KJCEM 24. 5. 035~043
September 30, 2023

www.kicem.or.kr (pISSN 2005-6095 / eISSN 2465-9703)
http://dx.doi.org/10.6106/KJCEM.2023.24.5.035

기계학습 알고리즘 기반 하자 정보 관리 시스템 개발 - 공동주택 전용부분을 중심으로 -

박다슬[1] · 차희성[2*]
[1]아주대학교 스마트융합건축학과 석사과정 · [2]아주대학교 스마트융합건축학과 교수

A Developing a Machine Leaning-Based Defect Data Management System For Multi-Family Housing Unit

Park, Da-seul[1], Cha, Hee-sung[2*]
[1]Grauduate Student, Department of Architectural Engineering, Ajou University
[*2]Professor, Department of Architectural Engineering, Korea University

Abstract : Along with the increase in Multi-unit housing defect disputes, the importance of defect management is also increased. However, previous studies have mostly focused on the Multi-unit housing's 'common part'. In addition, there is a lack of research on the system for the 'management office', which is a part of the subject of defect management. These resulted in the lack of defect management capability of the management office and the deterioration of management quality. Therefore, this paper proposes a machine learning-based defect data management system for management offices. The goal is to solve the inconvenience of management by using Optical Character Recognition (OCR) and Natural Language Processing (NLP) modules. This system converts handwritten defect information into online text via OCR. By using the language model, the defect information is regenerated along with the form specified by the user. Eventually, the generated text is stored in a database and statistical analysis is performed. Through this chain of system, management office is expected to improve its defect management capabilities and support decision-making.

Keywords : Apartment Buildings, Defect Management, Machine Learning, OCR, Generative Information Extraction

1. 서론

1.1 연구의 배경 및 목적

1980년대 이후 경제 성장과 함께 도시 인구가 급증하면서 공동주택은 주요 주택 유형으로 자리 잡았다(이동우, 2022). 국내 공동주택(아파트·연립주택·다세대주택) 연도별 주거 비율은 2010년 57.2%에서 2021년 63.3%로 꾸준히 증가하고 있는 양상을 보인다(통계청, 2021).

공동주택의 증가와 함께 하자 분쟁 건수 또한 증가하고 있다. '하자 심사·분쟁조정위원회'에 접수된 분쟁 기록에 따르면, 공동주택 하자 분쟁 건수는 2010년 69건에서 2019년 4,290건으로 약 62배 증가하였다. 이로 인해 하자관리의 중요성이 점차 커지고 있다.

이에 따라 공동주택에 대한 체계적인 하자 관리가 필요하다. 특히 공동주택 전용부분의 경우, 사적인 영역으로 구분되어 관리체계가 별도로 마련되어 있지 않다. 발생하는 대부분의 하자가 동일 또는 유사한 특징을 가지는 데도 이를 개별 건으로 처리하고 개별 세대와 하자 분쟁에 초점을 맞추어 관리한다. 이는 하자 분쟁을 관리의 대상이 아닌 대응해야 하는 대상이라고 보기 때문이다.

또한 관리사무소 중심의 하자 정보 관리 시스템이 도입되어야 한다(Yang et al., 2022). 현행 하자관리 시스템은 건설사 중심으로 운영된다. 따라서 관리사무소는 건설사에 내용을 단순히 전달 및 보고하는 데 그친다. 이는 건설사에 대한 관리사무소의 의존성을 높이고, 관리사무소의 하자관리 역량을 저해하는 요소로 작용한다.

따라서 본 논문에서는 관리사무소를 중심으로 전용부분 특성에 맞는 하자 정보 관리 시스템을 제안하고자 한다. 접

* **Corresponding author:** Cha, Heesung, Department of Architectural Engineering, Ajou University, 206 World Cup Road, Wonchun-dong, Yeongtong-gu, Suwon, Korea
E-mail: hscha@ajou.ac.kr
Received June 20, 2023: **revised** -
accepted September 11, 2023

수된 하자를 자동으로 전산화하고, 특정 양식과 항목에 맞추어 정보를 제공할 것이다. 입주자를 통해 접수된 하자는 전산화가 필요할 경우, OCR 모듈을 활용하여 전산화 과정을 거친다. 이후 하자 정보는 관리사무소가 원하는, 또는 건설사가 요청하는 양식에 맞게 변형된다. 이를 위해 딥러닝 기반의 사전학습 언어모델이 사용된다. 사전학습 언어모델을 기반으로 하자 정보를 특정한 형태로 변형하여 저장한다. 이후 추가로, 빈도 등을 분석하고 대시보드에 나타내어 관리사무소가 정보를 활용할 수 있게 한다.

이러한 시스템의 도입과 운영을 통해, 관련 업무의 질적 향상을 도모하고 건설사·하자 관련 업체와의 원활한 의사소통을 할 수 있을 것으로 기대한다.

1.2 연구의 범위 및 방법

본 연구는 관리사무소를 중심으로 하자 정보 관리 시스템을 제안한다. 이를 위해 하자 정보 취득 과정에서 발생하는 정보의 전반적인 내용을 기술하며, 시스템 적용 및 검증을 통해 시스템의 실효성을 검증할 것이다.

연구는 세 단계로 나누어 단계별로 진행되었다. (1) 먼저, 관련 선행연구와 각 건설사의 입주자 사전점검 리스트를 통해 시스템 분류 체계를 구축한다. (2) 다음으로 수립된 분류 체계를 바탕으로 기계학습 기반의 광학문자인식(OCR)·자연어처리(NLP) 방법론을 활용하여 시스템을 설계한다. 또한 하자유형 분류·주요 하자 키워드 도출 두 가지 기능을 추가로 달성한다. (3) 마지막으로 제안된 시스템을 구현하여 실현 가능성을 평가한다. 이를 요약하여 나타낸 흐름도는 〈Fig. 1〉과 같다.

