리 기법 중 하나인 개체명 인식(Named Entity Recognition, NER)을 다산콜 민원텍스트 데이터에 적용하여 그 활용 가능성을 제고하는 것이다. 이 과정에서 기존에 배포된 사전 훈련모형을 사용하지 않고 국립국어원 '모두의 말뭉치' 데이터로 미세 조정하여 공간정보 추출에 특화된 개체명 인식 모델을 새롭게 개발하였다.

기존 NER 모델의 경우 최대 14개 이하의 개체 범주만 인식할 수 있어 세부적인 지역 구분이 어렵다는 한계가 있다. 지역 정보에는 국가, 도시, 행정 구역 등 다양한 위계가 있음에도 불구하고 이 모든 지역 정보를 'LC(location)' 태그로만 인식하기 때문이다.

따라서 본 연구는 총 150개의 개체, 그리고 지역 정보의 경우 다양한 지역 위계를 세분류할 수 있는 국립국어원 모두의 말뭉치의 '개체명 분석 말뭉치'를 학습용 데이터로 활용하는 새로운 NER 모델을 개발하고, 그 과정 및 결과를 공유함으로써 도시계획 분야의 자연어 처리 활용 가능성에 기여하고자 하였다. 이를 통해 국어학, 인공지능학, 도시계획 및 정책 등 각 분야에 산재하여 연구되고 있는 도시데이터 자원과 자연어 처리 기법을 종합한 분석을 수행할 수 있었다.

Ⅲ. 데이터 및 분석방법

1. 연구범위 및 설계

본 연구는 2022년 1-6월 동안 발생한 다산콜 불법주정차 민원을 분석하였다. 불법주정차 민원은 상담내용에 대체로 주소정보를 포함하고 있어 공간정보 추출을 위한 개체명 분석에 적합한 자료다.[3] 다산콜은 서울시의 종합민원상담센터이므로 공간적 범위는 서울특별시로 한정하였다.

연구 설계의 핵심은 기존에 개발된 개체명 인식(NER) 모델을 미세 조정하여 행정동 이하의 공간정보를 인식할 수 있도록 학습시키는 것이다. 데이터 품질이 좋고 민원 경위가 자세히 적혀 있는 '답변' 변수에 새로운 NER 모델을 적용하여 도로명주소에 해당하는 공간정보를 추출하였다. 다음으로 네이버 API를 통한 지오코딩(geocoding)을 수행하여 도로명주소를 좌표로 변환함으로써 서울 전역에 걸쳐 다산콜 불법주정차 민원 포인트 데이터를 구득하였다. 이처럼 자연어 처리 모델링 및 데이터 전처리 과정을 거쳐 무결점 데이터를 구축하고 이를 활용한 기초 분석 및 시각화를 수행하였다.

2. 연구 자료

1) 다산콜 민원텍스트

현재 다산콜은 상담원들이 상담 이후 해당 민원내역을 텍스트로 변환 및 구축하는 시스템을 갖추고 있으며, 이 과정에서 지나친 강성·악성 민원을 정제하는 작업을 수행한다. 본 연구에서는 기관 간 협의를 통해 민원인의 이름과 전화번호가 삭제된 다산콜 민원텍스트 데이터를 구득한 후 분석을 수행하였다.

다산콜 데이터는 크게 '상담시간', '상담분야', '상담유형', '이관부서', '이관횟수', '질문', 그리고 '답변' 변수로 구성된 자료다. 이 항목들은 크게 상담 일자·분야·유형 및 이관정보 등 정제된 키워드로 코딩된 정형 데이터와 질문 및 답변 항목과 같이 그 내용과 형식이 통일되지 않고 구어체, 특수문자, 불용어 등이 혼재한 비정형 데이터로 구분된다.

이중 민원 데이터의 공간정보와 직간접적으로 연관된 변수는 '이관부서'와 '질문'이다. 우선, '이관부서' 변수의 경우 다산콜 민원 접수 시스템 특성상 결측치가 많다는 특징이 있다. 현재 다산콜 데이터의 경우 상담사가 민원을 담당부서로 접수시키는 '민원접수'와 상담사가 시민의 질의에 직접 답변을 진행하는 '행정상담'으로 구분되는데 연도별 평균 약 70%의 데이터가 이관부서 정보가 결측된 '행정상담'인 것으로 확인되었다. 나머지 약 30%의 '민원접수' 데이터는 대체로 "서울특별시 강남구 도시관리공단"과 같이 도시명, 자치구명, 그리고 기관명이 포함된 값을 가진다. 다만, '민원접수' 데이터 역시 최소 공간단위로 자치구 수준까지만 파악할 수 있다는 한계를 지닌다.

이러한 데이터 특성은 크게 두 가지 시사점을 제공한다. 첫째, 약 70% 정도의 이관부서가 결측된 '행정상담' 데이터의 위치정보를 구득할 수 있다면, 최소 수십만 건의 데이터를 공간적으로 분석 가능하다는 것을 의미한다. 둘째, 이관부서 정보 유무와 별개로 행정동, 지번 주소, 도로명주소 등 자치구 미만의 공간정보를 인식할 수 있는 NER 모델링이 가능하다면 미세한 공간분석 단위 수준의 도시 연구가 가능하다는 것이다. 좌표와 같이 행정동 이하의 NER 모델링이 가능하다는 것은 행정동 기반의 민원 분포 현황이나 여타 도시정책변수 간의 관계를 분석하는 것을 넘어, 포인트 기반의 민원 예측 모델링 또는 더욱 미시적인 도시 연구가 가능함을 의미한다.

본 연구는 결측된 이관부서 정보를 대리하여 공간정보를 추출할 수 있는 변수로 답변 텍스트를 활용하였다.[4] 이 변수들에는 실제 민원인과 상담원이 묻고 답한 상담 내용이 기록되어 있다. 특히 불법주정차와 같은 교통 민원의 경우 신고 내용과 함께 위치정보를 포함한 경우가 많은 동시에 그 표본 수 또한 충분하다는 점을 확인할 수 있었다. 이에 본 연구는 공간정보가 대체로 양호하게 기록되어 있는 불법주정차 데이터만을 대상으로 한 NER 모델링을 설계하였다.

불법주정차 데이터를 선별하기 위해 '상담분야(중)' 변수의 '불법주정차(구도로)', '불법주정차(자동차전용도로)', '불법주정차(시도로)' 값과 '상담유형(소)' 변수의 '불법주정차' 값을 추출, 병합 후 중복치를 제거하였다. 이후 '답변' 변수에 대해 한글, 공백, 숫자, 그리고 '-' 특수문자를 제외한[5] 모든 문자를 제거하는 텍스트 전처리를 추가적으로 수행하였다.

