검색 증강 생성(RAG) 기술에 대한 최신 연구 동향

이은빈[1], 배호[2,3]
[1]이화여자대학교 인공지능융합전공 석사과정
[2]이화여자대학교 사이버보안학과 교수
[3]이화여자대학교 인공지능융합전공 교수
eunbinlee@ewha.ac.kr, hobae@ewha.ac.kr

A Survey on Retrieval-Augmented Generation

Eun-Bin Lee[1], Ho Bae[2,3]
[1]Dept. of Artificial Intelligence Convergence, Ewha Womans University
[2]Dept. of Cyber Security, Ewha Womans University
[3]Dept. of Artificial Intelligence Convergence, Ewha Womans University

요 약

글로벌 시장에서 Large Language Model(LLM)의 발전이 급속하게 이루어지며 활용도가 높아지고 있지만 특정 유형이나 전문적 지식이 부족할 수 있어 일반화하기 어려우며, 새로운 데이터로 업데이트하기 어렵다는 한계점이 있다. 이를 극복하기 위해 지속적으로 업데이트되는 최신 정보를 포함한 외부 데이터베이스에서 정보를 검색해 응답을 생성하는 Retrieval- Augmented Generation(RAG, 검색 증강 생성) 모델을 도입하여 LLM의 환각 현상을 최소화하고 효율성과 정확성을 향상시키려는 연구가 활발히 이루어지고 있다. 본 논문에서는 LLM의 검색 기능을 강화하기 위한 RAG의 연구 및 평가기법에 대한 최신 연구 동향을 소개하고 실제 산업에서 활용하기 위한 최적화 및 응용 사례를 소개하며 이를 바탕으로 향후 연구 방향성을 제시하고자 한다.

1. 서론

최근 글로벌 시장에서의 Large Language Model (LLM)의 발전이 급속하게 이루어지면서 여러 산업에서의 활용도가 높아지고 있다. 하지만 LLM은 대규모 데이터셋에서 훈련되었음에도 특정 유형이나 전문적 지식이 부족할 수 있어 일반화하는 데 한계가 있다[1]. 또한 모델의 훈련이 완료된 이후에는 새로운 데이터나 최신 정보로 즉각 업데이트되기 어렵다. 이를 극복하기 위해 지속적으로 업데이트되는 최신 정보를 포함한 외부 데이터베이스에서 정보를 검색해 LLM의 응답을 생성하는 모델 Retrieval-Augmented Generation(RAG, 검색 증강 생성)[2] 모델이 제안되었다. RAG는 LLM이 최신 정보를 정확하고 빠르게 제공할 수 있도록 보장하며, 외부 데이터베이스에 전문 지식을 추가할 경우 정확한 응답을 얻을 수 있어 LLM의 환각 현상(Hallucination)을 최소화할 수 있다. 특히 장문의 문장에서 맥락 처리를 할 때 기존의 방법론인 파인튜닝(Fine-Tuning)이나 문맥 창(Context Window)보다 우수한 성능을 보인다는 점에서 LLM의 효율성과 응답 품질을 향상시킬 때 RAG는 더욱 유용하게 사용될 수 있다[3].

본 논문에서는 LLM의 검색 기능을 강화하기 위한 RAG의 연구 및 평가기법에 대한 최신 연구 동향을 소개한다. 또한 실제 산업에서 RAG를 활용하기 위한 성능 최적화 및 응용 사례들을 소개한다. 이를 바탕으로 향후 RAG 모델을 발전시킬 수 있는 연구 방향성을 제안하고자 한다.

2. RAG 연구동향

2.1 LLM 검색 강화를 위한 RAG

사용자로부터 질의 쿼리를 입력 받을 때, 사전학습된 LLM에 포함되어 있지 않은 외부지식을 참고해야할 수도 있다. 이처럼 모델의 관점에서 파라미터화되지 않은 외부지식이 필요한 문제들을 지식집약적(Knowledge-Intensive) 문제라고 할 수 있다. 지식집약적 문제를 해결하기 위해 외부지식으로부터 참조하려는 시도는 Lewis, Patrick, et al.[2], Guu, Kelvin, et al.[4]에서 이루어졌으며, 이는 RAG 연구의 초기 흐름과 연결된다.

