스토어 오염 완화를 위한 분산 스토어 및 중재자 기반 멀티에이전트 시스템

어호선[1], 양경식[2]
[1]고려대학교 인공지능융합학과 석사과정
[2]고려대학교 컴퓨터학과 조교수

redysj0517@korea.ac.kr, g_yang@korea.ac.kr

Multi-agent System by Distributed Stores and Mediators for Mitigating Store Contamination

Hosun Eoo[1], Gyeongsik Yang
[1]Dept. of AI Convergence, Korea University
[3]Dept. of Computer Science and Engineering, Korea University

요 약

대규모 언어 모델(LLM)은 다양한 자연어 과제에서 뛰어난 성능을 보이지만, 단일 스토어 기반 검색 과정에서 발생하는 스토어 오염과 단일 에이전트 의존성으로 인한 한계가 존재한다. 본 연구는 이러한 문제를 해결하기 위해, MMDS(mediator-based MAS with distributed store)라는 새로운 멀티 에이전트 프레임워크를 제안한다. MMDS는 중재자 에이전트가 복합 질의를 분해하고, 관련성이 높은 도메인 에이전트만 토론에 참여하도록 조율하여, 불필요한 응답을 제거하고 보다 신뢰성 있는 결과를 도출한다. 금융, 과학, 의학, 영양 등 다양한 데이터셋에 대한 복합 질의를 기반으로 평가한 결과, 제안한 MMDS는 타 기법들에 비해 추론 과정에 필수적인 정답 문서의 검색 정확도를 최대 8배 개선함을 확인하였다.

1. 서론

대규모 언어 모델(LLM)은 다양한 자연어 과제에서 뛰어난 성능을 보이나, 한정된 데이터셋으로 훈련된 정적 파라미터를 사용하는 과정에서 사실적 환각(hallucination)과 최신성 부족 문제가 발생한다[1]. 이를 보완하기 위해 최근 제안된 RAG(retrieval-augmented generation)는 LLM을 에이전트로 구동하고, 에이전트의 추론 과정에 외부의 추가 정보 및 근거를 검색하여 추론 및 결과 생성 과정에 통합함으로써, 환각 문제를 효과적으로 완화하였다[1].

초기의 RAG는 이질적 도메인 문서를 하나의 저장소(스토어)에 모두 함께 적재하는 방식을 사용하였는데, 이는 도메인 특화 신호를 희석시키고 검색 품질과 근거 신뢰도가 저하하는 스토어 오염이 발생하였다[2]. 또한, LLM의 사실 환각 한계를 줄이기 위해 추론(reasoning)과 외부 도구 호출이나 검색과 같은 행동(acting)을 교차적으로 수행하여 정보를 보강하는 ReAct 방식이 제안되었다. 그러나 이 방식은 여전히 단일 에이전트에 의존하므로, 역할 분담이나 복잡한 과제의 수행 등에 낮은 정확도와 효율성을 보인다[3].

따라서, 최근에는 여러 에이전트를 활용하는 MAS(multi-agent system)이 적극적으로 연구되고 있다. MAS는 에이전트 간 추론 과정에서의 역할을 분담하고, 에이전트 간 토론 등을 거쳐 생성되는 답변을 분해하여 복잡도를 줄이고, 상호 검증을 통해 환각을 완화한다. 현재까지 제안된 MAS 연구들은 주로 에이전트 간 어떻게 문제를 분해하고 상호 협업을 할 지에 집중하고 있다[4].

본 연구는 MAS 구조에서 단일 스토어 사용에 따른 환각 문제에 주목한다. 기존의 MAS는 주로 문제 분해와 협업 방식에 초점을 두었으나, 복합 질의 환경에서 단일 스토어를 사용할 경우 도메인 간 경계가 흐려져 검색 품질과

응답 신뢰도가 저하되는 한계가 존재한다.

