맥락 품질을 고려한 검색 증강 생성: 한국어 기반 자동화 팩트체킹을 중심으로

송선영[1], 윤예준[1]⁰, 박건우[2,1]
[1]숭실대학교 지능형반도체학과, [2]숭실대학교 AI 융합학부
kismk8743@gmail.com, yeayen789@gmail.com, kunwoo.park@ssu.ac.kr

Retrieval-Augmented Generation Considering Context Quality: Focusing on Automated Fact-Checking in Korean

Seonyeong Song[1], Yejun Yoon[1]⁰, Kunwoo Park[2,1]
[1]Department of Intelligent Semiconductors, Soongsil University
[2]School of AI Convergence, Soongsil University

요 약

검색 증강 생성은 언어 모델이 학습하지 않은 지식을 참조하여 추론할 수 있도록 하는 데이터 중심 방법으로 지식 집약적 작업에 효과적이다. 하지만, 관련성이 떨어지는 정보가 맥락으로 제공될 경우 추론 성능에 악영향을 줄 수 있다. 이 연구는 한국어 기반의 팩트체킹을 중심으로, 맥락으로 제공된 근거의 품질을 고려하여 추론할 수 있도록 하는 프롬프트 및 지시문 조정 방법을 제안한다. 사실 검증에 있어 근거의 관련성을 언어 모델 추론 과정에서 고려하도록 한 제안 방법은 베이스라인 검색 증강 생성방법 대비 macro f1 기준 최대 0.161 향상된 성능을 보였다. 이 연구는 굳건한 팩트체킹 및 검색 증강 생성 시스템 구축에 있어 맥락 및 근거 품질을 고려하는 것이 중요함을 시사한다.

1. 서 론

웹과 인터넷 등 정보 기술의 발전으로 온라인 환경에서 많은 정보를 쉽고 빠르게 공유할 수 있게 되었다. 하지만, 진위성 검증 절차의 부재 또는 부족으로 거짓 정보가 범람하게 되며, 정보의 정확성을 검증하는 팩트체킹의 중요성이 커지고 있다. 전문가 팩트체커가 정보의 진위성을 판단하는 방법은 정확도가 높지만 끊임없이 생성되는 수많은 정보를 모두 검증하기에는 확장성이 낮다. 이에 따라, 정보 진위성 자동 검증을 다루는 자동화 팩트체킹이 활발히 연구되었다. 신뢰할 수 있는 외부 지식 데이터베이스로부터 주장과 관련된 문서를 검색한 후, 그 중 주장을 뒷받침할 근거 문장을 찾아 주장의 진위성을 판단하는 단계별 과업으로 다루어진다. 문제의 중요성에도 불구하고, 검증 대상 주장과 근거가 한국어로 작성된 경우에 대해서는 연구가 부족하다.

이 연구는 한국어 기반 팩트체킹을 위한 언어 모델의 검색 증강 생성방법을 다루며, 주장 검증에 있어 맥락으로 제공된 근거의 관련성을 고려하여 검증 성능을 높이는 방법을 연구한다. 검색 증강 생성은 언어 모델이 학습하지 않은 정보를 검색해 참조하는 데이터 중심 방법이다. 팩트체킹 등 지식 집약적 과업에서 높은 성능을 보이는 것으로 알려져 있으나[1], 추론 시 관련성이 떨어지는 맥락 정보가 제공될 경우 언어 모델 추론 성능에 악영향을 줄 수 있다는 한계[2]가 있다. 이를 보완하기 위해, 이 연구는 언어 모델이 근거의 관련성을 고려하여 진위성을 판단할 수 있도록 지시문 조정 학습하여 팩트체킹 파이프라인의 예측 성능을 높이는 방법을 제시한다. 구체적으로, 각 검색 근거에 대해 주장과의 관계에 대한 다른 모델의 추론 결과를 프롬프트에 증강하는 방법, 언어 모델에게 직접 근거 관련성을 판단하도록 하는 프롬프트 방법을 제시한다. 한국어 팩트체킹 데이터셋에 대한 성능 평가 실험에서 일반적인 검색 증강 생성방법 대비 macro f1 기준 최대 0.161 향상된 결과를 관측하였으며, 검색 증강 생성 및 팩트체킹 시스템 구축에 있어 근거 관련성을 고려하는 것이 중요함을 시사한다.

