Journal of the Korean Association of Geographic Information Studies, 26(3) 2023, pp.37-51 ISSN 1226-9719(Print)
https://doi.org/10.11108/kagis.2023.26.3.037 ISSN 2287-6952(Online)

AI를 활용한 비정형 문서정보의 공간정보화*

윤상원[1] · 박정우[2] · 남광우[3※]

Spatialization of Unstructured Document Information Using AI*

Sang-Won YOON[1] · Jeong-Woo PARK[2] · Kwang-Woo NAM[3※]

요 약

도시현상의 해석을 위해 공간정보는 필수적이다. 위치정보가 부족한 도시정보를 공간정보로 변환하기 위한 공간정보화 방법론이 꾸준히 개발되어왔다. 정형화된 주소정보나 지명 등을 이용한 Geocoding이나 이미 위치정보가 있는 공간정보와의 공간결합, 참조데이터를 활용한 수작업 형태 등이 대표적이다. 그러나 아직도 행정기관에서 작성되는 수많은 문서정보들은 비정형화된 문서형태로 인해 공간정보화의 수요가 있음에도 그동안 깊이 있게 다루어지지 못하였다. 본 연구는 자연어 처리 모델인 BERT를 활용하여 도시계획과 관련된 공개문서의 공간정보화를 진행한다. 주소가 포함된 문장 요소를 문서로부터 추출하고, 이를 정형화된 데이터로 변환하는 과정을 중점적으로 다룬다. 18년 동안의 도시계획 고시공고문을 학습 데이터로 사용하여 BERT 모델을 학습시켰으며, 모델의 하이퍼파라미터를 직접 조정하여 성능을 향상시켰다. 모델 학습 후의 테스트 결과, 도시계획시설의 유형을 분류하는 모델은 96.6%, 주소 인식 모델은 98.5%, 주소 정제 모델은 93.1%의 정확도를 보였다. 결과 데이터를 GIS 상에 맵핑하였을 때, 특정 지점의 도시계획시설에 관한 변경 이력을 효과적으로 표출할 수 있었다. 본 연구로 도시계획 문서의 공간적 맥락에 대한 깊은 이해를 제공하며, 이를 통해 이해관계자들이 더욱 효과적인 의사결정을 할 수 있게 지원하기를 기대한다.

주요어 : 비공간 데이터, BERT, 공간정보, 자연어 처리, 개체명 인식

ABSTRACT

Spatial information is essential for interpreting urban phenomena. Methodologies for

2023년 07월 17일 접수 Received on July 17, 2023 / 2023년 08월 19일 수정 Revised on August 19, 2023 / 2023년 08월 25일 심사완료 Accepted on August 25, 2023
* 본 연구는 국토교통부/국토교통과학기술진흥원의 지원으로 수행되었음(과제번호 RS-2022-00143404).
1 경성대학교 도시공학과 석사과정 / Graduate Student. Dept. of Urban Planning & Engineering, Kyungsung University
2 경성대학교 학술연구교수 / Research Professor, Kyungsung University.
3 경성대학교 도시공학과 교수 / Professor, Dept. of Urban Planning & Engineering, Kyungsung University.
※ Corresponding Author E-mail : kwnam4794@gmail.com

spatializing urban information, especially when it lacks location details, have been consistently developed. Typical methods include Geocoding using structured address information or place names, spatial integration with existing geospatial data, and manual tasks utilizing reference data. However, a vast number of documents produced by administrative agencies have not been deeply dealt with due to their unstructured nature, even when there's demand for spatialization. This research utilizes the natural language processing model BERT to spatialize public documents related to urban planning. It focuses on extracting sentence elements containing addresses from documents and converting them into structured data. The study used 18 years of urban planning public announcement documents as training data to train the BERT model and enhanced its performance by manually adjusting its hyperparameters. After training, the test results showed accuracy rates of 96.6% for classifying urban planning facilities, 98.5% for address recognition, and 93.1% for address cleaning. When mapping the result data on GIS, it was possible to effectively display the change history related to specific urban planning facilities. This research provides a deep understanding of the spatial context of urban planning documents, and it is hoped that through this, stakeholders can make more effective decisions.

KEYWORDS : Non-Spatial data, BERT, Spatial Information, Natural Language Processing, Named Entity Recognition

서 론

사물인터넷을 비롯한 정보통신기술이 발달함에 따라 카드데이터나 교통데이터와 같은 위치기반 빅데이터의 생산과 활용이 가속화되고 있다. 그러나 빅데이터의 양적 증가에도 불구하고 의사결정을 지원하기 위해서는 여전히 의사결정과 연관된 다양한 데이터셋이 요구된다. 이에 따라 비정형 데이터인 행정문서에서 추가적인 정보의 추출은 데이터의 포괄성(comprehensiveness)을 확보하기 위한 중요한 수단이며, 추출된 정보를 활용하여 공간데이터화가 가능하다면 다양한 의사결정에 도움을 줄 수 있을 것이다.

특히, 스마트시티 개념에서 디지털 트윈을 기반으로 한 공간의사결정이 도시행정의 필수적인 경쟁력 요소로 기대되고 있는 가운데, 지자체들은 데이터 센터 구축을 통해 통합 데이터 포털을 개설하여 도시 전반의 사회경제적인 지표를 기준으로 다양한 데이터를 제공하고 있다. 공공기관 홈페이지나 통합 데이터 포털에서 제공되는 70% 이상의 행정정보는 공간정보의 성격이지만 정확한 위치정보를 포함한 정보는 매우 제한적이다. 정형화된 통계자료로부터의 위치정보 추출을 통해 공간정보로 변환하는 과정은 이미 활용되지만, 연구보고서, 공고문 등의 비정형 문서에서 정보를 추출하여 정형화된 공간 데이터로 변환하는 연구는 부족한 실정이다. 또한, 공공포털에는 도시계획시설 변경 및 결정에 관한 고시 관련 자료는 존재하지만, 도시계획 변경 및 도시지역의 개발 변천을 확인할 수 있는 공간적 차원의 이력에 대한 자료는 존재하지 않는다.

통계자료와는 다르게 보고서, 사업계획서, 고시공고문 등의 문서자료에는 위치정보뿐 아니라 해당 문서자료의 작성 목적에 따른 도시계획과 관련된 공간적 차원의 정성적인 정보를 다수 포함하고 있다. 이는 정형화된 단순 통계자료에서는 얻을 수 없는 의미있는 정보이다. 이에 본 연구는 공공기관이 공개한 문서에서 공간정보와 연계가 가능한 주소를 추출하고 정형화된 데이터를 수집하는 AI 적용 방법론을 만드는 데 목적이 있다. 이를 위해 문서 내 공간정보로 변경

이 가능한 주소자료를 추출하는 AI 모델로 자연어 처리 모델인 BERT를 활용하고자 한다.

이론적 배경 및 선행연구 고찰

1. 이론적 배경

한국어는 교착어로, 정보를 접미사와 어미로 표현하므로 형태소 분석은 의미 파악에 중요하다. 따라서 문장 내 의미를 파악하고 구조화하기 위해 한국어 형태소 분석을 지원하는 모델이 필요하다. 대표적인 모델로 BiLSTM-CRF, Transformer 등이 있으며, 최근에는 BERT, GPT 등의 사전 학습된 언어 모델을 활용하여 한국어 자연어 처리 성능을 향상시키는 연구가 활발히 진행되고 있다.

