Con-RAG: 기여도 기반 검색 필요성 분석을 통한 효율적인 검색 증강 생성

*김경현[1], 김영화[1], 진교훈[2], 김영빈[1, 2]
[1]중앙대학교 AI학과, [2]중앙대학교 첨단영상대학원
e-mail: khyun8072@cau.ac.kr, movie112@cau.ac.kr,
fhzh123@cau.ac.kr, ybkim85@cau.ac.kr

Con-RAG: Contribution-based Analysis of Retrieval Necessity for Efficient Retrieval Augmentation Generation

*Kyeonghyun Kim[1], Yeonghwa Kim[1], Kyohoon Jin[2] and YoungBin Kim[1, 2]
[1]Department of Artificial Intelligence, Chung-Ang University,
[2]Graduate School of Advanced Imaging Science, Multimedia & Film,
Chung-Ang University

Abstract

Large language models have significantly improved response generation capabilities but often generate inaccurate answers due to their reliance on internal knowledge. Retrieval augmented generation incorporates retrieved information into response generation but does not adequately address unnecessary retrievals. To tackle this issue, we propose contribution-based retrieval augmented generation (Con-RAG). Con-RAG determines the necessity of retrieval by evaluating the contributions of questions and retrieved information using the Shapley value. By training a classifier on these contributions, Con-RAG classifies questions based on whether retrieval is necessary, thus mitigating unnecessary retrievals and reducing computational costs. Experimental results show that Con-RAG outperforms existing methods by improving accuracy while using fewer parameters, highlighting its efficiency in reducing unnecessary retrievals. This demonstrates Con-RAG's effectiveness in enhancing performance while maintaining computational efficiency.

I. 서론

대규모 언어 모델 (Large Language Models, LLM)의 발전은 방대한 학습 파라미터와 지식을 기반으로 답변 생성 능력을 크게 향상시켰다[1]. 그러나 LLM은 내재된 지식에 의존하기 때문에 최신 정보 및 특정 분야에 대해 부정확한 답변을 생성하는 문제점이 있다. 따라서 다양한 질문에 대응하기 위해 검색 증강 생성 (Retrieval Augmented Generation, RAG)[2]이 제안되었다.

RAG는 질문과 관련된 정보를 검색하고 이를 답변 생성 과정에 반영하여 정확도를 높일 수 있다. 하지만 LLM에 내재된 일반 상식과 같이 검색이 불필요한 경우에도 검색을 수행하는 문제가 있다. 이는 긴 텍스트로 인한 추가적인 연산을 발생시키거나 답변 생성의 성능 저하를 초래할 수 있다.

이를 해결하기 위해, 주어진 질문에 대한 검색 필요성을 판단함으로써 불필요한 검색 문제를 개선하는 방법론이 연구되고 있다. Self-RAG[3]는 특수 토큰을 사용하여 검색 필요성을 판단하지만, 이를 위해 LLM에 대한 end-to-end 학습이 필요하여 높은 연산량이 요구된다. Adaptive-RAG[4]는 Chain-of-Thought[5]를 활용하여 답변 생성 과정에서 필요한 검색 횟수를 도출함으로써 검색 필요성을 판단한다. 그러나 다량의 토큰 생성을

