Multi-hop 질의응답을 위한 Five WS 기반 RAG Retrieval 기법

정일균[1]
[1]고려대학교 SW·AI 융합대학원

jik9210@korea.ac.kr

Five Ws-Enhanced Retrieval for Multi-hop Question Answering in RAG

Ilgyun Jeong[1]
[1]Korea University Graduate School of SW·AI Convergence

요 약

본 연구는 RAG(Retrieval-Augmented Generation)의 Multi-hop 질의응답 성능 향상을 위해 5W1H(Who, What, When, Where, Why, How) 태그 기반 검색 방식을 제안한다. 기존 임베딩 기반 검색은 복합적 문맥 연결성을 단일 벡터로 표현하는 데 한계가 있으며, 지식그래프 기반 방법은 연산 비용이 높아 실용성이 떨어진다. 이에 따라 GPT 를 활용해 질의와 문서 조각에서 5W1H 태그를 자동 생성하고, 문장 유사도와 태그 유사도를 결합한 검색 점수를 산출하였다.
HotpotQA 데이터셋을 활용한 실험 결과, 제안 방식은 Baseline 대비 F1-score 가 14.3% 향상되었고, Precision@2 와 Recall@2 또한 각각 10%와 14.3% 개선되었다. 특히 태그 가중치가 0.2~0.3 일 때 최적 성능을 보였으며, 초기 검색 정확도(P@1, MRR)는 유지되었다. 이는 5W1H 태그가 Multi-hop 추론에서 효과적인 보조 신호로 작용함을 보여준다.

Ⅰ. 서 론

대규모 언어모델(LLM)과 결합된 RAG(Retrieval-Augmented Generation) 기법은 외부 지식을 활용하여 환각을 줄이고 답변의 신뢰도를 높이는 방식으로 널리 사용되고 있다.[1] 그러나 실제 질의응답 환경에서는 단일 사실만을 요구하는 Single-hop 질의보다, 여러 문서와 단서를 종합해야하는 Multi-hop 질의응답이 빈번하게 등장한다. 기존의 임베딩 기반 검색 방식은 이러한 Multi-hop 상황에서 질의의 복합적인 문맥 연결성을 단일 벡터로 표현하는데 한계가 있어 성능 저하가 발생한다.
이를 해결하기 위해 엔티티 관계 추론이나 지식그래프 기반 검색이 제안되었으나, 이러한 접근은 높은 연산 비용과 복잡한 사전 처리 과정을 요구하는 실용적 한계를 지닌다. [2]
따라서 보다 단순하면서도 Multi-hop 질의에 효과적으로 대응할 수 있는 새로운 검색 방식을 제안한다.
본 연구에서는 이에 대한 대안으로 Five WS(Who, What, When, Where, Why, How) 기반 태그 정보를 활용한 Retrieval 방법을 제안한다. Five Ws 는 사실을 구조적으로 표현하는 기본적인 프레임워크로, 질의와 문서 조각을 의문사 단위에서 매칭할 수 있도록 한다. 예를 들어 “루즈벨트는 언제 사망했고 언제 대통령이 되었는가?”라는 Multi-hop 질의는 단순 임베딩 비교만으로는 정보 간의 연결성을 파악하기 어렵다. 하지만 [Who: 루즈벨트], [When: 사망 시점], [When: 대통령 취임 시점]과 같이 태그를 부여하면, 검색 시스템이 필요한 정보 조각 간의 관계를 더 명확히 인식할 수 있다.
따라서 본 연구의 목표는 Multi-hop 환경에서 RAG 의 Retrieval 성능을 향상시키기 위해 Five Ws 기반 태그 정보를 결합한 검색 방식을 설계하고, 실험을 통해 그 효과를 정량적으로 검증하는 데 있다.

Ⅱ. 본론

2.1 데이터셋

본 연구에서는 Multi-hop 질의응답 태스크를 평가하기 위해 HotpotQA 데이터셋을 사용하였다. HotpotQA 는 단일 문서에 근거하는 질문뿐만 아니라, 두 개 이상의 문서를 종합해야 정답을 도출할 수 있는 Multi-hop 질문을 포함하고 있어 본 연구의 목적에 적합하다.

2.2 태그 생성

검색 성능 향상을 위해 각 질의와 문서 조각에 대해 Five Ws(Who, What, When, Where, Why, How) 태그를 GPT-4.1 nano 를 통해 자동 생성하였다. 태그 생성 과정은 다음과 같다.

프롬프트 기반 추출
LLM 에게 구조화된 프롬프트를 제공하여, 입력 텍스트에서 5W1H 요소를 JSON 형식으로 추출하도록 하였다.

- who : 인물, 조직, 주체
- what : 사건, 사물, 행위, 상태
- when : 시간, 날짜, 기간
- where : 장소, 위치, 국가
- why : 이유, 목적, 원인
- how : 방법, 과정, 수단

2.3 방법론

태그 기반 검색 기법
본 연구에서는 각 질문과 문서 컨텍스트에서 Five Ws(Who, What, When, Where, Why, How) 태그를 추출하여 검색에 활용하였다. 검색 점수를 산출하기 위해 두 가지 임베딩 모델을 사용하여 비교하였다:
1) all-mpnet-base-v2 (Sentence-Transformers, 768 차원)
2) text-embedding-3-small (OpenAI, 1536 차원)
이들 모델은 질문과 문서 컨텍스트, 그리고 태그를 임베딩 벡터로 변환하는 데 사용되었으며, 변환된 벡터를 기반으로 코사인 유사도를 계산하였다.
검색 기법은 다음 두 가지 방식으로 구분된다.
1) Baseline: 질문과 문서 컨텍스트 전체를 임베딩하여 문장

유사도(SentenceSim)만으로 검색을 수행한다.
2) Proposed: Baseline 의 문장 유사도에 더해, 질문 태그와 문서 태그 간의 태그 유사도(TagSim)를 결합하여 최종 검색 점수를 산출한다.
계산식은 다음과 같다.

$$Score = (1 - \lambda) \cdot SentenceSim + \lambda \cdot TagSim$$

여기서 λ는 태그 유사도의 가중치이며, 실험에서는 0.0, 0.2, 0.3, 0.4, 0.5 값을 적용하여 비교하였다.

