검색모델과 LLM의 상호작용을 활용한 사고사슬 기반의 한국어 질의응답

박민준°, 심묘섭, 민경구, 최주영, 정해민, 최정규
LG AI 연구원
{minjun.park,myoseop.sim,kyungkoo.min,jooyoung.choi,haemin.jung,stanleyjk.choi}@lgresearch.ai

Interactions of Retriever and LLM on Chain-of-Thought Reasoning for Korean Question Answering

Minjun Park°, Myoseop Sim, Kyungkoo Min, Jooyoung Choi, Haemin Jung, Stanley Jungkyu Choi
LG AI Research

요 약

최근 거대언어모델(LLM)이 기계 번역 및 기계 독해를 포함한 다양한 문제들에서 높은 성능을 보이고 있다. 특히 프롬프트 기반의 대규모 언어 모델은 사고사슬 방식으로 적절한 프롬프팅을 통해 원하는 형식의 답변을 생성할 수 있으며 자연어 추론 단계에서도 높은 정확도를 보여주고 있다. 그러나 근본적으로 LLM의 매개변수에 질문에 관련된 지식이 없거나 최신 정보로 업데이트 되지 않은 경우 추론이 어렵다. 이를 해결하기 위해, 본 연구는 검색문서와 생성모델의 상호작용을 통해 답변하는 한국어 질의응답 모델을 제안한다. 검색이 어려운 경우 생성형 모델을 통해 질문과 관련된 문장을 생성하며, 이는 다시 검색모델과 추론 과정에서 활용된다. 추가로 "판단불가"라는 프롬프팅을 통해 모델이 답변할 수 없는 경우를 스스로 판단하게 한다. 본 연구결과에서 GPT3를 활용한 사고사슬 모델이 63.4의 F1 점수를 보여주며 생성형 모델과 검색모델의 융합이 적절한 프롬프팅을 통해 오픈-도메인 질의응답에서 성능의 향상을 보여준다.

주제어: 한국어 질의응답, LLM, 검색, 생성형 AI, 프롬프팅

1. 서론

거대언어모델(Large Language Models, LLM)은 최근에 기계 번역 및 기계 독해를 포함한 다양한 문제들에서 높은 성능을 보여주고 있다. LLM은 적절한 프롬프팅을 통해 원하는 형식의 답변을 도출할 수 있으며 생성형 인공지능의 특성상 프롬프트에 따라 결과가 달라질 수 있다 [1, 2]. 특히 질의응답의 영역에서 LLM에서 기계독해 방식으로 내포한 문맥이 제공되거나 사전에 학습이 되었을 경우, 높은 확률로 정확한 답변을 제공할 수 있다 [3]. 그러나 Open-domain question answering (ODQA) 의 경우 지문이 따로 존재하지 않고 사전학습이 되어 있지 않을 경우 LLM이 잘못된 답변을 하는 경우가 많다. 이를 해결하기 위해 사전에 구축된 외부 지식(Knowledge resource)을 활용하여 질문에 답할 수 있는 문서를 검색하는 방법이 효과적이다 [4, 5].

한번의 검색으로 질문에 대한 관련 지식을 효과적으로 보완할 수 있지만 [6], 이 방법은 더 복잡한 사고사슬 추론 질문에 대해서는 명확한 한계가 있다. 복잡한 질문의 경우, 부분적인 지식을 검색하고 부분적인 추론을 수행한 후, 현재까지 수행된 추론의 결과를 기반으로 관련 문서를 검색하고 보완해야 한다. 이러한 방법을 사고사슬 (Chain-of-Thoughts) 방식이라 칭하며 단계별 자연어 추론을 통해 복잡한 질문에 답할 수 있다 [7, 8]. 본 연구는 LLM의 생성된 답변을 활용하는 멀티 스텝 검색 및 생성 모델을 제안하며 이는 단일 검색 모델보다 우수한 성능을 보여준다.

2. 관련 연구

2.1 ODQA 에서의 프롬프팅 방법

2.1.1 BM25 기반의 검색 모델

본 논문은 모델에 사용되는 프롬프트의 일부로 BM25의 검색결과를 활용한다 [9]. BM25는 검색된 passage의 순위를 결정하며 쿼리와 문서에서 각 단어의 빈도를 고려하여 쿼리 용어가 문서 내에서 자주 나타나는 경우 해당 문서에게 가중치를 준다. 이전 TF-IDF(Term Frequency-Inverse Document Frequency) 접근 방식과 달리 BM25는 문서 모음 전체에서 용어의 희귀성을 고려하고 IDF 점수를 정규화하기 위해 로그 함수를 사용하여 매우 희귀한 용어도 과도한 중요도 부스팅을 받지 않도록 한다.