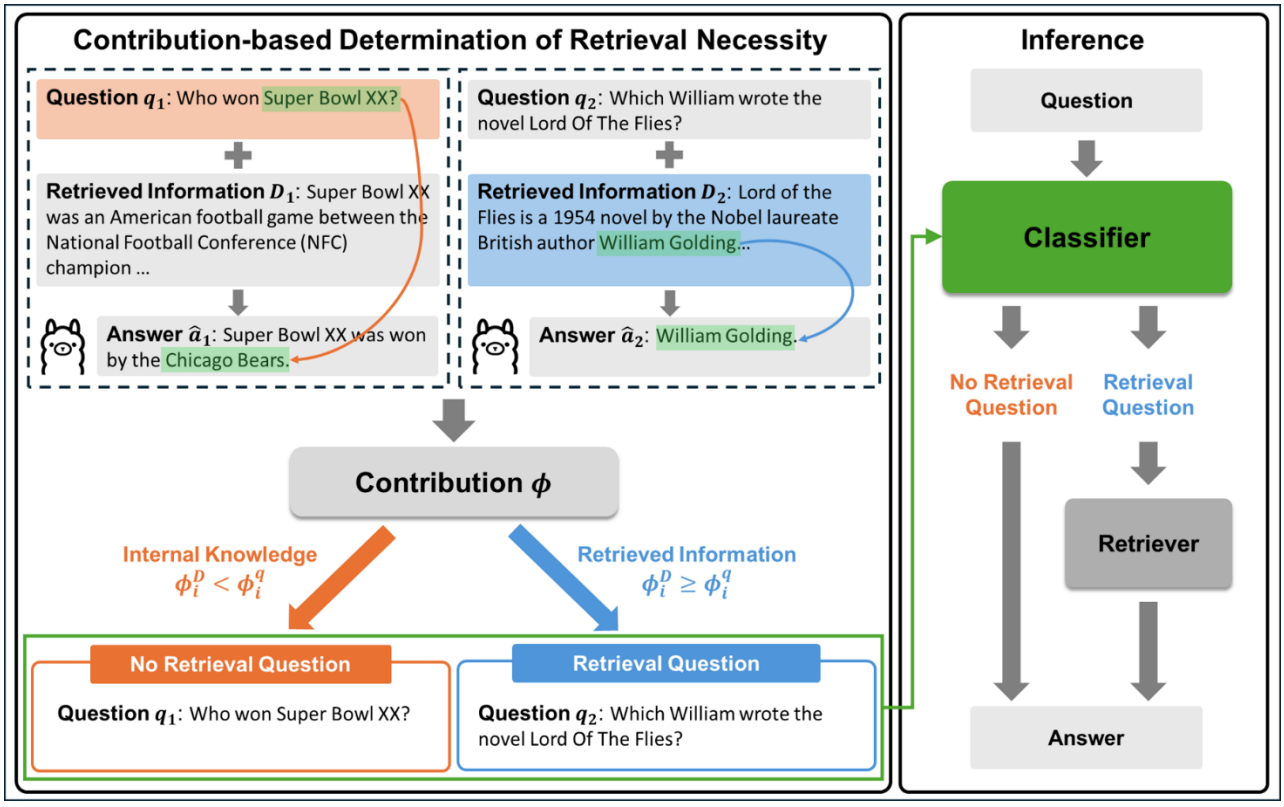

그림 1 기여도 기반 검색 필요성 분류기 학습 및 답변 생성

요구하므로 여전히 높은 연산량을 필요로 한다. 따라서 이를 해결하기 위한 효율적인 방법론의 필요성이 대두되고 있다.

본 논문은 질문과 검색된 정보의 기여도를 기반으로 검색 필요성을 판단하는 Contribution-based Retrieval Augmented Generation (Con-RAG)을 제안한다. Con-RAG는 Shapley value[6]를 활용하여 질문과 검색된 정보의 기여도를 각각 평가하고, 이를 기반으로 검색 필요성 분류기를 학습하여 불필요한 검색 문제를 해결하고 효율적으로 답변을 생성할 수 있다. 또한 Con-RAG는 적은 수의 학습 파라미터를 지닌 분류기를 활용함으로써 검색 필요성 판단을 위해 필요한 높은 연산량 문제를 해결하여 검색 연산 효율성을 향상시킨다.

질의응답 태스크에 대한 실험 결과, Con-RAG는 Natural Questions[7]에서 Self-RAG 대비 4.74%p의 성능 향상을 보이며, TriviaQA[8]에서 Self-RAG 대비 4.89%p, Adaptive-RAG 대비 3.69%p 높은 정확도를 보인다. 따라서, 제안 방식이 기여도 기반의 검색 필요성 분류기를 통해 성능을 향상시키는 동시에, 기존 방법론에 비해 적은 수의 학습 파라미터를 사용하여 효율적인 검색을 수행할 수 있음을 확인할 수 있다.

II. 본론

본 논문은 Shapley value를 활용하여 검색 필요성을 판단하는 Con-RAG를 제안한다. Con-RAG는 LLM이 생성한 답변에 대해 질문과 검색된 정보 각각의 기여도를 비교하여 검색 필요성을 판단하고, 이를 기반으로 검색 필요성 분류기를 학습한다. Con-RAG의 전체적인 구조는 그림 1과 같다.

2.1 기여도 기반 검색 필요성 판단

학습 데이터셋 $T = \{q_i^{train}, a_i^{train}\}_{i=1}^{|T|}$은 질문 $q_i^{train} \in Q_{train}$ 와 그에 대한 답변 $a_i^{train} \in A_{train}$ 쌍으로 구성된다. 질문 q_i^{train}과 관련된 상위 k개의 정보 $D_i = \{d_1, d_2, \dots, d_k\}$를 검색한 뒤, q_i^{train} 와 D_i 를 입력으로 받아 LLM이 생성한 답변을 $\hat{a}_i = LLM(q_i^{train}, D_i)$ 로 정의한다. 이때 LLM이 생성한 ai가 정답인 경우로 한정하여, LLM이 정답을 맞힌 질문 q_i로 구성된 집합 $Q_c \subseteq Q_{train}$으로 학습 데이터셋을 한정한다.