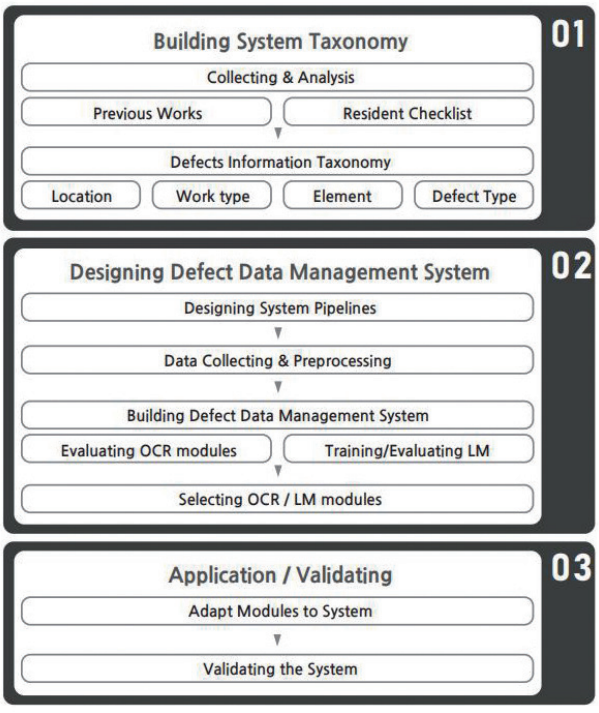

Fig. 1. Research Flowchart

2. 선행연구 분석

공동주택 하자 관리를 위해 분류 체계 및 프로세스 개선, 데이터 전산화, 시스템 자동화와 같은 다양한 관점에서 연구가 진행되었다.

우선, 프로세스 개선을 위해 거주자 또는 전문가의 요구를 반영하여 개선점을 도출하려는 시도가 있었다. 기존의 하자 정보 관리시스템을 고찰하고 거주자가 요구하는 기능을 적용하여, 개선모델의 구조와 데이터 흐름 계통도를 제시하는 연구가 있었다(강현욱 외, 2019). 이를 통해 거주자가 하자 정보를 정확하게 입력하고 건설회사와 정보를 공유할 수 있도록 하였다. 또한 서로 다른 공동주택 하자 체크리스트의 사용으로 인해 품질 점검에 대한 어려움이 발생함을 지적하고, 전문가들을 대상으로 설문조사를 진행하여 개선방안을 도출한 연구가 있었다(유병재 외, 2022).

반면 데이터 전산화 관점에서 입력 기기나 변환 시스템을 이용하여 문서를 전산화하는 연구가 진행되었다. 디지털펜을 활용하여 문서 작성 단계에서 전산화를 실현하고자 하는 연구가 있었다(손봉기, 2010). 이는 장치를 기반으로 근원적인 전산화를 실현할 수 있다는 이점이 있었지만, 접수자의 서체에 따른 인식률 차이 등의 문제점을 또한 보여주었다.

이와 달리 사용자가 작성한 문서를 정확하게 인식하고자 하는 연구도 있었다. 주로 업무자가 작성한 정보를 이미지로 간주하여 이미지 내에서 텍스트를 추출하는 광학문자인식(OCR)이 연구되었다. 광학 문자인식 기술을 활용하여 다양한 문서의 텍스트를 추출하고 이를 대규모 건설 프로젝트에 적용하려 시도하였다(Narges et al., 2019).

마지막으로 하자 접수 과정에서 발생하는 많은 양의 비정형 데이터 기반의 하자 정보를 처리하는 자동화 시스템에 관한 연구 또한 진행되었다. 이미지 형태의 하자 문서를 건설사의 하자 관리 시스템에 재입력해야 하는 불편을 해소하기 위해, 기계학습 기반의 하자 정보 자동 분류 시스템을 제안하였다(김은혜 외, 2021). 제안된 시스템은 BoW, TF-IDF 두 가지 단어 임베딩 방법과 서포트 벡터 머신(Support Vector Machine), 랜덤 포레스트(Random Forest) 두 분류기를 통해 하자보수 시설 공사 중 마감 공사의 세부 공종을 분류하였다. 또한 주거 건물의 하자를 자동으로 식별하기 위해 언어모델 기반의 텍스트 분류 모델이 활용되었다 (Jeon et al., 2022; Yang et al., 2022). 사전학습된 언어모델 중 하나인 BERT (Bidirectional Encoder Representations from

Transformers)를 활용하였다. 입력된 하자 정보 텍스트 내 각 단어 또는 토큰들을 설정한 하자 관련 분류체계에 따라 분류하도록 하였다. 이를 통해 입력된 하자 문서 내에 단어 또는 어절에 해당하는 하자 유형을 자동 예측하였다.

이처럼 여러 연구가 다양한 층위에서 진행되었지만, 아쉬운 점 또한 존재한다. 우선 공동주택 관리자 측면보다 건설사를 중심으로 하자 관리 시스템에 관한 연구가 이루어졌다는 점이다. 위 연구들은 부분적인 분류 시스템 또는 기능에 초점을 맞추어 연구되었다. 이를 현장의 관리사무소에서 사용하기에는 사용성이 떨어지고 관리사무소의 요구를 반영하지 못한다. 특히 건설사가 보유한 하자 관리 시스템의 하자 데이터는 관리자와 소유자에게 공유되지 않고 있다. 이는 데이터가 공동주택의 품질관리에서 활용되지 않아, 2차 하자가 발생하는 등 부정적인 영향을 미친다.

또한 많은 연구가 공동주택의 공용부분에 초점을 맞추고 있다. 이는 전용부분은 사적인 영역으로 구분되기 때문이다. 하지만 전용부분에 대한 하자 접수는 주로 비전문가인 입주자로부터 발생하기 때문에 체계적인 관리가 필요하며, 이를 위한 연구가 필요하다.

따라서 본 연구에서는 관리사무소가 필요로 하는 기능을 구현하여 자체적으로 하자 정보를 다룰 수 있도록 시스템을 제안한다. 또한 상대적으로 연구가 덜 진행된 전용 부분에 대해 다룰 것이다.

