의료 객관식 질의응답을 위한 동적 지식 그래프 생성과 선택지별 근거 판단 기반 RAG 프레임워크*

서성원[O1], 김영민[1,2†]
[1]한양대학교 산업데이터엔지니어링학과, [2]한양대학교 기술경영전문대학원
finale22@hanyang.ac.kr, yngmnkim@hanyang.ac.kr

A RAG Framework with Dynamic Knowledge Graph Construction and Option-Wise Evidence Judgment for Medical Multiple-Choice Question Answering

Seongwon Seo[O1], Young-Min Kim[1,2†]
[1]Department of Industrial Data Engineering, Hanyang University
[2]Graduate School of Technology & Innovation Management, Hanyang University

요 약

의료 객관식 질의응답은 전문적인 의학 지식과 복잡한 추론을 요구하며 검색 증강 생성(Retrieval-Augmented Generation, RAG)은 이러한 문제를 해결하기 위한 방법으로 제안되어 왔다. 그러나 검색 문서를 그대로 활용하는 바닐라(vanilla) RAG 방식은 여러 문서에 흩어진 정답 근거를 구조적으로 연결하지 못하고 각 선택지에 대한 근거를 명시적으로 판단하지 못한다. 따라서 정답 예측 성능을 저하시킬 수 있다는 한계가 존재한다. 본 연구는 이러한 한계를 해결하기 위해 의료 객관식 질의응답을 위한 새로운 프레임워크를 제안한다. 제안 방법은 두 단계의 지도식 미세 조정을 통해 학습된다. 1단계에서 모델은 검색 문서에서 핵심 개체와 관계를 추출하여 동적 지식 그래프를 생성하도록 학습된다. 2단계에서는 모델이 검색 문서와 동적 지식 그래프를 기반으로 선택지별 정답 근거를 판단하고 추론 전략을 결정한 뒤 최종 정답을 예측하도록 학습된다. MedQA와 MedMCQA에 대한 실험 결과, 제안 방법은 바닐라 RAG보다 각각 5.03%p와 5.96%p 높은 정확도를 기록하였다. 이 결과는 의료 객관식 질의응답에서 검색 문서의 단순 활용보다 동적 지식 그래프 생성과 선택지별 근거 판단을 결합한 방식이 더 효과적임을 보여준다.

1. 서 론

사전 학습된 언어 모델은 주목할 만한 텍스트 생성 능력을 보여주지만, 학습하지 않은 도메인 지식에 관한 질의에는 부정확한 응답을 한다[1]. 이러한 문제를 해결하고자 검색 증강 생성(Retrieval-Augmented Generation, RAG)이 제안되었다. RAG는 외부 코퍼스(corpus)에서 질의와 관련된 문서를 검색하고 언어 모델이 검색 문서를 활용하여 응답을 출력하는 프레임워크이다[2]. RAG는 언어 모델의 부족한 도메인 지식을 외부 문서로 보완함으로써 응답 성능 향상 가능성을 보였다.

의료 도메인의 객관식 질의응답은 의학 시험, 연구 등을 포함하여[3] 전문적인 의학 지식과 단순하지 않은 추론을 요구한다. 최근 연구는 의료 객관식 질의응답에도 RAG를 적용하여 성능 향상을 시도하고 있다[4]. 그러나 언어 모델이 검색 문서를 입력받아 이를 문맥으로 직접 활용하는 바닐라(vanilla) 방식은 오히려 정답 예측 성능을 떨어뜨리기도 한다[4]. 따라서 의료 객관식 질의응답에서 여러 문서에 흩어져 있는 정답 근거를 연결하고 선택지별로 정답 근거를 명시적으로 판단하는 과정이 필요하다.

본 연구는 RAG에서 언어 모델이 검색 문서를 지식 그래프로 구조화하고 선택지별 정답 근거를 판단하여 정답을 예측하는 새로운 프레임워크를 제안한다. 프레임워크는 두 개의 학습 단계로 구성되어 있으며 각 단계에서 지도식 미세 조정(supervised fine-tuning)을 활용한다. 1단계에서 언어 모델은 입력받은 검색 문서에서 정답 판별을 위한 핵심 개체(entity)와 관계(relation)를 추출함으로써 검색 문서을 지식 그래프로 구조화하는 능력을 학습한다. 2단계에서는 언어 모델이 입력받은 검색 문서와 생성한 동적 지식 그래프에서 선택지별 정답 근거를 찾고 평결을 선택한다. 다음으로 평결에 따라 정답 추론을 위한 지식 활용 방식을 결정한다. 학습된 모델은 동적 지식 그래프 생성부터 선택지별 근거 판단, 지식 활용 전략 결정, 정답 예측까지 일괄적으로 수행한다.