Transformer 모델은 2017년 'Attention is all you need' 를 통해 Google에서 공개되었다. 트랜스포머 모델은 기계 번역, 질의응답, 감정 분석, 요약, 모델링 생성 등의 다양한 자연어 처리 작업에서 우수한 성능을 보인다(Vaswani *et al.*, 2017). BERT 모델은 "Bidirectional Encoder Representations from Transformers"의 약자로, Google에서 2018년에 발표한 자연어 처리(NLP)를 위한 딥러닝 모델이다. 기존에 활용되는 단방향 자연어 처리 모델을 보완한 양방향 학습이 가능한 모델이다. BERT 모델은 사전 학습된 레이블이 없는 방대한 텍스트 데이터를 이용하여 모델을 학습시키고 분류 신경망을 추가하는 전이 학습 방법으로서, 레이블이 있는 다른 데이터의 추가 훈련(파인 튜닝, Fine-tuning)을 진행하여 하이퍼파라미터를 재조정하면 성능을 높일 수 있는 특징이 있다(Devlin *et al.*, 2018).

2. 선행연구 고찰

딥러닝 및 자연어 처리 모델과 문서의 분류, 문장 내 관심단어의 추출과 유사한 연구를 대상으로 선행연구를 고찰하였다.

첫 번째로, 딥러닝 및 자연어 처리 모델 관련 연구는 Noh *et al.*(2021), Kim *et al.*(2020), You *et al.*(2021), Lee *et al.*(2022) 등의 연구가 있었다. Noh *et al.*(2021)는 뉴스 텍스트 데이터를 기초로 한 카테고리 분류 작업을 위하여 LSTM 모델을 활용하였다. RNN, BiLSTM 모델과의 성능을 비교하였을 때, 각 카테고리별 분류에 적합한 모델이 다른 것으로 나타났다. 해당 연구에서 활용한 LSTM 모델의 정확도는 높게 나타났으나, 학습 3회 시 과적합이 발생하는 문제가 있었다. Kim *et al.*(2020)은 사전 및 규칙 기반 추출 방법과 Bidirectional LSTM-CRF 모델을 결합한 개체명 인식 모델을 구현하였다. 이 모델은 범죄 사칭수법이 태깅된 비정형 데이터를 바탕으로 기관, 이름, 직급 등의 개체명을 인식하는 데 우수한 성능을 보였다. You *et al.*(2021)는 한국어 개체명 인식을 위하여 사전 훈련된 모델인 BERT-Base, Multilingual Cased를 Fine-tuning하는 방식으로 실험을 설계하였다. 모델의 성능을 개선하는 방향으로 학습 데이터의 구성과 후처리에 따른 실험, 토큰화 방식 등 데이터 전처리에 대한 중요성을 강조하였다. Lee *et al.*(2022)는 한국어 형태소 분석을 위하여 Transformer Encoder와 Decoder를 결합한 BERT-Fused Transformer 모델을 활용하였다. 입력 문장의 길이가 길어져도 Encoder에서 입력 문장을 인코딩하여 고정된 표현 벡터를 출력하기 때문에 문장의 길이에 영향을 받지 않는 장점이 있으나, 복잡한 모델 구조로 인하여 학습 및 추론 시간이 오래 걸리는 단점이 있다.

두 번째로, 문서의 분류와 문장 내 관심단어의 추출과 관련된 연구는 Kim *et al.*(2022), Lee *et al.*(2004), Song *et al.*(2014), Colin *et al.*(2017) 등의 연구가 있었다.

Kim *et al.*(2022)은 수집한 학술지 논문에 대해 사전 학습 모델인 BERT를 활용하여 학술문헌 13가지 범주로 분류 작업을 진행하였다. 학습 데이터의 크기와 품질에 따른 분석을 진행하였을 때, 성능의 차이가 확연히 나타나는 것을 확인하였다. Lee *et al.*(2004)는 주소 정제를 위해 먼저 국내에서 사용되는 주소 체계를 파악하였다. 비정형 및 비표준 주소 텍스트 데

이터를 정형화하는 방안으로 형태소 분석 개념을 적용한 주소요소 분석 기법을 도입하였다. 또한, 주소의 위계를 설정하는 룰 기반의 주소 보정 시스템을 설계하였다. 이를 통해 형태소 분석에서 발생하는 요소원형 복원과 결합성 문제를 보완하였다. 그 결과, 표준화된 주소 입력 및 검색 방법을 제시할 수 있었다. Song *et al.*(2014)은 행정구역의 위계정보와 각 구역 명칭들의 유사도를 계산하여 행정구역 명칭의 줄임말, 주소정보 누락, 오타 등 오류가 발생하는 경우에도 지번 주소를 도로명주소로 변환하는 방법론을 제시하였다. Colin *et al.*(2017)은 텍스트 문서에서의 지명 인식을 위하여 자연어 처리 도구인 OpenCalais와 오픈 소스 툴인 Geodict를 활용하였다. 이를 통해 텍스트 데이터에서 지리적인 컨텍스트를 파악하고, 지리적 위치와 관련된 정보를 추출하여 다양한 응용 분야에서 GIS와의 연계를 통한 활용 가능성을 제시하였다.

유사 연구들은 모델의 정확도 향상 및 유형 분류를 위한 단순 딥러닝 모델의 적용이라는 한계를 가지며, 주소 관련 연구의 경우 주어진 주소의 정제 및 Geocoding에 그쳤다. 그러나 본 연구에서는 비정형 데이터인 도시계획 고시공고문을 기반으로 복합적인 유형의 분류 작업을 수행하고, 문장 내 주소 정보를 추출하는 데 딥러닝 모델을 적용하였다. 이는 기존의 연구와는 차별화된 접근 방식을 보여준다.

연구 방법론

1. 연구의 범위와 방법

본 연구에서는 부산광역시 홈페이지에 게재된 도시계획시설 변경 및 결정 고시 자료의 주요 내용 및 주소 정보를 기초 데이터로 수집하여 전처리하고, 각 주제에 따른 모델을 구축하기 위하여 세 가지의 학습 데이터셋을 생성하였다. 내부적으로 BERT 모델의 동작 구조와 같이 입력 데이터를 문장 또는 단어 단위로 분리하여 토큰화 작업을 수행하고, 토큰화된 데이터는 BERT 모델의 입력 형식에 맞게 변환하였다. 각 토큰은 고유한 정수 인덱스로 맵핑되고, 각 모델에 적합한 입력 시퀀스의 최대 길이를 설정하여 패딩이 진행되도록 하였다. 또한, 입력 문장에 대해 각 토큰을 사전 학습된 임베딩 벡터로 변환하는 'Token Embeddings', 문장 세그먼트 정보를 임베딩하는 'Segment Embeddings', 문장 내 단어의 상대적인 위치 정보를 임베딩하는 'Position Embeddings' 이 수행되었다. 이