검색 필요성을 판단하기 위해, 제안하는 Con-RAG는 Shapley value를 활용하여 가능한 입력 조합의 성능 평균을 통해 특정 입력의 기여도를 계산한다. 특정 입력 m 의 기여도 $\phi^m(v)$ 는 다음과

같이 정의된다.

$$\phi^m(v) = \sum_{S \subseteq M \setminus \{m\}} \frac{|S|!(|M|-|S|-1)!}{|M|!} [v(S \cup \{m\}) - v(s)] \quad (1)$$

여기서 M은 모든 입력의 집합을 나타낸다. S는 특정 입력 m을 제외한 M의 부분 집합이며, $v(S)$는 S에 대한 모델의 출력이다. 기여도 $\phi^m(v)$는 S에 대한 m의 유무를 고려하여, m의 영향에 대한 가중 평균을 의미한다.

이를 활용하여 Con-RAG는 LLM이 생성한 답변 $\hat{a}_i$에 대한 질문 q_i와 검색된 정보 D_i 각각의 기여도 ϕ_i^q, ϕ_i^D를 계산한다. 모든 입력의 집합을 $M = \{q_i, D_i\}$으로 정의하고, M에 대한 부분 집합 $\emptyset, \{q_i\}, \{D_i\}, \{q_i, D_i\}$을 구성한다. $v(\cdot)$는 $\hat{a}_i$의 정답 토큰에 대한 확률의 합을 나타내며, $\hat{a}_i$에 대한 두 기여도 ϕ_i^q, ϕ_i^D를 다음과 같이 정의한다.

$$\phi_i^q := \phi_i^q(v) = \frac{1}{2}[\big(v(\{q_i, D_i\}) - v(\{D_i\})\big) + (v(\{q_i\}) - v(\emptyset))] \quad (2)$$

$$\phi_i^D := \phi_i^D(v) = \frac{1}{2}[\big(v(\{q_i, D_i\}) - v(\{q_i\})\big) + (v(\{D_i\}) - v(\emptyset))] \quad (3)$$

ϕ_i^q는 D_i에 대하여 q_i의 추가적인 영향 $v(\{q_i^*, D_i\}) - v(\{D_i\})$과 q_i의 독립적인 영향 $v(\{q_i^*\}) - v(\emptyset)$ 간의 평균을 의미한다. 이와 유사하게, ϕ_i^D는 D_i의 추가적인 영향과 독립적인 영향의 평균으로 계산한다.

두 기여도 ϕ_i^q와 ϕ_i^D 간의 비교를 기반으로 정답을 맞힌 질문 집합 Q_c에 대한 검색 필요성을 판단한다. $\phi_i^D \geq \phi_i^q$인 경우, 생성한 답변에 대해 검색된 정보의 기여도가 높으므로 Q_{ex}로 구분한다. 반면 $\phi_i^D < \phi_i^q$인 경우, 검색된 정보보다 LLM의 내재된 지식을 활용하므로 Q_{in}로 구분한다. 두 질문 집합 Q_{ex}, Q_{in}은 다음과 같다.

$$Q_{ex} = \{q_i | q_i \in Q_c \; s.t \; \phi_i^D \geq \phi_i^q\} \quad (4)$$

$$Q_{in} = \{q_i | q_i \in Q_c \; s.t \; \phi_i^D < \phi_i^q\} \quad (5)$$

이처럼 기여도 비교 기반 검색 필요성 판단은 검색이 필요한 질문에 대해서만 검색을 수행함으로써 불필요한 검색 문제를 해결하여 효율적으로 검색을 수행할 수 있다.