2) 국립국어원 모두의 말뭉치

국립국어원(National Institute of Korean Language, NIKL)에서 관리하는 '모두의 말뭉치'는 한국어 말뭉치를 제공하는 데이터 플랫폼이다(국립국어원, 2021). 언어처리 산업 및 연구 분야 인공지능 학습에 필요한 한국어 빅데이터를 장기적으로 생산 및 관리하기 위해 2018년 구축되었다(김한샘 외, 2018). 현재 2023년 기준 51건의 다양한 한국어 말뭉치 데이터를 제공하고 있다. 문어체[6], 구어체, 인터넷 기반 등 다양한 어체로 구성되었을 뿐만

아니라 분석 목적에 따라 추론, 요약, 교정, 감성 분석 등에 최적화된 한국어 말뭉치를 신청 및 구득할 수 있다.

본 논문은 개체명 인식을 수행하기 위해 2019년[7]부터 2022년까지의 '개체명 분석 말뭉치'(이하 개체명 말뭉치)를 학습용 데이터로 활용하였다(국립국어원, 2020, 2022a, 2022b, 2023). 이 데이터들은 최종적으로 150개의 세부 개체명으로 구성되었고 각각 최소 총 300만 어절 이상의 분량으로 이루어져 있다(표 1). 모든 파일은 UTF-8로 인코딩된 JSON 형식으로 제공된다.

개체명 말뭉치의 데이터 구조는 〈표 2〉와 같다. 모든 문장에 대하여 'id'를 부여하였고 개별 문장은 원시 텍스트가 적힌 'form', 토큰별 정보가 있는 'word', 개체명에 관한 정보가 담긴 'NE' 속성들로 구성된다. 단, 'word'와 'NE'의 토큰 수가 서로 일치하지 않는다는 점을 주의해야 한다. 모든 토큰에 대해 개체명 분석을 수행한 것이 아니라 150개의 의미 분류 체계에 따른 표지에 부합하는 토큰만 별도로 'NE' 속성(attribute)에 추출 및 정리되었기 때문이다. 'word'와 'NE'는 공통으로 토큰별 id, 원시 토큰(form), 시작 위치(begin), 종료 위치(end)에 대한 속성값을 가진다. 한편, 'NE' 속성에는 'word'와 달리 개체명으로 선별된 토큰에 한하여 'label' 값이 추가된 것을 확인할 수 있다.

Table 1. Comparison of train data for NER modeling

Year	The no. of NE	Written vol.	Spoken vol.	Web vol.	Total vol.
2019	15*	2M	1M	-	3M
2020	150	-	-	5M	5M
2021		1M	2M	-	3M
2022		3M	1M	1M	5M

Note: It was re-categorized with 150 entities in 2021.

Table 2. An example of Korean NE corpus data

{ "id":	"NWRW1800000029.315.2.3",	
	"form"	"멕시코(Mexico) 파키스탄(Pakistan) 방글라데시(Bangladesh) 베트남(Vietnam) 역시(also) 해외파견(overseas-dispatch) 근로자의(of workers) 송금이(remittance) 한몫을 한다(does its part).",
	"word": [	{ "id": 1, "form": "멕시코(Mexico)", "begin": 0, "end": 3 },
		{ "id": 2, "form": "파키스탄(Pakistan)", "begin": 4, "end": 8 },
		...
		{ "id": 10, "form": "한다(does).", "begin": 40, "end": 43 }],
	"NE": [	{ "id": 1, "form": "멕시코(Mexico)", "label": "LC", "begin": 0, "end": 3 },
		{ "id": 2, "form": "파키스탄(Pakistan)", "label": "LC", "begin": 4, "end": 8 },
		...
		{ "id": 5, "form": "근로자(workers)", "label": "CV", "begin": 27, "end": 30 }] }

Note: The number of NEs is not equal to that of words.
This example was edited from NIKL (2020).

이 개체명 말뭉치는 NER 모델링을 위한 학습 데이터로 사용된다. 개체명 인식과 같은 태그 작업은 일종의 분류 모델링 중 하나이기에 지도 학습(supervised learning)에 속한다(유원준·안상준, 2022). 이는 특정한 단어가 어떠한 개체 범주에 속할지 사전 학습된 모델을 바탕으로 분류해내는 작업이기 때문이다. 이러한 텍스트 분류 작업을 수행하기 위해서는 문장을 토큰 단위로 분절한 입력 데이터와 각 단어의 순서에 상응하는 개체명 정보가 순차적으로 부여된 레이블 데이터가 쌍을 이루는 시퀀스 데이터(sequence data)[8]가 필요하다.

이 모델 학습용 데이터를 구축하기 위해 개체명 말뭉치를 JSON 파일 형태인 〈표 2〉에서 시퀀스 데이터 형태인 〈표 3〉으로 변환하는 전처리 작업을 수행하였다. 한국의 수도인 서울은 'LCP_CAPITALCITY', 마포구·용산구 등의 자치구는 'LCP_COUNTY', 홍익대학교와 같은 특정 건물 및 장소명은 'LC_OTHERS', 한강대로와 같은 도로명주소는 'AF_ROAD'로 세부적으로 개체명이 태그된 것을 확인할 수 있다. 그 밖에도 2행과 같이 여타 150개의 개체로 태그된 다양한 표본들이 존재한다.

다만, 〈표 3〉의 2번 행과 같이 토큰이 어느 개체에도 속하지 않는다는 의미의 'O' 태그만으로 구성된 표본(이하 미태그 데이터)의 비중이 문어체와 구어체·웹기반 데이터에서 각각 최소 약 30%와 80%를 차지하는 것을 확인하여 이에 대한 별도의 전처리가 필요할 것으로 판단하였다.

2019년부터 2022년까지 연도별로 제공되는 개체명 분석 말뭉치

Table 3. Transformed Modu NE corpus as sequence data

No.	Text data	Named entity (NE)
1	19일(19th) 낮(afternoon) 서울(Seoul) 마포구(Mapo-gu) 어울마당로(Eoulmadang-ro) 홍익대(Hongik University) 인근(nearby).	DT_DAY O O LCP_COUNTY O LC_OTHERS O
2	한국도(South Korea) 과거(past) 개발연대 시절(the industrialization period) 서독에(West Germany) 나간 광원과(miner) 간호사(nurse)...	O O O O O O O O O..
3	전설의(legend) 마꾸도나루도(MacDonald) 모르나?(Don't you know)	O OGG_FOOD O
...		
n	5일(5th) 서울(Seoul) 용산구(Yongsan-gu) 한강대로(Hangang-daero) 용산역(Yongsan station) 인근에 있는(nearby)...	DT_DAY LCP_CAPITALCITY LCP_COUNTY AF_ROAD...

Note: This example was extracted from NIKL (2020).

데이터를 어체별로 재구성하여 데이터셋을 구축한 결과는 〈표 4〉와 같다. 문어체 데이터는 총 390,568건이며 태그와 미태그 데이터의 비율이 각각 약 67%와 33%인 것을 확인하였다. 반면 구어체와 웹 데이터는 상대적으로 대화체 또는 단문 위주로 구축되어 표본 수 자체는 많으나 미태그 데이터의 비율이 최소 약 80%를 상회하는 것으로 나타났다.

