Memory-aware RAG 기반 접근을 통한 언어모델의 장기 기억 강화

장효석[1,2], 조한얼[3], 이상철[1], '김찬수[1,2,*]

1 한국과학기술연구원 인공지능·정보·추론 연구실,

2 과학기술연합대학원대학교 AI-로봇

3 세종대학교

*Correspondence should be addressed to eau@ust.ac.kr

Enhancing Long-Term Memory in Language Models via a Memory-Aware RAG Framework

Hyoseok Jang[1,2], Haneol Cho, Sangchul Lee, Chansoo Kim[1,2,*]

1 AI·Information·Reasoning (AI/R) Laboratory, Korea Institute of Science and Tech. (KIST)

2 AI-Robot Department, University of Science and Technology (UST)

3 Sejong University, Republic of Korea

요 약

소형 언어모델(SLM)은 우수한 추론 능력에도 불구하고 장기 기억 부족으로 실제 응용에서 한계를 보인다. 본 연구는 외부 메모리와 적응형 검색을 결합한 메모리 인식 RAG(Memory-aware RAG) 프레임워크를 제안한다. 제안된 접근법은 단기 버퍼와 장기 저장소로 구성된 계층적 메모리 구조, 시간 가중치 기반 검색, 하이브리드 인덱싱, 사용자 피드백을 통한 자가 갱신 메커니즘을 포함한다. 이를 통해 SLM은 장기간의 상호작용에서도 맥락적 일관성과 개인화를 유지할 수 있다. 실험 계획에는 성능 비교, 사용자 기반 평가, 효율성 분석이 포함되며, 다양한 도메인에서 소형 모델의 실질적 활용 가능성을 검증할 예정이다. 본 연구는 소형 언어모델의 활용성을 크게 확장하고, 경량화된 대화형 에이전트 구현을 위한 새로운 길을 제시한다.

Ⅰ. 서론

최근 소형 언어모델(Small Language Models, SLMs)은 상대적으로 적은 파라미터 규모에도 불구하고 합리적인 추론 능력과 빠른 응답 속도로 주목받고 있다 [4]. 이들은 엣지 디바이스, 자원 제약 환경, 맞춤형 응용 시스템 등에서 특히 활용 잠재력이 크다. 그러나 소형 언어모델의 근본적인 제약은 기억력의 부족이다. 제한된 모델 크기 때문에 장기간의 대화 맥락을 유지하거나, 누적된 사용자 상호작용을 일관되게 반영하는 능력이 현저히 떨어진다. 그 결과, SLM은 실제 대화형 에이전트나 장기적 개인화 서비스에서 제대로 활용되지 못하고 있는 실정이다.

Retrieval-Augmented Generation(RAG)은 외부 지식베이스를 활용해 모델의 한계를 보완하는 대표적 기법이다 [1]. 이 기법은 사용자 질의에 관련된 정보를 외부 데이터베이스에서 검색한 후, 이를 언어모델의 컨텍스트로 함께 제공하여 답변의 근거를 보강하는 방식으로 작동한다. 이를 통해 모델의 환각(hallucination) 현상을 완화하고 사실적 정확성을 높이는 데 크게 기여했습니다. 대부분의 관련 연구는 이처럼 정적인 외부 지식을 얼마나 효과적으로 검색하고 활용하는지에 집중되어 왔습니다. 그러나 기존 RAG는 사실 검색(factual retrieval)과 정적 지식 grounding에 치중되어 있으며, 시간에 따라 변화하는 개인화된 맥락이나 연속적 상호작용의 축적을 다루는 데 한계가 있다 [2][3]. 이에 본 연구는 소형 언어모델의 부족한 장기 기억을 보완하기 위한 메모리 인식 RAG(Memory-aware RAG, M-RAG) 프레임워크를 제안한다.

Ⅱ. 본론

M-RAG는 소형 언어모델의 한계를 보완하기 위해 계층적 메모리 구조와 적응형 검색 전략을 결합한다.

1. 계층적 메모리 구조(Hierarchical Memory)

- 단기 버퍼: 최근 대화 맥락을 저장하여 즉각적인 응답 일관성 유지.
- 장기 저장소: 의미적 임베딩과 이벤트 기반 인덱스를 활용해 사용자 과거 상호작용을 보존 [5]

2. 적응형 검색(adaptive Retrieval)

- 시간 가중치와 중요도 학습을 통해 오래된 정보와 최신 정보 간 균형을 유지하며, 맥락적으로 중요한 기억을 우선적으로 검색.

3. 하이브리드 인덱싱(Hybrid Indexing)

- 의미적 유사도 기반 검색과 이벤트 기반 검색을 결합해, 단순 텍스트 매칭을 넘어 사용자 경험 전반을 반영.