2.2 검색 필요성 분류기 학습

검색 필요성 판단을 위한 분류기를 학습하기 위해 질문 집합 Q_{ex}와 Q_{in}을 사용하여 정답 레이블 y_i를 다음과 같이 정의된다.

$$y_i = \begin{cases} 1, & if \; q_i^* \in Q_{ex} \\ 0, & if \; q_i^* \in Q_{in} \end{cases} \quad (6)$$

분류기 학습에 사용하는 손실 함수는 이진 크로스 엔트로피(Binary Cross Entropy, BCE)이며, 이는 분류기의 예측 오류를 최소화하는 방향으로 학습된다. q_i에 대한 검색 필요성 분류기의 출력 $\hat{y}_i$에 대하여, 손실 함수의 식은 다음과 같다. 손실 함수의 식은 다음과 같다.

$$Loss_{BCE} = -\frac{1}{|Q_c|}\sum_{k=1}^{|Q_c|}(y_k \log(\hat{y}_i) + (1-y_k)\log(1-\hat{y}_i)) \quad (7)$$

이에 따라, 검색 필요성 분류기는 Q_c에 대하여 검색이 필요한 질문 집합 Q_{ex}과 검색이 불필요한 질문 집합 Q_{in}으로 분류함으로써 검색 필요성을 판단한다.

2.3 검색 필요성 판단 분류기 기반 답변 생성

추론 시, 학습된 검색 필요성 분류기를 사용하여 질문 q_t에 대해 검색 필요성을 판단한다. 이를 기반으로, 검색 필요성에 따른 LLM의 답변 생성 방식은 다음과 같다.

$$\hat{a}_t = \begin{cases} LLM(q_t, D_t), & if \; \hat{y}_t = 1 \\ LLM(q_t), & if \; \hat{y}_t = 0 \end{cases} \quad (6)$$

이와 같이 검색 필요성을 기준으로 검색을 수행함으로써 효율적인 답변 생성이 가능해진다. 또한, 불필요하게 검색된 정보로 인한 추가적인 연산을 방지하여 답변 생성 과정에서의 연산 효율성을 향상시킬 수 있다.

Ⅲ. 구현

3.1 데이터셋 및 평가지표

본 논문에서는 실험을 위해 질의응답 데이터셋으로 다양한 유형의 질문을 포괄적으로 다루는 Natural Questions와 TriviaQA를 사용한다. Natural Questions는 Google 검색 엔진을 통해 수집된 질문으로 구성되어 있으며, 최대 5개의 토큰으로 이루어진 간단한 답변을 포함한다. TriviaQA는 퀴즈 및 퀴즈 리그 웹사이트에서 수집된 질문으로 이루어져 있다. 실험에서는 모든 데이터셋에 대해 질문과 답변 쌍만 사용하고, 독해 작업을 위한 참고 문서는

제외하여 사용한다[9]. 성능 비교를 위한 평가 지표로는 정확도를 사용하여 LLM이 생성한 답변이 정답을 포함하는지를 평가한다[4].

3.2 실험 환경

질문과 관련된 정보를 검색하기 위해 txtai [1] 라이브러리를 이용하여 사전에 인덱싱된 위키피디아 코퍼스 기반 데이터베이스에서 검색(k=5)을 수행한다. Shapley value를 활용하기 위해 Captum [2] 라이브러리를 사용하며, 답변 생성 모델로 Llama-2-Chat(7B)[10]을 사용한다. 검색 필요성을 판단하는 분류기로는 T5-small[11]을 사용하고, HuggingFace [3] 라이브러리에서 제공하는 사전 학습된 가중치를 활용한다. 분류기 학습을 위한 optimizer로는 Adam을 사용하며, 학습률은 5e-5로 지정한다. 모든 실험은 단일 NVIDIA RTX 3090 환경에서 진행한다.

3.3 실험 결과

Model	Natural Questions	TriviaQA
Self-RAG	33.60	57.00
Adaptive-RAG	**44.60**	<u>58.20</u>
Con-RAG (Ours)	<u>38.34</u>	**61.89**

표 1 답변 생성 정확도 비교

제안하는 Con-RAG와 기존 방법론에 대해 검색 필요성 판단을 통한 답변 생성 정확도를 비교한 결과는 표 1과 같다. Natural Questions에서 Con-RAG는 Self-RAG와 비교하여 4.74%p의 성능 향상을 보인다. 다중 검색 기반의 Adaptive-RAG와 달리 제안하는 Con-RAG는 기여도 기반 검색 필요성 분류기를 통해 단일 검색만을 활용한다. 또한, TriviaQA에서 Con-RAG는 Self-RAG 대비 4.89%p, Adaptive-RAG 대비 3.69%p 정확도 향상으로 가장 우수한 성능을 보인다. 이를 통해 제안하는 Con-RAG는 기존 방법론들과 비교하여 효과적인 검색 필요성 판단을 수행함을 입증한다.