어체별 데이터를 탐색한 결과, 문어체 데이터셋이 한국어 말뭉치와 그 태그 품질이 우수한 것을 확인하였다. 구어체 또는 웹 기반 데이터는 상대적으로 추임새, 유행어, 신조어 등 연구의 모델링의 학습에 적합하지 않은 일상적 용어를 상당수 차지하고 있기 때문이다. 본 연구는 앞선 개체명 말뭉치 데이터의 구득, 전처리 및 변환 과정을 통해 얻은 2019-2022년도 문어체 데이터셋을 기존 NER 모델을 미세 조정하기 위한 학습용 데이터로 사용하였다.

3. 분석 기법

1) 개체명 인식

개체명 인식(Named Entity Recognition, NER)이란 현실 세계에서 실제 이름이 있는 개체(named entity)가 사람, 장소, 조직 등 어떤 단어 유형에 속하는지 인식하는 작업을 의미한다(루이스 턴스톨 외, 2022; 유원준·안상준, 2022). 즉, 텍스트 내의 단어를 미리 정의해놓은 개체 범주에서 인식, 추출하는 기법이므로(한국정보통신기술협회, 2023) NER 모델은 특정 단어가 어떠한 범주에 속하는지 식별 및 할당하는 분류 모형에 해당한다(Li et al., 2022).

본 연구에서 사용한 NER 모델은 기존 BERT(Bidirectional Encoder Representation from Transformers)를 다양한 한국어 말뭉치 빅데이터로 사전 학습시킨 언어모델 중 하나인 KLUE-BERT이다(Park et al., 2021). 유원준·안상준(2022)을 참고하여 KLUE-BERT를 이용한 개체명 인식 모델링을 수행하였다.

단순히 공간정보를 추출하는 작업은 키워드 검색 기반의 접근법으로도 가능하다. 그러나 검색 기법은 키워드 사전에 포함되지 않는 단어를 추출할 수 없다. 더욱이 자연어는 하나의 개체를 다양한 이름으로 부를 뿐만 아니라 오탈자, 약어가 섞여 있고 띄어쓰기, 문법 등이 통일되지 않아 복잡한 형태를 가진다. 데이터 규모가 클수록 검색 기반의 접근은 한계에 봉착하며 정확성도 떨어진다.

이와 달리 NER은 문맥을 고려하여 개체명을 손쉽고 정확하게 판별할 수 있다. 예를 들어 "국토계획로 1길~100길"까지의 정보가 주어졌을 때, 키워드 검색 기법은 데이터셋의 일치 여부를 고려하므로 국토계획로 101길을 공간정보로 인식할 수 없다. 그러나 NER을 적용한다면 주어진 데이터를 학습하여 기존 데이터셋에 없는 국토계획로 101길을 위치정보로 인식할 수 있도록 모델을 개발할 수 있다. 데이터의 규모와 복잡성이 커질수록 스스로 학습 규칙을 찾아낼 수 있는 인공지능 모델의 유용성은 더욱 커진다.

기존의 NER 모델들은 한국해양대학교 자연언어처리연구실(2019)[9]과 네이버·창원대학교 적응지능연구실(2019)[10]이 각각 구축해 놓은 한국어 개체명 데이터로 학습되었다. 그러나 각 연구실이 구축해 놓은 개체명 학습 데이터는 개체명이 10개, 그리고 14개에 불과하여 개체명의 세분류가 어렵다는 한계가 있다(〈표 5〉 참고). 가령, 국가, 도시, 행정구역, 건물 등 다양한 위계의 공간정보를 오직 'LOC' 태그만으로 인식한다(그림 1). 그러므로 이러한 기존 모델을 세부적인 공간정보를 필요로 하는 도시학 분야에 그대로 적용하기에는 분석의 한계가 따른다.

2) 사전 학습과 미세 조정

본 연구는 이러한 기존 NER 모델을 도시연구 분야에 적합하도록 미세 조정하는 방법을 선택하였다. 미세 조정(fine-tuning)이란 전이 학습 방법 중 하나다. 트랜스포머의 등장으로 모델링 분야가 크게 발전하였으나 이를 적용할 수 있는 텍스트 빅데이터를 구득하고 대규모의 레이블링 작업을 수행하는 것은 현실적으로

Table 4. Sample size and ratio of Modu NER corpus

Dataset	Sample size	Tagged ratio	Non-tagged ratio
Written	390,568	67%	33%
Spoken	838,372	15%	85%
Web-based	1,569,959	20%	80%

Table 5. Named entity categories of Naver NER model

No.	Named Entity category	Tag	Definition
1	PERSON	PER	People names in real or virtual world
2	FIELD	FLD	Academical discipline, theory, law, technoloy
3	ARTIFACTS_WORKS	AFW	Artifacts created by humans
4	ORGANIZATION	ORG	Institution, organization, meeting/conference
5	**LOCATION**	**LOC**	**Name of regions and administrative districts**
6	CIVILAZATION	CVL	Terms related civilazation and culture
7	DATE	DAT	Date
8	TIME	TIM	Time
9	NUMBER	NUM	Number
10	EVENT	EVT	Event, accident, celebration
11	ANIMAL	ANIM	Animal
12	PLANT	PLT	Plant
13	MATERIAL	MAT	Metal, rock, chemical
14	TERM	TRM	Medical, IT terms, etc.

Source: Naver and CNU (2019)

```
result_list

[[('120다산콜재단은', 'ORG-B'),
  ('서울', 'LOC-B'),
  ('동대문구', 'LOC-B'),
  ('신설동에', 'LOC-B'),
  ('위치한', 'O'),
  ('서울시의', 'LOC-B'),
  ('종합민원상담센터이다', 'AFW-B')]
```

[('120 Dasancall foundation', 'ORG-B'), ('Seoul', 'LOC-B'), ('Dongdaemun-gu', 'LOC-B'), ('Sinseol-dong', 'LOC-B'), ('located in', 'O'), ('Seoul', 'LOC-B'), ('Complaint Center', 'AFW-B')]
※ This translation is provided for understanding but may not be precise.

Figure 1. The result of existing NER modeling

어렵다(루이스 턴스톨 외, 2022). 이러한 데이터의 현실적인 문제를 해결하기 위해 등장한 것이 전이 학습이다.

전이 학습(transfer learning)이란 "한 가지 문제에 대해 학습한 기능을 가져와서 비슷한 새로운 문제에 활용"하는 모델링 기법을 의미한다(Chollet, 2020). 예를 들어, 진돗개를 식별할 수 있는 모델로 시바견을 식별하는 데 사용하는 것이 전이 학습의 일종이라고 볼 수 있다. 수백만 개의 이미지 데이터로 학습하는 컴퓨터 비전(computer vision) 분야에서 빈번히 활용되는 방법으로, 대량의 데이터셋으로 미리 모델을 학습시키는 사전 학습(pre-training)과 이 사전학습모형(pre-trained model)을 다른 세부적 또는 새로운 작업에 적용하는 미세 조정(fine-tuning) 과정으로 크게 구성된다(루이스 턴스톨 외, 2022).

