JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 29 No. 9, pp. 25-37, September 2024
https://doi.org/10.9708/jksci.2024.29.09.025

Development of an Automated ESG Document Review System using Ensemble-Based OCR and RAG Technologies

Eun-Sil Choi*

*Student, Dept. of SW and AI Convergence, Korea University, Seoul, Korea

[Abstract]

This study proposes a novel automation system that integrates Optical Character Recognition (OCR) and Retrieval-Augmented Generation (RAG) technologies to enhance the efficiency of the ESG (Environmental, Social, and Governance) document review process. The proposed system improves text recognition accuracy by applying an ensemble model-based image preprocessing algorithm and hybrid information extraction models in the OCR process. Additionally, the RAG pipeline optimizes information retrieval and answer generation reliability through the implementation of layout analysis algorithms, re-ranking algorithms, and ensemble retrievers. The system's performance was evaluated using certificate images from online portals and corporate internal regulations obtained from various sources, such as the company's websites. The results demonstrated an accuracy of 93.8% for certification reviews and 92.2% for company regulations reviews, indicating that the proposed system effectively supports human evaluators in the ESG assessment process.

▸ **Key words**: ESG Assessment, Supply Chain Management, Document Review Automation, OCR, RAG

[요 약]

본 연구는 ESG 서류 검토 과정의 효율성 향상을 위해, 광학 문자 인식(OCR)과 검색 증강 생성(RAG) 기술을 융합한 새로운 자동화 시스템을 제안한다. 제안된 시스템은 OCR 프로세스에 앙상블 모델 기반의 이미지 전처리 알고리즘과 하이브리드 정보 추출 모델을 적용하여 텍스트 인식의 정확도를 향상시키며, RAG 파이프라인에 레이아웃 분석 알고리즘과 재순위화 알고리즘, 앙상블 검색기 등을 적용하여 정보 검색과 답변 생성의 신뢰성을 최적화한다. 시스템의 성능을 평가하기 위해 온라인 포털에 게시된 인증서 이미지와 기업 웹사이트 등에 공개된 회사 내규를 사용하여 테스트를 진행한 결과, 인증서 검토에서 93.8%, 회사 내규 검토에서 92.2%의 정확도를 달성하며, 제안된 시스템이 ESG 평가 과정에서 인간 평가자를 효과적으로 보조할 수 있음을 보여주었다.

▸ **주제어**: ESG 평가, 공급망 관리, 서류 검토 자동화, OCR, RAG

• First Author: Eun-Sil Choi, Corresponding Author: Eun-Sil Choi
*Eun-Sil Choi (kkie89@gmail.com), Dept. of SW and AI Convergence, Korea University
• Received: 2024. 07. 30, Revised: 2024. 08. 13, Accepted: 2024. 08. 26.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

기후변화와 책임경영의 중요성이 대두됨에 따라 ESG가 기업경영의 핵심 요소로 부상하고 있다. 소비자들은 기업에게 환경친화적이고, 윤리적인 제품을 제공할 것을 요구하고 있으며, 투자자들은 ESG 경영을 통해 장기적인 비용 절감과 효율성 향상을 실현하는 것은 물론, 기업의 사회적 책임을 다할 것을 요구하고 있다. 기업의 가치와 지속가능성을 판단하는 기준이 재무 성과에서 환경(E), 사회(S), 지배구조(G)와 같은 비재무적 성과로 옮겨가고 있는 것이다. 이를 뒷받침하듯 ESG 공시나 성과평가를 위한 기준이 다양한 이니셔티브와 기관으로부터 연일 발표되고 있으며, 전 세계적으로 ESG에 대한 규제가 강화되고 있다.

국내 기업들도 시장에 발표된 주요 기준을 토대로 ESG 경영전략을 수립하고, 전담 인력을 배치하는 등 변화의 움직임을 보이고 있다. 다만, 실제 성취한 ESG 성과를 지속가능경영보고서를 통해 공개 중인 기업은 2023년 기준 161개 사로 여전히 미흡한 수준이다. ESG 정보 공시를 위해서는 해당 기업과 종속회사는 물론, 거래 중인 구매회사와 협력회사에 대한 데이터 확보가 필요하다 보니, 대다수의 기업이 선뜻 시작하지 못하고 있는 것이다.

이런 와중에 지난 4월, 유럽의회에서 EU 공급망 실사 지침(Corporate Sustainability Due Diligence Directive , CSDDD)이 가결되었다. 협력업체에 대한 ESG 성과 측정이나 정보 관리는 물론, 환경과 인권, 노동 분야에 대한 점검이 규제화된 것이다. 이에 따라 최근 대기업을 중심으로 협력업체에 대한 ESG 평가 도입이 가속화되고 있으며, 신속하고 정확한 공급망 평가를 위해 기업 간 다양한 논의가 이루어지고 있다.

공급망 관리를 위한 ESG 평가에는 서류 검토와 현장 실사, 결과 집계, 개선 방안 수립 등 여러 단계가 포함된다. 특히, 서류 검토의 경우 회사 내규와 같이 표준화되지 않은 자료들로 인해 많은 시간이 소요되며, 검토자의 실수로 인해 세부 항목에 대한 평가가 잘못될 가능성도 존재한다.

이에 본 연구에서는 광학 문자 인식(Optical Character Recognition, OCR)[1] 기술과 검색 증강 생성(Retrieval-Augmented Generation, RAG)[2] 기술을 활용하여 ESG 서류 검토를 자동으로 수행하는 모듈을 개발하고, 해당 모듈의 정확도를 검증함으로써, AI 공급망 관리시스템 구현의 가능성을 확인하고자 한다.

이어지는 2장에서는 공급망 관리를 위한 ESG 평가 모형과 서류검토 자동화의 이점에 대해 간략하게 정리하고, 서류검토 자동화에 사용될 OCR 기술과 RAG 기술에 대해 알아본다. 3장에서는 앞장에서 살펴본 기술의 한계점을 소개하고, 이를 개선하여 본 연구에 최적화하기 위한 방법을 제시한다. 4장에서는 자동화 모듈을 생성 및 평가하는데 사용될 ESG 지표와 증빙서류에 관해 설명하고, 3장에서 제안한 내용을 토대로 자동화 모듈을 구현하여 적용한 결과를 분석한다. 마지막으로 5장에서는 본 연구의 결론과 향후 과제에 대해 기술한다.

II. Theoretical Background

공급망 관리를 위해서는 ESG 평가 모형을 토대로 기업의 현 수준을 진단하고, 개선 방안을 마련하여 피 평가기업에 제공하는 과정이 필요하다. 그러나 ESG 평가를 위한 서류검토 작업에 많은 시간과 노력이 소요되다 보니, 목표기간 내에 전체 협력사에 대한 평가를 마무리하는 것이 녹록지 않은 실정이다.

이러한 문제에는 광학 문자 인식(Optical Character Recognition, OCR) 기술과 검색 증강 생성(Retrieval-Augmented Generation, RAG) 기술을 활용한 '서류검토 자동화'가 좋은 해결책이 될 수 있다.

본 장에서는 공급망 관리에 적용되는 ESG 평가 모형의 기본 개념과 서류검토 자동화의 필요성을 정리하고, OCR 및 RAG 기술에 대한 선행 연구를 살펴보고자 한다.

1. ESG Assessment

1.1 ESG Assessment Model

ESG 평가 모형이란 기업의 비재무적 리스크와 지속가능성을 진단하기 위해 고안된 모형으로, 기업이 작성한 설문지와 증빙서류, 각종 공공기관으로부터 수집된 데이터를 기반으로 ESG 평가 등급을 산출하는 것을 목적으로 한다.

ESG 평가 영역은 크게 환경, 사회, 지배구조 3가지로 나뉘며, 영역별 평가지표는 글로벌 표준과 기타 ESG 관련 법규 및 지침 등을 고려하여 개발된다. 각 영역별 주요 평가 내용과 목표는 다음과 같다.

먼저, 환경 부문에서는 에너지 사용이나 폐기물 관리, 오염물 통제 등 기업의 환경적 영향과 관행을 평가하며, 기업이 환경적 책임을 얼마나 잘 관리하고 있는지 측정하는 것을 목표로 한다. 다음으로, 사회 부문에서는 노동 관행과 안전보건 경영 체계, 지역사회 참여 등을 평가하며, 기업의 주요 이슈와 사회적 책임 이행 수준을 파악하는 것을 목표로 한다. 마지막으로, 지배구조 부문에서는 기업의 의사결정 구조와 정보공시 현황, 윤리경영 체계 등을 평가

하며, 기업이 투명하고 윤리적으로 경영되고 있는지 판단하는 것을 목표로 한다.