본 연구가 제안하는 프레임워크의 성능을 의료 객관식 질의응답 벤치마크(benchmark)인 MedQA와 MedMCQA에 대해 평가하였다. 또한, 백본(backbone)의 내부 지식 기반 추론 성능 및 바닐라 RAG를 활용한 추론 성능과 비교하였다. 본 연구의 기여는 제안 프레임워크가 검색 문서로 인해 성능이 떨어지는 문제를 해결할 수 있음을 보였다는 점이다.

* 이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (RS-2026-25492127)

† 교신 저자(Corresponding Author)

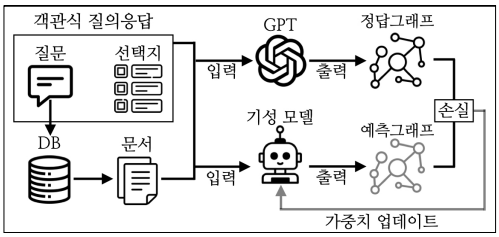

그림 1. 1단계 학습 방법

2. 동적 지식 그래프 생성

동적 지식 그래프는 개체(entity)와 관계(relation) 두 가지 구성 요소를 가지며 각 요소는 다음과 같은 튜플(tuple)로 표현된다.

- 개체 튜플: ("개체", 개체명, 개체 유형, 설명, 근거)
- 관계 튜플: ("관계", 출발, 관계 유형, 도착, 설명, 근거)

동적 지식 그래프 생성은 다음 다섯 가지 원칙을 따른다: 1) 개체 및 관계 추출 시 선택지를 구별하는 데 기여하는 정보만을 선별적으로 추출한다. 2) 개체는 검색 문서에 근거해야 하며 관계는 오직 문서에 의해 명시적으로 지지될 때만 추출된다. 이는 모델이 내부 지식에 의존하여 요소를 추출하는 것을 방지하기 위함이다. 3) 모든 개체와 관계는 출처 문서 인덱스(index)를 근거에 명시하도록 한다. 이는 후속 단계에서 선택지별 근거를 찾을 때 추적이 가능하도록 만들어 근거 기반 판단을 강제할 수 있다. 4) 개체 유형 및 관계 유형은 UMLS(Unified Medical Language System)의 의미 유형(semantic type) 체계를 참고하여 각각 15개 범주로 제한한다. 이러한 제약은 중복 추출을 줄일 수 있다[5]. 5) 검색 문서가 질문과 무관하거나 정답 판정에 도움이 되지 않는 정보만을 담고 있는 경우 모델은 빈 지식 그래프를 출력할 수 있다. 이는 무관한 문서로부터 억지로 정보를 추출하여 노이즈(noise)를 증폭시키는 현상을 방지하기 위함이다.

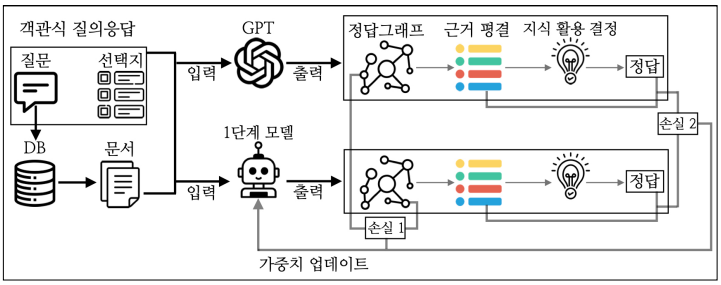

그림 2. 2단계 학습 방법

3. 선택지별 근거 판단

기존 RAG에서 언어 모델은 검색 문서를 그대로 입력받아 이를 문맥(context)으로 활용하여 정답을 생성한다[1]. 그러나 이러한 방식은 의료 객관식 질의응답에서 언어 모델이 정답과 무관한 정보에 의존하게 할 수 있다. 본 연구는 모델 스스로 검색 정보의 활용 방식을 판단하는 것을 목표로 한다. 이를 위해 각 선택지에 대한 정답 근거 판단과 그 결과에 따른 지식 활용 전략 결정을 명시적으로 출력하도록 학습시킨다.

모델은 검색 정보와 동적 지식 그래프에서 각 선택지에 대한 정답 근거를 찾는다. 이때 문서 인덱스와 개체명, 관계 인덱스를 참조한다. 그리고 이를 기반으로 정답 근거가 선택지를 지지하는지, 반박하는지, 불충분한지 설명한다. 다음으로 근거를 기반으로 평결(verdict)을 선택한다. 평결은 다음의 세가지 범주로 구성된다.