이를 해결하기 위해, 본 연구에서는 분산 스토어와 중재자 에이전트(mediator)를 결합한 새로운 MAS 프레임워크인 MMDS(mediator-based MAS with distributed store)를 제안한다. MMDS 는 사용자의 복합 질의를 중재자 에이전트가 분해하여 도메인별 스토어와 연결하고, 관련성이 높은 도메인 에이전트만 토론에 참여하도록 조율한다. 이후 에이전트 간 토론 과정에서 발언의 연관성을 분석하여 불필요한 응답을 제거함으로써, 정보 오염과 환각 문제를 최소화한다.

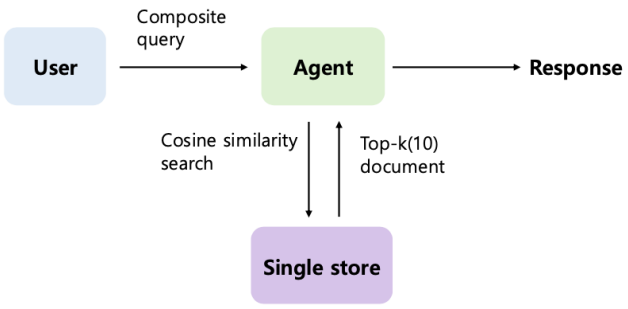
MMDS 를 검증하기 위해, 우리는 대부분의 연구에서 활용하는 BEIR 데이터셋을 활용하여 금융, 과학, 의학, 영양의 네 가지 이질적 도메인 환경을 구성한다. 제안한 MMDS 를 타 MAS 방식과 비교한 결과, 기준 정확도에서 8 배 이상의 효과적인 개선을 입증하였다.

2. 배경 및 제안방식

본 장에서는 대표적인 MMS 방식과 새롭게 제안하는 MMDS 구조를 설명한다.

2.1 MMS 구조

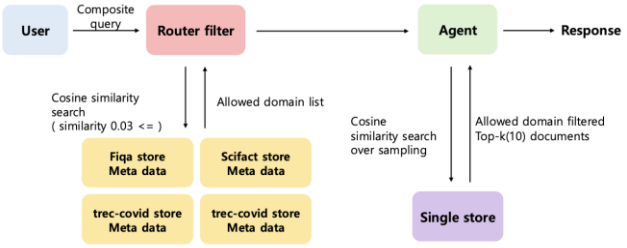
첫째, Baseline 은 그림 1 과 같이 가장 기본적인 방식으로서, 사용자(그림의 user) 질의를 받은 단일 에이전트가 단일 스토어에서 유사도 검색을 한 후 유사도가 높은 상위 외부 지식(문서)을 반환 받아서 답변 생성에 활용하는 구조이다. [1] 등이 본 구조를 활용한다.



<그림 1. baseline 구조도>

둘째, filtered router 는 그림 2 와 같이 사용자의 질의를 특정 분야(도메인)에 적합하게 한정(제한)하기 위한 필터를 사용하는 구조이다. 구체적으로, 사용자 질의가 속할 수 있는 각 도메인별로 도메인의 특징과 대표의미를 요약하는 메타데이터 벡터를 사전에 구축한다. 이 메타데이터 벡터는 상호 간 크기 차이를 제거하기 위해 L2 정규화를 거친다. 이후 사용자가 입력된 질의를 필터에서 임베딩한 후, 해당 질의 벡터와 도메인별 메타데이터 벡터 간의 유사도를 계산한다. 높은 유사도를 보이면 허용 도메인으로 분류된다.

이 단계가 끝나면, 단일 에이전트는 단일 스토어에서 추론 과정에 사용 가능한 외부문서를 검색하고, 검색된 결과에서 허용 도메인에 속하는 문서만을 선별하여 질의와 무관한 외부 문서를 제거한다. 마지막으로 상위 문서들을 택하여, 추론에 사용한다.