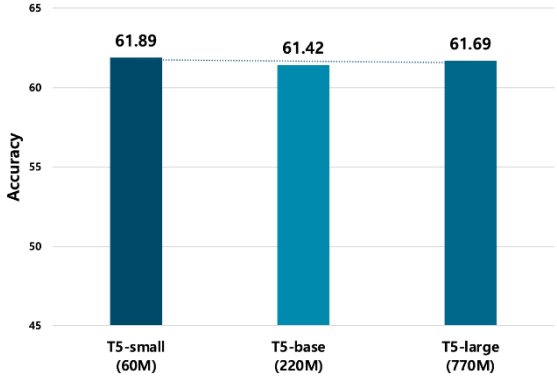

그림 2 학습 파라미터 수에 따른 정확도 성능 비교

제안하는 Con-RAG의 검색 필요성 분류기의 학습 파라미터 수에 따른 성능 비교를 위해, 다양한 크기의 T5를 사용하여 TriviaQA에 대한 정확도를 비교한 결과는 그림 2와 같다. 실험을 위한 T5-small, T5-base, T5-large의 각 학습 파라미터 수는 60M, 220M, 770M이다. T5-small은 적은 수의 학습 파라미터를 사용하면서도 T5-large와 T5-base 대비 각각 0.20%p, 0.47%p의 성능 향상을 보인다.

Ⅳ. 결론 및 향후 연구 방향

본 논문은 기여도를 기반으로 검색 필요성을 판단하는 Con-RAG를 제안한다. Con-RAG는 Shapley value를 활용한 기여도 평가를 통해 검색 필요성을 판단하고, 적은 수의 학습 파라미터를 지닌 분류기를 학습함으로써 불필요한 검색 및 높은 연산량 문제를 해결할 수 있다. 질의응답 태스크에 대한 실험 결과, 제안하는 Con-RAG는 Natural Questions와 TriviaQA에 대해 각각 기존 방법론 Self-RAG 대비 4.89%p, 4.74%p 향상된 성능을 보인다. 또한 Adaptive-RAG와 비교하여 TriviaQA에서 3.69%p 향상된 성능을 보인다. 이를 통해 Con-RAG가 기존 방법론의 학습에 필요한 높은 연산량 문제를 해결하고, 검색 필요성 판단을 기반으로 불필요한 검색 문제를 해결함으로써 효율적인 검색을 수행함을 입증한다.

본 논문에서는 기여도를 기반으로 검색 필요성을 판단하는 Con-RAG의 장점을 확인할 수 있다. 향후 연구에서는 LLM에 내재된 지식의 범위와 특성에 따라 모델의 답변 성능이 달라질 수 있음을 고려하여

[1] https://neuml.github.io/txtai/

[2] https://captum.ai/

[3] https://huggingface.co/

다양한 도메인 분야에서 강건성을 향상시키기 위한 방식을 설계하고자 한다.

Acknowledgements

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원(NRF-2022R1C1C1008534)과 정보통신기획평가원의 지원 (2021-0-01341, 인공지능대학원지원(중앙대학교))의 지원을 받아 수행된 연구임.

참고문헌

[1] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., and Amodei, D., "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877-1901, 2020.

[2] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., and Kiela, D., "Retrieval-augmented generation for knowledge-intensive NLP tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459-9474, 2020.

[3] Asai, A., Wu, Z., Wang, Y., Sil, A., and Hajishirzi, H., "Self-RAG: Learning to retrieve, generate, and critique through self-reflection," in *International Conference on Learning Representations*, 2023.

[4] Jeong, S., Baek, J., Cho, S., Hwang, S. J., and Park, J. C., "Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity," in *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2024.

[5] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., and Zhou, D., "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24824-24837, 2022.

[6] Rozemberczki, B., Watson, L., Bayer, P., Yang, H. T., Kiss, O., Nilsson, S., and Sarkar, R., "The Shapley value in machine learning," in *International Joint Conferences on Artificial Intelligence*, pp. 5572-5579, 2022.

[7] Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., and Petrov, S., "Natural questions: A benchmark for question answering research," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 453-466, 2019.

[8] Joshi, M., Choi, E., Weld, D. S., and Zettlemoyer, L., "TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension," in *Association for Computational Linguistics*, pp. 1601-1611, 2017.

[9] Min, S., Chen, D., Zettlemoyer, L., and Hajishirzi, H., "Knowledge guided text retrieval and reading for open domain question answering," *arXiv preprint arXiv:1911.03868*, 2019.

[10] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., and Scialom, T., "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[11] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., and Liu, P. J., "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1-67, 2020.