1.2 Benefits of Automated Document Review

서류검토 자동화란 OCR이나 자연어 처리(NLP) 등 여러 컴퓨팅 기술을 활용하여 문서나 이미지 파일에 포함된 텍스트 정보를 자동으로 추출 및 검증하는 프로세스를 의미한다. 이러한 자동화 시스템의 도입으로 서류검토 작업 중 발생하는 각종 인적 오류를 최소화할 수 있으며, 서류검토에 소요되는 시간과 비용을 절감함으로써 전반적인 운영 효율을 향상시킬 수 있다.

2. Technology Overview

2.1 OCR (Optical Character Recognition)

OCR은 이미지 내 텍스트를 디지털 형식으로 변환하는 기술로, 1960년대에 처음 도입되었으며 1980년대 이후 컴퓨터 기술이 발전됨에 따라 실용화 단계에 진입하게 되었다. 초기에 주로 은행 등 금융산업에 국한되어 사용되었던 OCR은 딥러닝 알고리즘의 출현으로 텍스트 인식률이 향상됨에 따라 점차 그 사용 범위가 확대되었으며, 오늘날에는 의료, 공공, 도소매 등 다양한 산업에서 데이터 처리를 위한 핵심 도구로 사용되고 있다[1].

OCR은 크게 글자 영역을 검출하는 '텍스트 탐지' 부분과 검출된 영역의 글자를 인식 및 판별하는 '텍스트 인식' 부분으로 나뉜다. 사용자는 텍스트 탐지에 앞서 이미지 전처리를 진행함으로써 OCR의 텍스트 인식률을 향상시킬 수 있으며, 텍스트 인식 이후에 맞춤법 검사나 문맥 분석과 같은 후처리를 진행함으로써 최종 결과물의 품질을 개선시킬 수 있다.

아래 그림 1은 OCR을 통해 이미지에서 텍스트를 추출하는 과정을 도식화한 것으로, 각 세부 과정에 대한 설명은 다음과 같다.

- **전처리(Pre-processing):** 문자 인식률 향상을 위해 입력된 이미지를 최적화하는 과정으로, 노이즈 제거부터 이진화, 밝기 및 대비 조정, 기울기 교정 등 다양한 방법을 적용할 수 있다.
- **텍스트 탐지(Text Detection):** 이미지에서 텍스트가 포함된 영역을 식별하는 과정으로, 텍스트의 위치를 나타내는 경계 상자(bounding box)를 예측하는 방법을 이용한다. 주로 객체 탐지 알고리즘을 활용하여 텍스트 영역을 탐지하며, CNN, Faster R-CNN, YOLO, SSD 등의 딥러닝 기반 기법을 활용할 수 있다.
- **텍스트 인식(Text Recognition):** 탐지된 텍스트 영역 내의 문자를 실제 디지털 텍스트로 변환하는 과정으로, 각 문자의 이미지 데이터를 분석하여 문자를 결정한다. 문자 인식에는 CNN, LSTM, Transformer 등 다양한 딥러닝 기반 알고리즘을 활용할 수 있다.
- **후처리(Post-processing):** 인식된 텍스트의 정확성을 향상시키기 위한 과정으로, 맞춤법 검사나 구문 분석, 문맥 기반 수정 등을 통해 최종 텍스트의 품질을 개선할 수 있다.

OCR 기술의 도입으로 디지털 시스템의 효율성이 크게 향상됨에 따라, 다양한 기업에서 OCR을 위한 오픈소스 라이브러리나 상용 서비스를 제공하고 있다.

상용 OCR 서비스는 높은 정확도와 다국어 지원, 지속적인 모델 업데이트 등 여러 가지 장점을 제공한다. 그러나 이러한 서비스는 주로 클라우드 기반으로 제공되기 때문에 망 분리 환경에서는 사용이 제한적이다. 또한, 데이터 프라이버시 문제와 비용 문제가 발생할 수 있으며, 특히 미세 조정(fine-tuning)에 대한 제약이 많아 연구 목적에 맞는 최적화가 어려운 경우가 많다. 이에 본 연구에서는 오픈소스 라이브러리 중 하나인 Tesseract를 활용하여 OCR 작업을 수행하고자 한다.

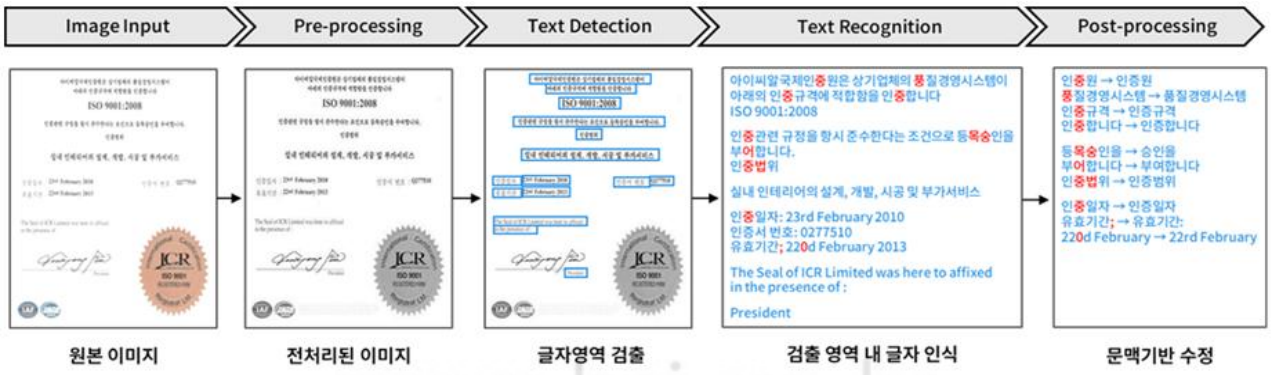


Fig. 1. OCR Process

2.2 RAG (Retrieval-Augmented Generation)

GPT나 Llama 시리즈와 같이 사전 훈련된 거대 언어 모델(Large Language Model, LLM)은 매개변수에 내재된 암시적 지식을 활용하여 외부 메모리에 접근하지 않고도 질의응답, 요약, 번역 등 다양한 자연어 처리 작업을 수행할 수 있다[3]. 이에 최근 연구 결과들은 이러한 LLM이 특정 작업에서 인간을 능가하는 성능을 보이며, 다양한 산업 분야에서 혁신적인 역할을 수행할 것으로 전망하고 있다.

그러나 이러한 잠재력에도 불구하고 LLM에는 여러가지 한계점이 존재한다. 학습 데이터에 기반한 LLM의 특성상, 지식의 동적 갱신이 어렵고, 추론 과정의 투명성이 부족하며, 사실과 다른 정보를 제공하는 환각(Hallucination) 현상을 보이기도 한다[4]. 이를테면, 기존에 경험하지 못한 지식에 대해 질의가 들어올 경우, 답변이 제한되거나 실재하지 않는 내용을 적당히 지어내서 답변하는 문제가 발생하는 것이다.

이러한 문제를 해결하기 위해 매개변수에 내재된 사전 지식과 외부 데이터베이스에 존재하는 신규 지식을 결합하는 하이브리드 모델[5-7]에 대한 연구가 지속적으로 진행되어 왔다. 2021년에 발표된 RAG 역시, 앞서 언급한 선행연구에 영향을 받아 발전된 것으로, 사전 훈련된 언어모델의 매개변수화된 지식과 외부 지식 베이스를 결합하여 LLM의 성능을 향상시키는 것을 목적으로 한다. 이를 위해 검색기는 외부 데이터에서 관련 정보를 찾는 역할을 수행하며, 생성기는 해당 정보를 토대로 최종 답변을 생성하는 역할을 수행한다[2].

따라서 RAG 기술을 적용할 경우, LLM은 사전 학습한 지식 외에도 최신 정보나 특정 도메인에 대한 세부 정보를 보강하여 답변할 수 있으며, 답변 시 참고한 정보의 출처를 함께 제공할 수 있어 시스템에 대한 정확도와 신뢰도를 향상시킬 수 있다. 아래 그림 2는 RAG의 작동 과정을 도식화한 것으로, 세부 단계에 대한 설명은 다음과 같다.

- **데이터 수집(Data Gathering):** LLM에 최신 지식이나 특정 도메인에 대한 심층 지식을 전달하기 위해 데이터를 수집하는 단계로, PDF, TXT, CSV, 웹 URL 등 다양한 형식의 원천 데이터를 사용할 수 있다.
- **데이터 로드(Data Loading):** 데이터 수집 단계에서 확보한 다양한 형식의 데이터를 시스템으로 불러오는 단계로, 원천 데이터의 구조를 파악하고 불필요한 텍스트를 제거하는 과정 등이 포함된다.
- **데이터 분할(Data Splitting):** 시스템에 로드된 데이터를 청크(Chunk) 단위의 작은 조각으로 나누는 단계로, 이를 통해 LLM에 불필요한 정보가 제공되는 것을 최소화할 수 있다. 다만, 청크 사이즈를 너무 작게 지정할 경우 핵심 정보가 누락되거나 텍스트의 맥락이 모호해지는 문제가 발생할 수 있어 주의가 필요하다.
- **데이터 임베딩(Data Embedding):** 청크 단위로 분할된 텍스트를 고차원의 숫자 벡터로 전환하는 단계로, 이를 통해 텍스트의 의미론적 특성을 보존하면서 LLM이 처리할 수 있는 데이터 형태로 변경할 수 있다. 임베딩에 사용되는 모델은 매우 다양하나, 우수한 성능을 위해서는 원천 데이터에 사용된 언어로 충분히 학습한 임베딩 모델을 선정하는 것이 필요하다.