<그림 2. filtered router 구조도>

2.2 제안구조: MMDS

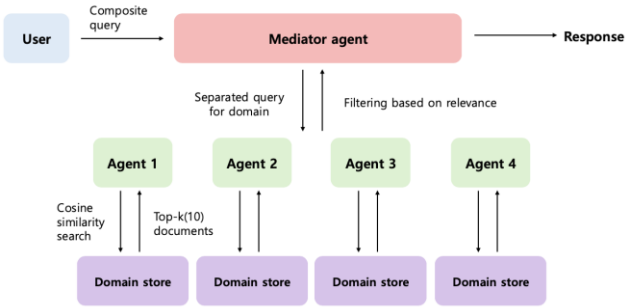
본 연구에서 제안하는 MMDS 는 복합 질의 상황에서 도메인별 전문성을 유지하고, 에이전트 간 협업을 조율하여 스토어 오염 문제를 개선하는 프레임워크이다. 본 연구에서는 12 개의 LLM 에이전트를 구축하며, 각 에이전트는 서로 다른 도메인에 대한 전문지식을 도메인 스토어(domain store)로 지니고 있다. 에이전트의 수는 변경 가능하다.

먼저, 사용자의 질의는 중재자 에이전트(그림 3 의 mediator agent)가 수신한다. Mediator agent 는 질의의 길이에 따라 길고 복잡한 질의를 여러 에이전트에 나누어 처리하도록 하기 위해, 다수의 질의 세그먼트(segment)로 나눈다. 이렇게 분해된 세그먼트는 각 에이전트로 전달되고, 에이전트마다 구축된 개별 도메인 스토어(domain store)와 비교된다.

이 때, 각 세그먼트와 도메인 스토어 간의 유사도가 사전에 정의된 임계값을 초과할 경우, 이는 질의가 해당 에이전트의 도메인에 적합하다는 것을 뜻한다. 따라서, 해당 에이전트가 추론 및 토론 과정에 참여하도록 활성화한다.

그 후, 복수의 에이전트가 참여하는 토론 과정이 진행된다. 각 에이전트는 자신들의 검색 결과를 근거로 응답을 제안하며, Mediator agent 는 제안된 응답과 사용자의 최초 질의 사이의 연관성을 평가한다. 연관성이 낮거나 불필요하다고 판단되는 응답은 무시되고, 최종적으로 가장 관련성이 높은 응답을 조합한다.

즉, MMDS 는 단순한 다중 스토어 접근을 넘어, 1) 질의 분해, 2) 관련 도메인 선별, 3) 에이전트 토론 및 조율의 절차를 거쳐 보다 신뢰성 있는 결과를 도출한다.



<그림 3. MMDS 구조도>

3. 실험 및 평가

3.1 실험 방법

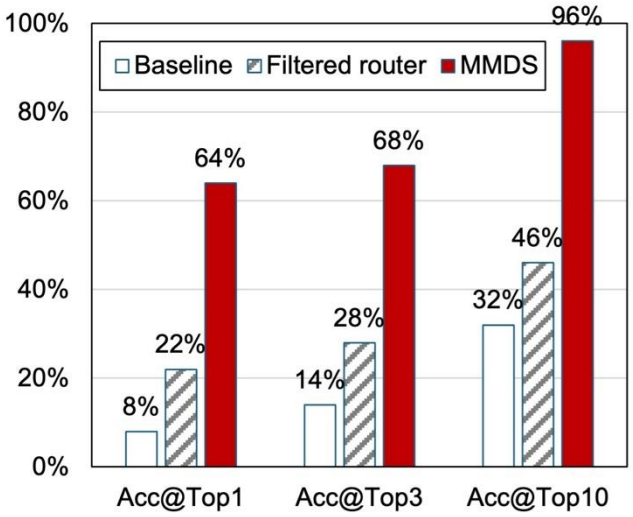
제안 기법을 평가하기 위해, MMS 평가의 표준 벤치마크인 BEIR (Benchmarking Information Retrieval)을 기반으로 금융, 과학, 의학, 영양에 대한 네 가지 도메인 데이터셋을 활용한다[1]. 구체적으로 1) FiQA 는 금융 영역의 의견/질의 응답에 대한 데이터, 2) SciFact 는 과학 논문 초록 기반 사실 검증에 대한 데이터, 3) TREC-COVID 는 COVID-19 관련 과학 문헌에 대한 의학 정보에 대한 데이터, 4) Nfcorpus 는 환자들의 실제 영양 관련 데이터를 지니고 있다. 각 도메인 데이터들에는 질문에 따른 추론에 필수적인 관련 문서가 어떤것인지 ground-truth 값을 포함하고 있다.