Yu, Wenhao, et al.[5]는 RAG의 견고성을 개선하

기 위해 RAG에서 관련 정보를 검색한 후 각 문서에 대하여 순차적인 리딩 노트(reading note)를 생성하여 일관성 있는 응답으로 통합하는 Chain-of-Note (CoN) 방법론을 제안하였다. 이는 잠재적으로 결함이 있는 외부 데이터에 대한 의존성을 크게 감소시켰으나 외부 데이터베이스로부터 정보를 검색하고 처리하는데 상당한 계산 비용과 시간이 소요될 수 있다는 단점이 있다.

RAG 기반 LLM의 효율성을 향상시키기 위해 He, Zhenyu, et al.[6]는 처리 속도에 초점을 맞춘 REST(Retrieval-Based Speculative Decoding) 모델을 제안하였다. 이는 코퍼스에서 구축된 문맥-계속성(context-continuation) 쌍이 포함된 외부 데이터베이스를 사용한다. 이는 주어진 문맥에서 연결될 가능성이 높은 토큰을 문맥-계속성 쌍으로부터 검색하기 위해 쿼리하는 과정에서 추론 과정을 간소화하였다.

RAG 모델의 효율성을 향상시키기 위한 다른 방법론으로 Wang, Yile, et al.[7]은 LLM이 자신이 알고 있는 것과 모르는 것을 평가하고 지식 공백에 대해서만 외부 정보를 검색하도록 하여 불필요한 검색 행동을 방지하는 Self-Knowledge Guided Retrieval Augmentation (SKR) 방법을 제안하였다. 직접적인 질문과 문맥 학습을 통해 필요할 때만 검색 방법을 사용함으로써 모델의 효율성을 확보하면서도 정확성과 관련성을 개선시켰다.

이후 RAG로부터 생성된 응답의 품질과 사실적 무결성을 향상시키기 위해 자기 비평(self-critique) 메커니즘을 통합한 Self-RAG을 Asai, Akari, et al.[8]에서 제안하였다. Self-RAG는 검색한 내용을 비평할 수 있는 리플렉션 토큰(reflection token)을 사용하여 입력 쿼리에 대해 검색된 정보의 관련성과 유용성을 평가한다. 이를 통해 원래 모델의 창의성과 다양성을 감소시키지 않으면서 응답의 품질과 사실성을 향상시키며, 모델이 자가 조절할 수 있는 메커니즘이라는 장점이 있다. 그러나 여전히 생성된 결과물 중 모든 내용이 인용된 자료와 완전히 부합하지 않거나 인용된 자료로 완전히 뒷받침되지 않는 정보를 포함할 수 있다는 한계점이 존재한다.

2.2 RAG 최적화 연구 방향

관련성이 낮거나 정확하지 않은 정보를 생성할 수 있는 RAG의 단점을 극복하기 위해 다양한 최적화 연구가 진행되고 있다.

Asai, Akari, et al.[9]는 관련성이 있는 증거적 층면에서 정보를 평가하고 선택하는 과정을 생성 프로세스에 추가함으로써 RAG 모델의 정확성을 높이기 위한 방법을 제안한다. 하지만 이는 단편적인 출력을 생성할 때에만 효과적이며, 전체적인 문맥 차원에서는 관련성이 낮은 응답을 생성할 수 있다는 한계가 있다.

이를 위해 Wang, Zhiruo, et al.[10]는 문맥을 효과적으로 필터링할 수 있도록 타당성 추론, 어휘 중첩, 조건부 교차 상호 정보(Conditional Cross-Mutual Information)의 3가지 필터링 전략을 사용하여 유용한 문맥과 그렇지 않은 문맥을 구별시켜 RAG를 최적화하는 방법을 제안한다. 그러나 매우 복잡하거나 추상적인 주제에 대한 검색 결과에서는 이와 같은 필터링 기술의 제한이 나타날 수 있다는 한계점이 있으므로 일반화하기 어려울 수 있다.