3. 배경지식

3.1 광학문자인식

광학문자인식(Optical Character Recog-nition; OCR)은 이미지 내의 문자를 인식하여 디지털 형태의 텍스트로 변환하는 기술을 말한다. 현재 다양한 기업들이 광학 문자인식 서비스를 모델 또는 API 형태로 제공하고 있으며, 대표적으로 TesseractOCR, EasyOCR, ClovaOCR, Google Vision AI 등이 있다.

3.2 사전학습 언어모델

최근 자연어처리 과제에서 사전학습에 기반한 언어모델이 광범위하게 사용되고 있다. 사전학습(pre-training)은 인공 신경망 기반의 언어모델이 개별 과제를 학습하기 전에 언어의 보편적 특징 및 표현을 먼저 학습하는 것을 의미한다.

사전학습은 비지도의 방법으로 대용량의 데이터를 사용한다는 특징을 가지고 있다. 웹 텍스트를 수집하여 이를 기반으로 임의의 단어를 마스킹 후 그 마스킹한 토큰이 무엇인지를 맞추는 과제를 구성할 수 있다. 이처럼 사람이 직접 수작업하지 않아도 많은 데이터를 입력과 레이블의 쌍으로 구축할 수 있으며, 모델이 이를 학습하여 언어의 임베딩 벡터 공간을 구축하는 것이 사전학습이다.

사전학습은 어떤 과제, 어떤 구조의 모델로 학습하느냐에 따라 성능과 그 활용이 다양하다. 최근에 연구되고 있는 사전학습 모델은 대부분 트랜스포머(Transformer) 구조를 변형 또는 일부만 사용하고 있으며, 대표적으로 BERT, GPT, BART, T5 등이 있다.

그중 본 논문에서는 T5 모델을 사용하여 학습 및 평가한다. T5는 사전학습 언어모델 중 하나로, 여러 과제를 '텍스트-투-텍스트(Text-To-Text)' 과제로 치환하여 사전학습한 것이 특징이다. 트랜스포머 기반의 인코더-디코더(Encoder-Decoder) 구조를 바탕으로 텍스트를 입력받아, 그에 대응되는 텍스트를 생성하도록 설계되었다.

4. 하자 정보 관리 시스템(DDMS)

4.1 시스템 하자 정보 분류 체계 구성

본 연구에서는 입주자 사전점검 리스트를 기반으로 분류체계를 구성하였다. 우선 위치, 공종, 부위, 하자 유형 네 가지의 큰 카테고리를 설정하였다. 이후 각 카테고리에 대한 세부 범주를 설정하여, 계층적 분류 체계를 구성하였다.

위치는 127m² 아파트의 일반적인 구조를 기준으로 현관, 거실, 주방, 침실, 화장실, 드레스룸, 발코니, 다용도실 8개 항목으로 설정하였다. 공종은 건축, 전기, 소방, 설비 4개 항목으로 설정하였다. 부위별 분류표는 바닥, 벽체, 천장, 문, 창호, 가구 6개 항목으로 설정하였다. 마지막으로 하자 유형은 균열, 파손, 누수, 들뜸, 오염 등 33개 항목으로 설정하였다. 추가로, 설정한 하자 유형별 분류표에 해당하지 않는 유형은 기타에 분류되도록 설정하였다. 이에 대한 자세한 내용은 〈Table 1〉에 첨부되어 있다.

Table 1. Defect Information Taxonomy

Category	Sub-category
Location	Porch, Living room, Dressing room, Kitchen, Toilet, Utility room, Balcony, Bedroom
Work type	Architecture, Electric, Fire fighting, Equipment
Element	Floor, Ceiling, Door, Wall, Window, Furniture
Defect Type	Crack occurrence, falling off phenomenon, leakage phenomenon, deformation phenomenon, breakage phenomenon, scratching phenomenon, lifting phenomenon, contamination phenomenon, fixing failure, condensation phenomenon, sagging phenomenon, discoloration phenomenon, poor finish, stamping phenomenon, noise generation, dent occurrence, horizontal failure, vertical defect, stripe defect, caulking defect, product defect, specification defect, opening and closing defect, draft defect, drainage defect, water pressure failure, painting defect, odor occurrence, Non-operational, micro-construction, malfunction, misconstruction, etc.

4.2 Defect Data Management System; DDMS

본 논문에서 제안하는 시스템은 이미지 형태의 문서를 디지털 텍스트로 변환하고, 이를 다양한 표현의 텍스트로 재생성하는 것을 목표로 한다. OCR 모델을 기반으로 이미지 형태의 하자 관련 문서를 디지털 텍스트로 변환한다. 이를 사용자가 원하는 텍스트 양식과 함께 언어모델에 입력하여 요약 및 정리된 텍스트를 생성한다. 최종적으로 생성된 텍스트는 내부 데이터베이스에 저장하고, 관리자가 의사결정에 활용할 수 있도록 시각화 등의 절차를 거치게 된다.

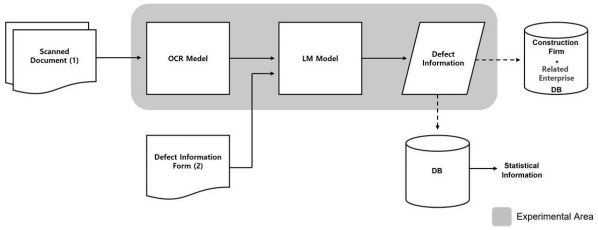

Fig. 2. System Framework

OCR 모델은 상용화된 API를 사용한다. 상용화된 무료 및 유료 API는 Tesseract-OCR, EasyOCR, Clova OCR, Google Vision AI 등이 있다. 본 연구에서는 공동주택 전용 부분에서 발생하는 하자 정보 문서를 텍스트로 변환하기 위해, 모델 선택의 과정을 거친다. 자체 구축한 문서를 네 개의 OCR 모델에 입력하여 성능을 평가하고, 그중 가장 성능이 뛰어난 API를 시스템 구조에 적합한 최적의 모델로 선정할 것이다.

