스마트팜 분야에서 RAG 기반 컨설팅 시스템 개발에 관한 초기 연구

김진우, 최문택*
성균관대학교
robotjinu25@skku.edu, *mtchoi@skku.edu

Early Development of a Retrieval-Augmented-Generation-Based Consulting System for Smart Farming

Jinwoo Kim, Mun-Taek Choi*
Sungkyunkwan Univ.

요 약

본 논문은 스마트팜 환경에서 검색 증강 생성(RAG) 구조 기반 컨설팅 시스템의 효과를 분석하기 위해 RAG 구조를 적용한 질의응답 시스템을 제안한다. 농촌진흥청 등에서 발간한 작물 재배·생육 관리 매뉴얼을 기반으로 벡터 데이터베이스를 구축하고, 이를 대규모 언어 모델과 결합한 RAG 아키텍처를 설계하였다. 본 초기 연구에서는 명확한 평가를 위해서 단답형 위주의 QA 평가 쌍을 LLM을 이용해서 구축했고, 이를 기반으로 순수 언어모델 방식과 RAG 구조 간의 정량적 성능 비교 실험을 수행하였다. 실험 결과, RAG 기반 방식은 Exact Match, ROUGE, BERTScore, SBERT similarity 등 모든 평가 지표에서 순수 언어모델 대비 뚜렷한 성능 향상을 보였으며, 특히 SBERT similarity 0.937과 EM 0.770을 기록하였다. 이러한 결과는 스마트팜과 같은 전문 도메인에서 문서 근거 기반 RAG 구조가 신뢰도 높은 자연어 컨설팅 및 의사결정 지원 도구로 활용될 수 있음을 보여준다.

Ⅰ. 서 론

최근 스마트팜 환경에서는 재배 조건의 복잡화와 정보의 증가로 인해, 농업인이 다양한 재배 의사결정을 효율적으로 지원받을 수 있는 컨설팅 시스템의 필요성이 커지고 있다. 작물 선택, 재배 관리, 환경 대응과 같은 의사결정은 다수의 농업 지식과 문헌 정보를 종합적으로 고려해야 하며, 이를 자연어 형태로 직관적으로 질의하고 활용할 수 있는 시스템이 요구된다. 그러나 기존 스마트팜 의사결정 지원 시스템은 주로 센서 데이터나 기상 정보를 기반으로 한 규칙 기반 또는 통계적 모델에 의존하고 있어, 농업인의 자연어 질의에 유연하게 대응하는 데 한계가 있다.

한편 대규모 언어 모델(Large Language Model, LLM)은 우수한 자연어 처리 능력을 바탕으로 다양한 분야에서 활용되고 있으나, 스마트팜과 같은 전문 도메인에서는 도메인 특화 지식과 최신성 측면에서 한계를 지닌다. 이를 보완하기 위한 방법으로 검색 증강 생성(Retrieval-Augmented Generation, RAG) 아키텍처가 제안되었으며, 이는 외부 문서를 검색 모듈과 연계함으로써 자연어 기반 응답의 정확성과 신뢰성을 향상시키는 접근법이다[1]. 그러나 스마트팜 분야에서 농업 문서를 기반으로 한 자연어 컨설팅 시스템에 RAG를 적용하고 그 효과를 실증적으로 분석한 연구는 아직 제한적인 상황이다. 이에 본 연구에서는 스마트팜 환경을 대상으로 대규모 언어 모델과 RAG 아키텍처를 결합한 문서 기반 자연어 컨설팅 시스템을 구축하고, 스마트팜 분야에서 RAG 구조의 효과를 정량적으로 평가하고자 한다.

Ⅱ. 본론

2.1 원시 데이터 개요

본 연구에서 활용하는 원시 데이터는 농촌진흥청 및 관련 농업 연구기관에서 발간한 작물 재배 및 생육 관리 매뉴얼의 PDF 형식 문서 데이터이다. 해당 데이터는 딸기를 포함한 특정 작물의 생육 단계별 특성, 환경 관리 기준(온도, 습도, 광, 이산화탄소 등), 재배 작업 일정, 병해충 관리 방법, 품질 및 수량성 평가 기준 등을 전문가 조사 및 현장 실증 결과에 근거하여 체계적으로 서술한 자료로, 표·텍스트·이미지·수치 범위 등 다양한 형태의 정보가 혼재된 비정형 및 반정형 데이터의 특성을 가진다.