4. 자가 갱신(self-updating)

- 사용자 피드백을 반영하여 메모리 중요도를 재조정함으로써 변화하는 상황에 지속적으로 적응.

실험

본 연구는 다음과 같은 시험을 통해 M-RAG의 효과성을 검증한다.

- 성능 비교: 기존 RAG 및 단순 SLM 대비 대화 일관성, 사실 정확도, 개인화 반영도를 측정 [3][5].
- 사용자 기반 평가: 실제 사용자가 M-RAG 기반 SLM과 상호작용 후 설문을 통해 기억 지속성·만족도·개인화 체감을 평가 [7].
- 효율성 분석: 메모리 저장 비용, 검색 지연(latency), 모델 크기 대비

성능 향상을 측정하여 경량 환경 적합성을 검증.

기대 결과

BenchmarkQED [8] 점수 기준 베이스 소형모델 대비 향상된 벤치마크 점수 획득

Ⅲ. 결론

소형 언어모델은 우수한 추론 능력에도 불구하고, 장기 기억 부족으로 실제 응용에서는 활용성이 크게 제한되어 왔다. 본 연구가 제안하는 M-RAG 프레임워크는 외부 메모리와 적응형 검색 메커니즘을 통해 이러한 한계를 보완하고, 소형 언어모델이 장기적 상호작용에서도 일관성과 개인화를 유지할 수 있도록 한다. 이를 통해 자원 제약 환경에서도 신뢰할 수 있는 대화형 에이전트를 구현할 수 있으며, 모바일, 교육, 금융, 헬스케어 등 다양한 도메인에서 소형 모델의 실질적 활용 가능성을 크게 확장할 수 있을 것이다.

ACKNOWLEDGMENT

This research was funded by the grant Nos. 2023-00262155; 2024-00339583; 2024-00460980; and 2025-02304717 (IITP) funded by the Korea government (the Ministry of Science and ICT).

참 고 문 헌

[1] Lewis, P., Perez, E., Piktus, A., et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS, 2020.

[2] Zhao, X., et al. MemoRAG: Global Memory Enhanced Retrieval for Long Contexts. arXiv preprint arXiv:2409.05591, 2024.

[3] Zhuang, Z., et al. From RAG to Memory: HippoRAG 2. arXiv preprint arXiv:2502.14802, 2024.

[4] Anil, C., et al. Scaling Down: Small Language Models with Surprising Reasoning Capabilities. arXiv preprint arXiv:2310.xxxxx, 2023.

[5] Chen, X., et al. LongMem: Augmenting Language Models with Long-Term Memory. arXiv preprint arXiv:2306.07174, 2023.

[6] Ma, X., et al. MemLong: Memory-Augmented Retrieval for Long Text Modeling. arXiv preprint arXiv:2408.16967, 2024.

[7] Li, M., et al. LoCoMo: Evaluating Very Long-Term Conversational Memory of LLM Agents. arXiv preprint arXiv:2402.17753, 2024.

[8] Edge, D., Trinh, H., Morales Esquivel, A., & Larson, J. (2025, June 5). *BenchmarkQED: Automated benchmarking of RAG systems*. Microsoft Research. Retrieved August 29, 2025, from https://www.microsoft.com/en-us/research/blog/benchmarkqed-automated-benchmarking-of-rag-systems/

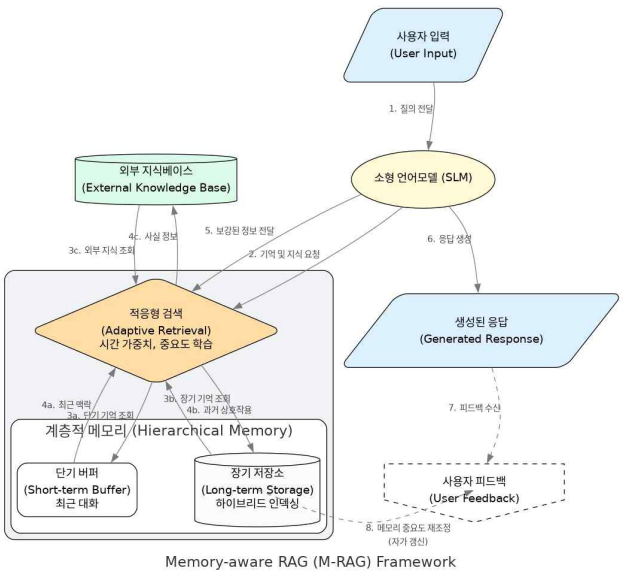

그림 1 M-RAG 구조도