디지털 텍스트로 변환된 하자 정보를 사용자가 원하는 양식에 맞게 재생성하는 것은 생성형 언어모델 중 하나인 T5를 사용한다. 생성형 언어모델은 입력이 주어지면, 그에 맞는 결과를 '생성'한다. 기존에는 분류 모델을 사용하여 입력 텍스트를 미리 정의된 범주에 맞게 분류하고 그 분류 범주에 해당하는 정보로 표시되었다. 이는 정보의 처리와 관리가 편리하지만, 범주가 추가되거나 체계를 재구성할 경우에 모델을 재학습해야 하는 불편함이 있었다. 따라서 본 논문에서는 이러한 문제점을 극복하기 위해, 미리 정의된 범주를 예측하는 것이 아닌 언어를 출력하는 생성형 언어모델을 사용하여 유연하게 정보를 이용할 수 있도록 하였다.

또한 다양한 생성형 언어모델 중 T5를 사용한 이유는 프롬프트 친화적인 모델이기 때문이다. 본 논문에서 제안하는 하자관리 시스템은 OCR을 통해 전산화된 텍스트를 언어모델을 사용하여 일정한 양식에 맞추어 재생성한다. 이때 제공하는 입력 양식을 고정된 형태가 아닌, 관리자가 설정할 수 있게 구성한다. 이를 언어모델의 관점에서 보면, 임의의 프롬프트에 따라 적절한 언어적 변환 기능을 제공해야 하는 것과 같다. 따라서 프롬프트 기반으로 사전학습한 언어모델 중 대표적인 T5를 사용하였다.

이를 위해서 하자 정보와 사용자가 변환하고자 하는 양식, 출력 레이블에 대한 학습/평가 데이터셋이 구축되어야 한다. 따라서 본 연구에서는 하자 정보, 사용자 양식, 출력 레이블을 자체적으로 구축하고 실험 및 평가를 진행한다.

마지막으로, 재생성된 하자 정보 텍스트는 데이터베이스에 저장되고 설정한 범주에 따라 분류 및 시각화되어 표현한다. 분류 및 시각화하기 위해서 공종, 부위, 위치, 하자 유형 각각의 범주에 대하여 간단한 인공신경망 기반의 분류 모델을 구성 및 사용할 것이다.

4.3 시스템의 적용

많은 인력과 전산화 시스템을 가진 건설사와 달리, 관리사무소의 하자 정보 관리는 비효율적으로 운영된다. 하자 정보에 대한 전문적인 지식 없이 건설사와 입주자 간의 의사소통을 단순 전달한다. 그렇기에 하자에 대한 빠른 대응과 종합적인 관리를 위한 의사결정에 어려움을 겪는다.

따라서 4.1에서 구성한 시스템을 관리사무소 하자 관리 업무를 위해 사용할 수 있다. 입주자가 접수한 하자 문서를 전산화하고, 이를 특정한 형식에 맞추어 일관된 표현으로 재생성할 수 있다. 또한 이를 기반으로 통계적 정보를 추출하고 대시보드에 표시하여 공통된 하자가 무엇이며, 여러 세대에서 발생하는 하자의 발생 패턴을 쉽게 파악할 수 있다. 또한 특정 기간 이후 건설사가 공동주택에 파견한 하자관리 인력이 부재하게 되는데, 이때 발생하는 하자관리 공백을 메울 수 있다.

5. 시스템 실험 및 검증

5.1 광학문자인식

5.1.1 광학문자인식 평가

본 실험에 활용된 데이터셋은 광학문자인식 모델을 위한 데이터셋과 언어모델을 위한 데이터셋으로 나뉜다. 두 데이터셋 모두 공동주택 하자 접수 관련 문서를 바탕으로 자체 구축하였다. 먼저 광학문자인식 데이터셋은 일자, 세대 명, 위치, 공종, 부위, 하자 내용 총 6개의 항목으로 구성된 문서 양식을 기반으로 인쇄체와 필기체로 정보를 구성하였다. 필기체는 다양한 서체에 대한 인식률을 높이기 위해 5명의 서로 다른 글씨체를 가진 데이터 주석자(data labeler)를 대상으로 구축되었다. 추가로, 광학문자인식 모델은 별도의 학습 과정을 거치지 않기 때문에 평가 데이터만 구축하였다.

데이터는 총 1,900개의 단어와 문장으로 되어 있으며 인쇄체 400개와 필기체 1,500개로 되어 있다. 데이터의 길이는

1자에서 10자 사이의 길이를 갖고 있으며, 평균 문자 길이는 5이다. 모든 문서는 광학문자인식 모델이 텍스트를 인식하고 추출할 수 있도록, 스캔 후 PDF 형식 또는 이미지로 변환하여 평가를 진행하였다.

5.1.2 언어모델 학습 및 평가

언어모델의 학습과 평가에 사용된 데이터셋은 직접 구축한 하자 접수 데이터셋과 KLUE 벤치마크·모두의 말뭉치 데이터셋이 있다. 우선 본 연구를 위해 구축한 하자 접수 데이터셋은 일자, 세대 명, 위치, 공종, 부위, 하자 내용 총 6개의 항목으로 입력을 구성하였다. 각 하자 내용에 따른 위치, 공종, 부위, 하자 유형 항목은 3명의 작업자가 교차 검증을 통해 구축하였다. 반면 일자, 세대 명 항목은 비지도 방법으로 생성하였다.

총 4,000개의 데이터셋을 구축하였으며, 그중 3,600개를 학습 데이터로, 400개는 평가 데이터로 사용하였다. 구축한 데이터셋은 언어모델을 조정 학습 및 평가하는 데에 활용되었다.