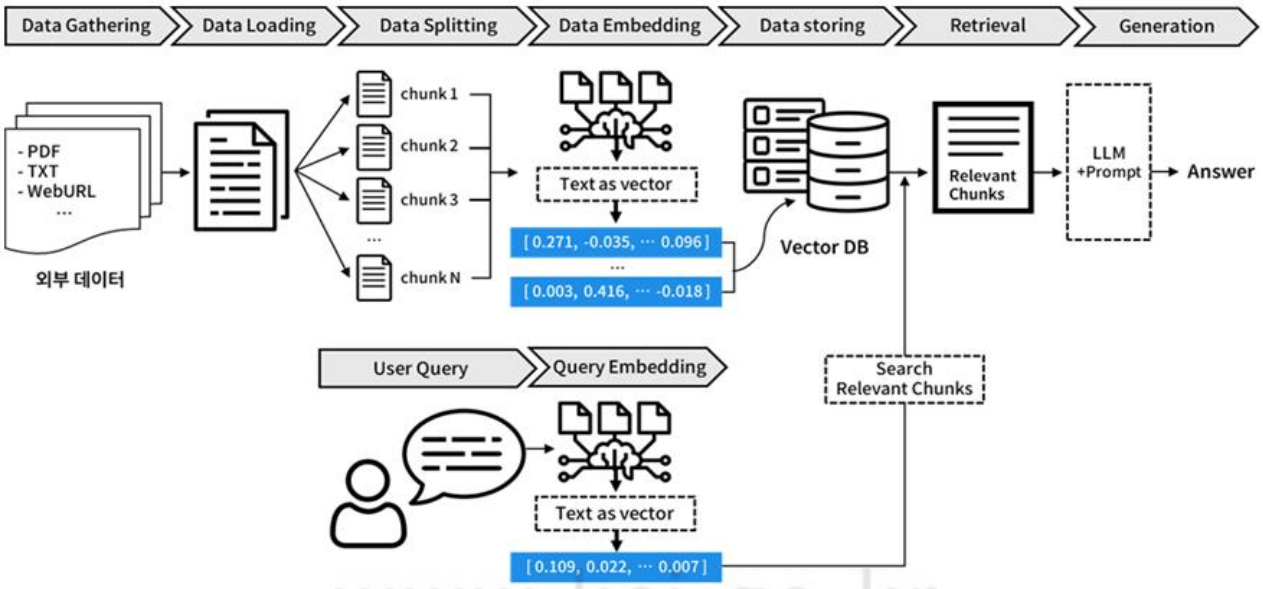


Fig 2. RAG Process

- **데이터 저장(Data Storing):** 임베딩된 청크와 메타 정보 등을 저장하는 단계로, 해당 데이터에 대한 저장소로는 주로 벡터 데이터베이스를 사용한다. 벡터 데이터베이스는 근사 최근접 이웃(Approximate Nearest Neighbor, ANN)[8] 알고리즘을 사용하여 고차원 벡터를 효율적으로 인덱싱하고 검색하기 때문에 사용자의 질문이 들어왔을 때 관련 정보가 포함된 청크를 신속하게 찾을 수 있다.
- **문서 검색(Retrieval):** 사용자의 질문과 관련된 정보를 검색하는 단계로, 먼저 사용자 질문을 벡터화한 후 코사인 유사도 등의 메트릭을 이용하여 가장 관련성 높은 청크를 찾는 방식으로 진행된다. 이 과정에서 키워드 기반 검색을 병행하는 등 앙상블 기법을 적용함으로써, 검색 결과의 정확도를 더욱 향상시킬 수 있다.
- **답변 생성(Generation):** 사용자의 질문과 문서 검색 단계에서 조회된 연관 정보를 토대로 최종 답변을 생성하는 단계로, 이때 Few-shot learning[9]이나 기타 특화된 프롬프트 템플릿 등을 사용하게 되면 답변의 품질을 향상시킬 수 있다. 한편, 최근 LLM에 대한 관심이 높아짐에 따라, 자체적인 LLM을 구축하거나 기존 모델을 미세조정(fine-tuning) 하여 사용하는 기업이 증가하고 있으나, 해당 작업에는 적지 않은 시간과 자원이 소요되는 바, 프롬프트 엔지니어링(Prompt Engineering)을 통해 오픈소스 LLM을 최적화하는 방식을 먼저 검토하는 것이 바람직하다.

RAG에 대한 효용성이 입증됨에 따라 LLM 애플리케이션 구축을 위한 오케스트레이션 프레임워크가 빠르게 발전하고 있다. 오케스트레이션 프레임워크를 사용하여 시스템을 구축할 경우, RAG에 포함된 다양한 컴포넌트를 일관된 방식으로 관리할 수 있으며, 새로운 기능 추가나 컴포넌트 교체가 용이해져 확장성이 향상된다. 또한, 각 컴포넌트의 상태와 성능을 모니터링할 수 있어 문제를 빠르게 감지하고 대응할 수 있다.

RAG 시스템 구축에 활용되는 대표적인 오케스트레이션 프레임워크에는 LangChain[10]과 LlamaIndex[11]가 있다. 먼저 LangChain은 모듈화된 구조와 풍부한 통합 기능을 제공하여, 복잡한 LLM 애플리케이션의 개발을 용이하게 만드는 특징이 있다. 특히, 다양한 LLM과 임베딩 모델, 벡터 저장소에 대하여 높은 호환성을 제공하며, 체인(Chain)과 에이전트(Agent)라는 개념을 토대로 복잡한 작업을 효율적으로 구현할 수 있어, 가장 빈번하게 사용되고 있다. 한편, LlamaIndex는 대규모 데이터셋에 대한 효율적인 인덱싱과 쿼리 기능에 특화된 프레임워크로, 다양한 데이터 소스에 대한 커넥터를 제공한다. 또한, 고급 쿼리 엔진 옵션을 통해 정확하고 신속한 정보 검색을 지원하는 특징이 있어, RAG 시스템 구현 시 자주 사용되고 있다.

본 연구에서는 커뮤니티 지원이 활발하고 유연한 확장성을 갖춘 LangChain 프레임워크를 채택하여 RAG 시스템을 구현하고자 한다.

III. The Proposed Scheme

1. Limitations of Existing Technologies

1.1 Limitations of OCR Technologies

OCR이 데이터 입력을 위한 새로운 도구로 자리매김함에 따라 최근 몇 년간 괄목할 만한 수준의 기술적 발전을 이루었다. 그러나 이러한 진보에도 불구하고, OCR 기술은 여전히 몇 가지 한계점을 가지고 있다.

가장 대표적인 한계점은 OCR의 정확도가 이미지의 품질에 따라 크게 영향을 받는다는 점이다. 예를 들어, 과도하게 밝거나 어두운 이미지, 해상도가 낮거나 노이즈가 심한 이미지 등은 정상적인 이미지 대비 현저히 낮은 텍스트 인식률을 보인다. 따라서, OCR 시스템의 안정적인 성능을 담보하기 위해서는 주어진 이미지에 적합한 전처리 과정이 반드시 선행되어야 한다.

그러나 적절한 전처리 기법을 선택하는 것은 결코 쉬운 일이 아니다. 이미지마다 요구되는 최적의 처리 방식이 다르며, 전처리에 사용되는 방법과 적용 단계도 매우 다양하기 때문이다. 일례로 이미지 전처리 단계에서는 입력된 이미지의 크기, 밝기, 대비 등을 세부 단위로 조정하는 것은 물론, 노이즈 제거나 흑백 전환, 샤프닝 등의 방식을 적용할 수 있다. 이 때문에 OCR을 활용한 시스템을 개발할 때는 이미지 전처리를 위한 알고리즘 구현에 많은 시간을 할애하게 되며, 이 과정에서 다양한 시행착오를 경험하게 된다.

OCR의 또 다른 주요 한계점은 문서의 구조적 특징을 효과적으로 해석하지 못한다는 점이다. 예를 들어, 다단 구조의 문서나 표(Table) 등이 포함된 문서에 대해 OCR을 진행할 경우, 텍스트 추출의 정확성이 크게 저하되는 문제가 발생한다. 이는 대다수의 OCR 엔진이 텍스트를 선형 방식으로 처리하도록 설계되어 있어, 문서의 레이아웃을 정확하게 파악하지 못하기 때문이다.