2.1.2 활용한 ODQA

기존의 사고사슬 기반의 연구에서는 반복 검색을 활용하였다. 신경망 쿼리 표현을 사용하여 검색한 기계독해 모델의 결과를 바탕으로 업데이트하는 반복 검색 방법이 있었으며 [10] GPT3를 사용하여 긴 질문에 사고사슬로 답변하는 방법이 있다 [11]. 그러나 제안된 방법들은 지도 학습에 의존하기에 적은 데이터를 다룰 때 확장되기 어렵다.

2.2 LLM을 활용한 질의응답

2.2.1 Self-prompting

해당연구는 LLM의 매개변수에 저장된 방대한 지식과 문맥 이해 능력을 활용한 프롬프팅 방법을 제안하였으나 사전학습

시 지식이 저장되지 않으면 모델이 질문에 정확하게 답할 수 없다는 한계가 있다 [12].

2.2.2 Recitation-augmented 언어모델

해당연구는 외부 말뭉치에서 검색하지 않고 더 정확한 사실적 지식을 생성할 수 있는 모델을 제안한다 [13]. 이 모델은 먼저 자체 학습된 데이터에서 샘플링을 하여 한개 이상의 관련 문서를 검색 및 활용하여 최종 답변을 생성한다. 그러나 해당 모델은 모델 내부에 답변에 대한 지식을 보유하고 있음을 가정한다.[1]

3. 제안방법

본 논문은 검색모델과 LLM을 활용한 질의응답 시스템을 제안한다. 전체 시스템에는 네개의 구성요소가 있다: (i) 질문을 기반으로 한 검색모델을 실행한다, (ii) "판단불가"로 답변된 경우 LLM은 검색된 단락을 사용하여 답변을 시도한다, (iii) 생성된 문장과 LLM이 이전에 생성된 문장 및 검색된 단락을 입력값으로 사용하여 새로운 단락을 검색하고 기존 단락들을 함께 활용하여 LLM이 질문에 답한다, 마지막으로 (iv) LLM이 다시 한번 답변을 "판단불가"로 예측하면 최대 반복수에 도달하거나 "판단불가"가 아닌 다른 답변이 나올 때까지 프로세스를 반복한다.

3.1 검색모델

시스템의 첫번째 단계로 BM25를 기반으로 질문과 가장 유사한 단락을 검색한다. 검색모델을 위한 말뭉치를 구축하기 위해 Korquad 버전1의 학습 및 검증 데이터 [14] 중 중복된 단락들을 제거하고, 총 10,566개 단락의 말뭉치로 구성했다.

3.2 LLM 질의응답 모델

질문에 가장 유사한 단락을 검색한 후, LLM은 주어진 맥락을 기반으로 질문에 대한 답변을 한다. 그림 1의 추론(Reasoning) 부분에서 모델이 문서를 기반으로 결정할 수 없을 경우 "판단불가"로 답변하도록 지시하기 위해 구체적인 사례를 프롬프트에 추가했다. 추론 단계에서 이전에 생성된 문장과 검색된 단락을 활용하기 위해 문서에 들어갈 문장으로 G1, P1, P2, Q를 모두 합친다. 표 1은 문서-질문-답변 형식으로 구성된 프롬프트이며 마지막 줄의 빈 칸에 답변을 하도록 유도된다.

3.3 LLM 문장 생성

시스템이 질문에 대답하지 못하는 경우, LLM은 특별한 프롬프팅 없이 훈련된 매개변수로만 질문에 답변한다. 표 2의 프롬프트를 사용해 모델은 문서없이 질문 기반으로 답변을 생성하도록 하며, 모델의 내부에 저장된 지식을 활용한다.

[1] https://github.com/Edward-Sun/RECITE

표 1. LLM 질의응답 시 사용되는 프롬프트

아래 예시를 참고해서 마지막 질문에 대한 정답을 알려줘.