또한 실험에서는 언어모델을 조정 학습하기 전에 벤치마크 데이터셋을 기반으로 학습하였다. 조정학습과 유사한 유형의 과제를 설정하고 이를 위한 벤치마크 데이터셋을 선정하였다. '대화 상태 추적'과 '바꿔 쓰기' 과제를 하자 접수 문서 생성과 유사한 과제로 설정하였다. 본 실험에서 언어모델은 접수된 하자 관련 텍스트와 생성하고자 하는 양식인 프롬프트(prompt)를 입력으로 하여, 하자 관련 텍스트에서 필요한 정보를 추출 및 재표현(representation)하는 것을 목적으로 한다. 이는 대화 텍스트에서 주어진 슬롯(slot)에 대해 그에 맞는 값(value)을 생성하는 대화 상태 추적 과제와 유사하다. 슬롯이 프롬프트, 값이 재생성된 텍스트로 치환될 수 있다. 따라서, 실험에서는 KLUE 벤치마크 중 대화 상태 추적을 위해 구축된 WoS 데이터셋을 하자 접수 관련 데이터에 대하여 조정 학습하기 전에 학습할 것이다. 이러한 전이 학습(transfer learning)을 통해 구축한 적은 수의 데이터에 대해서 좋은 성능을 산출하는 것을 목표로 한다.

바꿔 쓰기 과제 또한 본 실험의 과제와 유사하므로, 벤치마크 데이터 중 하나를 선정하여 전이 학습을 하였다. 본 과제는 하자 접수 문서를 입력받아, 다른 형태로 재생성하는 과제이다. 이때 입력 텍스트와 재생성 텍스트는 의미상 변화가 거의 없다는 것이 특징이다. 이는 입력 문장과 의미는 같지만, 표현은 같은 문장을 생성하는 바꿔 쓰기 과제와 유사하다. 이에 실험에서는 국립국어원에서 제공하는 모두의 말뭉치 중, 유사 문장 말뭉치를 하자 접수 데이터셋을 조정 학습하기 전 학습하기 위한 데이터로 사용하였다.

추가로 언어모델을 통해 생성된 텍스트를 시스템 분류 체계에 맞게 분류하기 위한 데이터셋을 구축하였다. 시스템 분류 체계는 4장에서 설정한 분류 체계를 의미하며, 위치, 공종, 부위, 하자 유형 총 4개의 범주에 대하여 각각 분류할 수 있도록 설정하였다. 위치는 총 8개의 값을 가지며, 공종·부위·하자 유형은 각각 4개, 6개, 33개의 값을 가진다. 따라서 하나의 하자 접수 텍스트에 대해서 위치·공종·부위·하자 유형으로 각각 분류될 수 있도록 데이터셋을 병렬적으로 구성하였다.

5.2 평가 척도

5.2.1 OCR 평가 척도

본 연구에서는 Tesseract OCR, Easy OCR, Google Vision AI, Clova AI, 네 가지 OCR 모델의 성능을 평가하기 위해 편집 거리(Levenshtein Distance)를 기반으로 메트릭을 구성하였다. 편집 거리는 두 문자열 간의 유사도를 측정하는 알고리즘으로, 서로 다른 두 문자열이 같아지기 위해 몇 번의 편집을 해야 하는지를 계산한다. 여기서 편집은 삽입, 삭제, 대체, 세 가지 연산을 말하며, 최소 연산 수를 계산하여 두 문자열 간의 차이를 측정한다. 두 문자열 중 하나를 기준 문자열로 선택하고, 다른 문자열의 각 문자를 차례대로 비교하며 연산 수를 계산한다. 이를 통해 문자열의 각 위치에 해당하는 값들을 편집 거리 행렬(Levenshtein Distance Matrix)로 나타내게 된다.

5.2.2 언어모델 평가 척도

본 실험에서는 언어모델의 성능을 평가하기 위해 BLEU, ROUGE 두 가지 평가지표를 사용하였다. 먼저, BLEU (Bilingual Evaluation Understudy)는 입력이 순서 정보를 가진 단어들로 이루어져 있고, 레이블 또한 단어들의 시퀀스로 이루어진 번역 등과 같은 생성 과제에 사용되는 평가지표이다. BLEU는 길이에 대한 페널티 부분과 n-gram에 대한 각각의 precision에 대한 부분으로 구성되어 있다. 여기서 n-gram은 연속된 단어의 시퀀스를 나타낸다.

다음으로 ROUGE (Recall-Oriented Understudy for Gisting Evaluation)는 생성 과제를 평가하기 위한 평가지표 중 하나이다. ROUGE는 재현율 중심(Recall-Oriented)으로 평가지표가 설계되어 있어, 모델이 추출한 요약문에서 실제 요약 대상 문서의 중요한 내용을 얼마나 재현하였는지를 측정하는 요약 과제에 주로 사용이 된다. BLEU와 마찬가지로 ROUGE 점수는 0 에서 1 사이의 값을 가지며, 1에 가까울수록 예측값과 레이블값이 유사하다는 것을 의미한다.

5.3 모델

실험은 OCR 모델과 언어모델에 대하여 정량적, 정성적 평가를 진행하였다. 실험에 사용된 OCR 모델은 Tesseract OCR, Easy OCR, Google Vision AI, Clova AI이다. Tesseract

OCR, Easy OCR는 파이썬 라이브러리로 배포된 모델이며, Google Vision AI, Clova AI는 API 서비스로 제공되는 모델이다.

언어모델은 배포된 한국어 T5 모델 중 KoT5를 사용하였다. 여러 T5 모델 중 본 모델을 사용한 이유는 단어 수가 적게 구성되어 다른 모델보다 크기가 가볍기 때문이다. KoT5(https://huggingface.co/psyche/KoT5) 모델의 주요 파라미터는 〈Table 2〉와 같다.

Table 2. Model Parameters

Parameters	Value
feedforward dimension	3072
key-value dimension	64
model dimension	768
number of positions	512
number of decoder layers	12
number of heads	12
vocab size	32128

본 연구에서는 T5 사전학습 모델을 그대로 조정 학습 및 평가하는 것뿐만 아니라, 전이 학습을 통해 추가적인 성능 향상을 도모하였다. 본 실험과 유사하다고 판단되는 두 개의 벤치마크 과제를 각각 학습 후, 조정 학습하도록 하였다. 본 실험은 입력 텍스트와 프롬프트 텍스트를 주고, 입력 텍스트 내에 프롬프트 항목에 해당하는 텍스트를 찾아서 결합하여 생성하는 것을 목표로 한다. 이는 대화 속에서 특정 항목에 해당하는 슬롯을 예측하는 대화 상태 추적(Dialogue state tracking)과제와 유사하다. 또한 주어진 입력을 내용은 변함없이 재생성하는 바꿔쓰기 과제와 유사하다.