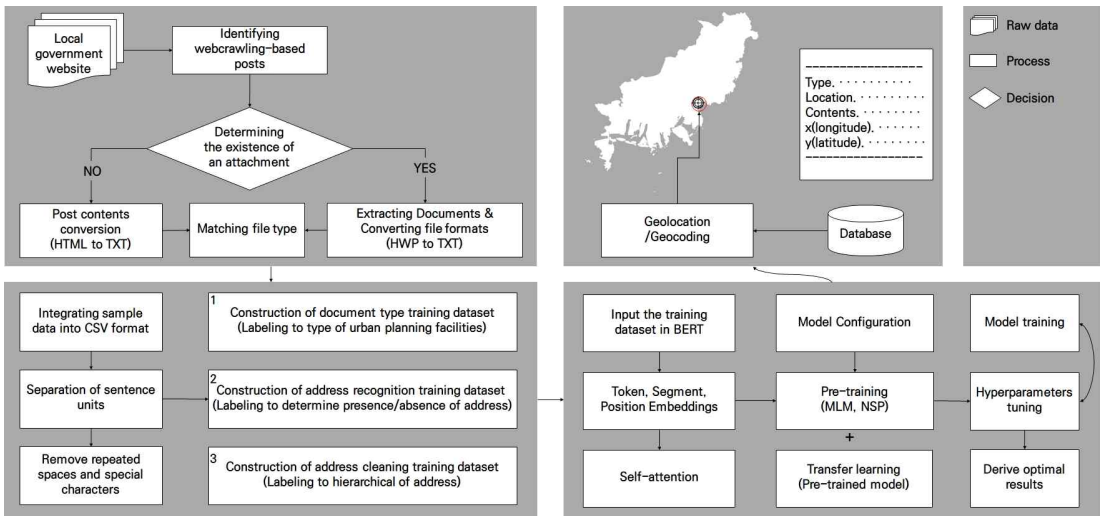

FIGURE 1. The overall process of this study

후 Self-attention이 포함된 양방향 인코더로 입력 시퀀스의 문맥 정보를 추출하였다. 대규모 텍스트 데이터를 가진 사전 학습된 BERT 모델로 학습시키고, 하이퍼파라미터 재조정으로 각 주제별 모델의 성능을 개선할 수 있었다. 세 가지 모델을 활용하여 추출된 결과들은 DB에 저장되며, 이는 Geocoding을 통해 변환된 좌표와 함께 해당 지점에서 발생한 변경 및 결정 고시에 대한 내용으로 출력하였다.

개발 환경은 파이썬 3.7.0 버전을 기준으로 Visual Studio Code(VSC) 개발자 환경에서 작성하였으며, PyTorch 1.12.1과 PyTorch의 BERT 0.6.2 버전을 딥러닝 모듈로 사용하였다. 필요에 따라 TensorFlow-gpu 2.10.0 버전으로 연산하여 학습 및 분석의 시간을 단축하였다. 본 연구의 흐름은 그림 1과 같다.

2. 학습 데이터셋 구축

1) 기초 데이터 선정

본 연구에서는 도시계획 변경 및 결정 고시의 이력자료 구축 및 공간화를 위하여 부산광역시 홈페이지에 게재된 2005년부터 2022년까지의 도시계획시설 변경 및 결정 고시 자료를 기초 데이터로 활용하였다. 기초 데이터를 자동으로 수집하기 위해 Python 라이브러리인 Selenium의 webdriver 모듈을 활용하여 페이지 이동 및 반복적인 다운로드를 진행할 수 있었으며, 웹 페이지 내 특정 요소나 속성에 접근하기 위한 XPath 및 CSSselector를 활용하여 본문의 내용과 첨부파일을 수집하였다. 첨부파일이 없는 경우에는 수집된 본문 내용을 기초 데이터로 활용하였다.

2) 데이터셋 생성

자연어 처리 모델인 BERT를 학습하기 위해 기초 데이터들을 딥러닝에 적합한 학습자료 형태로 가공하는 과정이 요구되었다. 웹 크롤링을 통해 다운로드한 첨부파일 형식은 HWP, CSV, PDF 등 다양한 형식으로 제공되기 때문에 Pyhwp 라이브러리를 활용하여 TXT 형식으로 변환 후, 파일명 형식('공고일자_도시계획시설 사업명_고시공고번호')을 일치화하였다. Python으로 문서 내용을 학습 데이터셋으로 구축하였으며, 문장 단위로 분리하고, 공백과 불필요한 문자를 제거하여 총 세 가지의 학습 데이터셋을 구성하였다. 본 연구에서 다루는 비정형 데이터인 도시계획 고시공고문 내 도시계획시설의 유형과 주소 정보를 포함하고 있으며, 이러한 정보는 문서 유형의 자동 분류와 주소 정보의 추출을 위한 주요 핵심 키워드가 되었다. 이러한 데이터셋의 특징을 잘 활용하여 모델의 학습 성능을 극대화하도록 하였다.

수집한 도시계획 고시공고문의 샘플을 그림 2와 같으며, 고시공고문 내용을 문장 단위로 분리하고, 필요에 따라서는 분리한 문장을 토큰 단위로 추가 분리하는 작업을 진행하였다. 본 연구에서 구축한 학습 데이터셋은 세 가지로 '문서 유형 분류', '주소 인식', '주소 정제' 이다. 첫 번째로, '문서 유형 분류 학습 데이터셋' 을 구축하기 위한 고시공고문 내 유형 도출 방안을 검토하였다. 그림 2의 'A-2', 'A-3', 'A-8' 의 문장은 도시계획시설 유형을 판단할 수 있는 단어인 '유원지' 가 포함되어 있는 문서 유형 분류가 가능한 문장이다. 그림 2의 'A-2', 'A-3', 'A-8' 문장을 제외한 항목들은 모두 도시계획시설 유형 내용을 포함하고 있지 않으며, 이는 '유형 분류 예외' 로 분류하였다. 따라서, 도출된 유형 키워드 중 빈도 수가 높은 것을 문서의 유형으로 선정하는 방식으로 진행하였고, '유형 분류 예외' 항목은 모델 학습에는 활용되었으나, 결과 도출 과정에서는 제외하였다.

두 번째로, '주소 인식 학습 데이터셋' 을 구축하기 위한 주소 포함 여부 확인 방안을 검토하였다. 그림 2의 'A-9', 'A-12', 'A-16', 'A-17' 문장과 같이 문장 내에 '시/도 - 시/군/구 - 읍/면/동/리 - 지번', '시/군/구 - 읍/면/동/리 - 지번' 또는 '읍/면/동/리 - 지번' 의 주소를 포함한 경우를 제외하고는 주소에 해당하지 않는 것으로 구분하였다.

세 번째로, '주소 정제 학습 데이터셋' 은 그

A-1	부산광역시 고시 제2022-388호
A-2	**도시계획시설사업(민락유원지 보상사업) 실시계획(변경)인가 고시**
A-3	부산광역시 고시 제2006-471호(2007. 1. 2.)로 도시계획시설(민락유원지) 결정 및 조성계획 결정되고, 부산광역시 고시 제2007-453호(2007.11.28)로 조성계획 결정(변경), 부산광역시 고시 제2009-306호(2009. 8. 5.)로 조성계획 결정(변경), 부산광역시 고시 제2013-97호(2013. 3.13.)로 조성계획 결정(변경) 및 지형도면 고시, 부산광역시 고시 제2014-515호(2015. 1. 1.)로 조성계획 결정(변경) 및 지형도면 고시, 부산광역시 고시 제2017-207호(2017. 7. 5.)로 조성계획 결정(변경) 및 지형도면 고시, 부산광역시 고시 제2017-239호(2017. 7. 19.)로 조성계획 결정(변경) 및 지형도면 고시, 부산광역시 고시 제2020-387호(2020. 10. 14.)로 고시된 도시계획시설(민락유원지)에 대하여「국토의 계획 및 이용에 관한 법률」제86조, 제88조, 제91조 및 같은 법 시행령 제96조, 제97조, 제100조 규정에 의거 아래와 같이 도시계획시설사업 실시계획(변경) 인가 고시합니다.
A-4	관계도서는 부산광역시(공원운영과 ☎888-3824) 및 수영구(일자리경제과 ☎610-4532)에 비치하고 이해관계인 및 일반인에게 보입니다.
A-5	2022년 10월 12일
A-6	**부 산 광 역 시 장**
A-7	**1. 사업의 종류** : 도시계획시설사업
A-8	**2. 사업의 명칭** : 민락유원지 보상사업
A-9	**3. 사업의 위치** : 부산광역시 수영구 민락동 106-1번지 일원
A-10	**4. 사업의 면적 및 규모(변경있음)**
A-11	가. 사업면적 : (당초) A=41,670㎡ (변경) A=41,676㎡
A-12	나. 규 모 : 보상수용(토지및지장물 등), 민락동 106-1번지 외 15필지
A-13	**5. 사업시행자의 성명 및 주소(변경있음)**
A-14	가. 성 명 : 티아이홀딩스그룹㈜
A-15	나. 주 소
A-16	(당초) 서울특별시 중구 세종대로 39, 6층(남대문대로4가, 대한서울상공회의소)
A-17	(변경) 서울특별시 중구 세종대로 39, 13층(남대문대로4가, 대한서울상공회의소)
A-18	**6. 사업의 착수 및 준공예정일(변경있음)**
A-19	(당초) 2020. 10. 14 ~ 2022. 10. 13
A-20	(변경) 2020. 10. 14 ~ 2024. 10. 13(증24개월)
A-21	**7. 사업비 :** (당초) 9,701,000,000원 (변경) 9,703,342,000원(증 2,342,000원)
A-22	**8. 수용 또는 사용할 토지 소유권 및 소유권이외의 권리자 조서(변경)** : 붙임 조서 참조