미세 조정은 크게 다른 분석 목적에 맞게 모델 용도를 변경하거나(repurpose) 모델의 가중치를 전체적으로 갱신하는(full fine-tuning) 방법으로 작업할 수 있다. 전자의 경우 사전학습된 가중치를 유지하므로 컴퓨팅 비용을 절감할 수 있으나 변경된 분석 목적에 맞는 새로운 학습용 데이터를 구득하는 것이 필요하고 후자는 가중치를 갱신하는 방법이기에 연산 비용이 크고 그 과정이 복잡할 수 있다(Dickson, 2023). 그러므로 본 연구의 미세 조정 과정은 분석 목적에 맞게 모델 용도를 변경하는(repurpose) 방법에 해당한다.

본 연구는 최대 150개, 지역 정보의 경우 13개의 세부 개체명이 태그된 국립국어원 모두의 말뭉치의 '개체명 분석 말뭉치(이하 개체명 말뭉치)' 데이터(표 6)로 미세 조정함으로써 세부적인 공간정보의 식별이 가능한 새로운 NER 모델(이하 Modu 모델)을 개발하였다.

이러한 작업을 통해 Modu 모델은 기존 모델과 달리 세부적인 위치정보를 인식할 수 있도록 새롭게 개발되었다. 요약하자면, Modu 모델은 불법주정차 민원의 공간정보를 식별하기 위해 자치구/행정동, 도로명주소, 건물번호를 서로 다르게 인식할 수 있다. Modu 모델링을 통해 다산콜 데이터를 도로명주소 기반의 좌표 포인트로 구축하는 것이 가능해졌다. 이로써 기존에 불가능하였던 다산콜 민원의 포인트 데이터 구축이 가능해지며 이를 기반으로 특정 민원 지점 또는 이를 집계한 행정동 단위의 도시 연구를 수행할 수 있다.

Table 6. Named entity categories of Modu NER corpus

No,	Category	Tag
1	PERSON(PS)	PS_NAME, PS_CHARACTER, PS_PET
2	STUDY_FIELD(FD)	FD_SCIENCE, FD_SOCIAL_SCIENCE, FD_MEDICINE, FD_ART, FD_HUMANITIES, FD_OTHERS
		...
5	**LOCATION(LC)**	LCP_COUNTRY, LCP_PROVINCE, LCP_COUNTY, LCP_CITY, LCP_CAPITALCITY, LCG_RIVER, LCG_OCEAN, LCG_BAY, LCG_MOUNTAIN, LCG_ISLAND, LCG_CONTINENT, LC_SPACE, LC_OTHERS
		...
15	TERM(TM)	TM_COLOR, TM_DIRECTION, TM_CLIMATE, TM_SHAPE, TM_CELL_TISSUE_ORGAN, TMM_DISEASE, TMM_DRUG, TMI_HW, TMI_SW, TMI_SITE, TMI_EMAIL, TMI_MODEL, TMI_SERVICE, TMI_PROJECT, TMIG_GENRE, TM_SPORTS

Source: NIKL (2022)

Ⅳ. 결과 및 논의

1. Modu NER 모델의 개발 과정 및 평가

Modu NER 모델 개발의 목표는 행정동 이하의 위치정보를 인식하는 것이다. 이를 위해 최대 150개의 개체명, 그리고 지역정보는 총 13개의 세부 개체명으로 태그된 모두의 말뭉치 플랫폼의 개체명 말뭉치 데이터로 KLUE-BERT를 미세조정하였다. 유원준·안상준(2022)에서 제공하는 코드와 가중치 설정을 유지하되 학습 및 훈련 데이터를 문어체 개체명 말뭉치로 대체하였다. 모델 학습에 적합한 데이터 형태로 변환하기 위해 JSON 형식의 원시 개체명 말뭉치를 시퀀스 데이터로 변환하였다.

그 결과, 기존 NER 모델링 결과(그림 1)와 달리 다양한 위계의 공간정보를 서로 다른 태그로 인식하도록 미세 조정된 것을 확인할 수 있었다. 〈그림 2〉와 같이 Modu 모델은 수도인 서울시를 'LCP_CAPITALCITY', 동대문구·신설동 등의 자치구 및 행정동 정보는 'LCP_COUNTY', 난계로28길은 도로명을 뜻하는 개체인 'AF_ROAD', 건물번호를 뜻하는 숫자는 'QT_ADDRESS'로 달리 분류하였다.

```
result_list

[[('120다산콜재단은', 'O'),
  ('서울특별시', 'LCP_CAPITALCITY'),
  ('동대문구', 'LCP_COUNTY'),
  ('신설동', 'LCP_COUNTY'),
  ('소재의', 'O'),
  ('서울시의', 'O'),
  ('종합민원상담센터다', 'O')],
 [('120다산콜재단은', 'O'),
  ('서울', 'LCP_CAPITALCITY'),
  ('동대문구', 'LCP_COUNTY'),
  ('난계로28길', 'AF_ROAD'),
  ('23', 'QT_ADDRESS'),
  ('소재의', 'O'),
  ('서울시의', 'O'),
  ('종합민원상담센터다', 'O')]]
```

[('120 Dasancall foundation', 'O'), ('Seoul', 'LCP_CAPITALCITY'), ('Dongdaemun-gu', 'LCP_COUNTY'), ('Nangye-ro 28-gil', 'AF_ROAD'), ('23', 'QT_ADDRESS'), ('located in', 'O'), ('of Seoul', 'O'), ('Complaint Center', 'O')]
※ This translation is provided for understanding but may not be precise.

Figure 2. The result of Modu NER modeling

이처럼 Modu 모델을 검증한 결과 'LCP_COUNTY', 'AF_ROAD', 'QT_ADDRESS'를 활용하여 도로명주소 정보를 추출하는 방향으로 연구 설계를 확정하였다. 정리하자면, '자치구-행정동-도로명-건물번호' 순서로 주소 정보가 나열될 수 있도록 'LCP_COUNTY', 'AF_ROAD', 'QT_ADDRESS' 태그들을 차례대로 추출하여 온전한 도로명주소 정보를 구득하고자 하였다.

물론 Modu 모델이 150개의 모든 개체명에 대하여 성능이 좋은 것은 아니다. 그러나 본 연구에서는 도로명주소 추출에 필요한 세 개의 태그가 주요 분석대상이므로 추가적인 미세조정은 진행하지 않고 현재 수준의 모델 성능으로 학습을 완료하였다. 세 개의 태그에 대하여 정밀도(precision), 재현율(recall), 그리고 이 둘의 조화 평균인 F1-score 등 모델 평가 지표에서 준수한 성능을 확보하였다(그림 3). 또한 모델링의 에포크가 증가할수록 손실값이 점차 유의미하게 낮아지는 것을 확인하였다(그림 4).