따라서 5.1장에서의 방법으로 벤치마크 학습을 위한 데이터셋을 구축하고, 이를 학습한 후, 자체 구축한 데이터셋에 대한 조정학습을 진행한다.

추가로 조정 학습한 언어모델을 통해 생성된 텍스트를 분류기로 4장에서 설정한 체계에 따라 분류하기 위해 BERT 기반의 텍스트 분류 모델을 구성하였다. 사전학습된 BERT 모델 중 'klue/bert-base' 모델에 각 분류 체계의 범주에 맞는 완전 연결 계층(fully-connected layers)을 추가하여 조정학습 및 평가하였다.

5.4 실험 환경 및 결과

5.4.1 실험 환경

본 연구는 AWS (Amazon Web Services)에서 제공하는 리눅스 서버에서 실험을 진행하였다. GPU는 NVIDIA의 A100 40GB 그래픽 카드 1장을 사용하였다. 학습과 평가에서 배치 크기(Batch Size)는 각각 16, 32로, 모델의 입력과 출력 시퀀스의 최대 길이는 입출력의 길이를 고려해 모두 256으로 설정하였다. 이외에도 학습 시 데이터 순서를 무작위로 섞어서 실험하였다

5.4.2 실험 결과

1) OCR 모델 평가

본 연구에서는 시스템 구조에 적합한 최적의 모델을 선정하기 위해 Tesseract OCR, Easy OCR, Google Vision AI, Clova AI, 네 가지 OCR 모델의 성능을 평가하였다. 위에서 살펴본 OCR 평가 데이터를 바탕으로 평가하였으며, 그 결과는 〈Table 3〉과 같다.

Table 3. Quantitative Results for OCR

Model	Levenstein Score (↑)
Tesseract OCR	0.3635
Easy OCR	0.5013
Google Vision AI	0.6333
Clova AI	**0.9441**

결과를 보면, 라이브러리 형태의 모델보다 API 형식의 서비스를 제공하는 모델이 더 높은 성능을 보인다. 파이썬 라이브러리인 Tesseract OCR, Easy OCR에 비해 API 형식의 서비스 모듈인 Google Vision AI, Clova OCR이 더 높은 인식률을 보인다. 이는 OCR 모델에서는 API 서비스 모델의 전처리, 후처리 기술 및 노하우가 요인으로 작용한 것으로 보인다. OCR 모델은 모델 자체 성능뿐만 아니라 전처리, 후처리 등의 과정도 중요하다. API 형태의 OCR 서비스는 자체 기술과 노하우를 가진 전처리, 후처리 모듈을 OCR 모델과 함께 사용한다. 그렇기에 API형태의 모델들이 전/후처리 모듈이 없거나 적은 파이썬 라이브러리에 비해 높은 성능을 보인 것으로 판단된다.

또한, 한국어에 대해 더 많은 도메인 정보를 가진 모델이 좋은 성능을 보인다는 것을 알 수 있다. Clova OCR은 한국어, 영어, 일본어에 특화된 서비스로, 해당 언어 도메인에서 좋은 성능을 보인다. 반면, Google Vision AI는 다국어 도메인에 대하여 학습된 모델로, 영어나 다른 언어에 대해서는 높은 성능을 보인다. 이에 한글 필기체로 구성된 하자 접수 정보에 대한 인식률이 Google Vision AI에 비해 Clova OCR이 높게 나타났다.

따라서 본 논문에서는 프레임워크 구성 시 Clova OCR을 시스템의 OCR 모델로 설정할 것이다. 추가로, 이후 성능이 향상된 모델 또는 서비스가 있으면 이를 프레임워크 내의 OCR 모듈로 대체할 수 있도록 구성할 것이다.

2) 언어모델 평가

구축한 하자 접수 데이터셋을 평가한 결과는 〈Table 4〉와

같다. 'FT'는 하자 접수 데이터셋에 대한 조정학습을 의미하며 'PG'는 바꿔 쓰기 데이터셋에 대한 학습, 'DST'는 대화 상태 추적 데이터셋에 대한 학습을 의미한다. '+'는 언어모델이 해당 학습을 하였다는 것을 나타낸다. 또한 'AUG'는 데이터 증강을 통해 데이터를 늘려 학습한 것을 말한다. 데이터 증강은 언어모델을 바꿔 쓰기 학습한 모델을 기반으로 조정 학습의 입력 데이터를 변형하여 새로운 데이터로 등록하는 절차를 통해 구현하였다.

결과를 살펴보면 증강되기 전, 후 모두 조정 학습만 한 모델보다 벤치마크 데이터셋에 대하여 학습 후 조정 학습한 모델이 높은 성능을 보여주었다. 증강 전에는 BLEU 점수에 관해서는 대화 상태 추적 데이터셋에 대하여 학습 후 조정 학습한 'T5(+DST+FT)' 모델이 성능이 높았다. 반면 ROUGE-1에 대하여 바꿔 쓰기 과제를 학습하고 하자 접수 데이터셋에 대해 조정 학습한 모델이 높은 성능을 보였다.

데이터 증강 후 조정 학습한 결과도 이와 비슷하다. 대화 상태 추적 데이터셋에 대하여 학습하고 하자 접수 데이터셋에 대하여 조정 학습 몇 평가한 'T5(+DST+AUG+FT)' 모델이 BLEU, ROUGE 모두에 대하여 다른 방법으로 학습한 모델 대비 뛰어난 성능을 보여주었다.