실험은 각 데이터셋에서 질의를 무작위로 샘플링하고, 네 도메인의 질의를 하나로 합친 복합질의(composite query)를 총 100 개 구성한다. 이는 단일 스토어 환경에서 발생할 수 있는 스토어 오염 상황을 시뮬레이션하기 위함이다. 정답 문서 집합은 각 데이터셋의 공개 qrels (test relevance judgments)를 사용한다. 모든 실험은 동일한 임베딩 모델과 인덱싱 기법 하에서 진행하여, 비교군 간의 차이가 프레임워크 구조에서 비롯되도록 통제한다.

본 연구에서는 제안한 MMDS 를 2 장에서 설명한 baseline 및 filtered router 와 비교한다. 각 기법은 LLM 에이전트로서 OpenAI 에서 제공한 1536 차원의 `text-embedding-3-small` 모델을 사용한다. 또한, 에이전트의 스토어로는 FAISS `vector store` 를 사용한다. 실험 진행 시 상위 10 개의 외부 문서를 선택하여 활용한다.

평가 지표로는 제안한 구조가 얼마나 스토어 오염을 개선하여, 주어진 복합질의가 주어졌을 때 추론에 필수적인 외부 문서를 찾아내는 지 평가한다. 즉, 검색된 상위 K 개의 문서에 관련 문서가 포함되면 1, 아니면 0 으로 계산하는 Acc@TopK (Accuracy at Top-K)를 사용한다

3.2 실험 결과



<표 1> 성능 비교표

표 1 은 세가지 모델에 대한 문서검색 성능을 비교한다. 구체적으로, 제안한 MMDS 는 기존 기법대비 Acc@Top1 에서 baseline 과 filtered router 에 비해 8 배, 2.9 배 높은 정확도를 보인다. Acc@Top3 에서는 baseline 과 filtered router 에 비해 4.85 배, 2.42 배 높은 정확도를 보인다. 특히, 선별한 10 개 문서 내에 관련문서가 포함될 확률인 Acc@Top10 에서 MMDS 는 96%의 정확도로 관련 문서를 포함하는데, 이는 대부분의 복합 질의에 대해 관련 문서를 잘 검색하고 활용할 수 있음을 드러낸다.

4. 결론

본 연구는 단일 스토어 기반의 MAS 가 지니는 정보오염 한계를 극복하는 MDSS 를 제안하였다. 중재자 에이전트를 통한 복합질의의 분해 및 특화 도메인별 에이전트, 분산 스토어 활용을 통해 MDSS 는 기존 기법 대비 최대 8 배 이상 개선된 외부 문서 활용 능력을 보였다.

향후 연구에서는 복합 질의의 동적인 분해 및 프레임워크 파라미터 설정의 기법, 선택된 외부 문서가 최종 답변의 정확도 및 품질 개선에 미치는 영향, 추론 및 토론 과정의 해석가능성을 높이기 위한 연구를 수행할 예정이다.

Acknowledgement

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원(RS-2024-00336564)과 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(RS-2024-00405128)을 받아 수행된 연구임.

참고문헌

[1] Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N., Küttler H., Lewis M., Yih W-t., Rocktäschel T., Riedel S., Kiela D. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv, 2020.

[2] Parshin Shojaee, Sai Sree Harsha, Dan Luo, Akash Maharaj, Tong Yu, Yunyao Li “Federated Retrieval Augmented Generation for Multi-Product Question Answering” arXiv, 2025

[3] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, Yuan Cao “ReAct: Synergizing Reasoning and Acting in Language Models” arXiv, 2022

[4] Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., Awadallah, A. H., White, R. W., Burger, D., Wang, C. “AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation.” arXiv, 2023.

[5] Thakur N., Reimers N., Rücklé A., Srivastava A., Gurevych I. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. arXiv, 2021.