2. 다산콜 민원데이터의 기초 분석 및 시각화

2022년 1-6월 다산콜 불법주정차 민원 약 19만 건에 대하여 모델링, 지오코딩, 그리고 전처리를 수행하여 분석용 데이터를 확보하였다. 답변 텍스트에 Modu NER 모델을 적용하여 민원텍스트에 포함된 주소정보를 인식하고 관련 태그를 추출하여 도로명주소

Named Entity	Precision	Recall	F1-score	Support
LCP_COUNTY	0.84	0.90	0.87	1075
AF_ROAD	0.75	0.79	0.77	200
QT_ADDRESS	0.92	0.92	0.92	37

Figure 3. Validation results of main named entity tags

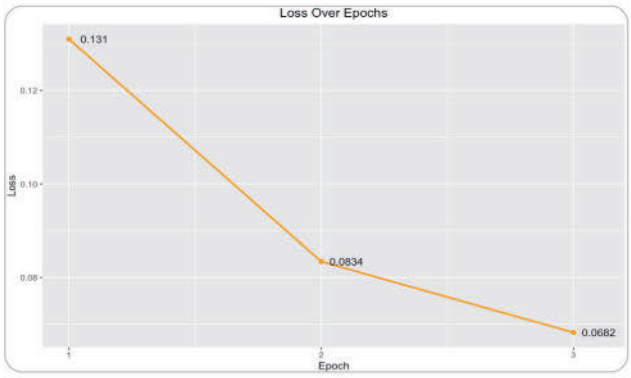

Figure 4. The loss graph of Modu NER model

정보를 확보하였다. 이 도로명주소에 지오코딩을 수행하여 각 주소정보에 해당하는 위도 및 경도로 이루어진 좌푯값을 구득하였다. 주소정보가 온전하지 않거나 모델링의 오작동으로 발생한 결측치를 제외하는 전처리 작업을 끝으로 최종적으로 약 13만 4천 건의 무결점 데이터를 확보하였다. 그 결과는 〈그림 5〉와 같다.

이처럼 Modu 모델링을 통해 기존 모델로는 불가능하였던 좌표 단위의 공간 데이터 구축이 가능해졌다. 본 연구는 새롭게 개발한 Modu 모델로 추출한 다산콜 민원데이터를 활용한 시범적 분석을 수행하였다. 〈그림 5〉와 같이 좌표 단위의 데이터이므로 포인트 자체 또는 이를 집계구와 같이 더 미시적인 단위에서 집계가 가능하다. 본 연구에서는 이를 활용한 시범적 분석사례를 제시하기 위해 〈그림 6〉과 같이 최종 다산콜 무결점 데이터를 행정동 단

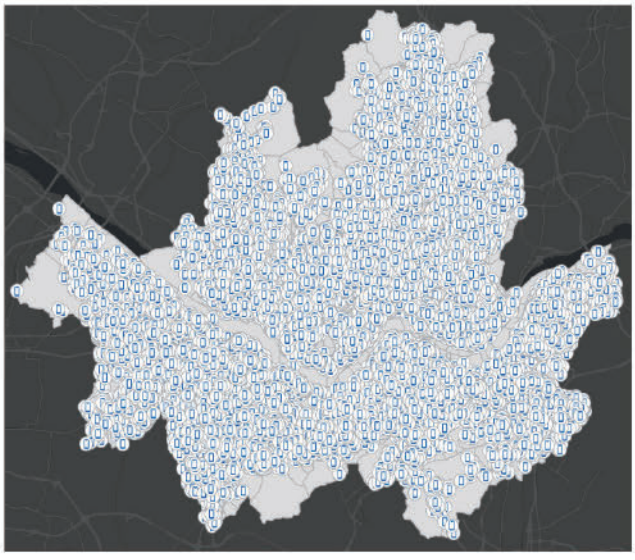

Figure 5. The results of Modu Modeling

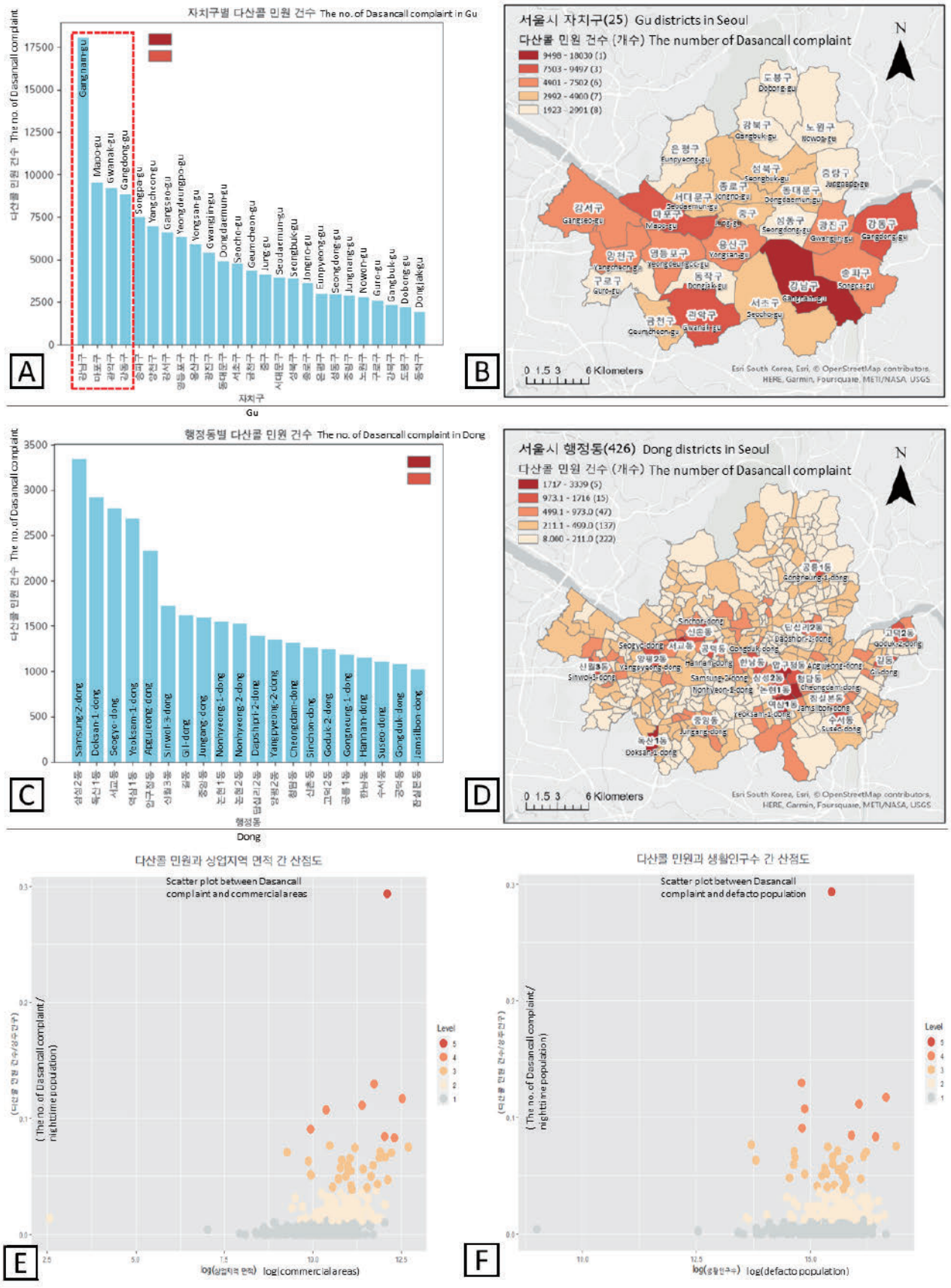

Figure 6. Basic analysis and visualization using Dasan Call data and urban environmental variables

위에서 집계 및 탐색하는 기초 분석을 포함하였다.