FIGURE 2. Urban planning public announcement document sample

TABLE 1. Types of urban planning facilities for document classification

No.	Urban planning facility type	Number of sentence	No.	Urban planning facility type	Number of sentence
1	Public Facility	327	8	Amusement Park/Recreation Area	155
2	Park	677	9	Medical Facility	35
3	Plaza/Square	38	10	Parking Lot & Bus Terminal/Bus Stop	58
4	Miscellaneous Supply Facilities	149	11	Railway/Port/Airport	30
5	Green Space/Greenery	40	12	Sewage/Waste/Stormwater/Water Pollution Prevention Facilities	64
6	Road	783	13	School	179
7	Plural Type	122	14	Classification Exception	11,406

림 2의 'A-9', 'A-12', 'A-16', 'A-17' 문장과 같이 주소를 포함한 문장이 지번 주소일 경우 '시/도 - 시/군/구 - 읍/면/동/리 - 지번' 등의 단어로, 도로명주소는 '시/도 - 시/군/구 - 도로명 - 건물번호' 의 규칙으로 라벨링하였다.

문서 유형 분류 학습 데이터셋의 유형별 샘플 수는 표 1과 같다. 도시계획시설의 유형을 분류하는 주제의 특성상 다중 분류를 진행하기 위하여 다양한 유형을 포함하였으나, 샘플 수가 낮은 유형은 병합하여 항목을 구성하였다. '공공시설(Public Facility)', '공원(Park)', '광장(Plaza/Square)', '기타공급설비(Miscellaneous Supply Facilities)', '녹지(Green Space/Greenery)', '도로(Road)', '복수유형(Plural Type)', '유원지(Amusement Park/Recreation Area)', '의료시설(Medical Facility)', '주차장 및 자동차정류장(Parking Lot & Bus Terminal/Bus Stop)', '철도/항만/공항(Railway/Port/Airport)', '하수/폐기물/우수/수질오염방지시설(Sewage/Waste/Stormwater/Water Pollution Prevention Facilities)', '학교(School)', '유형 분류 예외(Classification Exception)', 총 14개의 유형으로 분류할 수 있었다.

첫 번째는 도시계획시설 변경 및 결정 관련 내용과 그에 따른 도시계획시설 유형의 라벨링을 진행한 '문서 유형 분류 학습 데이터셋(표

TABLE 2. Document classification training dataset example

No.	Contents	Labels
A-2	Announcement of the Implementation Plan (Amendment) Authorization for the Urban Planning Facility Project (Compensation Project for Minrak Amusement Park)	Amusement Park/Recreation Area
A-3	Pursuant to Busan Metropolitan City Notice No. 2006-471 (Jan 2, 2007), the urban planning facility (Minrak Amusement Park) was determined and its establishment plan was decided. Following this, the establishment plan was modified according to Busan Metropolitan City Notice No. 2007-453 (Nov 28, 2007), No. 2009-306 (Aug 5, 2009), No. 2013-97 (Mar 13, 2013), and the topographical drawings were announced. Further modifications to the establishment plan and the announcement of topographical drawings were made under Busan Metropolitan City Notice No. 2014-515 (Jan 1, 2015), No. 2017-207 (Jul 5, 2017), No. 2017-239 (Jul 19, 2017), and No. 2020-387 (Oct 14, 2020) for the urban planning facility (Minrak Amusement Park). In accordance with Articles 86, 88, and 91 of the "Land Planning and Utilization Act" and Articles 96, 97, and 100 of the same Act's Enforcement Decree, the following is hereby announced as the authorization notice of the urban planning facility project execution plan (modified)	Amusement Park/Recreation Area
A-8	2. Project Name: Minrak Amusement Park Compensation Project	Amusement Park/Recreation Area
A-22	8. Document (modification) of land ownership and rights other than ownership to be acquired or used: Refer to the attached document	Classification Exception

TABLE 3. Address recognition training dataset example

No.	Contents	Labels
A-4	The related documents are available at Busan Metropolitan City (Parks Management Department ☎888-3824) and Suyeong-gu (Employment and Economy Department ☎610-4532) for stakeholders and the general public to view	not_in
A-9	3. Location of the project: 106-1, Minrak-dong, Suyeong-gu, Busan Metropolitan City	in
A-12	4-b. Specifications: Compensation acquisition (land and fixtures, etc.), Minrak-dong 106-1 and 15 other plots	in
A-16	(Originally) 39 Sejong-daero, Jung-gu, Seoul, 6th Floor (Namdaemun-ro 4-ga, Korea Seoul Chamber of Commerce)	in

TABLE 4. Address cleaning training dataset example

No.	Contents	Labels
A-9	3. Location of the project : Area of 106-1, Minrak-dong, Suyeong-gu, Busan Metropolitan City	O O O O O O O O O loc_jibun O loc_dong O loc_sigungu O loc_sido
A-12	4-b. Specifications : Compensation acquisition (land and fixtures, etc.), Minrak-dong 106-1 and 15 other plots	O O O O O O O O O O O O O O O O O O O loc_dong loc_jibun O O O O
A-16	(Originally) 39 Sejong-daero, Jung-gu, Seoul, 6th Floor (Namdaemun-ro-4-ga, Korea Seoul Chamber of Commerce)	O O O loc_num loc_rn O loc_sigungu O loc_sido O O O O loc_dong O O O O O O O

2)' 으로, 14,063개의 샘플을 생성하였다. 그림 2의 'A-2' , 'A-3' , 'A-8' 은 문장 내 '민락유원지' 라는 키워드를 포함하여 '유원지' 로 구분하였고, 'A-22' 는 도시계획시설 유형과 관련된 키워드를 포함하지 않으며, 이는 '유형 분류 예외(Classification Exception)' 로 구분하였다. 또한 '복수유형(Plural Type)' 은 게시글의 제목이나 첨부파일 내에서 두 종류 이상의 도시계획시설을 포함하는 내용으로 지칭하였다.