전반적으로 'DST'를 학습한 모델이 좋은 성능을 보였는데, 이는 과제 특성에 기인한 것으로 판단된다. 대화 상태 추적은 하자 접수 데이터에 대한 조정 학습과 유사하게 '텍스트+프롬프트'의 형태로 구성된다. 반면 바꿔 쓰기의 경우, 과제 기술(task description)을 제외한 별도의 프롬프트가 없다. 따라서 이와 같은 과제 유사성이 학습 성능에 영향을 미친 것으로 보인다.

Table 4. Quantitative Results for LM

Model	BLEU	BLEU-1	ROUGE-1
T5(+FT)	0.6499	0.8793	0.3290
T5(+PG+FT)	0.8454	0.9513	0.4574
T5(+DST+FT)	0.8477	0.9485	0.4374
T5(+AUG+FT)	0.8595	0.9248	0.4920
T5(+PG+AUG+FT)	0.8709	0.9569	0.4707
T5(+DST+AUG+FT)	**0.9608**	**0.9869**	**0.5006**

6. 결론

6.1 연구의 결과

본 논문에서는 광학문자인식과 언어모델을 활용하여 통합적인 하자 정보 관리 시스템을 구축하였다. 이미지 형태의 하자 접수 문서를 전산화하고 관리사무소 중심의 하자 관리를 위해 텍스트를 재생성하여 통계적 정보를 제공하는 기능을 각각 광학문자인식 모델과 언어모델로 구현하였다. 광학문자인식을 통해 스캔 된 문서를 전산화하고, 미리 설정한 양식 프롬프트와 함께 언어모델에 입력하여 원하는 문서로 재생성하고 이를 데이터베이스에 저장하였다. 이후 저장된 데이터는 빈도를 기반으로 그래프 등을 대시보드에 표현하였다.

또한 시스템의 성능을 기능별로 평가하였다. 광학문자인식의 성능을 평가하기 위해, 사용 가능한 라이브러리 및 API를 조사하였다. 조사를 통해 선정된 4개의 라이브러리와 API를 구축한 데이터를 기반으로 인식의 정확도를 평가하였다. 결과적으로 Clova OCR API가 가장 좋은 성능을 보였으며, 이를 시스템의 광학 문자인식 모듈로 설정하였다.

반면 언어모델은 T5를 기반으로 여러 학습 방법을 적용하여 그 성능을 평가하였다. 사전학습된 T5 언어모델 중 하나를 기본 모델로 선정하였다. 이후 구축한 데이터셋을 기반으로 조정 학습을 한 모델, 전이 학습 후 조정 학습한 모델로 나누어 평가하였다. 여기서 전이 학습은 조정 학습과 유사하지만, 다른 도메인의 데이터셋을 학습하는 방법을 말한다. KLUE 벤치마크 데이터셋 중 WoS 데이터셋을 기반으로 학습한 모델, 모두의 말뭉치의 바꿔쓰기 데이터셋을 기반으로 학습한 모델을 평가하였다. 추가로 조정 학습 데이터셋의 개수가 적어, 데이터를 증강하여 평가하였다.

평가 결과는 WoS 데이터셋을 기반으로 학습을 하고, 증강된 데이터로 조정 학습한 모델이 가장 좋은 성능을 나타내었다. 이는 전이 학습이 본 실험에서는 유효하였으며, 대화 상태 추적과 프롬프트 기반 하자 문서 변환 과제의 유사성에 기인한 것으로 판단한다.

6.2 한계점 및 향후 연구 방향

본 연구에서는 하자 접수 관련 문서와 레이블을 직접 구축하여 실험 및 평가하였다. 라벨링 작업에서 교차 검증을 통해 오류를 줄이고, 데이터 증강 기법을 통해 실제적인 데이터 수를 늘리는 등의 데이터 개선 작업을 하였다. 그런데도 언어모델을 조정 학습하기에는 그 수가 현저히 부족하고 다양성도 떨어진다. 따라서 강건(robust)하고 여러 프롬프트를 반영하는 모델을 위해서, 더욱 다양하고 많은 데이터가 필요할 것으로 보인다.

또한 시스템의 실증적인 검증과 평가가 필요하다. 본 연구에서는 관리사무소를 위한 기계학습 기반의 하자 정보 관리 시스템을 제안하였다. 시스템의 실효성을 평가하기 위해서 실제 현장에 적용할 필요가 있다. 시스템 적용을 통해, 현장에서 발생하는 기능적 문제점을 파악하고 개선하여야 한다.

추가로, 향후 세부공종 분류에 대한 실험 및 연구가 진행되어야 한다. 본 논문에서는 생성된 텍스트에 대해, 건축·전기·소방·설비 총 네 가지 대공종으로 분류하였다. 대공종에

는 여러 세부 공종들이 포함되어 있으며, 각각의 세부 공종들은 통계적으로 유의미한 정보로서 가치를 지닌다. 따라서 향후 세부공종 분류에 관한 연구가 진행되어야 한다. 이를 통해 하자관리 시스템의 효율성을 높이고, 이해관계자들 간의 원활한 의사소통에 기여할 수 있을 것이다.

이러한 한계점 및 미결 과제에도 불구하고, 본 논문에서 제안한 시스템은 범용적인 공동주택 하자 관리 시스템을 구축하는 데에 활용될 수 있을 것이다. 이미지 형식의 하자 접수 문서를 자동으로 전산화하고, 생성형 언어모델을 사용하여 관리자가 원하는 양식으로 정보를 재구성한다. 이는 비단 관리사무소의 전용부분 관리뿐만 아니라, 도메인 특성에 따라 다른 양식으로 정보를 표현 및 관리해야 하는 다양한 관리 영역에 활용될 수 있을 것으로 기대한다.

References

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A, Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B. Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). "Language models are fewshot learners." arXiv preprint arXiv:2005.14165.

Devlin, J., Chang, M.W., Lee, K., and Toutanova, K. (2019). Bert: Pretraining of deep bidirectional transformers for language understanding, In NAACL-HLT (1), arXiv:1810.0480.

Go, S.S., and Lee, H.M. (2016). "Modeling of Apartment Defect Management System applying UML." *Journal of the Architectural Institute of Korea Structure & Construction*, 22(7), pp. 123-130.