먼저, 자치구 및 행정동 수준에서 다산콜 불법주정차 민원 분포 현황을 살펴보았다. 두 공간 단위에서 민원 건수를 집계하였고 Natural breaks를 기준으로 단계 구분도를 작성하였다. 〈그림 6〉

의 B와 D와 같이 각각 자치구와 행정동 단위에서 다산콜 민원 건수를 집계하였다. 또한 민원 건수가 많은 상위 2개 집단에 포함되는 행정구역에 대해 그래프를 제작하였다.

자치구 수준에서는 강남구, 마포구, 관악구, 강동구 순서로 민원 건수가 많다는 것이 확인되었다(그림 6의 A). 한편, 행정동 수준의 기초 그래프는 양상이 약간 상이하였다. 전반적으로 민원 건수가 많은 행정동이 위 4개 자치구에 포함된 경우가 많았다. 그러나 독산1동처럼 행정동 중 두 번째로 민원이 많으나 자치구 수준의 그래프에서 드러나지 않은 경우도 발견하였다(그림 6C). 행정동 수준에서는 삼성2동, 독산1동, 서교동, 역삼1동, 압구정동, 신월3동, 길동, 중앙동, 논현1동, 논현2동, 답십리2동, 양평2동, 청담동, 신촌동, 고덕2동, 공릉1동, 한남동, 수서동, 공덕동, 잠실본동 순서로 민원 건수가 많았다.

민원의 공간적 분석과 함께 행정동 단위에서 결합할 수 있는 여타 도시환경변수와 다산콜 데이터 간의 관계를 탐색하는 분석도 수행하였다. 〈그림 6〉의 E와 F는 상주인구로 나눈 다산콜 민원 건수와 상업지역 면적 및 생활인구수 사이의 산점도다. 분석 대상이 불법주정차 민원이므로 불법주정차에 영향을 미칠 수 있는 도시환경변수를 취합 후 그 관계를 탐색적으로 분석하고자 하였다. 그 결과 상업지역의 면적이 넓고 생활인구수가 많을수록 민원 건수가 많은 행정동들이 상위 순위에 위치하는 것을 확인하였다.

V. 결 론

본 연구에서는 기존 NER 모델의 한계를 극복하기 위해 국립국어원 개체명 말뭉치 데이터를 미세조정하여 도시계획 및 정책 분야에 적합한 새로운 NER 모델을 개발하였다. 이를 통해 기존 모델에서는 구현 불가능했던 행정동 이하 단위의 민원데이터 공간정보화가 가능해졌다. 이를 활용하여 행정동, 집계구, 나아가 포인트 단위의 데이터를 활용한 응용분석이 가능할 것으로 예상한다. 새롭게 개발한 Modu NER 모델은 다산콜 데이터뿐만 아니라 다양한 종류의 텍스트 데이터에 적용할 수 있다. 개체명에 따라 성능 편차가 있으나 분석 목적에 따라 최대 150개의 여타 개체명 인식도 가능하므로 그 유용성은 더 클 것으로 기대한다.

연구의 의의는 크게 두 가지다. 첫째, 다산콜을 비롯한 국내 민원데이터의 활용 가능성을 높일 수 있는 새로운 개체명 인식을 개발하였다. 이 과정에서 도시학, 인공지능학, 국어학 등 다양한 부문의 분석기법 및 데이터셋을 접목한 융복합 연구의 가능성을 제시할 수 있었다. 둘째, 도시계획 및 정책 분야에 정형적으로 이루어지는 텍스트 분석 패턴을 벗어나 자연어 처리 기법을 적용한 연구를 수행하였다. 본 모델을 추후 깃허브(GitHub) 등 오픈소스 플랫폼에 배포하여 도시계획 및 정책 분야의 자연어 처리 활용사례 개발에 기여하고자 한다. 셋째, 도시정책변수로서의 민원데이터를 재발견하였다. 민원은 시민의 적극적인 의사와 목소리가 반영된 도시 데이터다. 특히 다산콜 데이터는 도시의 시급한 현안과 시민들의 정책 수요를 직접적으로 파악할 수 있는 중요한 자료이다. Modu NER 모델을 활용한 다산콜 데이터의 공간정보화는 행정 당국의 효율적이고 실효성 있는 정책 집행에 필요한 근거자료를 제공할 수 있다.

다만, 본 연구에서는 기술적 한계로 인해 가중치를 갱신하여 종합적인 모델 성능을 향상시키는 미세 조정은 수행하지 못하였다. 현재 Modu NER 모델은 공간정보를 비롯한 일부 개체명에 한하여 준수한 성능을 보인다. 후속 연구를 통해 충분한 컴퓨팅 파워와 기술적 지원이 더해진다면 추가적인 미세 조정이 이루어질 것으로 기대한다. 또한, 본 연구는 최종 무결점 데이터를 활용하여 기초적인 분석만 수행하였으므로 활용사례를 충분히 제시하지 못하였다는 한계를 수반한다. 이를 보완하기 위해 향후에는 불법주정차 이외의 민원에 대한 분석과 더불어 특정 민원과 관련된 도시환경변수 간의 관계를 심층적으로 분석하는 후속 연구를 수행함으로써 민원 활용사례 발굴에 이바지하고자 한다.

주1. 문장을 이러한 토큰 기준으로 분리 및 추출하는 과정을 토큰 처리 또는 토큰화(tokenization)라고 하고 이러한 토큰화 작업을 수행할 수 있도록 제작된 모델이나 프로그램을 토큰 처리기 또는 토크나이저(tokenizer)라고 한다(박건숙, 2022).

주2. RNN은 입력값과 이전 은닉 상태를 통해 도출한 파라미터로 결괏값과 현재 은닉 상태를 반환하는 입출력 구조를 가지고 있어 과거의 정보 또는 문맥 정보를 반영해야 하는 자연어 처리에 널리 쓰이게 되었다(임희석·고려대학교 자연어 처리연구실, 2020).

주3. 불법주정차 민원은 '상담분야(대)' 변수의 '교통' 부문 '상담분야(중)'의 '불법주정차' 항목을 이용하여 추출되었고 총 표본 크기는 약 19만 건이다.

주4. '질문' 변수에는 약 27%의 데이터에 '외부 민원 접수', '현장 민원 접수' 등과 같은 형식적인 행정 용어가 적혀 있고 '답변' 변수에 민원의 경위가 자세히 기록되어 있는 것을 확인하였다.

주5. 숫자와 '-' 특수문자는 도로명주소 인식에 필수적인 정보이므로 제외하지 않았다.

주6. 모든 연도에 걸쳐 문어체 데이터는 신문을 뜻한다.