Ⅲ. 실험결과

본 절에서는 제안한 5W1H 태그 기반 접근법의 효과를 검증하기 위해 수행한 실험 결과를 제시한다.
성능 평가는 아래의 4 개 지표를 사용하였다.
1) Precision@k (P@k): 상위 k 개 문서 중 정답 문서의 비율
2) Recall@k (R@k): 정답 문서가 상위 k 개에 포함되는 비율
3) F1 Score: Precision 과 Recall 의 조화 평균
4) Mean Reciprocal Rank (MRR): 정답 문서의 평균 역순위

3.1 Weight 변화에 따른 성능 분석

표 1 은 Weight 값을 0.0 에서 0.5 까지 변화시켰을 때의 성능 변화를 보여준다.

- F1-score 는 Weight=0.2 및 0.3 에서 0.800 으로 최대치를 기록하였으며, Baseline(Weight=0.0)의 0.700 대비 14.3% 향상되었다.
- Precision@2 와 Recall@2 또한 동일한 Weight 구간에서 0.800 을 기록하며 Baseline(0.700) 대비 각각 10% 및 14.3% 개선되었다.
- 반면, Weight 가 과도하게 증가한 경우(0.4 이상)에는 성능이 다시 하락하는 역 U 자 형태의 곡선이 관찰되었다. 이는 태그 가중치가 부족하거나 과도할 경우 모두 성능 저하로 이어짐을 시사한다.
- Precision@1 과 MRR 은 Weight 값에 관계없이 각각 0.900, 0.950 으로 안정적으로 유지되었다. 이는 제안 기법이 초기 검색 결과의 품질에는 영향을 주지 않으면서, 다중 hop 검색 단계에서 성능을 집중적으로 개선함을 의미한다.

표 1. Weight 변화에 따른 성능 비교

Weight	F1	MRR	P@1	P@2	R@2
0.0	0.7	0.95	0.9	0.7	0.7
0.2	0.8	0.95	0.9	0.8	0.8
0.3	0.8	0.95	0.9	0.8	0.8
0.4	0.7	0.95	0.9	0.7	0.7
0.5	0.65	0.95	0.9	0.65	0.65

3.2 Baseline 대비 5W1H 태그 적용 효과

표 2 은 Baseline(Weight=0.0)과 5W1H 태그 적용 결과(Weight=0.2~0.3 구간)를 비교한 결과를 요약한 것이다.
- Precision@1 은 두 방법 모두 0.900 으로 동일하여 초기 검색 정확도는 유지되었다.
- Precision@2 는 0.700 에서 0.800 으로 10% 향상되었으며, Recall@2 와 F1-score 는 모두 0.700 에서 0.800 으로 증가하여 14.3%의 개선 효과를 보였다.

표 2. Baseline 대비 5W1H 태그 적용 성능 개선

Metric	Baseline	5W1H Tags	Improvement
P recision@1	0.900	0.900	0%
Precision@2	0.700	0.800	+10%
Recall@2	0.700	0.800	14.3%
F1-Score	0.700	0.800	14.3%

3.3 결과 해석

실험 결과, 5W1H 태그는 초기 검색 품질을 저해하지 않으면서 multi-hop 추론 성능을 개선하는 효과를 나타냈다. 특히 Weight=0.2~0.3 구간에서 최적의 성능 향상이 관찰되었으며, 이는 적절한 수준의 태그 정보가 문서 검색의 맥락적 일관성을 강화하는 데 기여함을 보여준다. 반면, 과도한 가중치는 불필요한 잡음을 유발하여 오히려 성능을 저하시켰다. 이러한 결과는 5W1H 태그가 단일 문서 검색보다 다중 hop 기반의 추론 문제에 특화된 보조 신호로 작동함을 시사한다.

Ⅳ. 결론

본 연구는 Multi-hop 질의응답에서 RAG 의 한계를 극복하기 위해 5W1H 기반 태그 정보를 활용한 검색 방식을 제안하였다. HotpotQA 실험 결과, 제안 기법은 Baseline 대비 F1-score 와 Recall@2 에서 최대 14.3% 향상을 보였으며, 초기 검색 정확도는 유지되었다. 이는 5W1H 태그가 Multi-hop 상황에서 문맥적 일관성을 강화하는 효과적 신호임을 입증한다. 다만, 태그 품질이 LLM 성능에 의존하고, 데이터셋이 제한적이라는 한계가 존재한다. 향후 연구에서는 다양한 도메인 검증과 동적 가중치 조정 방식을 탐구하여 제안 기법의 범용성과 실용성을 높일 예정이다.

참 고 문 헌

[1] P. Lewis, E. Perez, A. Piktus, et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” Advances in Neural Information Processing Systems (NeurIPS), 2020.

[2] X. Zhu, Y. Xie, Y. Li, and W. Hu, “Knowledge Graph-Guided Retrieval Augmented Generation,” arXiv preprint arXiv:2502.06864, 2025.