Jang, H.S. (2013). "A Study on the Building of Defect Information DB Management System of Apartment House for Defect Prevention." *The Korean Society of Disaster Information*, 9(3), pp. 300-314.

Jeon, K., Lee, G., Yang, S., and Jeong, H.D. (2022). "Named entity recognition of building construction defect information from text with linguistic noise." *Automation in Construction*, 143, 104543.

Kang, H.W., Park, Y.H., and Kim, Y.S. (2019). "Improvement Model of Defect Information Management System for Apartment Buildings." *Korean Journal of Construction Engineering and Management*, KICEM, 20(4), pp. 13-21.

Kang, J.Y. (2021). "Residents' Needs for Housing Unit Maintenance in a Apartment Building." *Journal Of The Korean Housing Association*, 32(6), pp. 111-118.

Kim, B.Y. (2022). Hybrid-Convolution Neural Network-Loss Distribution Approach-Association Rule Mining (HCLA) Defect Management Model for Automatic Classification and Characteristic Analysis of Apartment Defects. a domestic doctoral dissertation. Hanyang University.

Kim, E.H., Ji, H.G., Kim, J.N., Park, E.I., and Ohm J. Young. (2021). "Classifying Sub-Categories of Apartment Defect Repair Tasks: A Machine Learning Approach." KIPS Transactions on Software and Data Engineering, KTSDE, 10(9), pp. 359-366.

Kim, O.G., and Choe, J.H. (2010). "Current Status and Problems of Housing Defective Disputes." *Korean Journal of Construction Engineering and Management*, KICEM, 11(2), pp. 5-8.

Kim, S.W. (2019). "Legal Matters on the Construction Defect and Management Neglect of Apartment Houses." *The Korean Society Of Property Law*, 36(1), pp. 185-206.

Lee, D.W. (2022). Analysis of Defect Types and Management Methods of Apartment Houses Through Case Studies, a domestic master's thesis. Hanyang University.

Lee, J.B. (2006). "Methods for improving efficiency in Apartment Maintenance Management using on Analytic Hierarchy Process." *Korean Journal of Construction Engineering and Management*, KICEM, 7(4), pp. 69-77.

Lee, J.H. (2019). "Global research trends based on natural language processing (NLP) using unstructured text data in the construction industry." *Korean Journal of Construction Engineering and Management*, KICEM, 20(2), pp. 62-66.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. ICLR 2020.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). "BART: Denoising sequence-tosequence pretraining for natural language generation, translation, and comprehension." *In Association for Computational Linguistics*, ACL, pp. 7871 - 7880.

Oh, J.H., Song, Y.W., Choi, Y.K., and Lim, H.C. (2019). "Business Process Improvement of Defect Management in Apartment Housing Project." *Korean Journal of Construction Engineering and Management*, KICEM, 10(5), pp. 16-27.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu. P.J. (2020). "Exploring the limits of transfer learning with a unified text-to-text transformer." JMLR, arXiv:1910.10683.

Sajadfar, N., Abdollahnejad, S., Hermann, U., and Mohamed, Y. (2019). "Text detection and classification of

construction documents. In ISARC." *Proceedings of the International Symposium on Automation and Robotics in Construction*, IAARC, 36, pp. 446-452.

Son, B.K. (2010). "A Handwritten Document Digitalization Framework based Defect Management System in Educational Facilities." *Korean Institute of Sustainable Design and Educational Environment*, 9(2), pp. 1-11.

Subramani, N., Matton, A., Greaves, M., and Lam, A. (2020). A survey of deep learning approaches for ocr and document understanding. arXiv, arXiv:2011.13534.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., and Polosukhin. I. (2017). "Attention is all you need," Advances in neural information processing systems, 30, pp. 5998-6008.

Yang, D.U., Kim, B.Y., Lee, S.H., Ahn, Y.H., and Kim, H.Y. (2022). AutoDefect: Defect text classification in residential buildings using a multi-task channel attention network. Sustainable Cities and Society, 80(5), 103803.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J.G., Salakhutdinov, R., and Le, Q.V. (2019). "XLNet: Generalized autoregressive pretraining for language understanding." In Advances in Neural Information Processing Systems.

Yu, B.J, Bang, H.S., and Kim, O.K. (2022). "Improvement of Defect Checklists for the Quality Inspection of Apartment Houses." *Proceedins of the Korean Institute of Building Construction Conference*, 22(1), pp. 33-34.

Yun, S.H., Park, C.W., and Jeong, S.Y. (2008). "A Study on the Integrated Construction Defects Assessment of Apartments for Systematic Asset Management." *Journal of the Architectural Institute of Korea Structure & Construction*, JAIK, 24(8), pp. 179-186.

요약 : 공동주택 하자 분쟁의 증가와 함께, 하자관리의 중요성 또한 커지고 있다. 그러나 기존의 연구는 '공용 부분'에 초점을 맞추어 진행되었다. 또한 하자관리의 주체인 '관리사무소'를 위한 시스템 연구도 부족한 실정이다. 이는 관리사무소의 하자관리 능력의 부족과 관리 품질의 저하를 초래한다. 따라서, 본 논문에서는 관리사무소를 위한 기계학습 기반의 하자 정보 관리 시스템을 제안한다. OCR과 NLP 모듈을 사용하여 관리상의 불편한 점을 해소하는 것을 목표로 한다. OCR을 통해 수기로 작성된 하자 정보를 디지털 문서로 변환한다. 이후 언어모델을 이용하여 사용자가 지정한 양식과 함께 하자 정보를 재생성한다. 최종적으로 생성된 텍스트를 데이터베이스에 저장하고 이를 기반으로 통계적 분석을 실행한다. 이러한 일련의 과정을 통해, 관리사무소의 하자관리 역량을 향상할 수 있도록 돕고, 의사결정을 지원할 수 있을 것으로 기대한다.

키워드 : 공동주택, 하자관리, 기계학습, 광학문자인식, 생성형 정보추출