따라서, 복잡한 구조의 이미지에 대해 OCR 작업이 필요할 때는 문서의 레이아웃을 분석하고 세그먼트 간의 관계를 파악할 수 있는 고급 알고리즘의 적용이 필요하다.

OCR의 마지막 한계점은 학습된 폰트의 다양성에 따라 텍스트 인식률이 크게 달라진다는 점이다. 이는 대다수의 OCR 모델에서 공통적으로 나타나는 특성으로, 미리 학습한 폰트에 대해서는 높은 정확도를 보이지만, 그렇지 않은 폰트에 대해서는 미흡한 수준의 성능을 보이게 된다. 따라서 안정적인 OCR 결과를 얻기 위해서는 다양한 폰트를 추가로 학습하는 미세조정(fine-tuning) 과정이 필수적이다.

1.2 Limitations of RAG Systems

RAG는 거대 언어 모델(LLM)의 생성 능력과 외부 지식베이스의 확장성을 결합하여 답변의 정확도와 신뢰도를 개선하는 혁신적인 방법으로 주목받고 있다. 그러나 이러한 RAG 시스템에도 몇 가지 중요한 한계점이 존재한다.

가장 대표적인 한계점은 RAG 시스템의 성능이 구성 요소들의 개별 성능에 크게 의존한다는 점이다. 예를 들어, 데이터 분할 시 청크 사이즈가 부적절하게 설정되면 검색기(retriever)와 LLM 등 RAG 시스템 전반에 악영향을 줄 수 있으며, 임베딩 모델의 품질이 좋지 않으면 의미적 유사성을 제대로 포착하지 못해 연관 정보가 올바르게 검색되지 않을 수 있다. 또한, 검색기의 성능이 낮으면 관련 없는 정보가 LLM에 전달되어 부정확한 답변이 생성될 수 있고, LLM의 생성 능력이 부족하면, 전달된 정보의 품질에 상관없이 부정확한 답변이 생성될 수 있다.

따라서, RAG 시스템을 구축할 때는 각 구성 요소에 대한 개별 성능을 먼저 확인한 후, 이들 간의 시너지를 최적화하는 방향으로 시스템 고도화를 진행해야 한다.

RAG 시스템의 또 다른 주요 한계점은 '중간정보 소실' 문제가 발생할 수 있다는 점이다. LLM이 답변을 생성하기 위해서는, 검색기가 찾은 연관 정보와 사용자의 질문, 시스템 프롬프트 등이 전달되어야 하는데, 이러한 컨텍스트가 길어지게 되면 LLM의 정보처리 작업에 오류를 일으킬 수 있다. 전체 정보를 균형 있게 처리하지 못하고, 주로 시작과 끝부분에 위치한 정보에 집중하게 되는 것이다[12]. 이로 인해 중간에 위치한 중요 정보가 무시되거나 소실되는 현상이 발생할 수 있으며, 종단에는 잘못된 답변을 생성하는 결과로 이어질 수 있다. 따라서, 복잡하고 긴 컨텍스트를 다룰 때는 정보 소실 문제를 완화하기 위한 별도의 알고리즘 적용이 필요하다.

RAG 시스템의 마지막 한계점은 LLM을 사용하는 모든 시스템이 그렇듯, 환각(Hallucination) 현상이 발생할 수 있다는 점이다. 물론, RAG를 통해 외부 지식베이스에서 찾은 연관 정보를 LLM에 제공할 경우, 정보 부족으로 인한 환각 현상은 상당 부분 해소될 수 있다. 그러나 LLM이 검색 정보를 잘못 해석하거나 누락하여 발생하는 환각 현상은 기존과 동일한 빈도로 발생할 수 있다.

따라서 안정적인 RAG 시스템을 구축하기 위해서는 LLM의 응답에 대해 추가적인 검증 과정을 도입하는 것이 필요하다.

2. Improvement and Optimization Methods

2.1 Proposed method for OCR Technologies

본 연구에서는 기존 OCR 기술의 한계점을 극복하고, 텍스트 인식률을 개선하기 위해 다음과 같은 방법론을 제안한다.

첫째, 앙상블 모델 기반의 이미지 전처리

이미지 전처리에는 만능 해결책이 존재하지 않는다. 입력값으로 제공되는 이미지마다 최적의 전처리 방식이 다르기 때문이다. 이러한 특성으로 인해 일부 연구에서는 컨볼루션 신경망(Convolutional Neural Network, CNN)을 활용하여 이미지 전처리 기법을 자동으로 선택하는 알고리즘[13]을 개발하기도 했으며, 구조 측정 연산자(Structure Measure Operator, SMO)를 기반으로 과잉 전처리 된 이미지를 탐지하는 접근법[14]을 제안하기도 했다.

하지만 이러한 방식도 결국 절대적인 해결책이 되지 못한다. 위 접근법이 모든 이미지에 대해 완벽한 전처리 방법을 제시한다고 보장할 수 없으며, 이미지의 전체적인 특성과 지역적인 특성이 다를 경우, 단일 전처리 결과로는 충분한 수준의 텍스트 인식률을 담보할 수 없기 때문이다.

따라서, 본 연구에서는 다양한 전처리 기법을 병렬적으로 적용한 후, 각 전처리 결과에 대해 개별적으로 OCR을 진행하고자 한다. 이후, 필요한 정보를 추출하는 단계에서는 모든 OCR 결과를 종합하는 앙상블 기법을 사용하여, 단일 전처리 방법의 한계를 극복하고 OCR의 정확도를 향상시키고자 한다.

둘째, 한글 폰트에 대한 미세조정(fine-tuning)

현재 OCR 작업에 널리 사용되는 여러 오픈소스 엔진(Tesseract, EasyOCR, PaddleOCR)은 각각 미국, 태국, 중국 등 해외에서 개발되어 한글 폰트에 대한 학습이 충분하지 않다. 이 때문에 기본으로 제공되는 OCR 엔진을 사용하여 한글 문서에 대한 텍스트 추출 작업을 진행할 경우, 만족스러운 결과물을 얻는 것이 어려운 상황이다.

따라서, 본 연구에서는 기본적인 Tesseract 엔진에 한국에서 자주 사용되는 폰트를 추가로 학습시켜, 한글 텍스트에 대한 인식률을 개선하고자 한다.

셋째, 하이브리드 정보 추출 모델 적용

OCR 기술을 통해 추출한 텍스트에는 종종 오탈자나 형식에 맞지 않는 내용이 포함되어 있어, 정확한 정보 추출에 어려움을 초래한다. 전통적으로 이러한 OCR의 오류는 사람의 개입을 통해 검토 및 수정되어 왔으나, 최근 거대 언어 모델(Large Language Models, LLM)이 발전됨에 따라 이 과정의 자동화가 가능해졌다.

LLM은 방대한 텍스트 데이터로 학습되어 있어 OCR로 추출된 텍스트의 오류를 문맥에 맞게 교정할 수 있다. 또한, OCR 결과에서 기업명이나 만료일과 같은 특정 정보를 식별하고 추출하는 작업도 수행할 수 있어, 서류 검토 과정의 효율성을 크게 향상시킬 수 있다. 하지만 LLM이 제공하는 정보를 그대로 신뢰하기에는 무리가 있다. LLM을 활용한 텍스트 교정 및 정보 추출 과정에서 환각 현상이 발생할 수 있기 때문이다.

따라서 본 연구에서는 LLM을 활용한 정보 추출과 규칙에 기반한 정보 추출을 병행하는 하이브리드 모델을 사용함으로써, LLM의 출력 결과를 검증하고, 전체 시스템의 정확성과 신뢰도를 제고하고자 한다.

2.2 Proposed method for RAG Systems

본 연구에서는 기존 RAG 시스템의 한계점을 극복하고, 전반적인 시스템 성능을 향상시키기 위해 다음과 같은 방법론을 제안한다.

첫째, 원본 문서의 레이아웃 분석을 위한 알고리즘 개발

문서에서 텍스트를 추출할 때, 레이아웃에 대한 고려가 선행되지 않으면, 원본 데이터의 무결성이 훼손되거나 텍스트의 맥락 이해가 어려워지는 문제가 발생할 수 있다. 예를 들어, 다단 구조 문서의 경우, 좌측 컬럼을 모두 추출한 후 우측 컬럼을 추출해야 하는데, 일반적인 문서처럼 선형 추출 방식을 적용하게 되면, 원본 문서와 상이한 결과물이 산출될 수 있다. 또한, 문서 내에 존재하는 헤더나 푸터가 본문 텍스트와 혼재되어 추출될 경우, 문맥의 의미가 모호해지는 문제가 발생될 수 있다. 이러한 문제들은 검색기의 성능을 저하시켜 사용자의 질의와 관련된 정보를 정확하게 찾아내는 데 부정적인 영향을 미치며, 결과적으로 LLM이 부적절한 답변을 생성할 확률을 높이게 된다.