문서: 통일독립촉진회에서 활동하던 중 김구가 암살되었다. 김구의 장례는 7월 5일 김규식이 절충하여 국민장으로 치러졌다.
질문: 김규식이 김구의 장례식과 관련하여 장례위원회 측과 정부측을 조율하여 결정한 장례 방식은?
정답: 국민장

문서: 뉴욕 양키스 외야수 조 디마지오가 53 게임 연속 안타를 쳐낸 1941년 7월 13일, 루 게릭의 일대기를 영화로 만들겠다는 계획이 발표되었다.
질문: 프랑스의 수상은 누구입니까?
정답: 판단불가

문서: 1839년 바그너는 괴테의 파우스트을 처음 읽고 그 내용에 마음이 끌려 이를 소재로 해서 하나의 교향곡을 쓰려는 뜻을 갖는다.
질문: 바그너는 교향곡 작곡을 어디까지 쓴 뒤에 중단했는가?
정답:

표 2. LLM 답변 생성시 사용되는 프롬프트

아래 문서를 참고해서 질문에 대한 정답을 알려줘.

문서: 노아의 방주는 히브리 경전 또는 구약성경에 기록된 설화에 등장하는 배로, 아브라함 계통의 종교의 전승기록 속에 등장하는 직육면체 모양에 문이 옆에 있고, 뚜껑이 위에 달린 물에 뜨는 구조물이다.
질문: 노아의 방주에 대해 기록하고있는 복음서는 무엇인가?
정답:

4. 실험 및 결과

4.1 시스템 구조에 따른 결과

본 연구는 세 가지 접근법을 비교한다: (i) 제로샷(Zero-shot)의 경우 LLM이 질문만을 기반으로 답변을 출력하는 방법, (ii) 단일 검색모델의 경우 BM25가 말뭉치에서 질문과 가장 유사한 단락을 검색하고 검색된 단락과 LLM의 지식을 기반으로

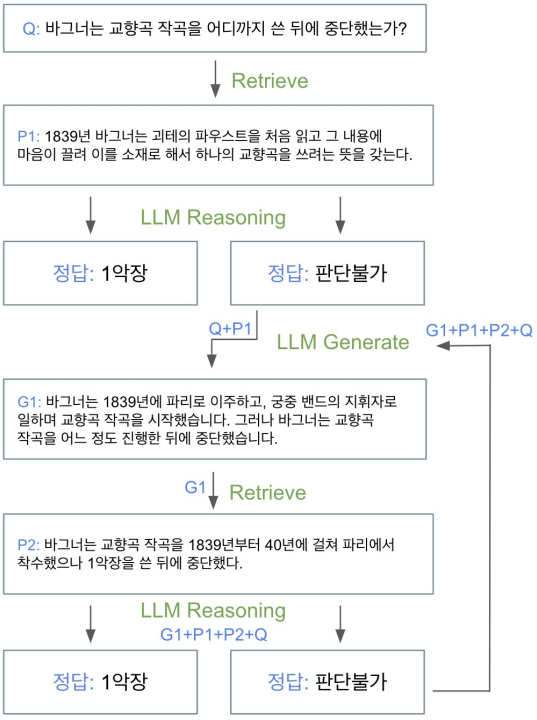

그림 1. 사고사슬 기반의 전체 검색 및 답변 시스템 구조

표 3. KorQuAD 1.0 검증데이터 성능 결과

	F1	EM
Zero-shot	14.8	2.9
Single Retriever	60.7	49.5
Multi-step CoT	**63.4**	**50.5**

질문에 답하는 방법, 마지막으로 (iii) 사고사슬의 경우, 1에 묘사되었듯이 검색과 생생모델을 모두 활용하는 방법이다. 표 3는 제로샷 모델의 경우 F1 점수가 14.8이었으며 질문이 매우 구체적이라 정확하게 예측하는 것이 매우 어려웠음을 보여준다. 검색모델을 사용 시 사고사슬 모델이 63.4의 F1 점수로 단일 검색모델보다 우수하였으며 생성된 문장들이 도움이 되었음을 보여준다.

4.2 생성모델의 결과 비교

본 논문은 다양한 한국어 LLM의 생성 능력을 비교했다. KoAlpaca[2]와 KoGPT[3]는 방대한 양의 한국어 데이터셋으로 학습되었다. 한국어에 특화된 LLM의 한국어 생성 능력을 평가하였으나 그림 4에 나타난 것처럼, GPT3.5-turbo 모델이 질문 답변에 가장 유용한 문장을 생성했으며 다른 한국어 LLM에 비해 성능이 소폭 증가하였다.

[2]https://github.com/Beomi/KoAlpaca
[3]https://github.com/kakaobrain/kogpt

표 4. 생성성모델에 따른 성능 비교

	F1	EM
KoAlpaca-5.8b	63.1	50.1
KoGPT-1.5b	62.0	50.0
GPT3.5-Turbo	**63.4**	**50.5**

표 5. 모델 Ablation study

	F1	EM
G1+P2+Q	63.0	**50.9**
G1+P1+P2+Q	**63.4**	50.5

또한, 질의응답 모델을 한국어 LLM으로 대체하는 시도를 했으나, 표 1의 프롬프트를 사용 시 한국어 LLM은 질의응답의 패턴을 인식하지 못하고 정확히 답변할 수 없었다. 따라서 한국어 LLM은 질의응답 모델로 사용하지 않고 문장 생성에만 사용되었다.