두 번째는 주소를 포함한 문장을 인식하기 위한 라벨링을 진행한 '주소 인식 학습 데이터셋(표 3)' 이며, 14,063개의 샘플을 생성하였다. 그림 2의 'A-9' , 'A-12' , 'A-16' 과 같이 문장 내에 주소 구조를 포함하면 주소에 해당하는 것으로 인식하고, 포함하지 않는 'A-4' 는 주소로 인식하지 않도록 라벨링을 진행하였다. 그 결과 1,020개의 'in' 샘플과 13,043개의 'not_in' 샘플로 구성된 학습 데이터셋을 구축하였다.

표 4에서는 '주소 정제 학습 데이터셋' 을 확인할 수 있다. 앞서 구축한 '주소 인식 학습 데이터셋' 의 각 샘플에서 단어 토큰 단위로 분리하여 해당 토큰에 대한 라벨링과 주소와의 연관성이 낮은 word를 생략하는 라벨링을 진행하였다. '시도' 는 'loc_sido' , '시군구' 는 'loc_sigungu' , '읍면동' 은 'loc_dong' , '리' 는 'loc_ri' , '지번' 은 'loc_jibun' , '도로명' 은 'loc_rn' , 건물번호는 'loc_num' 으로 구분하였다. 주소 정제 학습 데이터셋 특성상 복잡한 라벨 구조로 인하여 기존의 학습 데이터셋에서 확장하여 19,821개의 샘플을 생성하였다.

모델 학습 및 결과

1. 모델 학습

본 연구에서는 104개의 언어에 대해 사전 학습된 Hugging Face Transformers 라이브러리 bert-base-multilingual-cased 모델을 전이학습하여 고시공고문의 도시계획시설 유형 분류 및 주소 정제에 적합한 새로운 BERT 모델을 개발하였다. 학습 데이터셋별로 bert-base-multilingual-cased 모델의 토크나이저를 사용하여 문장 단위로 분리하는 토큰화를 진행하였다. 입력 데이터의 최대 길이 설정 및 패딩 처리 후, 도시계획시설 유형, 주소 포함 여부 및

정제 내용에 대해 라벨링된 column 데이터들을 출력 데이터로 변환하였다. 이후, 입력 데이터와 출력 데이터를 Tensor 형태로 묶어 데이터셋을 생성하고, DataLoader를 생성하여 읽어올 데이터의 batch size를 설정하는 등의 작업을 통해 BERT 모델의 형식에 맞게 데이터로 변환하였다. 본 연구에서는 다양한 optimizer 중 SGD, Adam, AdamW, BertAdam 등을 검토하였다. 특히, AdamW를 주로 활용하였는데, 이는 Adam의 핵심 알고리즘에 L2 regularization을 통합한 버전이다. Adam은 gradient의 지수 가중 평균을 활용하여 각 매개변수에 대한 학습률을 동적으로 조정하는 특성을 가지고 있다. 그러나, 기존의 Adam에서는 L2 weight decay 처리가 완전히 효과적이지 않았기 때문에, AdamW는 이 문제를 해결하기 위해 일반적인 L2 weight decay 방식을 적절히 통합하여 과적합 문제를 보다 효과적으로 방지할 수 있게 설계되었다. 본 연구에서는 통상적인 손실 함수에 의존하는 대신 gradient를 직접 계산하여 모델의 가중치를 업데이트하는 방식을 채택하였다. 이 방식으로 기존의 손실 함수 기반의 방식과는 다르게, 모델의 파라미터 업데이트를 직접 제어하며, 특정 목적이나 조건에 더욱 특화된 모델 학습이 가능하였다.

2. Optimizer 비교 및 학습 결과

표 5는 문서 유형 분류 모델 학습 시 각 옵티마이저의 성능을 비교한 것이다. AdamW가 최고의 성능을 보일 것이라는 예상과는 달리 다른 Adam 모델들과의 성능 차이가 크지 않았다. 하지만 가장 높은 정확도를 보여 AdamW를 옵티마이저로 선택하였다. 문서 유형 분류를 위한 14,063개의 샘플 데이터를 훈련, 검증, 테스트용으로 8:1:1 비율로 분리하였으며, 모델 학습은 총 4회의 epoch 동안 진행하였다. 학습 결과에서는 훈련 손실값(Train_loss), 훈련 정확도(Train_acc), 검증 손실값(Val_loss), 검증 정확도(Val_acc)를 나열하였으며(표 6), 이러한 학습 과정을 그래프로 시각화한 것은 그림 3과 같다. 문서 유형 분류 모델은 고시공고문 제목 및 내용에 도시계획시설 명칭만을 명확하게 포함할 경우에는 정확하게 구분할 수 있었지만, 명칭과 지명이 혼합된 내용(예: 가야유원지 불교조각공원)을 포함한 문서에서는 오분류 문제가 발생하였다.

표 7은 주소 인식 모델 학습 시 각 옵티마이저의 성능을 보여준다. 다른 옵티마이저보다 높은 정확도를 보인 AdamW를 선택하였다. 샘플 데이터는 '문서 유형 분류 모델' 과 동일한 14,063개로, 데이터 분리 비율과 epoch는 동일하게 유지하였다. 주소 인식 모델을 학습시킨 과정은 표 8과 그림 4와 같다. 수집된 고시공고문의 주소 정보 중 '광역시' 를 '시' 로 생략하여 작성하는 경우와 '시' 를 생략하여 '시군

TABLE 5. Comparison of learning performance by optimizer(Document classification)

Optimizer	Test_loss	Test_accuracy	Learning time
SGD	0.365	0.909	27:47
Adam	0.177	0.959	13:44
AdamW	0.181	0.966	13:43
BertAdam	0.226	0.962	15:42

TABLE 6. Document classification training' s Accuracy and loss results(AdamW)

Epoch	Train_loss	Train_acc	Val_loss	Val_acc
1	0.390	0.900	0.214	0.940
2	0.190	0.954	0.156	0.957
3	0.115	0.973	0.126	0.973
4	0.075	0.982	0.123	0.977

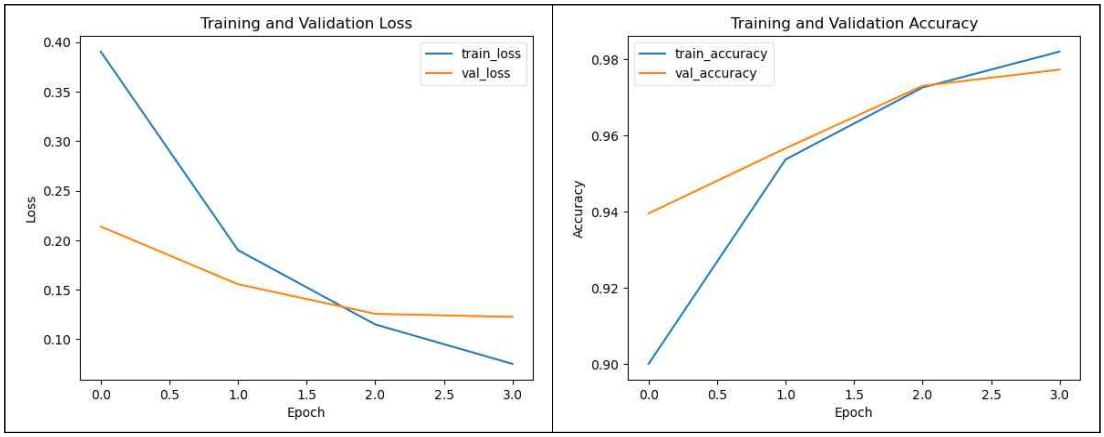

FIGURE 3. Result of document classification training

TABLE 7. Comparison of learning performance by optimizer(Address recognition)