따라서 본 연구에서는 문서의 레이아웃을 분석하는 알고리즘을 개발하고, 이를 텍스트 추출에 앞서 적용함으로써 원본 문서의 의미론적 구조를 보존하고, RAG 시스템의 정보 검색 및 답변 생성 성능을 향상시키고자 한다.

둘째, 앙상블 검색기(ensemble retriever) 구현

앙상블 검색기는 다양한 검색 모델을 결합하여 정보 검색의 정확성과 신뢰성을 향상시키는 기법으로, 각 모델이 서로 다른 알고리즘과 접근 방식을 활용하여 데이터를 처리하기 때문에, 개별 모델의 한계를 상호 보완하고 종합적인 성능 향상을 도모할 수 있다.

앙상블 검색기의 대표적인 구현 방식은 희소 검색기(Sparse Retriever)와 밀집 검색기(Dense Retriever)를 결합하는 것으로, 두 검색 방식의 상호 보완적 특성으로 인해 'hybrid search'라고도 불린다[15]. 이 방식의 효과성을 이해하기 위해 각 검색기의 특징을 살펴보면 다음과 같다.

희소 검색기는 키워드를 기반으로 원본 문서를 탐색하기 때문에 임베딩 과정이 불필요하고 검색 속도가 빠르나, 동의어나 유사어 인식에 취약하다는 단점이 있다. 반면, 밀집 검색기는 의미적 유사성을 기반으로 관련 정보를 검색하기 때문에 동의어나 유사어 등을 감안하여 검색 결과를 제공할 수 있으나, 고성능 임베딩 모델이 전제되지 않으면 연관문서 검색 성능이 크게 저하될 수 있으며, 때때로 핵심 키워드가 누락된 청크를 가장 관련성 높은 정보로 선별하는 오류를 일으키기도 한다.

따라서 본 연구에서는 이러한 특성을 고려하여 희소 검색기로는 BM25 알고리즘을 채택하고, 밀집 검색기로는 FAISS 기반의 벡터 검색을 채택하여 앙상블 검색기를 구현하고자 한다. 또한, BM25 검색기에 한국어 특성을 고려한 형태소 분석기를 적용함으로써 사용자 질의에 대한 연관 정보 검색 성능을 제고하고자 한다.

셋째, Re-ranking 알고리즘 적용

검색기를 통해 획득한 연관 정보들은 사용자 질의와 관련성 측면에서 차이를 보인다. 검색된 정보 중 일부는 사용자의 질의와 높은 관련성을 갖지만, 다른 일부는 그렇지 않을 수 있는 것이다. 뿐만 아니라 LLM의 경우, 다수의 연관 정보가 함께 전달되는 등 컨텍스트가 길어지게 되면, 시작과 끝부분에 위치한 정보에 집중하는 경향이 있어, 중간에 위치한 주요 정보들을 무시하거나 누락하는 등의 문제를 일으키기도 한다[9]. 이러한 특성으로 인해, 사용자 질의에 대한 연관 정보가 정확하게 제공되었음에도 불구하고, LLM이 부정확한 답변을 생성하는 상황이 발생되기도 한다.

따라서, 본 연구에서는 문서 검색 단계 이후에, ESG 서류검토 자동화에 최적화된 재정렬(Re-ranking) 알고리즘을 적용함으로써, 검색된 연관 정보의 순서를 효과적으로 재구성하고, RAG 시스템에 대한 전반적인 성능 향상을 도모하고자 한다.

넷째, 다층적 Fact-checking 매커니즘 도입

RAG 시스템의 신뢰성 제고를 위해서는 생성된 답변에 대한 체계적인 사실 검증 과정이 필수적이다. 이는 단순히 정확성을 향상시키는 것을 넘어, 시스템의 투명성과 설명 가능성을 증대시켜 사용자의 신뢰를 획득하는 데 중요한 역할을 한다.

시스템의 출력에 대한 사실 검증 방법은 크게 선행 검증과 후행 검증으로 분류된다. 선행 검증은 최종 결론 도출에 앞서, 앙상블 기법이나 프롬프트 엔지니어링 등을 활용하여 LLM의 답변을 내부적으로 검증하는 방식을 의미한다. 이 과정에서는 다중 모델 앙상블(Multi-model Ensemble) 기법을 활용하여 여러 모델에서 얻은 출력을 비교하거나, 자기 일관성 검사(Self-consistency Check)를 통해 동일 모델의 다양한 출력을 평가할 수 있다.

후행 검증은 LLM이 산출한 최종 결론에 대해 인간 전문가가 직접 답변의 정확성을 평가하는 방식을 의미한다. 이 과정에서는 답변의 사실적 정확성뿐만 아니라, 맥락 적절성, 일관성, 그리고 윤리적 측면까지 종합적으로 평가될 수 있다. 특히, 후행 검증 과정에서 적재된 정보들은 향후 인간 피드백 기반 강화 학습(Reinforcement Learning from Human Feedback, RLHF)[16]을 위한 학습 자원으로 활용될 수 있기 때문에, 시스템의 지속적인 발전을 위해서는 이러한 검증 과정을 포함하여 Fact-checking 매커니즘을 설계하는 것이 유리하다.

따라서, 본 연구에서는 프롬프트 엔지니어링 등을 통해 LLM의 답변과 답변 내용에 대한 정확도를 수치화하여 제시하도록 시스템을 설계하고, 나아가 LLM 답변에 활용된 연관 정보를 발췌하여 사용자에게 함께 제공함으로써, LLM 답변의 정확성을 다각도로 검증하고 이를 데이터베이스화할 수 있는 환경을 구축하고자 한다.

IV. Experiments and Results

1. Experimental Design

1.1 Selection of ESG Indicators

본 연구의 주요 목적은 ESG 서류검토 자동화의 실현 가능성을 검증하는 데 있다. 따라서, 다양한 ESG 평가지표 중 서류 검토에 상당한 시간이 소요되거나 증빙서류로부터 데이터를 추출하는 과정이 복잡한 지표들을 중심으로 하여 총 11개의 대상 지표를 선정하였다.

본 연구에서 실험을 위해 선정된 ESG 평가지표와 증빙서류의 종류는 아래 표 1과 같다.

Table 1. ESG Indicators and Supporting Documents

ESG Indicators	Documents
ISO14001 retention period	Certificate
ISO50001 retention period	
ISO9001 retention period	
ISO45001 retention period	
ISO37001 retention period	
Family-friendly certificate Possession Status	
Anti-Discrimination regulation	Corporate Internal Regulation
Workplace harassment regulation	
Statutory allowance regulation	
Violation reporting regulation	
Whistleblower protection regulation	

위 11개 지표에 대한 증빙서류는 크게 인증서와 회사 내규로 분류되며, 각 서류 유형별 검토 사항은 다음과 같다. 먼저, 인증서에서는 기업명과 인증서명, 유효기간 등을 추출함으로써 제출된 증빙서류가 피 평가기업의 인증서가 맞는지 확인하고, 업체가 설문지 상에 기재한 인증서 보유 현황과 일치하는지 검토해야 한다.

다음으로 회사 내규에서는 기업명과 규정명, 세부 규정 내용 등을 추출함으로써 증빙서류가 피 평가기업의 내부 규정이 맞는지 확인하고, ESG 경영을 위한 세부 규정이 회사 내규상에 정확하게 명시되어 있는지 검토해야 한다.

1.2. Data Collection Method

서류검토 자동화 시스템 개발에 필요한 데이터를 확보하기 위해, 공공기관 경영정보 공개시스템(ALIO)[17]과 각 기업별 공식 웹사이트에 공개된 회사 내규 문서를 수집하였으며, 각종 포털 사이트에 등록된 인증서 이미지를 다운받아 취합하였다.

이렇게 수집된 데이터의 종류와 규모는 아래 표 2와 같으며, 본 연구에서는 해당 데이터를 7:3의 비율로 나누어, 학습용 데이터셋(Train dataset)과 평가용 데이터셋(Test dataset)으로 각각 활용하였다.

Table 2. Scale and Composition of Collected Data

Document	Count
Certificate (ISO, Family-Friendly)	161
Corporate internal regulation	258
Total	419

2. Experimental Results

3장에서 제안한 방법론을 바탕으로 구현된 '서류검토 자동화 시스템'의 구체적인 운용 프로세스와 성능 평가 결과는 다음과 같다.

2.1 Reviewing Certification documents

인증서 검토 과정은 이미지 전처리부터 답변 생성까지 총 5단계로 구성되며, 전체 프로세스는 그림 3과 같다.