4.3 이전 문서의 보존여부 성능평가

본 연구는 "판단불가"의 답변이 나올 때, 이전에 생성된 문장과 검색된 단락들을 모두 보존 및 활용하는 것이 필요한지 실험하였다. 표 5의 결과는 보존하는 것의 유무는 큰 차이를 보여주지 않았다. 이는 문서의 길이가 모델의 성능을 대변하지 않는다는 것을 증명한다. 본 논문은 최종 모델로 F1 점수가 더 높은 보존방법을 사용하였다.

5. 결론

본 연구는 검색모델과 생성형 LLM을 결합한 사고사슬 프롬프트 형식의 질의응답 시스템을 제안하였다. 이 시스템은 매개변수에 학습되지 않은 질문에 대해서는 답변하지 못하는 LLM의 근본적인 한계를 극복하며 동시에 LLM로 생성된 문장들을 활용하여 기존 모델에 비해 더 높은 성능을 보여줬다. 또한, 한국어 LLM의 문장생성 능력과 프롬프트가 주어졌을 때 패턴 인식의 어려움의 한계를 다루었다. 추가실험으로 이전 문서의 보존 여부는 모델의 성능에 큰 영향을 미치지 않는 것으로 나타났으며, 이를 통해 모델이 문서의 길이에 강하게 의존하지 않음을 확인할 수 있다.

발전방향으로 검색모델에 검색랭킹(Re-ranking) 적용을 통해 더 정확한 문서 검색을 할 수 있다. 또한 한국어 말뭉치의 규모를 늘리거나 LLM에 더 많은 양의 데이터 학습을 통해 오픈-도메인 질문에 대해 정확한 답변을 할 수 있다. 본 연구결과는 사고사슬 질의응답 시스템의 가능성과 자연어 이해 분야에 기여할 수 있으며, 미래에 검색과 생성형 LLM이 융합된 모델의

기반을 제공할 것으로 기대한다.

참고문헌

[1] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, Vol. 33, pp. 1877–1901, 2020.

[3] S. Lee, H. Jang, Y. Baik, S. Park, and H. Shin, "Kr-bert: A small-scale korean-specific language model," *arXiv preprint arXiv:2008.03979*, 2020.

[4] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, "Dense passage retrieval for open-domain question answering," *arXiv preprint arXiv:2004.04906*, 2020.

[5] Y. Qu, Y. Ding, J. Liu, K. Liu, R. Ren, W. X. Zhao, D. Dong, H. Wu, and H. Wang, "Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering," *arXiv preprint arXiv:2010.08191*, 2020.

[6] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, Vol. 33, pp. 9459–9474, 2020.

[7] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, Vol. 35, pp. 24 824–24 837, 2022.

[8] H. Trivedi, N. Balasubramanian, T. Khot, and A. Sabharwal, "Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions," *arXiv preprint arXiv:2212.10509*, 2022.

[9] S. Robertson, H. Zaragoza *et al.*, "The probabilistic relevance framework: Bm25 and beyond," *Foundations and Trends® in Information Retrieval*, Vol. 3, No. 4, pp. 333–389, 2009.

[10] R. Das, S. Dhuliawala, M. Zaheer, and A. McCallum, "Multi-step retriever-reader interaction for scalable open-domain question answering," *arXiv preprint arXiv:1905.05733*, 2019.

[11] R. Nakano, J. Hilton, S. Balaji, J. Wu, L. Ouyang, C. Kim, C. Hesse, S. Jain, V. Kosaraju, W. Saunders *et al.*, "Webgpt: Browser-assisted question-answering with human feedback," *arXiv preprint arXiv:2112.09332*, 2021.

[12] J. Li, Z. Zhang, and H. Zhao, "Self-prompting large language models for open-domain qa," *arXiv preprint arXiv:2212.08635*, 2022.

[13] Z. Sun, X. Wang, Y. Tay, Y. Yang, and D. Zhou, "Recitation-augmented language models," *arXiv preprint arXiv:2210.01296*, 2022.

[14] S. Lim, M. Kim, and J. Lee, "Korquad: Korean qa dataset for machine comprehension," *Proceeding of the Conference of the Korea Information Science Society*, pp. 539–541, 2018.