2.2 스마트팜 컨설팅을 위한 RAG 아키텍처

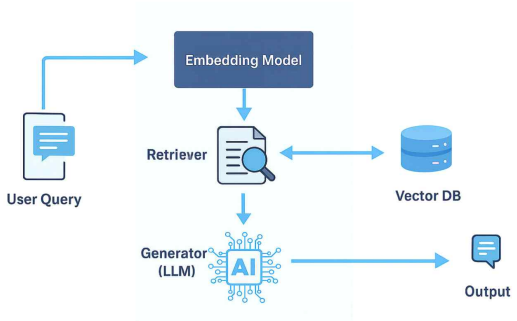

그림. 1. RAG 구조

본 연구에서는 그림 1에 제시된 전형적인 RAG 아키텍처를 따랐으며, 사용자 쿼리(질의)를 임베딩 공간으로 변환한 뒤 관련 문서를 검색하고, 해당 문서를 컨텍스트로 활용하여 생성 모델이 최종 응답을 생성하는 구조를 사용하였다. 답변 생성 모델(Generator)로는 Qwen2.5(32B) 모델을 채택하였다.[2] 이는 본 연구에 사용한 NVIDIA H100 80GB 환경에서 단일 GPU 기반 추론이 가능하면서도, 본 연구에서 검토한 로컬 오픈소스 모델 중 문맥 이해 및 응답 생성 성능이 가장 우수하고 한국어 처리 성능이 뛰어난 모델로 판단되었기 때문이다.[2] Retriever 단계에서는 임베딩된 사용자 질의를 기준으로 벡터 데이터베이스에서 관련 문서를 검색하고, 이를 생성 모델의 입력 컨텍스트로 활용한다. 본 연구의 사용자 질의는 스마트팜 관련 일반 문의를 대상으로 하며, 구체적인 예시는 2.5절에서 제시한다. 쿼리 임베딩을 위해 upskyy/bge-m3-korean 모델을 사용하였으며, 이는 한국어 질의에 대해 높은 의미 표현력을 제공한다. 한편, 검색 성능은 벡터 데이터베이스에 저장된 문서의 정확성과 정합성에 크게 의존하므로, 문서의 수집, 정제, 분할 및 메타데이터 구성과 같은 전처리 과정은 RAG 파이프라인 전반의 성능을 좌우하는 핵심적인 선행 조건이라 할

수 있다. 문서 전처리 과정은 OCR 기반 전처리를 통해 벡터 데이터베이스를 구축하는 절차로 구성된다. 먼저 PDF 형식의 원시 문서에 대해 PaddlePaddleOCR(0.9B)을 적용하여 텍스트를 추출하고, 후속 처리를 위해 이를 Markdown 형식으로 변환하였다. 해당 변환은 문서 가독성을 확보하고 LLM 입력 및 이후 처리 과정에서 구조적 일관성을 유지하기 위해 수행되었다.

2.3 RAG 시스템 성능 평가 방법

RAG 기반 질의응답 시스템에서 평가에 사용되는 QA쌍의 품질은 모델 성능 평가의 신뢰성과 직결되며, 통계적으로 유의미한 비교를 위해서는 충분한 규모의 QA쌍이 요구된다. 그러나 전문가 검증에 기반한 수작업 QA 생성 방식은 시간적·비용적 부담이 크다는 한계가 있다. 이에 본 초기 연구에서는 평가용 QA쌍 구축의 효율성을 확보하기 위해 LLM 기반 자동 QA 생성 방안을 채택하였다. 구체적으로, 원시 문서를 LLM 입력으로 활용해 초기 QA seed를 생성한 뒤, 고품질 질문 선별 과정을 거쳐 최종 GT(ground-truth) QA쌍을 구축하였다. QA 생성에는 GPT-4o-mini를 사용하였으며, 이는 평가 대상인 답변 생성 LLM(Qwen2.5 32B)보다 상위 수준의 언어 이해 및 생성 안정성을 확보한 모델을 답변 생성기로 활용함으로써, 평가 데이터의 신뢰성과 객관성을 보장하기 위함이다. 또한 본 연구는 평가의 단순성과 일관성을 확보하기 위해 단답형 질문·응답 쌍이 생성되도록 프롬프트를 설계하여 QA를 구성하였다.

LLM의 성능 평가는 생성된 응답과 정답 간의 표면적 일치도, 의미적 유사성, 그리고 검색 단계의 효율성을 종합적으로 고려하여 수행된다.[1] 본 연구에서는 표면적 및 의미적 유사도 평가를 위해 Exact Match(EM), ROUGE-1, ROUGE-L과 같은 n-gram 기반 지표와 함께 BERTScore 및 SBERT를 사용하였다. BERTScore는 bert-base-multilingual-cased를 기반으로 토큰 수준 의미 유사성을 평가하여, 한국어 질의·응답에서 조사 및 어미 변화로 인한 표면적 불일치를 완화하고 의미 보존 여부를 반영한다. 또한 문장 수준 의미 유사도인 SBERT snunlp/KR-SBERT-V40K-klueNLI-augSTS로 임베딩한 후 cosine similarity로 계산함으로써, 한국어 문장 의미 정합성을 정량화하였다.[4] EM은 생성된 응답이 정답과 정확히 일치하는지를 평가하는 지표이다. 또한 검색 단계의 성능 분석을 위해 retrieval 효율성을 나타내는 recall과 mean reciprocal rank(MRR)을 측정하였다. recall은 정답 문서가 검색 결과에 포함되는 비율을, MRR은 검색 결과 내에서 정답 문서의 평균 순위를 반영하는 지표이다.[1]