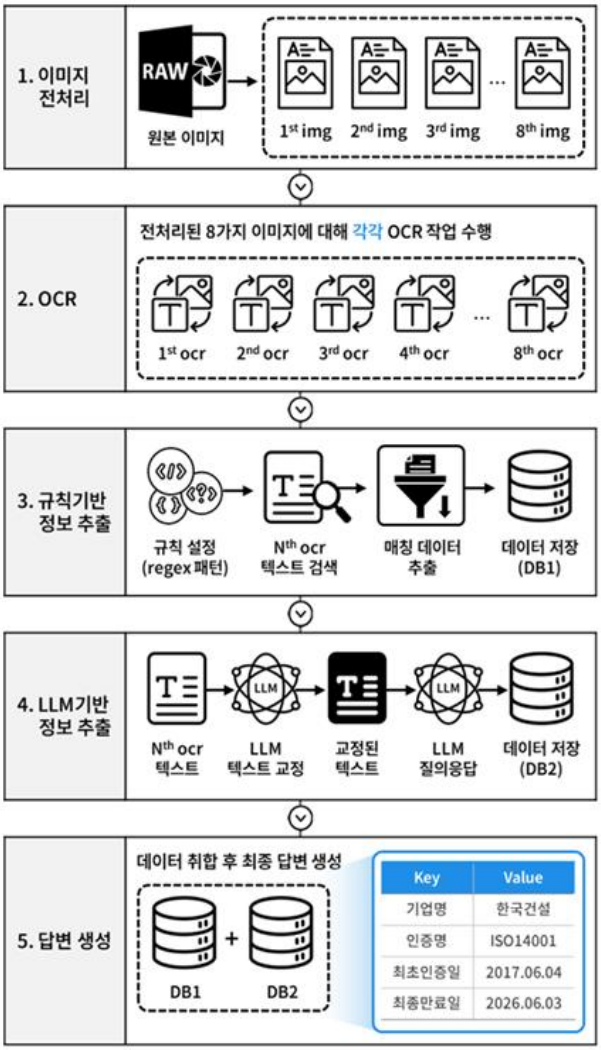


Fig. 3. Process of Reviewing Certification documents

먼저 이미지 전처리는 크기와 대비 조정 등을 활용하여 총 8가지 방식으로 수행되며, 이후 각 이미지에 대해 OCR 작업과 정보 추출 작업이 진행된다.

정보 추출 단계의 경우, 크게 2가지 방식이 병행되는데 먼저, 정규 표현식(Regular Expression)을 토대로 '규칙 기반 정보 추출'이 진행되며, 이어서 LLM을 활용하여 텍스트를 교정하고, 교정된 텍스트를 다시 LLM에 전달하여 필요 정보를 질의하는 'LLM 기반 정보 추출'이 진행된다.

최종 답변은 위 두 가지 결과 값을 취합하여 생성되며, 사용자에게 답변을 제공할 때는 그림 4와 같이 항목별 최빈값과 빈도수 정보를 함께 제공함으로써 답변에 대한 정확도를 사용자가 충분히 인지할 수 있도록 보장한다.

```
# 파일 경로
image_path = "[redacted]"

# 최종 결과값 추출
(company_name_ck, name_include_num, max_certification,
 max_certification_count, best_ocr_name, best_ocr_result,
 final_start_date, final_end_date) = determine_result(image_path, company_name)

# 결과 확인
print("="*60)
print("항목별 '최빈값(빈도수)' 산출 결과")
print("="*60)
print(f"기업명: {company_name_ck}({name_include_num}회)")
print(f"인증명: {max_certification}({max_certification_count}회)")
print(f"최초인증일 : {final_start_date}")
print(f"최종만료일 : {final_end_date}")
print("-"*60)
```

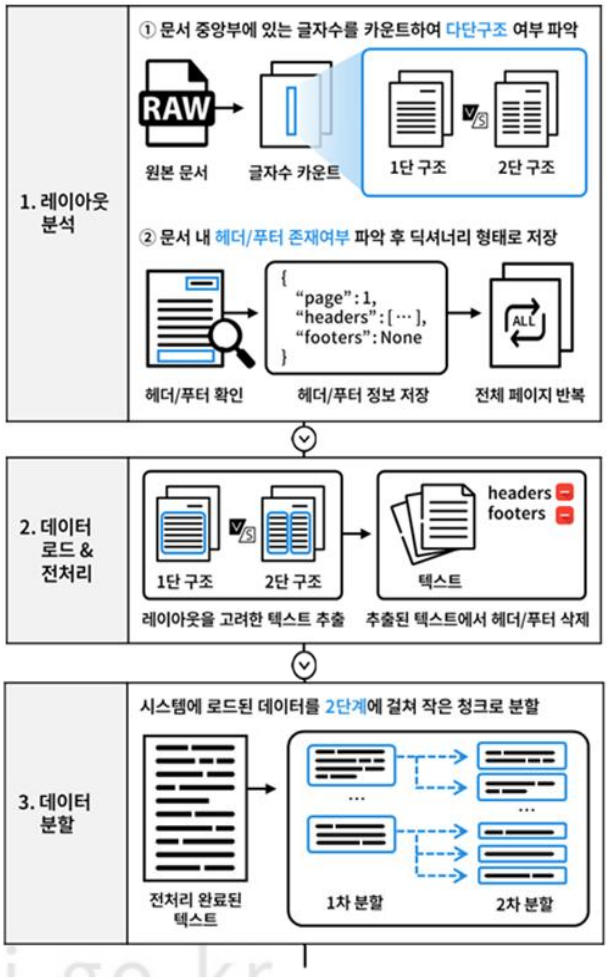
Last executed at 2024-07-21 16:01:12 in 17.63s

```
============================================================
항목별 '최빈값(빈도수)' 산출 결과
============================================================
기업명: [redacted](4회)
인증명: ISO14001(5회)
최초인증일 : 2020.06.23(7회)
최종만료일 : 2023.06.22(5회)
------------------------------------------------------------
```

Fig. 4. Result of Reviewing Certification documents

2.2 Reviewing Company Regulations

회사내규 검토 과정은 원본 문서에 대한 레이아웃 분석부터 답변 생성까지 총 6단계로 구성되며, 전체 프로세스는 그림 5와 같다.



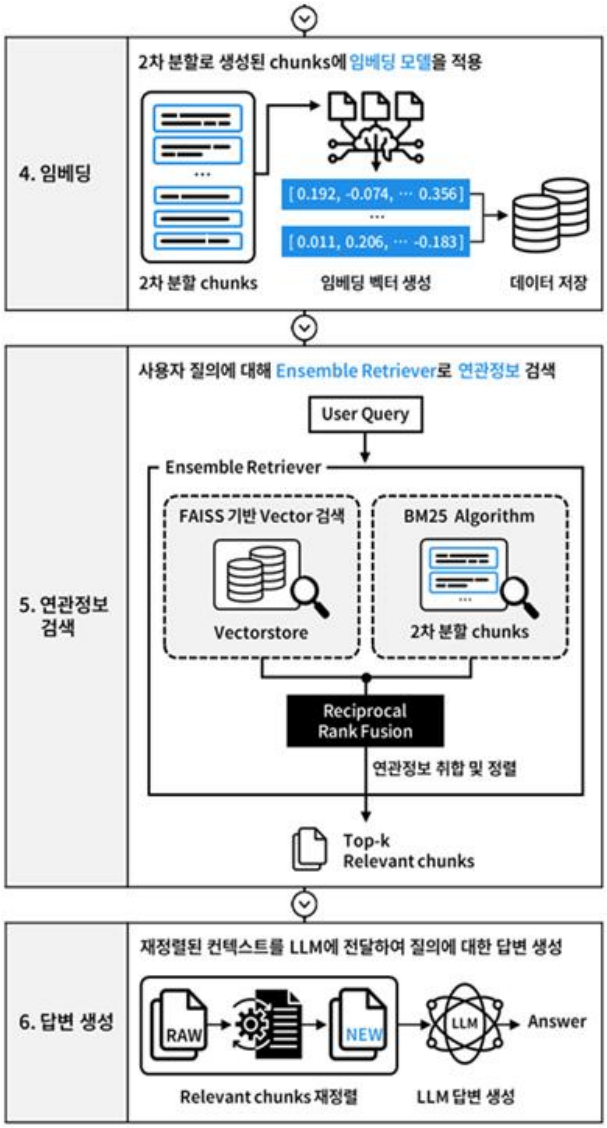


Fig. 5. Process of Reviewing Company Regulations

원본 문서에 대한 레이아웃 분석 작업은 크게 제출된 문서의 다단구조 여부를 파악하기 위한 작업과, 문서 내에 존재하는 헤더, 푸터 텍스트를 확인하는 작업으로 구분되며, 이를 위해 문서 중앙부에 위치한 글자 수를 파악하거나, 문서 내에 포함된 헤더, 푸터 텍스트를 딕셔너리 형태로 저장하는 작업 등이 수행된다.

데이터 로드 단계에서는 원본 문서의 구조적 특성을 감안하여 텍스트 추출이 진행되는데, 이때 앞서 생성된 헤더, 푸터 딕셔너리를 바탕으로 불필요한 텍스트를 삭제하는 작업이 함께 이루어진다.