2. 관련 연구

2.1. 검색 증강 생성

언어 모델의 매개변수에 내재된 지식을 바탕으로 요약, 기계 번역 등의 과업을 라벨 데이터 없이 수행할 수 있음을 보였다. 그러나, 팩트체킹과 같은 지식 집약적 과업을 수행하기 위해서 외부 지식을 검색하여 활용하는 것이 필요하다. 초기에는 REALM[1]과 RAG[3]와 같은 검색 증강 언어 모델이 연구되었으며, 검색기와 생성기를 종단간 방식으로 학습해 소수샷(few-shot) 추론 과업에서 우수한 성능을 보인

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 지역지능화혁신인재양성사업임(IITP-2024-RS-2022-00156360)

ATLAS[4], 블랙박스 언어 모델을 사용하며 검색기를 조정하는 방법을 제안한 REPLUG[5] 등이 연구되었다. 최근에는 맥락으로 활용하는 문서의 품질을 고려하는 시도들이 이루어지고 있으며[6,7], 이 연구에서 제안하는 방법은 이에 속한다. 최근 연구는 한국어 데이터셋을 대상으로 언어 모델 제로 샷 추론과 검색 증강 생성 방법의 성능을 비교하였다[8].

2.2. 자동화 팩트체킹

대표적인 자동화 팩트체킹 연구로 FEVER[9]가 있으며, 주어진 주장에 대해 위키피디아 문서 집합으로부터 근거 후보 문장을 검색하고, 이를 바탕으로 참, 거짓, 판단 불가로 예측하는 것을 목표한다. 이와 유사한 방법으로 생성된 한국어 팩트체킹 데이터셋이 제안되었다[10]. 한 연구에서는 그래프 기반의 증거 추론 신경망을 구축해 질의 응답에 대한 사실 여부 검증 방법을 제안하였으며[11], 계층적 구조 어텐션과 그래프 신경망 추론을 결합한 자동화 팩트체킹 모델을 제시하였다[12]. 최근 연구에서는, 자동화 팩트체킹을 위한 대조학습 기반 임베딩 검색 기술을 제안하였다[13]. 이 연구는 한국어 기반 팩트체킹 연구와 비교하여, 검색 증강 생성 방법을 다루며 맥락 품질 향상을 위한 생성기 학습에 집중하였다는 점에서 차별점을 지닌다.

3. 문제 및 데이터셋

3.1. 문제 정의

한국어로 작성된 주장에 대한 진위성 예측을 목표하며, 지식 저장소로부터 관련 문장을 검색하여 진위 여부를 예측하는 두 단계의 세부 과업으로 나뉜다.

(1) 근거 검색 (evidence retrieval): 문서 집합으로부터 주장검증에 관련된 상위 k개의 근거 후보 문장을 검색한다.

(2) 평결 예측 (verdict prediction): 주장과 근거 후보 문장을 활용해 주장의 진위 여부를 예측한다. 예측 목표 변수는 지지되는지 (supported), 부인되는지 (refuted), 충분한 정보가 없는지 (not enough information) 세 가지 레이블로 구성된다.

3.2. 데이터셋

79,783건의 데이터로 구성된 팩트체크 데이터셋[14]을 사용한다. 해당 데이터셋은 시청자미디어재단의 지원을 받아 서울대학교 연구팀이 구축한 데이터이다. 학습, 검증, 테스트 데이터는 각각 69,788건, 4,998건, 4,997건으로 구성된다. 총 1,866,795 건의 한국어 위키피디아 문서를 근거 검색을 위한 지식 저장소로 사용한다.

4. 연구 방법

4.1. 근거 검색

위키피디아 문서들을 약 180만개의 문장으로 분할하여 검색 대상 데이터로 사용한다. KorQuAD v1.0[1]에 미세조정된 KLUE-RoBERTa-base[2] 기반 한국어 DPR[15] 모델을 미세조정 후 임베딩 기반 FAISS 인덱스를 구축하고, 최대 내적 검색을 수행해 상위 5개의 문장을 근거 후보 문장으로 선택한다. 이 연구는 근거 품질을 고려하는 평결 예측 방법의 효과를 검증하는 것을 목표하며, 근거 검색 방법을 위와 같이 고정하여 실험한다.

