하위 주제 기반 질의 확장을 통한 RAG 응답 품질 향상

김현민, 유준*
*가천대학교
khm108@gachon.ac.kr, *joon.yoo@gachon.ac.kr

Improving RAG Answer Quality via Subtopic-Based Query Expansion

Kim Hyun-Min, Yoo Joon*
*Gachon University.

요 약

본 논문은 사용자의 추상적 질의로부터 의미 기반 하위 주제를 생성하고 이를 질의 확장에 활용함으로써 RAG 기반 질의응답 시스템의 검색 정확도와 응답 다양성을 향상시키는 방법을 제안한다. 확장된 질의는 정보 검색 단계에서 더 풍부한 문서 확보를 가능하게 하고, 생성 응답의 내용적 폭을 넓히는 데 기여함을 정량적·정성적 평가를 통해 확인하였다. 이와 같은 방법론은 질의의 표현력을 보완하고 응답 품질을 개선하는 데 효과적이며, 향후 하위 주제 반영 방식과 검색 모듈 조정을 통해 전체 응답 품질을 더욱 향상시킬 수 있을 것으로 기대한다.

Ⅰ. 서 론

최근 대규모 언어모델(Large Language Models, LLMs)의 발전은 질의응답(QA) 시스템의 응답 품질을 획기적으로 향상시켰으며, 다양한 산업 분야에서 실제 서비스에 적용되고 있다. 특히 Retrieval-Augmented Generation(RAG) 방식은 사전 학습된 언어모델에 외부 지식을 검색해 주입함으로써, 단순한 생성 기반 QA보다 더 우수한 성능을 발휘한다. 그러나 이러한 RAG 시스템은 검색 단계에서 질의의 표현력이 부족하거나 의미가 모호할 경우, 적절한 문서를 검색하지 못해 응답의 질이 저하되는 문제가 존재한다.

현실에서 사용자가 입력하는 질의는 종종 문맥이나 핵심 개념이 생략된 형태로 입력되기 쉽다. 이를테면 "우리 회사 복지에 대해 알려줘"라는 질문은 흔하면서도 추상적인 표현이며, 그 안에 포함될 수 있는 구체적인 항목들, 예를 들어 학자금 지원, 건강검진, 사내 복지시설, 유급 휴가와 같은 항목은 명시적으로 나타나지 않는다. 이처럼 하위 주제(subtopics)가 누락된 질의는 정보 검색 성능에 악영향을 미칠 수 있다. 따라서 본 연구는 사용자 질문에 명시되지 않은 의미적 하위 주제를 생성하고, 이를 기반으로 검색 영역을 확장하여 최종 응답의 다양성과 정확도를 향상하는 방법론을 제안한다.

1-1. 관련 연구

기존의 질의 확장(query expansion) 기술은 대부분 단어 수준의 유사도나 답변에 필요한 지식을 검색하기 위해 질의를 분할하는 방법을 기반으로 한다. 최근에는 LLM의 생성 능력을 활용해 의미론적 수준에서 질의를 확장하는 시도가 이루어지고 있으나, 이 역시 대개는 단순한 키워드 생성이나 답변 예측 수준에 머무른다.

예를 들어 Jagerman et al. (2023)은 다양한 프롬프트 전략을 통해 LLM이 확장 질의를 직접 생성하도록 하고 검색 성능을 비교하였으며[1], Lin et al. (2024)은 RQ-RAG를 통해 질의 재작성, 분해 등의 방법으로 질의를 재구성함으로써 모델을 개선하려 하였다. [2] 그러나 이들 접근법은 주로 정보 검색 정확도 향상에 초점을 두고 있어, 사용자의 압축된 질의로부터 누락된 의미적 하위 개념을 보완하는 데에는 제한적이다.

Ⅱ. 연구 목표 및 방법론

본 연구의 목표는 Retrieval-Augmented Generation(RAG) 기반 질의응답 시스템에서 사용자의 추상적이고 압축적인 질의로부터 의미론적으로 관련된 하위 주제를 자동으로 생성하고, 이를 검색 과정에 반영함으로써 검색 적합도 및 응답 생성 품질을 동시에 향상하는 것이다. 전체적인 파이프라인은 그림 1과 같다.