이후 진행되는 데이터 분할 단계에서는 2단계에 걸쳐 전처리 된 텍스트가 여러 개의 작은 청크로 분할되는데, 1차 분할에서는 조항번호 등을 구분 값으로 하여 비교적 큰 사이즈의 청크가 생성되며, 2차 분할에서는 1차 청크가 더 작은 단위의 청크로 분할되면서, 연관정보 검색에 최적화된 소형 청크가 생성된다.

임베딩 단계에서는 허깅페이스에 공개된 'ko-sroberta' 모델을 토대로, 2차 분할된 청크를 고차원 숫자 벡터로 변환하는 작업이 진행되며, 이때 생성된 임베딩 청크는 이후 벡터 검색을 위한 데이터로 활용되게 된다.

연관정보 검색 단계에서는 FAISS 기반 벡터 검색과 BM25 알고리즘이 결합된 '앙상블 검색기'를 활용하여 사용자 질의와 연관된 정보를 검색하는 과정이 진행되는데, 이렇게 검색된 문서는 바로 LLM에 전달되지 않고 re-ranking 과정을 통해 LLM의 답변 성능을 극대화할 수 있도록 재정렬되어 전달된다.

마지막으로 답변 생성 단계에서는 LLM 프롬프트와 재정렬된 연관 정보 등을 참고하여 '예/아니오' 형태로 최종 답변이 생성되는데, 사용자에게 답변을 제공할 때는 아래 그림 6과 같이 답변에 참고한 정보를 함께 제공함으로써, 답변에 대한 정확성과 신뢰성을 제고한다.

```
# 파일 경로
file_path = '[redacted]'

# 데이터 로드 및 전처리
cleaned_docs = preprocess_pdf_fn(file_path)

# 데이터 분할 - 1,2단계
parent_docs, child_docs = split_doc_fn(cleaned_docs)

# 참고문서 추출 및 LLM 답변 수신
relv_docs = kiwibm25_faiss_64.invoke(query)
texts, response_text, eval_text = generate_answer_fn(llm, relv_docs, query)

# 최종 결과값 출력
print(Style.BRIGHT, end="")
print(f"Q. {Question}\n", Style.RESET_ALL)
print(f": {eval_text}\n")
print('-'*77)
print(Style.BRIGHT, end="")
print("참고문서 1.\n",Style.RESET_ALL)
print(relv_docs[0].page_content)
print(f"\n출처: {policy_title}")
print('-'*77)
```

```
Last executed at 2024-07-23 05:03:54 in 5ms
Q. 차별금지 규정을 보유하고 있습니까?

: 예, 포함되어 있습니다.

-----------------------------------------------------------------------------
참고문서 1.

제 19 조 (공정한 대우) 부산항만공사는 교육, 승진 등에 있어서 임직원 개인의 능력과 자질에 따라 균등한 기회를 부여하고, 성과와 업적에 대해서는 공정하게 평가하고 보상하며, 성별·학력·연령·종교·출신지역·신체장애 등을 이유로 차별하지 않는다.

출처: 윤리강령
-----------------------------------------------------------------------------
```

Fig. 6. Result of Reviewing Company Regulations

3. Analysis of Results

본 연구에서 개발한 서류검토 자동화 시스템의 성능 평가를 위해, 사전에 구축한 테스트 데이터셋을 활용하였다. 이 데이터셋은 인증서와 회사내규 문서로 구성되어 있으며, 시스템의 자동 검토 능력을 객관적으로 측정하기 위해 학습 데이터와는 별도로 구분하여 보관되었다.

3.1 Analysis of Certificate Review Results

인증서 검토 과정에서는 4가지 핵심 데이터 필드(기업명, 인증서명, 최초 인증일, 만료일)를 대상으로 정보 추출을 진행하였으며, 시스템의 성능 평가는 이들 추출 데이터의 정확도(Accuracy)를 기반으로 실시되었다.

$$Accuracy = \frac{Number\,of\,Correct\,Predictions}{Total\,Number\,of\,Predictions} \times 100$$

그 결과, 모든 평가 항목에서 90% 이상의 정확도를 달성하며, 본 시스템의 높은 성능을 입증하였다. 각 데이터 필드별 세부 정확도는 표 3에 제시되어 있다.

Table 3. OCR Accuracy for Data Extraction (n=49)

Cert. Type	Data Fields	Accuracy (%)
ISO Cert. (n_1=33)	Company name	97.0 %
	Certificate name	100 %
	Initial issue date	93.9 %
	Expiration date	97.0 %
Family-Friendly Cert. (n_2=16)	Company name	93.8 %
	Certificate name	100 %
	Initial issue date	100 %
	Expiration date	100 %

표 3의 데이터를 분석한 결과, 데이터 추출의 정확도가 인증서 유형과 데이터 필드에 따라 유의미한 차이를 보이는 것으로 나타났다. 이중 주목할 만한 점은 가족친화 인증서의 정확도가 ISO 인증서에 비해 전반적으로 우수한 성능을 나타냈다는 것이다.

이러한 성능 차이의 주요 요인으로는 인증서 발급 기관의 다양성과 양식의 표준화 정도를 들 수 있다. 가족친화 인증서는 여성가족부에서 단일 양식으로 발급되는 반면, ISO 인증서는 다양한 국가와 인증기관에서 각기 상이한 형식으로 발급된다. 이로 인해 ISO 인증서의 경우, 데이터 필드의 위치와 표기 형식의 예측이 어려우며, 배경 이미지나 폰트 종류, 크기 등의 변수가 텍스트 인식 정확도에 더 큰 영향을 미칠 수 있다.

특히, ISO 인증서에서 '최초 인증일' 필드가 가장 낮은 정확도를 기록하였는데, 이는 일부 ISO 인증기관에서 최초 인증일에 대한 정보를 다른 정보에 비해 현저히 작은 폰트로 기재하는 경향에서 기인한 것으로 판단된다.

그림 7은 이러한 문제로 인해 최초 인증일이 부정확하게 인식된 ISO 인증서의 사례를 나타낸다. 그림에서 파란색 박스로 표시된 부분이 최초 인증일이 기재된 부분이다.

Fig. 7. Example of ISO Certificates with Inaccurately Recognized Initial Certification Date

제시된 두 이미지에서 최초 인증일은 기업명이나 인증서명, 만료일에 비해 현저히 작은 폰트로 기재되어 있으며, 일부 사례에서는 육안으로도 식별이 어려울 정도로 작게 표기되어 있음을 확인할 수 있다.

따라서, 향후 시스템 고도화를 진행할 때는 위 내용을 고려하여 이미지 전처리 단계를 강화하는 등의 노력이 필요할 것으로 보인다.

3.2 Analysis of Certificate Review Results

회사내규 검토 과정에서는 데이터 로드와 분할, 검색 등의 과정을 거쳐, 표 4에 명시된 규정에 대한 정보를 추출하였다. 이후, 해당 정보를 거대 언어 모델(LLM)의 입력값으로 제공함으로써, 각 규정이 회사 내규에 포함되어 있는지를 확인하였으며, LLM 응답의 정확도를 기준으로 시스템에 대한 성능 평가를 진행하였다.

Table 4. Detailed Regulations List

No	Detailed Regulations
1	Anti-Discrimination
2	Workplace harassment prevention
3	Statutory allowance payment
4	Violation reporting
5	Whistleblower protection

분석 결과, 각 규정에 대한 LLM의 응답이 평균 93.8%의 정확도를 보이며, 우수한 수준의 성능을 나타냈다. 각 세부 규정별 정확도(Accuracy)와 위양성(False Positive), 위음성(False Negative) 건수, 위음성 오류에 대한 세부 분석 결과는 아래 표 5~7에 제시되어 있다.

Table 5. Accuracy of RAG system (n=77)

Detailed Regulations	Accuracy (%)
Anti-Discrimination	94.8 %
Workplace harassment prevention	94.8 %
Statutory allowance payment	93.5 %
Violation reporting	93.2 %
Whistleblower protection	92.2 %

표 5의 데이터를 분석한 결과, LLM 응답의 정확도는 최소 92.2%에서 최대 94.8%로, 규정 종류와 무관하게 유사한 성능을 보였다.

Table 6. False Positive and False Negative Counts

Detailed Regulations	FP	FN
Anti-Discrimination	1	3
Workplace harassment prevention	3	1
Statutory allowance payment	0	5
Violation reporting	4	1
Whistleblower protection	5	1
Total	13	11

Table 7. Distribution of False Negative Errors

Error Source	Count	Ratio(%)
Retrieval Error	1	9.1%
LLM Generation Error	10	90.9%
Total	11	100 %

표 6과 7의 데이터를 분석한 결과, 전체 오류 중 45.8%(11개)가 위음성 즉, 실제로 회사 내규에 포함되어 있는 규정을 누락된 것으로 잘못 판단하여 발생한 오류로 파악되었으며, 위음성 오류의 약 90.9%가 검색 엔진이 관련 정보를 정확히 식별하여 제공했음에도 불구하고, LLM이 해당 정보를 잘못 해석하고 답변을 생성하여 발생한 것으로 확인되었다.