- Supported(지지): 문서 또는 그래프가 해당 선택지를 정답으로 지지하는 직접적 증거를 제공함.
- Contradicted(반박): 문서 또는 그래프가 해당 선택지가 오답임을 명시적으로 보임.
- Insufficient(불충분): 문서 또는 그래프만으로 해당 선택지의 정답을 판단할 근거가 부족함.

선택지별 평결이 결정되면 모델은 평결의 조합에 따라 세 가지 지식 활용 전략 중 하나를 결정한다. 이 결정은 검색 정보 및 그래프를 어떤 방식으로 활용할지를 명시적으로 분기한다. 지식 활용 전략은 다음의 세 가지 범주로 구성된다.

- Grounded(근거 기반): Supported 평결이 1개 존재하는 경우이다. 모델은 외부 지식을 기반으로 해당 선택지를 정답으로 선택한다.
- Elimination(제거): Supported 평결은 없으나 Contradicted 평결이 1개 이상 존재하는 경우이다. 모델은 Contradicted 선택지를 오답으로 소거한 뒤, 내부 지식을 활용하여 잔여 선택지 중에서 정답을 추론한다.
- Parametric(내부 지식): 모든 선택지가 Insufficient 평결인 경우이다. 외부 지식이 정답 추론에 충분하지 않음을 의미하며 모델은 내부 지식을 기반으로 정답을 추론한다.

이러한 지식 활용 전략 결정 방식은 검색 문서의 품질이 보장되지 않는 환경에서도 모델의 강건한 추론을 도울 수 있다.

4. 실험 및 분석

4.1 비교 모델 설정

본 연구는 Mistral-7B-Instruct-v0.3 모델을 백본으로 사용한다. 동일한 백본에 대해 질문과 선택지만을 입력하는 내부 지식 기반 추론과 검색 문서를 그대로 입력하는 바닐라 RAG 기반 추론, 그리고 제안 프레임워크 기반 추론의 성능을 비교하였다.

4.2 벤치마크 설정

본 연구의 실험에는 의료 객관식 질의응답 벤치마크인 MedQA와 MedMCQA를 사용하였다. 각 테스트셋은 MedQA 1,273 문항, MedMCQA 4,183 문항으로 구성된다. 평가 지표는 정확도(accuracy)로 계산된다.

4.3 모델 입력

RAG에서 공정한 성능 비교를 위해 각 질문에 대한 검색 문서를 사전에 구축하였다. 따라서 바닐라 RAG와 제안 프레임워크에 입력되는 검색 문서는 동일하다.

검색기 설정은 다음과 같으며 MedRAG[4] 툴킷(toolkit)을 활용하여 문서 검색을 수행하였다.

- 검색기(retriever): Contriever
- 코퍼스: PubMed + Textbooks
- 청킹(chunking): RecursiveCharacterTextSplitter, max 1000 chars
- 인덱스: FAISS IndexFlatIP

4.4 데이터 생성

본 연구에서는 강력한 모델의 동적 지식 그래프 생성 능력 및 정답 추론 능력을 가벼운 기성(off-the-shelf) 모델로 증류하고자 한다. 따라서 1, 2단계 학습 데이터는 모두 GPT-5.1을 활용하여 생성하였다. 1단계 학습 데이터는 질문, 선택지, 검색 문서를 입력하여 지식 그래프를 출력하여 구성하였다. 2단계 학습 데이터는 질문, 선택지, 검색 문서를 입력하여 지식 그래프, 선택지별 평결, 지식 활용 전략 결정, 정답을 순차적으로 출력하여 구성하였다.

2단계 학습 데이터는 여러 단계의 검증을 거친다. 먼저 그래프와 선택지 기반 추론을 구분하는 태그(tag)의 완결성, 필수 필드(field) 존재 여부, 평결과 결정의 정합성 등 구조적 제약을 검사한다. 이 검증을 통과한 데이터를 학습 데이터로 우선 채택한다. 검증을 통과하지 못한 샘플에 대해서는 실패 유형에 따라 두 가지 방식 중 하나로 데이터를 재생성한다: 1) 정답이 아닌 선택지가 Supported로 잘못 판정된 경우, 해당 선택지를 명시적으로 배제하도록 지시한 후 데이터를 재생성한다. 2) 정답이 Contradicted로 판정되었거나 평결 충돌이 발생한 경우, 정답을 직접 명시한 상태로 데이터를 재생성한다.

두 방식 모두 정답에 관한 정보가 명시적으로 입력되므로 재생성된 데이터는 정답 정보를 추론 과정에 노출시킬 위험이 있다. 이러한 힌트 누수(hint leakage)를 방지하기 위해 29개의 정규식 패턴을 활용하여 정답을 명시하는 표현을 탐지하고 제거한다. 재생성 검증을 통과한 데이터를 학습 데이터로 추가 채택한다.

