로컬 LLM 기반 분야 인식 질의 확장과 의미 중복 억제를 통한 효율적 전문 검색 프레임워크

한인종, 이종률*
충남대학교

haninjong@gmail.com, *jongryul.lee@cnu.ac.kr

A Field-Aware Query Expansion Framework with Semantic Redundancy Control for Local LLM-based Information Retrieval

Injong Han, Jong-Ryul Lee*
Chungnam Univ.

요 약

대형 언어 모델(LLM)을 활용한 질의 확장(Query Expansion)은 일반 도메인 문서 검색에서 효과를 보이고 있으나 전문 용어와 약어가 빈번히 사용되는 학술·산업 분야에서는 오확장과 의미 중복으로 인해 효과가 제한적이다. 특히 로컬 환경에서 경량 LLM을 사용하는 검색 증강 생성(Retrieval-Augmented Generation) 시스템에서는 이러한 문제가 검색 정합성과 효율성 저하로 더욱 두드러진다. 본 논문은 로컬 LLM 기반 환경에서 전문 도메인 검색의 효율성과 정확도를 동시에 향상시키기 위한 분야 인식 질의 확장(Field-Aware Query Expansion) 및 의미 중복 제거 (Semantic Redundancy Filtering) 프레임워크를 제안한다. 제안 방법은 질의를 분야 단위로 해석하여 해당 분야에 특화된 동의어와 약어만을 선택적으로 확장하고, 확장 결과의 의미적 중복을 체계적으로 억제함으로써 질의 표현 단계에서의 효율성과 정합성을 동시에 개선하는 로컬 RAG 환경에 적합한 구조적 해법을 제공한다.

Ⅰ. 서 론

최근 대형 언어 모델(Large Language Model)의 발전으로 질의 확장을 활용한 검색 성능 향상 연구가 활발히 이루어지고 있다. 기존 연구에 따르면 CoT, CoT/PRF 와 같은 LLM 기반 프롬프트 기법은 MS-MARCO 와 BEIR 등 일반 도메인 데이터셋에서 기존 PRF 방식보다 우수한 성능을 보이며, 모델 규모가 클수록 효과가 증가하는 경향을 보인다.[1] 또한 LLM을 활용해 질의로부터 다수의 질문을 생성하고 이를 정제하여 확장 표현을 선택하는 접근법도 제안되었다.[2] 그러나 이러한 방식은 의학, 과학, 공학과 같이 전문 용어와 약어가 빈번히 사용되는 도메인에서는 성능 향상이 제한적이며,[2] 질의와 직접적으로 관련되지 않은 표현의 포함이나 의미적으로 유사한 표현의 반복 생성으로 인해 검색 결과의 정합성과 다양성을 저해하는 구조적 한계를 드러낸다.

이 문제는 로컬 환경에서 경량 LLM 을 사용하는 검색 증강 생성(RAG) 시스템에서 더욱 심각하다. 보안, 비용, 응답 지연 등의 이유로 외부 상용 LLM 을 선택하지 않는 환경에서는 상대적으로 작은 모델을 활용해야 하며, 이 경우 생성 기반 질의 확장은 오히려 검색 품질 저하로 이어질 수 있다. 따라서 로컬 LLM 기반 검색 환경에서는 생성량을 늘리는 방식이 아니라, 질의 의미에 부합하는 표현을 선별하고 의미적 중복을 억제하는 구조적 접근이 필요하다.

본 논문은 이러한 문제의식을 바탕으로 전문 도메인 검색에 적합한 질의 확장 프레임워크로서 Field-Aware Query Expansion 과 Semantic Redundancy Filtering 을 결합한 구조를 제안한다. 제안 방법은 질의를 분야 인식 관점에서 해석하고, 확장 과정에서 발생하는 의미 중복을 체계적으로 제어함으로써 로컬 LLM 환경에서도 효율적이고 정합성 높은 검색을 가능하게 한다.

Ⅱ. 본론

1. Field-Aware Query Expansion

기존 LLM 기반 질의 확장 기법은 질의를 단일 의미 공간에서 처리하고, 생성된 표현을 포괄적으로 활용하는 경향이 있다. 그러나 전문 도메인에서는 동일한 용어라도 분야에 따라 의미가 달라질 수 있으며, 특정 분야에서만 사용되는 전문 용어와 약어가 존재한다. 이러한 특성을 고려하지 않은 확장은 오확장 가능성을 높이고 검색 정합성을 저해한다.

분야 인식 질의 확장은 질의를 일반적인 문장이 아닌 분야 조건부 의미 표현으로 해석한다. 이를 위해

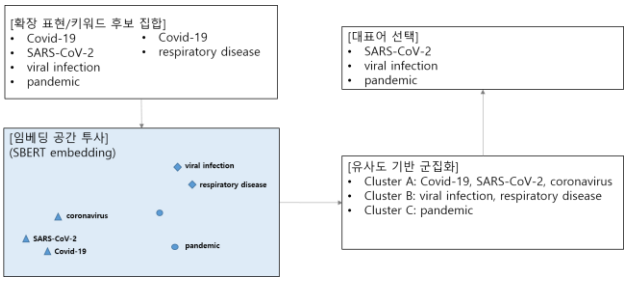
학술·산업 문서를 여섯 개의 대표 분야(①의학 및 보건과학, ②자연 및 물리과학, ③공학 및 컴퓨터과학, ④금융·경영 및 경제, ⑤법률 및 공공정책, ⑥사회 및 행동 과학)로 구분하여 질의가 속한 분야를 먼저 결정하고 중요도가 높은 표현을 추출한다. 이후 해당 분야에 대해 사전에 구축된 동의어 사전과 약어표만을 참조하여 질의의 핵심 키워드를 확장한다.