본 연구에서 개발한 서류 검토 자동화 시스템의 경우, 답변을 제공할 때, 그림 6과 같이 답변에 참고한 연관 정보를 함께 제공하도록 설계되어 있다. 이러한 설계는 사용자가 제시된 연관 정보를 기반으로 LLM의 오답을 효과적으로 식별하고 정정할 수 있도록 지원한다. 따라서 본 시스템을 ESG 평가 프로세스에 적용할 경우, 사용자가 체감하는 효율성은 측정된 정확도를 상회할 것으로 예상된다.

다만, 장기적으로 봤을 때 ESG 서류 검토에 투입되는 인적 자원을 최소화하는 것이 바람직한 바, 향후 LLM의 환각(hallucination) 현상을 감소시키기 위한 추가적인 연구가 필요할 것으로 판단된다.

V. Conclusions

최근 EU 공급망 실사 지침(CSDDD)이 유럽의회에서 가결됨에 따라, 대기업을 중심으로 협력업체에 대한 ESG 평가 도입이 가속화되고 있다. ESG 평가는 서류 검토, 현장 실사, 결과 집계, 개선 방안 수립 등 다양한 단계로 구성된다. 이 중 서류 검토 과정은 평가의 기초가 되는 중요한 단계임에도 불구하고, 과도한 시간 소요와 높은 인적 오류 발생 가능성 등 여러 문제점을 내포하고 있다. 이에 본 연구에서는 광학 문자 인식(OCR) 기술과 검색 증강 생성(RAG) 기술을 기반으로 ESG 서류를 자동으로 검토하는 시스템을 개발함으로써, 이러한 문제를 해결하고자 하였다. 또한 OCR 프로세스에 앙상블 모델 기반의 전처리 알고리즘과 하이브리드 정보 추출 모델 등을 적용하고, RAG 파이프라인에 레이아웃 분석 알고리즘이나 re-ranking 알고리즘 등을 구현하여 적용함으로써, 기존 기술이 가지고 있는 한계를 극복하고, 서류 검토 자동화 시스템의 정확도를 향상시키고자 하였다.

본 연구에서 제안한 서류 검토 자동화 시스템의 분석 결과를 종합하면 다음과 같다.

첫째, 개발된 시스템의 성능을 평가하기 위해 온라인 포털에 등록된 인증서와 기업 웹사이트 등에 공개된 회사 내규를 입력값으로 하여 LLM 답변의 정확도를 분석하였다. 그 결과, 인증서 검토와 회사 내규 검토에서 각각 93.8%, 92.2%를 상회하는 정확도를 보이며, 전반적으로 우수한 성능을 입증하였다.

둘째, 회사 내규 검토에서 세부 규정과 무관하게 유사한 수준의 정확도를 기록한 반면, 인증서 검토에서는 인증 유형에 따라 유의미한 수준의 정확도 차이가 관찰되었다. 구체적으로, 가족친화 인증서가 ISO 인증서보다 전반적으로 높은 정확도를 나타냈는데, 이는 인증서 양식의 표준화 정도와 발급 기관의 다양성에 기인한 것으로 분석되었다. 또한, ISO 인증서 내에서도 최초 인증일 필드가 가장 낮은 텍스트 인식률을 기록하였는데, 이는 일부 인증기관에서 최초 인증일에 대한 정보를 다른 정보에 비해 현저히 작은 폰트로 기재하는 경향에서 비롯된 것으로 확인되었다.

셋째, 회사 내규 검토 중 발생한 오류에 대해 심층분석을 진행한 결과, 전체 오류의 45.8%가 위음성으로 나타났으며, 이 중 90.9%는 검색 엔진이 관련 정보를 정확히 식별했음에도 불구하고, LLM이 정보를 잘못 해석하여 발생한 것으로 확인되었다. 이는 RAG 시스템과 re-ranking 알고리즘의 적용에도 불구하고 LLM의 환각 현상이 완전히 해소되지 않았음을 시사한다.

본 연구는 AI 기술을 활용한 ESG 서류 검토 자동화의 가능성을 입증함과 동시에 후속 연구의 방향성을 제시한다. 개발된 시스템은 모든 분야에서 90%를 상회하는 수준의 정확도를 보이며, 서류 검토 자동화 시스템이 ESG 평가 과정에서 인간 평가자를 효과적으로 보조할 수 있음을 보여주었다. 또한, LLM의 환각 현상과 인증서 유형에 따른 OCR의 성능 저하를 한계점으로 지적하며, 후속 연구에서 주목해야 할 과제를 제시하였다.

향후 연구에서는 다양한 인증서 양식에 대해 적응력을 높이고, LLM의 추론 능력을 개선하는 데 초점을 맞춰 기술 개선을 진행해야 한다. 또한, 본 연구에서 다루지 않은 ESG 관련 문서들, 예를 들어 지속가능경영보고서나 환경영향평가 보고서, 이사회 의사록 등에 대한 분석 기능을 추가함으로써, 시스템의 포괄성을 높이는 작업을 함께 진행해야 한다. 이러한 방향성을 토대로 추가적인 연구와 기술 개발이 이루어진다면, ESG 평가 프로세스의 효율성 향상은 물론, ESG 경영의 확산과 기업의 지속가능성 제고에 크게 기여할 수 있을 것으로 기대된다.

REFERENCES

[1] Raj, Aaryan, et al., "Revolutionizing data entry: An in-depth study of optical character recognition technology and its future potential," International Journal for Research in Applied Science & Engineering Technology, Vol. 11, No. 2, pp. 645-653, Feb 2023. DOI: 10.22214/ijraset.2023.49108

[2] Lewis, Patrick, et al., "Retrieval-augmented generation for knowledge-intensive nlp tasks," Advances in Neural Information Processing Systems, Vol. 33, pp. 9459-9474, 2020.

[3] Roberts, Adam, et al., "How much knowledge can you pack into the parameters of a language model?" arXiv preprint arXiv:2002.08910, 2020.

[4] Marcus Gary "The next decade in AI: four steps towards robust artificial intelligence," arXiv preprint arXiv:2002. 06177, 2020.

[5] Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, Yoav Shoham, “In-Context Retrieval-Augmented Language Models,” Transactions of the Association for Computational Linguistics, Vol. 11, pp. 1316-1331. Nov. 2023. DOI: 10.1162/tacl_a_00605

[6] Petroni, Fabio, et al., "How context affects language models' factual predictions," arXiv preprint arXiv:2005.04611, 2020.

[7] Guu, Kelvin, et al., "Retrieval augmented language model pre-training," International conference on machine learning, pp. 3929-3938, 2020.

[8] Indyk, Piotr, and Rajeev Motwani, "Approximate nearest neighbors: towards removing the curse of dimensionality," Proceedings of the thirtieth annual ACM symposium on Theory of computing, pp. 604-613, May 1998.

[9] Ye, Hongbin, et al., Ontology-enhanced Prompt-tuning for Few-shot Learning, Proceedings of the ACM Web Conference, New York, United States, 2022. DOI: 10.1145/3485447.3511921

[10] LangChain Introduction, https://python.langchain.com/v0.2/docs/introduction

[11] LlamaIndex, https://docs.llamaindex.ai/en/stable

[12] Liu, Nelson F., et al., "Lost in the middle: How language models use long contexts," Transactions of the Association for Computational Linguistics, Vol. 12, pp. 157-173, Feb 2024. DOI: 10.1162/tacl_a_00638

[13] Bui, Quang Anh, David Mollard, and Salvatore Tabbone, Selecting Automatically Pre-Processing Methods to Improve OCR Performances, 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), pp. 169-174, Kyoto, Japan, 2017. DOI: 10.1109/ICDAR.2017.36

[14] Cheng, H. D. and Zhang, Yingtao, Detecting of contrast over-enhancement, 2012 19th IEEE International Conference on Image Processing, pp. 961-964, Orlando, FL, United States, 2012. DOI: 10.1109/ICIP.2012.6467021

[15] Ensemble Retirever, https://python.langchain.com/v0.1/docs/modules/data_connection/retrievers/ensemble

[16] Ouyang, Long, et al., "Training language models to follow instructions with human feedback," Advances in neural information processing systems, Vol. 35, pp. 27730-27744, 2022.

[17] Public Institution Disclosure Site, https://www.alio.go.kr

Authors

Eun-Sil Choi received the B.A. degree in Statistics from Inha University, Korea, in 2013. She subsequently worked at Ecredible, a corporate credit and ESG assessment firm, where she spearheaded data analysis projects, developed evaluation models, and orchestrated process innovation initiatives. She is currently pursuing advanced studies at the Graduate School of SW and AI Convergence, Korea University. Her research interests include Big Data Analytics, Generative AI, Process Innovation and ESG Risk Management.