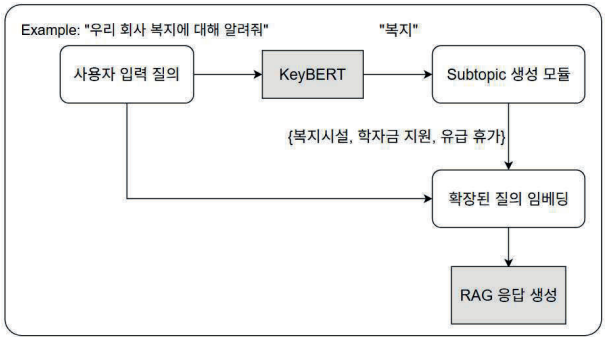

그림 1: 제안된 하위 주제 확장 기반 RAG 질의응답 구조

2-1. 하위 주제 추출

먼저 사용자 질의로부터 의미론적으로 관련된 하위 주제 키워드를 생성할 수 있도록 주제 단어를 추출한다. 주제 단어 추출에는 KeyBERT를 이용한다. KeyBERT란 주어진 문장에서 의미론적으로 전체 문서를 가장 잘 나타내는 키워드를 찾아주는 BERT 기반 키워드 추출 기법이다. [3] 사용자 질의에서 1-gram, 2-gram 단위의 짧은 주제 단어를 추출하여 질문에서 핵심이 되는 대표 단어를 선택한다. 이후 Llama 기반의 사전학습된 모델에 3-shot learning을 통해 주제 단어와 관련된 하위 주제 단어를 생성하도록 유도하여 최종적으로 검색 적합도를 높일 수 있는 의미 기반 하위 주제를 생성한다. [4]

2-2. 확장된 질의 임베딩 구성

생성된 하위 주제를 텍스트 기반 또는 임베딩 기반으로 기존 질의 문장

에 통합한다. 텍스트 기반 문장 결합은 단순한 문장 결합으로 생성된 하위 주제 단어들을 원본 질의 문장에 덧붙이는 방식을 통해 질의를 확장하며, 임베딩 기반 확장에서는 인코딩 된 질의에 하위 주제 단어 임베딩을 가중치에 따라 결합하는 Weighted Pooling 방식으로 수행하였다. 임베딩 기반 질의 확장을 수식으로 표현하면 다음과 같다.

$$e_{final} = \alpha \cdot e_q + (1-\alpha) \cdot e_s$$

Ⅲ. 실험 및 평가

3-1. 실험 설정

본 연구는 NarrativeQA 데이터셋의 테스트 split 중 상위 1,000개의 질문을 활용하여 실험을 수행하였다. [5] 각 질의에 대해 대응하는 답변과 문서를 함께 사용하였으며, 제안한 하위 주제 기반 질의 확장 기법의 성능을 기존 RAG 구조와 비교하였다. 기본 RAG 시스템은 Hugging Face 기반의 사전학습 언어모델과 FAISS 벡터 검색기를 결합하여 구현하였으며, 제안한 방식은 의미 기반 하위 주제 생성 및 질의 임베딩 확장 모듈을 추가하여 추론을 수행하였다. [6] 본 논문에서는 텍스트 기반 확장을 중심으로 평가하였다. 임베딩 기반 확장은 적용 방식에 따라 성능 편차가 크고 실험 설계가 복잡하여 이번 연구의 평가에서는 제외하였다. 향후 Weighted Pooling의 계수를 조정하는 방식으로 질의 표현을 정교하게 최적화하는 방안을 고려할 수 있다.