Optimizer	Test_loss	Test_accuracy	Learning time
SGD	0.580	0.940	7:21
Adam	0.059	0.982	3:54
AdamW	0.065	0.985	4:52
BertAdam	0.083	0.982	4:58

TABLE 8. Address recognition training’s Accuracy and loss results(AdamW)

Epoch	Train_loss	Train_acc	Val_loss	Val_acc
1	0.096	0.973	0.070	0.987
2	0.056	0.987	0.060	0.990
3	0.052	0.987	0.048	0.990
4	0.036	0.990	0.042	0.992

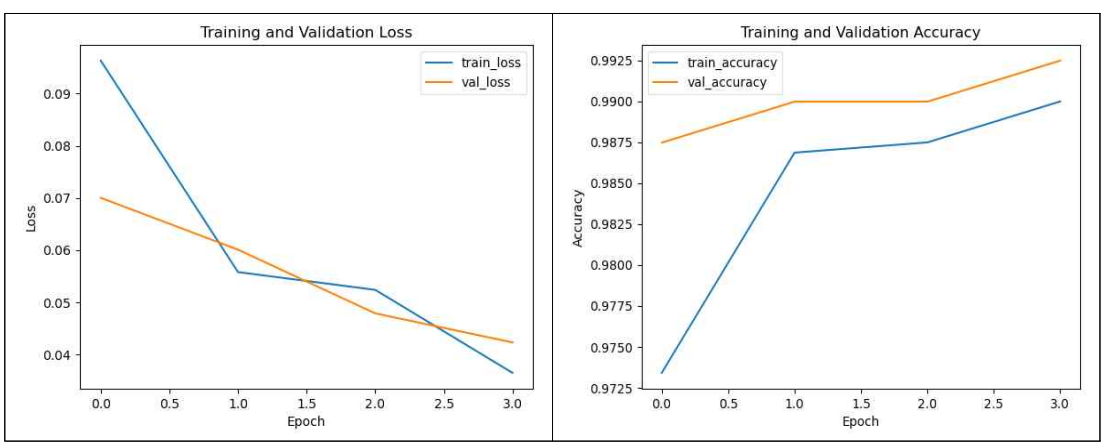

FIGURE 4. Result of address recognition training

TABLE 9. Comparison of learning performance by optimizer(Address cleaning)

Optimizer	Test_loss	Test_accuracy	Learning time
SGD	0.640	0.931	2:28:45
Adam	0.071	0.981	3:12:08
AdamW	0.062	0.985	2:42:33
BertAdam	0.083	0.992	4:20:05

TABLE 10. Address cleaning training’s Accuracy and loss results(SGD)

Epoch	Train_loss	Train_acc	Val_loss	Val_acc
1	1.905	0.335	1.515	0.531
2	1.477	0.507	1.190	0.700
3	1.254	0.637	1.000	0.796
4	1.108	0.723	0.869	0.852
5	1.006	0.781	0.778	0.888
6	0.933	0.818	0.709	0.911
7	0.879	0.845	0.662	0.926
8	0.854	0.856	0.653	0.929

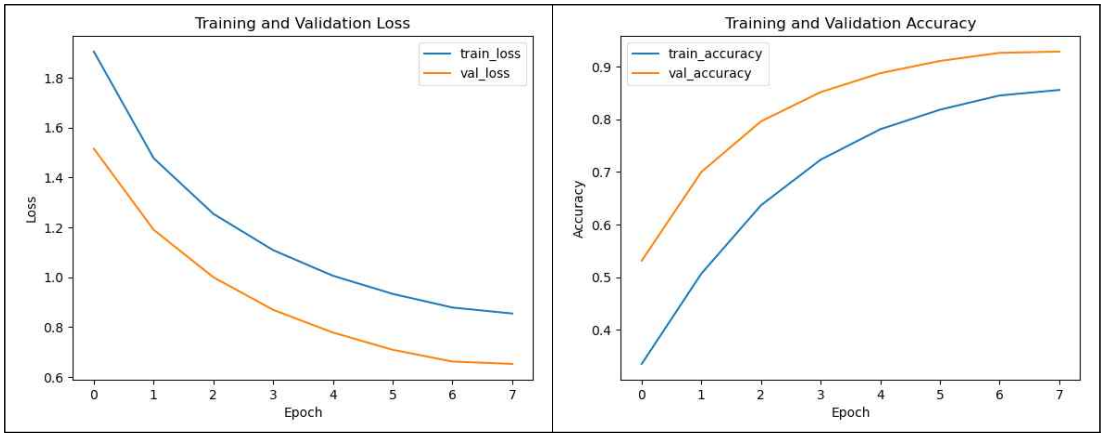

FIGURE 5. Result of address cleaning training

구’ 단위부터 주소가 기재된 내용, 그리고 기존 주소와 새주소를 혼용하여 기재하는 경우를 고려하여 학습을 진행하였다. 이러한 다양한 주소 표기 방식에 대한 학습을 통해 모델은 높은 성능을 보였다.

표 9는 주소 인식 모델 학습 과정에서 각 옵티마이저별 성능을 나타내며, 모델 학습은 19,821개의 샘플 데이터로 8회의 epoch 동안 진행하였고, 학습 과정은 표 10과 그림 5와 같다. SGD를 제외한 다른 옵티마이저는 높은 검증 정확도를 보였으나 과적합 문제가 발생하였다. 과적합은 학습 데이터셋의 특성, 모델의 복잡성 및 하이퍼파라미터 설정 등의 문제에서 발생할 수 있으며, BERT와 같은 대형 모델은 작은 규모의 데이터셋에서 과적합될 가능성이 더 높다는 사실을 확인하였다. 하지만, 주소 정제 학습 데이터셋은 문서 유형 분류 또는 주소 인식 학습 데이터셋보다 많은 샘플 데이터로 구성되어 있음에도 과적합이 발생하였다. 이는 학습 데이터셋의 구조적인 특성 문제와 Adam 관련

옵티마이저의 적응적 학습률 문제로 분석되었다. 적응적 학습률은 초기 학습 단계에서는 빠른 수렴이 가능하다는 장점이 있지만, 이것이 과적합을 초래할 수 있는 단점도 있었다. 반면, SGD는 고정된 학습률을 가지므로 이를 옵티마이저로 선택하였다. 예상과는 다르게 AdamW의 L2 weight decay 및 정규화 기능이 도입된 성능은 Adam, BertAdam과 큰 차이가 없었고, 특정 경우에서는 SGD를 사용한 모델이 더 우수한 결과를 보였다.

3. 문서 분류 및 주소 추출 결과물과 시각화

결과 데이터는 문서의 파일명, 유형 분류 결과, 정제된 주소를 포함하고 있으며, 이는 그림 6과 같다. 웹 크롤링과 전처리 시, 자동으로 다운로드된 문서의 제목을 '공고일자_도시계획시설 사업명_고시공고번호' 형식으로 설정하였다. 이 설정은 후속 단계에서 문서 오류 파악과 효율적인 문서 선별에 도움을 주었다.