복잡하거나 추상적인 작업을 처리할 때 LLM에 입력된 불필요한 데이터에 의해서도 성능이 저하될 수 있는데, 특히 Shi, Freda, et al.[11]은 수리 추론이나 논리적 판단과 같은 정밀도가 요구되는 작업에서 모델이 불필요한 정보를 필터링하거나 주어진 과제에 집중하는지 검증하고 평가하였다. [11]은 프롬프트에 불필요한 정보를 무시하도록 지시하는 명시적 내용을 추가하여 강건성을 유지할 수 있는 방법을 제안하였는데, 이는 명시적 지시가 모든 상황에 적용되지 않을 수 있으며, 추가적인 프롬프트에만 의존하므로 여전히 모델 자체의 개선이 필요하다는 단점이 있다.

환각 현상을 줄이기 위해 모델 자체를 개선한 방법론으로 Jiang, Zhengbao, et al.[12]은 동적 검색 시스템인 FLARE (Forward-Looking Active REtrieval augmented generation)을 제안하였다. 이는 미래의 필요를 예측하여 검색 쿼리를 생성하고 LLM의 출력을 직접 검색 쿼리로 사용하여 관련 정보를 가져오는 방식으로 특히 장문 텍스트 생성 시 모델의 유연성과 정확성을 크게 향상시켜 전반적인 환각 현상을 감소시켰다. 하지만 생성과 검색을 번갈아가는 과정에서 오버헤드와 생성 비용이 증가하며, 검색을 할 때마다 LLM이 여러 번 활성화되어야 한다는 한계가 존재한다.

2.3. RAG 기술 평가기법 연구 방향

RAG 기반 시스템을 평가하기 위해 Saad-Falcon, Jon, et al.[13]은 표준화된 방법론인 ARES(Automated RAG Evaluation System)을 제안하였다. 주어진 테스트 세트 안에서 RAG 시스템의 출력을 자동으로 평가하고, 성능 메트릭을 계산하는 방식이다. ARES는 정확도, 문서 유사도, 문서 재구성 등 다양한 메트릭을 통해 RAG 시스템의 성능을 평가한다. 또한 Es, Shahul, et

al.[14]은 충실도, 답변 관련성, 문맥 관련성 등의 메트릭을 통해 RAG 시스템을 평가하며, 이 세가지 측면에 대한 인간의 판단을 포함한 WikiEval 데이터셋을 도입하여 평가에 활용하였다.

하지만 특정 분야나 언어에 특화된 평가를 위해서는 추가적인 지표가 필요할 수 있어 향후 다층적인 관점에서 RAG를 평가할 수 있는 기법들에 대한 연구가 이루어져야 할 것이다.

2.4 RAG 기술 응용 연구 방향

RAG 모델은 빠르게 변화하는 최신 정보를 반영할 수 있고, 외부 데이터베이스에서 실시간으로 검색한 정보를 기반으로 응답을 생성할 수 있다. 이와 같은 장점을 활용하여 특정 분야나 산업에 특화시킨 응용 연구도 활발하게 이루어지고 있다.

Lozano, Alejandro, et al.[15]은 의료과학 문헌을 사용하여 임상 질문에 답할 수 있도록 RAG와 오픈소스 웹 어플리케이션을 통합한 시스템을 제안하였다. RAG 적용을 통해 의료 분야 연구자가 최신 연구 결과를 효율적으로 파악하고 의료 조언의 정확성을 높일 수 있도록 함으로써 RAG의 응용 분야를 확장하였다. 그러나 여전히 잠재적으로 RAG에서 잘못되거나 오해의 소지가 있는 정보를 생성할 수 있다는 한계가 있으므로 맹목적으로 신뢰하는 것은 위험할 수 있다는 한계점이 존재한다.