3-2. 평가

질의응답의 성능 평가는 생성된 응답과 정답 간의 의미 유사도를 측정하는 BERTScore(Precision, Recall, F1)를 기준으로 하였으며[7], 추가적으로 GPT-4 기반 모델을 활용하여 동일 질문에 대해 검색된 컨텍스트 간의 정보 다양성을 비교하는 Diversity 평가를 함께 수행하였다. [8]

BERT Score	Baseline	Proposal
Precision	0.8885	0.8646
Recall	0.9010	0.8831
F1	0.8942	0.8732

표 1 : BERTScore 기반 응답 품질 비교

BERTScore 평가 결과에서는 기존 RAG가 하위 주제를 이용한 질의 확장 방식보다 전반적으로 다소 높은 정량 평가 점수를 기록하였다. 이는 기존 질의가 정답과의 직접적인 표현 일치를 더 많이 포함하고 있어, 의미 유사도 기반 측정에서 더 유리하게 작용했기 때문이다. 반면 GPT-4를 활용한 Diversity 평가에서는 하위 주제 기반 질의 확장 방식이 856건으로 지식 베이스 검색 과정에서 기존 베이스라인보다 훨씬 다양한 정보를 포함한 문서를 찾은 것으로 나타났다.

Ⅳ. 결론

본 연구는 사용자의 추상적인 질의로부터 의미 기반 하위 주제를 생성하고 이를 검색 확장에 활용함으로써, RAG 기반 질의응답 시스템의 응답 품질과 정보 다양성을 향상시키는 방법을 제안하였다. 정량 평가에서는 기존 방식보다 소폭 낮은 성능을 보였지만, 정성 평가에서는 제안된 방식이 주어진 질의에서 더 다양한 정보를 포함한 문서를 찾아내는 것을 확인하였다. 이는 하위 주제 확장이 질의의 표현력을 보완하고, 보다 풍부한 응답을 유도할 수 있음을 시사한다. 향후 미세 조정된 T5 기반 모델을 활용하여 보다 안정적으로 하위 주제를 생성하거나, 임베딩 단계에서 질의를 확장하는 등 다양한 실험을 통해 질의응답 시스템의 표현력과 응답 품질 향상에 기여할 수 있을 것이다.

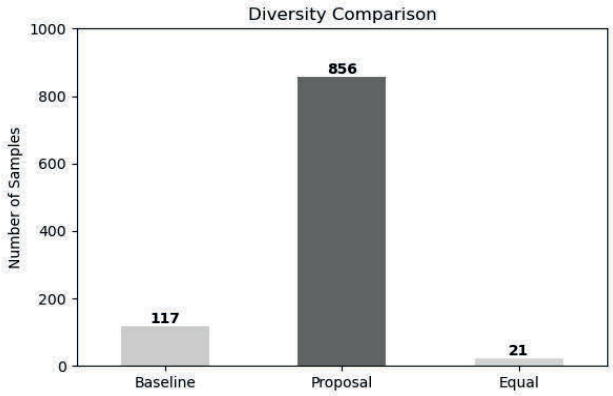

그림 2: 다양성 (Diversity) 비교 평가 결과

ACKNOWLEDGMENT

본 연구는 2021년도 과학기술정보통신부 이공분야기초사업의 지원(NRF-2021R1F1A1063640) 및 2020년도 과학 기술정보통신부 및 정보통신기획평가원의 SW중심대학사업 지원을 받아 수행하였음.

참 고 문 헌

[1] Jagerman, R., Zhuang, H., Qin, Z., Wang, X., & Bendersky, M. (2023). Query expansion by prompting large language models. arXiv preprint arXiv:2305.03653.

[2] Chan, C. M., Xu, C., Yuan, R., Luo, H., Xue, W., Guo, Y., & Fu, J. (2024). Rq-rag: Learning to refine queries for retrieval augmented generation. arXiv preprint arXiv:2404.00610.

[3] Grootendorst, M. (2020). KeyBERT: Minimal keyword extraction with BERT (Version 0.3.0). Zenodo. https://doi.org/10.5281/zenodo.4461265

[4] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

[5] Kočiský, T., Schwarz, J., Blunsom, P., Dyer, C., Hermann, K. M., Melis, G., & Grefenstette, E. (2018). The narrativeqa reading comprehension challenge. Transactions of the Association for Computational Linguistics, 6, 317-328.

[6] Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P. E., ... & Jégou, H. (2024). The faiss library. arXiv preprint arXiv:2401.08281.

[7] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675.

[8] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. arXiv preprint arXiv:2303.08774.