유형 분류 과정에서 '복수유형' 시설의 경우, 샘플 수가 더 많은 유형으로 분류되는 클래스 불균형 문제가 발생하였다. 예를 들어, 고시공고문에 '도시계획시설(○○유원지 내 ○○공원)' 과 같은 내용을 포함한다면, '유원지' 와 '공원' 중 샘플 수가 많은 유형으로 문서를 오분류하였다. 각 클래스별 샘플 수를 균등하게 설정하기 위해 샘플 수가 783개로 가장 많았던 '도로(Road)' 를 기준으로 샘플 추가 작업을 진행하였다. 추가된 샘플을 포함한 학습 데이터셋으로 모델을 학습시켰을 때, 기존 모델보다 낮은 96.0%의 정확도로 총 21,585개의 샘플 중 800여 건의 오분류가 발생하였다. 앞으로 학습 데이터셋의 불균형 문제를 효율적으로 해결해야 하며, 학습 데이터셋의 구조를 개선할 방안을 찾아야 한다.

주소의 예시로 '부산광역시 ○○○구 ○○○동 ○○-○○번지 일원' 의 경우, '일원' 이라는 키워드는 주소를 추출하여 활용하기까지 불용어이기에 이를 제거하였다. 또한, 주소 표기 방식을 다양하게 학습시킨 결과, 전체적인 주소 구조를 포함하지 않더라도 정상적으로 인식하고, 추출하는 것을 확인하였다. 그러나, 주소 추출 과정에서도 문장 내 두 가지 이상의 주소가 포함되거나 일반적인 주소 구조가 아닌 '부산광역시 동구 초량동 E-1-2, E-1-3, E-1-4' 와 같은 신규개발지의 블록명일 경우, 이를 수정하는 추가 작업이 필요하였다. 고시공고문이 사업시행지와 사업시행자의 주소를 모두 포함하는 경우도 있어, 이를 선별하는 과정도 필요하였다. 위의 사항들을 감안하였을 때, 정상적인 문서 유형 분류와 주소 추출 작업을 수행할 수 있었다.

앞서 정제된 도시계획시설 유형과 주소 정보를 GIS로 시각화한 모습은 그림 7과 같다. 도시계획시설 고시공고문의 주소를 정제하여 Geocoding을 통해 위도 및 경도로 변환하였고,

document_title	type_classification	address_cleaning
20120829_도시계획시설[항만]사업 실시계획인가 고시_2012-336	철도/항만/공항	부산광역시 동구 초량동 E-1-2, E-1-3, E-1-4
20150225_도시계획시설사업[학교_ 부경대학교 용당캠퍼스] 실시계획 변경인가 고시_2015-60	학교	부산광역시 남구 용당동 산100번지
20151209_도시계획시설사업[종합의료시설 _ 부산성모병원] 시행자 지정 및 실시계획인가 고시_2015-446	의료시설	남구 용호동 558-32번지
20180711_도시계획시설사업[가야유원지 주차장 조성사업] 실시계획 변경인가 고시_2018-257	주차장 및 자동차정류장	부산광역시 부산진구 가야동 491-5번지
20190116_도시계획시설사업[하수도] 공사완료 공고_2019-109	하수/폐기물/우수/수질오염방지시설	부산광역시 사하구 신평동 659-2번지
20190515_도시계획시설[도로,광장]사업 실시계획 변경고시_2019-132	복합	부산광역시 금정구 장전동
20191009_도시계획시설[광장_120호광장]사업 실시계획 고시_2019-282	광장	부산광역시 강서구식만동 211-2번지
20201216_도시계획시설사업[어린이대공원] 실시계획변경인가 고시_2020-499	공원	부산광역시 부산진구 초읍동 산79-7번지
20201230_도시계획시설사업[연구시설_ 수산과학연구소] 실시계획[변경] 고시_2020-523	연구시설	부산광역시 기장군 일광면 동백리 256번지
20201230_도시계획시설사업[황령산유원지 _ 산책로 정비] 실시계획 변경 고시_2020-530	유원지	부산광역시 부산진구 전포동 산50-1번지
20210105_도시계획시설사업[연구시설 _ 재난안전산업지원센터] 실시계획 고시_2021-6	연구시설	부산광역시 동래구 수안동 666-10번지
20210120_도시계획시설사업[화지공원 골프연습장] 실시계획변경인가 고시_2021-21	공원	부산광역시 부산진구 양정동 산73-28번지
20210210_도시계획시설[도로_대로1-8호선]사업 실시계획 변경 고시_2021-42	도로	부산광역시 사하구 감천동 33-11번지
20210210_도시계획시설[민락유원지]사업 시행자지정 변경고시_2021-40	유원지	부산광역시 수영구 민락동 110번지
20210224_도시계획시설[공원_사상근린공원]사업 실시계획 변경인가 고시_2021-62	공원	부산광역시 사상구 감전동 산1-8번지
20210414_도시계획시설사업[공원_수민어울공원] 실시계획[변경]인가 고시_2021-124	공원	동래구 낙민동 135-1번지 일원
20210505_도시계획시설사업[완충녹지 9,10] 실시계획인가고시_2021-148	녹지	부산광역시 북구 화명동 2289, 2291번지
20210505_도시계획시설사업[지사공원] 실시계획 변경인가 고시_2021-143	공원	부산광역시 강서구 지사동 1180번지
20210512_도시계획시설사업[월륜 경관녹지] 시행자지정 및 실시계획인가 고시_2021-161	녹지	부산광역시 연제구 연산동 2310번지