또한 Kang, Haoqiang, et al.[16]은 금융적 맥락에서 발생할 수 있는 환각 현상의 범위와 조건을 식별하였는데, 이를 통해 전문 분야에서 RAG를 사용할 때 발생가능한 환각 현상을 이해할 수 있는 경험적 증거를 제공하였다. 이처럼 금융 뿐 아니라 의료, 자율주행, 법률 등 다양한 전문 분야에 특화시켜 발생할 수 있는 환각 현상의 유형을 파악하고 완화하기 위한 연구가 필요하다.

3. 결론

정확하고 관련성 높은 응답을 생성하는 LLM을 위한 RAG의 연구가 활발하게 진행되고 있지만 아직 초기 단계에 불과하여 더욱 폭넓은 연구가 필요하다. 특히, 최근 생성형 AI에서 멀티모달 데이터의 중요성이 대두되고 있으므로 멀티모달 데이터를 처리할 수 있는 RAG의 연구가 필요하다[17].

또한 RAG를 적용하더라도 여전히 관련성이 낮거나 정확하지 않는 정보를 얻을 수도 있다는 문제점은 고질적인 한계점으로, 중요한 결정을 할 때 RAG에만 의존하는 것은 위험하다. 검색과 생성을 번갈아가며 하는 과정에서 오버헤드와 생성 비용을 증가시키는 것 역시 문제점이므로 이를 극복하기 위해 더욱 효과적으로 정보를 저장하고 검색할 수 있도록 RAG 모델을 개선하기 위한 연구가 필요하다.

4. 사사

이 논문은 2024년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.RS-2022-00155966, 인공지능융합혁신인재양성(이화여자대학교))

참고문헌

[1] MALLEN, Alex, et al. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. arXiv preprint arXiv:2212.10511, 2022.

[2] LEWIS, Patrick, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems, 2020, 33: 9459-9474.

[3] XU, Peng, et al. Retrieval meets long context large language models. arXiv preprint arXiv:2310.03025, 2023.

[4] GUU, Kelvin, et al. Retrieval augmented language model pre-training. In: International conference on machine learning. PMLR, 2020. p. 3929-3938.

[5] YU, Wenhao, et al. Chain-of-note: Enhancing robustness in retrieval-augmented language models. arXiv preprint arXiv:2311.09210, 2023.

[6] HE, Zhenyu, et al. Rest: Retrieval-based speculative decoding. arXiv preprint arXiv:2311.08252, 2023.

[7] WANG, Yile, et al. Self-knowledge guided retrieval augmentation for large language models. arXiv preprint arXiv:2310.05002, 2023.

[8] ASAI, Akari, et al. Self-rag: Learning to retrieve, generate, and critique through self-reflection. arXiv preprint arXiv:2310.11511, 2023.

[9] ASAI, Akari; GARDNER, Matt; HAJISHIRZI, Hannaneh. Evidentiality-guided generation for knowledge-intensive NLP tasks. arXiv preprint arXiv:2112.08688, 2021.

[10] WANG, Zhiruo, et al. Learning to filter context for retrieval-augmented generation. arXiv preprint arXiv:2311.08377, 2023.
[11] SHI, Freda, et al. Large language models can be easily distracted by irrelevant context. In: International Conference on Machine Learning. PMLR, 2023. p. 31210-31227.
[12] JIANG, Zhengbao, et al. Active retrieval augmented generation. arXiv preprint arXiv:2305.06983, 2023.
[13] SAAD-FALCON, Jon, et al. Ares: An automated evaluation framework for retrieval-augmented generation systems. arXiv preprint arXiv:2311.09476, 2023.
[14] ES, Shahul, et al. Ragas: Automated evaluation of retrieval augmented generation. arXiv preprint arXiv:2309.15217, 2023.
[15] LOZANO, Alejandro, et al. Clinfo. ai: An open-source retrieval-augmented large language model system for answering medical questions using scientific literature. In: PACIFIC SYMPOSIUM ON BIOCOMPUTING 2024. 2023. p. 8-23.
[16] KANG, Haoqiang; LIU, Xiao-Yang. Deficiency of Large Language Models in Finance: An Empirical Examination of Hallucination. arXiv preprint arXiv:2311.15548, 2023.
[17] CHEN, Wenhu, et al. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. arXiv preprint arXiv:2210.02928, 2022.