주7. 2019년도 개체명 분석 말뭉치의 경우 한국전자통신연 구원(ETRI)의 '세부분류 개체명 가이드라인 2018' 지침에 따라 15개 분석 표지만으로 분류됐었으나(국립국어원, 2020) 2021년도에 차정원·신서인(2021)의 보고서에서 정립한 150개 세부 분류체계에 따라 재분류하는 동시에 추가로 데이터를 보강하였다(국립국어원, 2021).

주8. 시퀀스 데이터란 용어 그대로 순서가 있는 자료로서 특정 시점에 상응하는 병렬 구조의 데이터를 뜻한다. 대표적으로 시퀀스 또는 텍스트 데이터가 있으며 씽을 이루는 데이터 간의 순차성이 핵심이다(Lendave, 2021).

주9. https://github.com/kmounlp

주10. https://github.com/naver/nlp-challenge

인용문헌
References

1. 120다산콜재단, 2022. "120다산콜센터 15주년, 재단 5주년 맞이 신비전 선포", 서울특별시.
   120 Dasan Cal, 2022. "The Fifteenth Anniversary of Dasan Call", Seoul.
2. 국립국어원, 2020.「국립국어원 개체명 분석 말뭉치(버전 1.0)」, 문화체육관광부.
   National Institute of Korean Language(NIKL), 2020. *NIKL Named Entity Analysis Corpus (Version 1.0)*, Ministry of Culture, Sports and Tourism.
3. 국립국어원, 2021. "한국인처럼 자연스럽게 대화하는 인공지능 개발 추진 -국립국어원, 에스케이텔레콤과 인공지능 한국어 모델 개발 업무협약-", 문화체육관광부.
   National Institute of Korean Language (NIKL), 2021. "Business Agreement for Developing AI Korean Model between NIKL and SKT", Ministry of Culture, Sports and Tourism.
4. 국립국어원, 2022a.「국립국어원 개체명 분석 말뭉치 2020(버전 2.1)」, 문화체육관광부.
   National Institute of Korean Language (NIKL), 2022a. *NIKL Named Entity Analysis Corpus 2020 (Version 2.1)*, Ministry of Culture, Sports and Tourism.
5. 국립국어원, 2022b.「국립국어원 개체명 분석 말뭉치 2021(버전 1.0)」, 문화체육관광부.
   National Institute of Korean Language (NIKL), 2022b. *NIKL Named Entity Analysis Corpus 2021 (Version 1.0)*, Ministry of Culture, Sports and Tourism.
6. 국립국어원, 2023.「국립국어원 개체명 분석 말뭉치 2022」 , 문화체육관광부.
   National Institute of Korean Language (NIKL), 2023. *NIKL Named Entity Analysis Corpus 2022*, Ministry of Culture, Sports and Tourism.
7. 김선재·이수기, 2020. "수도권 2기 신도시 주거환경만족도 요인 분석: 웹크롤링과 텍스트 마이닝을 활용하여",「국토계획」, 55(7): 5-20.
   Kim, S.J. and Lee, S.G., 2020. "Determinants of Residential Environment Satisfaction in the Second-Generation New Towns of the Seoul Metropolitan Area Using Web Crawling and Text Mining", *Journal of Korea Planning Association*, 55(7): 5-20.
8. 김한샘·강승식·김일환·김진웅·김평·김학수·남길임·류법모·문한별·박선우·박소영·박진호·송상헌·안의정·옥철영·윤영민·윤태진·이도길·이상곤·황은하, 2018.「2018년 국어 말뭉치 연구 및 구축」, 국립국어원.
   Kim, H.S., Kang, S.S., Kim, I.H., Kim, J.W., Kim, P., Kim, H.S., Nam, G.I., Ryu, B.M., Mun, H.B., Park, S.W., Park, S.Y., Park, J.H., Song, S.H., Ahn, Y.J., Ok, C.Y., Yun, Y.M., Yun, T.J., Lee, D.G., Lee, S.G., and Hwang, Y.H., 2018. *2018 Korean Corpus Research and Establishment*, NIKL.
9. 김현종·이태헌·유승의·김나랑, 2018. 민원 "분석을 위한 텍스트 마이닝 기법 연구: 계층적 연관성 분석",「한국산업정보학회논문지」, 23(3): 13-24.
   Kim, H.J., Lee, T.H., Yu, S.Y., and Kim, N.R., 2018. "A Study on Text Mining Methods to Analyze Civil Complaints:Structured Association Analysis", *Journal of the Korea Industrial Information Systems Research*, 23(3): 13-24.
10. 델립 라오·브라이언 맥머핸, 2021.「파이토치로 배우는 자연어 처리」, 박해선 역, 서울: 한빛미디어.
    Rao, D. and McMahan, B., 2021. *Natural Language Processing with PyTorch*, Translated by Park, H.S., Seoul: Hanbit Media.
11. 루이스 턴스톨·레안드로 폰 베라·토마스 울프, 2022.「트랜스포머를 활용한 자연어 처리」, 박해선 역, 서울: 한빛미디어.
    Tunstall, L., Von Werra, L., and Wolf, T., 2022. *Natural Language Processing with Transformers*, Translated by Park, H.S., Seoul: Hanbit Media.
12. 박건숙, 2022.「파이썬으로 시작하는 한국어 정보 검색과 자연어 처리파이썬으로 시작하는 한국어 정보 검색과 자연어 처리」, 서울: 노드미디어.
    Park, G.S., 2022. *Korean Information Searching and Natural Langauge Processing with Python*, Seoul: Node Media.
13. 박건철, 2020.「텍스트마이닝 기법을 활용한 민원 빅데이터 분석: 강남구 편」, 서울: 서울디지털재단 데이터혁신팀.
    Park, G.C., 2020. *Civil Complaint Bigdata Analysis Using Text Mining Techniques: Gangnam-gu Case*, Seoul: Seoul Digital Foundation.
14. 박건철·김장현·김병준·이겨레·장재연·김남균·강찬희, 2019.「도시데이터표준분석모델: 민원분석편」, 서울: 서울디지털재단.
    Park, G.C., Kim, J.H., Kim, B.J., Lee, G.R., Jang, J.Y., Kim, N.G., and Kang, C.H., 2019. *Urban Data Standard Analysis Model: Civil Analysis*, Seoul: Seoul Digital Foundation.
15. 박건철·백수진, 2018.「민원 빅데이터 분석결과 보고서: 구로구 민선6기 편」, 서울: 서울디지털재단 디지털정책팀.
    Park, G.C. and Baek, S.J., 2018. *Civil Bigdata Anlayis Report: Guro-gu Case*, Seoul: Seoul Digital Foundation.
16. 박영빈·이정민·강동헌·은선덕·박지영, 2022. "텍스트 마이닝 기법을 이용한 '장애인' 키워드 민원 데이터 분석: 시각장애인을 중심으로",「재활복지공학회 논문지」, 16(2): 57-70.
    Park, Y.B., Lee, J.M., Kang, D.H., Eun, S.D., and Park, J.Y., 2022. "Civil Complaint Data Analysis Using Text Mining: Focusing on the Visually Impaired", *Journal of Rehabilitation Welfare Engineering & Assistive Technology*, 16(2): 57-70.
17. 쇼홈 고시·드와이트 거닝, 2020.「예제로 배우는 자연어 처리 기초: NLP 알고리즘, 텍스트 분류와 요약, 감성 분석」, 김창엽·최민환 역, 서울: 에이콘출판사.
    Ghosh, S. and Gunning, D., 2020. *Natural Language Processing Fundamentals*, Translated by Kim, C.E. and Choi, M.H., Seoul: Acorn.
18. 이재혁·박홍준·김일권·권혁수, 2018. "텍스트 마이닝 분석을 통한 생태계서비스 수요-공급의 이슈 차이분석 -시흥시 민원과 도시계획 자료를 활용하여-",「농촌계획」, 24(3): 63-71.
    Lee, J.H., Park, H.J., Kim, I,G., and Kwon, H.S., 2018. "Issue Difference of Ecosystem Service Demand and Supply through Text Mining Analysis: Case Study of Shiheung using Complaints and Urban Planning Materials", *Journal of The Korean Society of Rural Planning*, 24(3): 63-71.
19. 임희석·고려대학교 자연어처리연구실, 2020.「자연어처리바이