FIGURE 6. Document classification and address extraction results

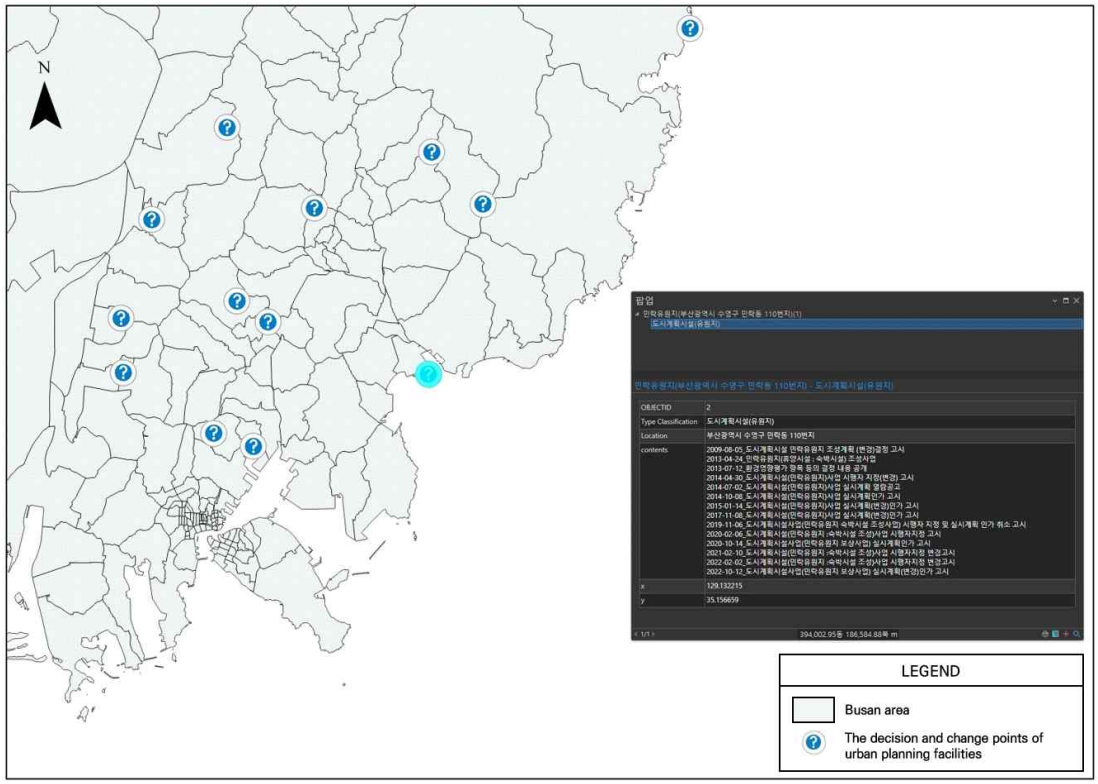

FIGURE 7. Visualizing GIS with Cleaned Data

도시계획시설 고시공고문 내용을 기반으로 도시계획시설 유형 분류 모델에서 추출된 결과와 정제된 주소, 해당 지점 도시계획시설 유형의 변경 및 결정 고시 이력이 DB에 축적되도록 하였다. 향후에는 축적된 데이터를 이용해 도시계획시설의 미래 변화나 필요성을 예측하는 모델을 개발하거나, 다른 분야의 데이터셋과 결합하면 다양한 관점에서의 공간적 분석과 같은 연구를 진행할 수 있을 것이다. 예를 들어, 교통, 환경, 사회경제적 데이터와 결합하면 도시계획의 영향을 다양한 측면에서 평가하고, 그 결과로 효과적인 계획안을 도출할 수 있을 것이다. 또한, 시각화된 데이터를 공개하여 도시계획 관련 이해관계자들이 데이터 기반의 합리적인 의사결정을 내리도록 지원할 수 있다. 이는 도시계획에 대한 시민들의 참여와 관심을 증진시킬 수 있는 플랫폼으로 제공할 수도 있다.

결 론

본 연구는 문서에 포함된 공간정보로 변경이 가능한 주소자료를 추출하기 위하여 세 가지의 새로운 BERT 모델을 구현하였다. 웹크롤링으로 부산광역시의 도시계획시설 고시공고문을 자동으로 수집하였으며, 샘플 데이터를 각 주제별 학습에 적합한 형태로 전처리 작업을 거쳐 데이터셋으로 구축하였다. Tensor 형태의 입력 및 출력 데이터를 사용하여 SGD와 AdamW 등의 Optimizer를 적용하였고, 일정한 epoch 간격으로 가중치를 업데이트하는 방식으로 모델을 설계하였다. 학습에는 14,063개의 문서 유형 분류 데이터, 주소 인식 데이터와 19,821개의 주소 정제 데이터를 구축하여 사용하였으며, 각 작업에 대해 적절한 파라미터 및 하이퍼파라미터를 설정하여 학습을 진행하였다. 또한, 정확도

및 손실값에 대한 그래프를 통해 학습 결과를 시각적으로 확인할 수 있었다.

모델의 학습 성능을 향상시키기 위해 전이학습과 하이퍼파라미터 재조정을 실시하였다. 학습 결과, 검증 정확도는 문서 유형 분류에서 96.6%, 주소 인식에서 98.5%, 주소 정제에서 93.1%로 확인되었다. 이 결과는 일반 문서를 활용한 공간 빅데이터 구축의 가능성을 시사하며, 이를 기반으로 공간계획의 이력을 활용한 다양한 시계열적 공간 분석을 진행할 수 있게 되었다.

최근 딥러닝 기반의 자연어 처리 기술이 많이 발전하였다. 이 기술은 검색, 챗봇, 요약, 음성인식 등의 분야에서 대용량 텍스트 데이터를 학습하여 높은 정확도의 결과를 도출하는 데 활용되고 있다. 현재도 다양한 딥러닝 모델이 계속해서 개선 및 개발되고 있으며, 본 연구를 통해 자연어 처리 분야의 연구 성과가 비정형 데이터의 공간 데이터 변환에도 적용될 수 있음을 확인하였다.

향후 연구에서는 BERT 모델이 아닌 다른 자연어 처리 모델들과의 성능 비교를 통해 어떤 모델이 특정 작업에 가장 효율적인지 평가하는 연구를 진행할 수 있을 것이다. 모델의 성능을 높이기 위해 추가적인 Fine-tuning과 고품질의 데이터셋 확보 방안을 탐색할 것이며, 다양한 유형의 문서에 대한 모델 최적화 연구도 진행될 것이다. 최근에는 TextRank와 같은 그래프 기반의 키워드 추출 알고리즘이 주목받고 있다. 이러한 알고리즘을 활용하여 문서의 핵심 키워드를 추출하는 것 외에도, 자연어 처리 기술과 결합하여 문서의 주요 내용을 빠르게 요약하거나 태그 처리를 진행할 수 있을 것이다. 이를 통해 이해관계자들이 문서의 핵심 내용을 빠르게 파악하고, 문서 관리 및 검색의 효율성을 높일 수 있을 것으로 기대된다. 또한, 공간정보를 포함한 문서 외에도 다양한 비정형 데이터와의 연계를 고려하여 통합적인 데이터 관리 및 분석 시스템 구축에 관한 연구도 진행할 수 있을 것이다. KAGIS

REFERENCES

Colin, R., Kevin, H. 2017. Spatial Context from Open and Online Processing (SCOOP): Geographic, Temporal, and Thematic Analysis of Online Information Sources. International Journal of Geo-Information. 6(193):1-15.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805v1 1-14.

Kim, H.D., Kim, J.Y., Hong, S.U., Kim, D.H. 2020. Analysis on Voice Phishing using Artificial Intelligence Named Entity Recognition Model for Information Search. The Journal of Police Science. 20(4):255-283 (김희두, 김종윤, 홍세은, 김대희. 2020. 인공지능 기반 개체명 인식 모델의 보이스피싱 여죄 분석 활용에 관한 연구. 경찰학연구 20(4):255-283).

Kim, I.H., Kim, S.H. 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 (김인후, 김성희. 2022. 딥러닝 기반의 BERT 모델을 활용한 학술 문헌 자동분류. 정보관리학회지 39(3):293-310).

Lee, C.J., Ra, D.Y. 2022. Korean Morphological Analysis Method Based on BERT-Fused Transformer Model. Journal of Information Processing Systems. 11(4):169-178 (이창재, 나동열. 2022. BERT-Fused Transformer 모델에 기반한 한국어 형태소 분석 기법. 정보처리학회논문지 11(4):169-178).

Lee, S.Y., Park, B.J. 2004. Design and Implementation of an Address correction

System for Standard Address. Master's Thesis, Univ. of Kwangwoon, Seoul, Korea. pp.1-58 (이상윤, 박병준. 2004. 표준주소 검색을 위한 주소보정 시스템 설계 및 구현. 광운대학교 대학원 석사학위논문. 1-58쪽).

Noh, Y.D., Cho, K.C. 2021. A Text Content Classification Using LSTM For Objective Category Classification. Journal of The Korea Society of Computer and Information 26(5):39-46.

Song, J.Y., Lim, H.C. 2014. A Korean road name address conversion model using hierarchical administrative division and word similarity. Master's Thesis. Korea University Graduate School of Computer and Information Technology, Seoul, Korea. pp.1-52 (송재용, 임해창. 2014. 행정구역 위계정보와 단어 유사도를 이용한 도로명주소 변환 모델. 고려대학교 컴퓨터정보통신대학원 석사학위논문. 1-52쪽).

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

You, H.J., Song, Y.S., Kim, M.S., Yun, G.H., Cheong, Y.N. 2021. Error Analysis and Evaluation of Deep-learning Based Korean Named Entity Recognition. Korean Journal of Linguistics. 46(3):803-828 (유현조, 송영숙, 김민수, 윤기현, 정유남. 2021. 딥러닝 기반 한국어 개체명 인식의 평가와 오류 분석 연구. 한국언어학회논문지 46(3):803-828).