2.4 평가 결과

	em	rouge1	rougel	BERT	SBERT	recall	mrr
LLM	0.065	0.196	0.015	0.706	0.352	-	-
RAG	0.770	0.890	0.699	0.946	0.937	0.885	0.646

표. 1. RAG 적용 여부에 따른 LLM 성능 비교 결과

실험은 순수 언어모델 기반 방식(LLM)과 검색 증강 생성 방식(RAG)을 비교하는 방식으로 진행되었다. 표 1은 RAG 적용 여부에 따른 LLM 성능 비교 결과를 나타낸다. 순수 언어모델 기반 방식은 EM이 0.065로 매우 낮게 나타났는데, 이는 동일한 의미를 전달하더라도 표현 차이와 확률적 생성 특성으로 인해 정답과의 완전 일치가 발생하기 어렵기 때문이다.

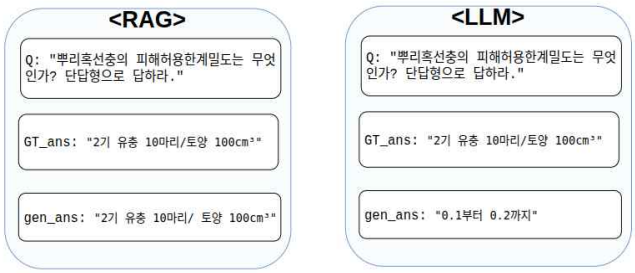

그림. 2. 단답형 질의(Q)에 대한 정답(GT_ans)과 RAG와 순수 LLM 응답(gen_ans) 비교 예시

실제로 그림 2의 예시에서 확인할 수 있듯이, 순수 LLM은 정답과 의미적으로 유사하지 않을 뿐만 아니라 어휘적 유사도 측면에서도 큰 차이를 보이는 응답을 생성하여, EM 점수가 0으로 평가되었다. 반면 RAG 기반 방식은 문서 검색을 통해 제공된 명시적 근거를 활용함으로써 단답형 질문에 대해 문서 내 정답 문구를 직접 참조하는 응답 생성이 가능해져 EM이 0.770으로 크게 향상되었으며, ROUGE-1, ROUGE-L, BERTScore 및 SBERT에서도 전반적으로 높은 성능을 보였다. 특히 SBERT는 0.937로, 의미적 정합성 측면에서 RAG의 우수성을 명확히 보여준다. 또한 recall(0.885)과 MRR(0.646)은 RAG 구조가 관련 문서를 효과적으로 검색하고 이를 응답 생성에 활용하고 있음을 나타내며, 이러한 결과는 단답형·사실 중심 질의 환경에서 문서 기반 근거 제공이 응답의 사실적 정확성과 의미적 정합성을 동시에 향상시킴을 시사한다.

Ⅲ. 결론

본 연구는 스마트팜 도메인에서 문서 기반 RAG 구조를 적용한 자연어 컨설팅 시스템을 제안하고, 실험을 통해 순수 언어모델 대비 문서 근거를 활용한 RAG가 질의 유형 전반에서 응답의 정확성과 신뢰성을 일관되게 향상시킴을 확인하였다. 향후 연구에서는 질의 유형에 따른 retrieval 전략의 차별화와 문서 표현 방식의 고도화를 통해 RAG 시스템의 검색 성능을 개선할 필요가 있다.[1] 또한 전문가 검수를 포함한 고품질 GT QA 구축 기준과 절차를 체계화하기 위한 연구가 요구된다. 이러한 개선을 통해 전문 도메인 환경에서 RAG 기반 질의응답 시스템의 신뢰성과 실용성을 더욱 향상시킬 수 있을 것으로 기대된다.

ACKNOWLEDGMENT

본 연구는 IT 컨버전스 프로그램으로부터 연구에 필요한 컴퓨팅 자원을 제공받았으며, 사용된 원시 데이터는 농업기술원에서 제공되었다.

참 고 문 헌

[1] P. Lewis et al., “Retrieval-Augmented Generation for Large Language Models: A Survey,”

[2] J. Yang et al., “Qwen Technical Report,” arXiv:2309.16609, 2023.

[3] T. Zhang et al., “BERTScore: Evaluating Text Generation with BERT,”

[4] S. Kim, J. Lee, S. Park, and J. Kang, “KR-SBERT: Sentence Embeddings for Korean,” arXiv preprint arXiv:2104.08027, 2021.

[5] A. Zelikman et al., “STaR: Bootstrapping Reasoning With Reasoning,”