4.2. 평결 예측

검색된 근거에 기반해 주장의 사실 여부를 판단하는 평결 예측 모델을 구축한다. gemma-7b-it[16] 모델을 초기 언어 모델로 사용하며, 주장과 5개의 근거 후보 문장을 입력으로 받아 사실 여부를 예측하도록 QLoRA[17] 기반 지시문 조정 학습을 수행한다. 프롬프트 형태에 따라 세 가지 방법으로 나뉜다.

(1) 기본 프롬프트 (그림 1): 주장과 검색된 5개의 근거후보 문장을 바탕으로 사실 여부를 예측하는 베이스라인 방법이다.

주장과 관련된 정보들이 주어집니다.
주어진 관련 정보들을 참고하여 주장의 사실여부를 답변해주세요.
주장의 사실여부는 '참', '거짓', '판단불가' 중 하나로 판단해주세요.

주장: *{주장}*

관련문장 1: *{검색된 관련문장 1}*
관련문장 2: *{검색된 관련문장 2}*
관련문장 3: *{검색된 관련문장 3}*
관련문장 4: *{검색된 관련문장 4}*
관련문장 5: *{검색된 관련문장 5}*

최종적으로, 주장의 사실 여부는 **{예측 라벨}** 입니다.

그림 1 기본 프롬프트. 기울임 초록색 텍스트는 주장과 맥락 내 샘플로 사용된 검색 문장, 굵은 빨간색 텍스트는 예측 라벨을 나타낸다.

실제 환경에서 검색된 5개의 근거 후보 문장들에는 주장과 관련되지 않은 문장이 포함되어 있을 수 있으며, 이들이 맥락 내 샘플로 사용될 경우 언어 모델의 평결 예측을 위한 추론 과정을 방해할 수 있다. 따라서, 이 연구는 주장에 대한 근거 관련성을 고려하는 두 가지 방법 (2-3)을 도입한다.

[1] https://korquad.github.io/category/1.0_KOR.html
[2] https://huggingface.co/klue/roberta-base

(2) 근거 관련성 라벨 증강 프롬프트 (그림 2): 주장으로부터 검색된 5개의 문장에 대해 도움을 주는지 여부를 다른 언어 모델로 판단하여 프롬프트 추가 맥락으로 주입한다. 이를 통해 관련성이 낮은 근거를 추론 시 고려하지 않도록 유도한다. SAIL[18]에서 착안한 방법이다.

주장과 관련된 정보들이 주어집니다.
주어진 관련 정보들을 참고하여 주장의 사실여부를 답변해주세요.
주장의 사실여부는 '참', '거짓', '판단불가' 중 하나로 판단해주세요.

주장: *{주장}*

관련문장 1: *{검색된 관련문장 1}*
관련문장 2: *{검색된 관련문장 2}*
관련문장 3: *{검색된 관련문장 3}*
관련문장 4: *{검색된 관련문장 4}*
관련문장 5: *{검색된 관련문장 5}*

관련문장 1은 **{유용한/방해되는}** 정보입니다. 관련문장 2는 **{유용한/방해되는}** 정보입니다. 관련문장 3은 **{유용한/방해되는}** 정보입니다. 관련문장 4는 **{유용한/방해되는}** 정보입니다. 관련정보 5는 **{유용한/방해되는}** 정보입니다.

최종적으로, 주장의 사실 여부는 **{예측 라벨}** 입니다.

그림 2. 근거 관련성 라벨 증강 프롬프트. 기울임 초록색 텍스트는 주장과 맥락 내 샘플로 사용된 검색 문장, 굵은 파란색 텍스트는 다른 모델로부터 예측된 관련성 라벨, 굵은 빨간색 텍스트는 예측 라벨을 나타낸다.

(3) 근거 관련성 자기 판단 프롬프트: 검색된 문장 중 관련 없는 문장을 언어 모델이 추론 과정에서 스스로 무시할 수 있도록 유도한다. 영문 기반의 수학 문제 풀이에 적용된 프롬프트[2]에서 착안하였다.