이 접근의 핵심은 LLM 을 새로운 표현을 대량 생성하는 역할이 아니라 분야에 적합한 표현을 선택·조정하는 역할로 활용한다는 점이다. 이를 통해 질의 확장은 의미적으로 일관된 범위 내에서 이루어지며, 전문 도메인에서 요구되는 용어 정확성과 맥락 적합성을 유지할 수 있다. 즉, LLM 은 표현 생성의 주체가 아니라, 분야 맥락에 부합하는 표현을 조정·선별하는 보조적 구성 요소로 기능한다.

2. Semantic Redundancy Filtering

질의 확장 과정에서 또 하나의 중요한 문제는 의미 중복이다. LLM 기반 확장은 의미적으로 매우 유사한 표현을 반복적으로 생성하는 경향이 있으며,[2] 이는 질의 길이를 불필요하게 증가시키고 검색 결과의 다양성을 감소시킨다. 단순한 문자열 기반 중복 제거만으로는 이러한 문제를 해결하기 어렵다.

SRF 는 확장된 표현 집합을 임베딩 공간에 투사한 후 의미적 유사도를 기준으로 중복을 억제하는 필터링 기법이다. 유사도가 높은 표현들은 하나의 군집으로 묶이고, 각 군집에서 대표 표현만을 선택함으로써 의미적으로 거의 동일한 표현의 중복 사용을 방지한다. 이 과정은 단순한 축약이 아니라, 질의에 새로운 문맥 정보를 제공하는 표현을 우선적으로 유지하도록 설계된다. 이러한 의미 중복 억제 접근은 정보 검색 분야에서 중복을 최소화하면서 정보 다양성을 유지하고자 한 기존 연구의 문제의식과도 맥을 같이 한다[3]. 나아가, 대표성 있는 정보의 선택을 통해 반복을 억제하려는 고전적 접근은 질의 표현 단계에서도 유사한 고려가 필요함을 시사한다[4].



[그림 1] SRF 의 개념적 구조: 확장된 질의 표현들을 임베딩 공간에 투사한 뒤, 의미적 유사도 기반으로 군집화하고 각 군집의 대표 표현만을 선택함으로써 의미 중복을 제거함.

SRF 를 적용하면 확장된 질의는 더 간결해지지만, 각 표현이 제공하는 정보량과 문맥적 다양성은 유지된다. 이는 검색 모델이 동일한 의미를 반복적으로 소비하는 대신, 질의의 다양한 해석 가능성을 균형 있게 반영하도록 돕는다.

3. 로컬 LLM 기반 검색 프레임워크에서의 의의

분야 인식 질의 확장과 의미 중복 제거는 로컬 LLM 기반 검색 환경에 특히 적합한 구조를 갖는다. 질의의 분야를 먼저 분류하고 분야별 사전과 약어표를 기반으로 한 확장은 LLM 호출 빈도를 최소화하며, 의미 중복 제거는 제한된 연산 자원을 보다 효율적으로 활용할 수 있게 한다. 이러한 구조는 GPU 자원이 제한된 작은 LLM 환경에서도 안정적인 질의 확장과 검색 품질 유지를 가능하게 한다.

Ⅲ. 결론

본 논문은 로컬 LLM 기반 RAG 환경에서 전문 도메인 검색의 구조적 한계를 해결하기 위한 Field-Aware Query Expansion 과 Semantic Redundancy Filtering 프레임워크를 제안하였다. 제안 방법은 질의를 분야 인식 관점에서 재구성하고, 확장 과정에서 발생하는 의미 중복을 체계적으로 억제함으로써 대형 LLM 에 대한 의존 없이도 효율적이고 정합성 높은 검색을 가능하게 한다. 이는 정보 보안과 비용 제약이 존재하는 실제 산업·연구 환경에서 실용적인 검색 확장 해법을 제공한다는 점에서 의의가 있다. 향후 연구에서는 문서 확장과 동적 분야 적응 기법을 결합하여 다양한 전문 도메인에 보다 유연하게 적용 가능한 범용 검색 프레임워크로 확장할 계획이다.

ACKNOWLEDGMENT

이 논문은 2025 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임. (No.RS-2022-00155857, 인공지능융합혁신인재양성(충남대학교)).

참 고 문 헌

[1] Rolf Jagerman et al., Query Expansion by Prompting Large Language Models, arXiv:2305.03653v1, 2023.

[2] Wonduk Seo et al., QA-Expand: Multi-Question Answer Generation for Enhanced Query Expansion in Information Retrieval, arXiv:2502.08557v1, 2025.

[3] Jamie Carbonell et al., The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries, SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, 335-336, 1998.

[4] Rodrygo L. T. Santos et al., "Search Result Diversification", Foundations and Trends® in Information Retrieval: Vol. 9: No. 1, pp 1-90, 2015.