블·핵심이론·응용시스템·딥러닝」, 서울: 휴먼싸이언스.
Im, H.S. and Korea University NLP&AI Lab, 2020. *Natural Language Processing Bible*, Seoul: Human Science.

20. 전창욱·최태균·조중현·신성진, 2022.「텐서플로 2와 머신러닝으로 시작하는 자연어 처리 (개정2판)」, 파주: 위키북스.
Jeon, C.U., Choi, T.G., Cho, J.H., and Shin, S.J., 2022. *Natural Langauge Processing with Tensorflow2 and Machine Learning*, Paju: Wikibook.

21. 차정원·신서인, 2021.「2021년 개체명 분석 및 개체 연결 말뭉치 연구 분석」, 국립국어원.
Cha, J.W. and Shin, S.I., 2021. *2021 Named Entity Analysis and Entity Connection Corpus Research*, NIKL.

22. 하재현·기동환·이수기·안동욱, 2019. "4차 산업혁명 요소 기술을 통해 대응 가능한 서울시 도시문제 및 이슈 도출 -텍스트 마이닝 분석과 델파이 조사의 적용을 중심으로-",「서울도시연구」, 20(4): 1-21.
Ha, J.H., Ki, D.H., Lee, S.G., and An, D.W., 2019. "Identification of Urban Problems and Issues Using 4th Industrial Revolution ElementTechnology in Seoul, Korea -Focusing on the Application of Text Mining Analysis and Delphi Survey", *Seoul Studies*, 20(4): 1-21.

23. Hong, S.G., Kim, H.J., and Choi, H.R., 2016. "An Analysis of Civil Traffic Complaints Using Text Mining" *International Journal on Information*, 19(11): 4995-5000.

24. Lee, E., Lee, S., Kim, K.S., Pham, V.H., and Sul, J., 2019. "Analysis of Public Complaints to Identify Priority Policy Areas: Evidence from a Satellite City around Seoul", *Sustainability*, 11(21): 6140.

25. Li, J., Sun, A., Han, J., and Li, C., 2022) "A Survey on Deep Learning for Named Entity Recognition", *IEEE Transactions on Knowledge and Data Engineering*, 34(1): 50-70.

26. Olivos, F., Saavedra, P., and Dammert, L., 2022. "Citizen Complaints as an Accountability Mechanism: Uncovering Patterns Using Topic Modeling", *Journal of Research in Crime and Delinquency*, 60(6): 740-780.

27. 네이버·창원대학교, 2019. 네이버, 창원대가 함께하는 NLP Challenge, https://github.com/naver/nlp-challenge
Naver and Changwon National University (CNU), 2019. NLP Challenge with Naver and CNU, https://github.com/naver/nlp-challenge

28. 서울연구원, 2014.01.20. "NYC 311 종합 서비스 (뉴욕市)", https://www.si.re.kr/node/47962
Seoul Institute, 2014, January 20. "NYC 311 Comprehensive Service", https://www.si.re.kr/node/47962

29. 유원준·안상준, 2022.「딥 러닝을 이용한 자연어 처리 입문」, 위키독스., https://wikidocs.net/book/2155.
Yu, W.J. and Ahn, S.J., 2022. *Introduction to Natural Language Processing Using Deep Learning*, Wikidocs, https://wikidocs.net/book/2155

30. 한국정보통신기술협회, "개체명 인식", 정보통신용어사전, 2023. 09.05. 읽음. http://word.tta.or.kr/dictionary/dictionaryView.do?word_seq=099981-5
Telecommunications Technology Association (TTA), 2023. September 5., "Telecommunications Dictionary", Named Entity Recognition, Accessed September 5. 2023, http://word.tta.or.kr/dictionary/dictionaryView.do?word_seq=099981-5

31. 한국해양대학교 자연언어처리연구실, 2019. "한국어 개체명 정의 및 표지 표준화 기술보고서와 이를 기반으로 제작된 개체명 형태소 말뭉치", https://github.com/kmounlp
Korea Maritime and Ocean University NLP Lab, 2019. "Technical Report on the Definition of Korean Named Entities and Standardization of Labels, and the Named Entity Morphological Corpus Created Based on It", https://github.com/kmounlp

32. Chollet F., 2020. "Transfer Earning & Fine-tuning", Keras. https://keras.io/guides/transfer_learning/#introduction

33. Dickson, B., 2023. July 10. "The Complete Guide to LLM Fine-tuning", TechTalks, https://bdtechtalks.com/2023/07/10/llm-fine-tuning/

34. Lendave, V., 2021. November 17. "A Tutorial on Sequential Machine Learning", Analytics India Magazine, https://analyticsindiamag.com/a-tutorial-on-sequential-machine-learning/

35. Park, S., Moon, J., Kim, S., Cho, W.I., Han, J., Park, J., Song, C., Kim, J., Song, Y., Oh, T., Lee, J., Oh, J., Lyu, S., Jeong, Y., Lee, I., Seo, S., Lee, D., Kim, H., Lee, M., Jang, S., Do, S., Kim, S., Lim, K., Lee, J., Park,K., Shin, J., Kim, S., Park, L., Oh, A., Ha, J.W., Cho, K., 2021. "KLUE: Korean Language Understanding Evaluation (arXiv:2105.09680; Version 1)", arXiv, http://arxiv.org/abs/2105.09680

36. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I., 2017. "Attention Is All You Need (arXiv:1706.03762; Version 1)", arXiv. http://arxiv.org/abs/1706.03762

37. Yang, A. (Director), 2021. 3. 25). A Chat with Andrew on MLOps: From Model-centric to Data-centric AI. https://www.youtube.com/watch?v=06-AZXmwHjo

Date Received	2024-05-27
Reviewed(1st)	2024-08-15
Date Revised	2024-08-24
Reviewed(2nd)	2024-09-08
Date Revised	2024-10-17
Reviewed(3rd)	2024-10-31
Date Accepted	2024-10-31
Final Received	2024-11-20