주장과 관련된 정보들이 주어집니다.
주어진 관련 정보들을 참고하여 주장의 사실여부를 답변해주세요.
주장의 사실여부는 '참', '거짓', '판단불가' 중 하나로 판단해주세요.
관련문장들 중 주장과 관련 없다고 판단되는 문장은 무시하세요.

주장: *{주장}*

관련문장 1: *{검색된 관련문장 1}*
관련문장 2: *{검색된 관련문장 2}*
관련문장 3: *{검색된 관련문장 3}*
관련문장 4: *{검색된 관련문장 4}*
관련문장 5: *{검색된 관련문장 5}*

최종적으로, 주장의 사실 여부는 **{예측 라벨}** 입니다.

그림 3. 근거 관련성 자기 판단 프롬프트. 기울임 초록색 텍스트는 주장과 맥락 내 샘플로 사용된 검색 문장, 굵은 빨간색 텍스트는 예측 라벨, 노란색 강조는 추가된 프롬프트를 의미한다.

5. 성능 평가 실험

5.1. 실험 세팅

DPR 검색모델 미세 조정 시 배치 크기는 128, 학습률은 1e-5, 옵티마이저는 Adam, 최대 epoch는 40으로 설정했다. 지시문 조정 시 배치 크기는 18, 학습률은 2e-4, 옵티마이저는 AdamW, 최대 epoch는 3으로 설정하였다. QLoRA는 rank는 6, alpha는 8, dropout은 0.05로 설정했고, "q_proj", "o_proj", "k_proj", "v_proj", "gate_proj", "up_proj", "down_proj" 레이어에 적용했다. 근거 관련성 라벨 증강을 위하 언어 모델로 Solar[19]를 사용했다.

5.2. 실험 결과

표 1. 평결 예측 성능

Model	Macro f1	Accuracy
(1) 기본 프롬프트	0.323	0.402
+ 생각의 사슬	0.322	0.4
(2) 근거 관련성 라벨 증강	0.363	0.45
+ 생각의 사슬	0.36	0.449
(3) 근거 관련성 자기 판단	0.46	0.499
+ 생각의 사슬	0.484	0.513

표 1은 각 프롬프트 및 지시문 조정 방법을 이용한 자동화 팩트체킹 시스템의 평결 예측 성능을 나타낸다. 성능 지표로 3진 분류에 대한 macro f1 및 정확도를 사용하였다. 근거 관련성을 고려하는 두 가지 제안 방법이 기본 프롬프트를 사용하여 지시문 조정 학습한 베이스라인 방법(1) 대비 대폭 개선된 성능을 보였다. 근거의 관련성 라벨 증강 방법(2)은 macro f1 기준 0.04, 정확도 기준 0.048 향상되었다. 근거 관련성 자기 판단 방법(3)은 방법(2) 보다 높은 성능을 보였으며, macro f1 0.46, 정확도 0.449를 기록하였다. 마지막으로, 각 방법에 생각의 사슬(Chain-of-Thought) 프롬프트[20]를 추가해 효과를 검증하였다. "차근차근 생각해보세요" 라는 한국어 프롬프트를 지시문

마지막에 추가하였다. 실험 결과, (1), (2) 방법에 대해서는 생각의 사슬을 적용하였을 때 성능 변화가 미미하였으나, 근거 관련성 자기 판단 방법의 경우 생각의 사슬 프롬프트 적용 시 macro f1 기준 0.016, 정확도 기준 0.014 향상되었다. 위 결과는 근거 관련성을 고려한 언어 모델 프롬프트 및 지시문 튜닝이 효과적인 팩트체킹에 있어 중요함을 보인다.

6. 결론

이 연구는 한국어 기반 자동화 팩트체킹을 중심으로, 맥락으로 제공된 근거의 품질을 고려해 추론하는 프롬프트 및 지시문 조정 방법을 제안하였다. 베이스라인 검색 증강 생성방법 대비 근거의 품질을 고려해 주장의 진위성을 예측하는 언어모델이 macro f1 기준 최대 0.161 향상된 성능을 달성하였다. 이는 검색 증강 생성에 있어 맥락 품질이 중요함을 시사하며, 품질을 고려하는 언어모델 프롬프트 및 지시문 조정을 통해 성능을 높일 수 있음을 보인다.