본 연구는 MedQA와 MedMCQA의 학습 데이터셋에서 각각 약 2,750개씩 총 5,500개의 샘플을 임의로 추출하여 데이터 생성을 위한 기반으로 사용하였다. 1단계에서는 5,500개(학습 5,000 / 검증 500) 데이터를, 2단계에서는 검증을 통과한 3,649개(학습 3,305 / 검증 344) 데이터를 최종 학습 데이터로 채택하였다.

4.5 모델 학습

1단계는 다음의 자기회귀 크로스 엔트로피(cross-entropy) 손실로 학습된다. 여기서 x는 질문과 선택지, 그리고 검색 문서를 의미하며 y는 응답 시퀀스(sequence)를 의미한다. T_C는 어시스턴트(assistant) 응답 구간의 토큰 인덱스 집합이다. 모델이 생성해야 하는 답변 부분만 손실을 계산하기 위해 학습 시 프롬프트는 손실 계산을 하지 않는다.

$$L_{S1}(\theta) = -\frac{1}{|T_C|}\sum_{t \in T_C} \log p_\theta(y_t | y_{<t}, x) \quad (1)$$

2단계에서는 응답 영역을 동적 지식 그래프 구간 T_g와 정답 추론 구간 T_a로 나누고 두 구간에 대해 가중 평균 크로스 엔트로피를 사용한다. 샘플마다 토큰 길이가 달라 손실 합의 크기가 달라지는 문제를 방지하기 위해 전체 응답 길이 $\lambda_g|T_g| + |T_a|$로 정규화하였다. 이때 λ_g는 그래프 구간의 학습 가중치를 조절하는 하이퍼파라미터(hyperparameter)이며 본 연구에서는 0.3으로 설정하였다.

$$L_{S2}(\theta) = -\frac{\lambda_g \sum_{t \in T_g} \log p_\theta(y_t | y_{<t}, x) + \sum_{t \in T_a} \log p_\theta(y_t | y_{<t}, x)}{\lambda_g |T_g| + |T_a|} \quad (2)$$

학습은 QLoRA(Quantized Low-Rank Adaptation)로 수행되었으며 epoch는 3으로 설정 후 best checkpoint를 최종 모델로 선택하였다.

4.6. 실험 결과 및 분석

표 1 실험 결과

추론 방식	MedQA	MedMCQA	Avg
내부 지식	0.5255	0.4626	0.4941
바닐라	0.4941	0.4114	0.4523
제안 프레임워크	**0.5444**	**0.4710**	**0.5077**

실험 결과, 제안 프레임워크는 바닐라 RAG 기반 추론보다 MedQA와 MedMCQA에서 각각 5.03%p, 5.96%p 높은 정확도를 보였다. 이는 본 연구가 지적했던, 검색 문서를 그대로 활용하는 경우 성능이 더 떨어지는 문제를 제안 방법으로 해결할 수 있음을 보여준다. 또한, 내부 지식 기반 추론에 비하여 1.89%p, 0.84%p 높은 정확도를 보이므로 제안 방법을 통해 RAG가 외부 지식을 효과적으로 활용할 수 있음을 보여준다.

5. 결론 및 향후 연구

본 연구는 의료 객관식 질의응답에 대해 RAG의 성능을 향상하기 위한 동적 지식 그래프 생성 및 선택지 기반 추론 방법론을 제안한다. 실험 결과를 통해 기존 바닐라 RAG에서 검색 문서가 입력되면 오히려 성능이 저하되는 문제를 제안 프레임워크로 해결할 수 있음을 확인하였다.

본 연구는 질의응답 도메인을 의료로 한정하여 연구 설계 및 실험을 수행하였다. 향후 연구는 법률과 같은 다른 전문 도메인의 객관식 질의응답으로 확장하고 다양한 검색기 및 백본 실험을 통해 제안 프레임워크의 일반화 가능성을 탐색할 필요가 있다.

참 고 문 헌

[1] Wei, Zhepei, Wei-Lin Chen, and Yu Meng. "InstructRAG: Instructing retrieval-augmented generation via self-synthesized rationales." arXiv preprint arXiv:2406.13629, 2024

[2] Lewis, Patrick, et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks." Advances in neural information processing systems, 33, 9459-9474, 2020

[3] Singhal, Karan, et al. "Toward expert-level medical question answering with large language models." Nature medicine 31.3: 943-950, 2025

[4] Xiong, Guangzhi, et al. "Benchmarking retrieval-augmented generation for medicine." Findings of the Association for Computational Linguistics: ACL 2024, 2024