이 연구는 몇 가지 한계점을 지닌다. 첫째, 하나의 데이터셋 및 언어모델에 대해 실험하여 다양한 모델에 대한 반복 실험이 필요하다. 둘째, 프롬프트 및 지시문 조정 방법의 효과를 검증하기 위해 검색 방법을 고정하였다. 검색 성능을 고도화하여 맥락 품질을 높이는 것도 가능한 연구 방향이다. 셋째, 실제 주장을 검증하는 시스템으로 나아가는 것이 바람직한 방향이나, 위키피디아를 지식 저장소로 사용한 하나의 팩트체킹 데이터셋에 대해 실험하였다. 보다 광범위한 실제 주장을 포괄하는 데이터셋 구축이 필요하다.

참고 문헌

[1] Guu, Kelvin, et al. "Retrieval augmented language model pre-training." International conference on machine learning. PMLR, 2020.

[2] Shi, Freda, et al. "Large language models can be easily distracted by irrelevant context." International Conference on Machine Learning. PMLR, 2023.

[3] Lewis, Patrick, et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks." Advances in Neural Information Processing Systems 33 (2020): 9459-9474.

[4] Izacard, Gautier, et al. "Atlas: Few-shot learning with retrieval augmented language models." Journal of Machine Learning Research 24.251 (2023): 1-43.

[5] Shi, Weijia, et al. "REPLUG: Retrieval-Augmented Black-Box Language Models." In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 2024.

[6] Lin, Xi Victoria, et al. "Ra-dit: Retrieval-augmented dual instruction tuning." The twelfth international conference on learning representations. 2024.

[7] Asai, Akari, et al. "Self-rag: Learning to retrieve, generate, and critique through self-reflection." The twelfth international conference on learning representations. 2024.

[8] 신중민, 박승렬, 김혜린, 이정훈. LLM 답변 향상을 위한 검색 기반 생성 기법: GPT3.5, GPT4의 Zero-shot, RAG 비교 연구. 한국정보통신학회 종합학술대회 논문집. 2023.

[9] Thorne, James, et al. "FEVER: a large-scale dataset for fact extraction and VERification." In Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Sheffield. 2018.

[10] 이종현, 나승훈, 신동욱, 김선훈, 강인호. 한국어 Fact 검증을위한 자동 Claim 데이터 생성. 한국정보과학회 학술발표논문집. 2021.

[11] 박은환, 나승훈, 신동욱, 김선훈, 강인호. 그래프 기반 증거 추론을이용한 질의응답에 대한 사실 여부 검증 연구. 한국정보과학회 학술발표논문집. 2021.

[12] 박은환, 나승훈, 신동욱, 전동`현, 강인호. 사실 확인을 위한 그래프 기반 추론 및 계층적 구조 어텐션 결합 모델. 한국정보과학회 학술발표논문집. 2021.

[13] 송선영, 안제준, 박건우. (2023). A Contrastive Learning Method for Automated Fact-Checking. Journal of KIISE, 50(8), 680-687, 10.5626/JOK.2023.50.8.680

[14] [Online]. Available: https://github.com/hongcheki/factcheck-ko-2021

[15] Karpukhin, Vladimir, et al. "Dense passage retrieval for open-domain question answering." In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6769-6781, Online. Association for Computational Linguistics. 2020.

[16]Team, Gemma, et al. "Gemma: Open models based on gemini research and technology." arXiv preprint arXiv:2403.08295 (2024).

[17] Dettmers, Tim, et al. "Qlora: Efficient finetuning of quantized llms." Advances in Neural Information Processing Systems 36 (2024).

[18] Luo, Hongyin, et al. "Search augmented instruction learning." Findings of the Association for Computational Linguistics: EMNLP 2023. 2023.

[19] Kim, Dahyun, et al. "Solar 10.7 b: Scaling large language models with simple yet effective depth up-scaling." arXiv preprint arXiv:2312.15166 (2023).

[20] Wei, Jason, et al. "Chain-of-thought prompting elicits reasoning in large language models." Advances in neural information processing systems 35 (2022): 24824-24837.