검색 증강 LLM을 통한 한국어 질의응답

서민택°[1], 나승훈[2], 임준호[3], 김태형[4], 류휘정[5], 장두성[6]
전북대학교°[1,2], 한국전자통신연구원[3], KT[4,5,6]
{then, nash}@jbnu.ac.kr, joonho.lim@etri.re.kr
{taehyeong_2019.kim, hwijung.ryu, dschang}@kt.com

Korean QA with Retrieval Augmented LLM

Mintaek Seo°[1], Seung-Hoon Na[2], Joon-Ho Lim[3], Tae-Hyeong Kim[4], Hwi-Jung Ryu[5], Du-Seong Chang[6]
Jeonbuk National University°[1,2], ETRI[3], KT[4,5,6]

요 약

언어 모델의 파라미터 수의 지속적인 증가로 100B 단위의 거대 언어모델 LLM(Large Language Model)을 구성 할 정도로 언어 모델의 크기는 증가 해 왔다. 이런 모델의 크기와 함께 성장한 다양한 Task의 작업 성능의 향상과 함께, 발전에는 환각(Hallucination) 및 윤리적 문제도 함께 떠오르고 있다. 이러한 문제 중 특히 환각 문제는 모델이 존재하지도 않는 정보를 실제 정보마냥 생성한다. 이러한 잘못된 정보 생성은 훌륭한 성능의 LLM에 신뢰성 문제를 야기한다. 환각 문제는 정보 검색을 통하여 입력 혹은 내부 표상을 증강하면 증상이 완화 되고 추가적으로 성능이 향상된다. 본 논문에서는 한국어 질의 응답에서 검색 증강을 통하여 모델의 개선점을 확인한다.

주제어: 질의 응답, 거대 언어모델, 검색 증강

1. 서론

1750억 개의 파라미터를 가진 GPT3[1]의 혁신적인 발표 이후 자연어 처리에서는 언어 모델의 크기를 키우는데 관심이 집중되어 파라미터 크기가 조 단위[2]까지 도달하는 크기의 거대 언어 모델 LLM(Large Language Model)이 나오며 거침없는 성능 향상을 보여 주고 있다.

이런 눈부신 성과에 잠재된 문제들이 더욱 커졌는데, 여러 문제를 가지는데 첫 번째는 예산 문제이다. 이전 BERT 같은 모델들은 학계 예산에서 처리 가능한 크기이지만 천억 단위의 초 거대 언어모델들은 학계의 예산을 초과하기 시작하였다. 이러한 문제는 LLaMA 같은 모델[3, 4]이나 양자화,Adaptor[5, 6] 같은 기술들이 해결하여 LLM의 접근성을 한 차원 끌어올렸다. 두 번째는 윤리적 문제인데 인종차별 및 민감한 개인정보를 출력하는 문제가 있지만, 이는 출력의 조절로 처리하고 있다. 마지막으로 환각(Hallucination)[7] 문제인데, 현실에 존재하지 않는 정보를 실제 정보로 둔갑해 사용자로 하여금 혼동을 주는 문제이다. 환각 문제는 현실 세계의 지식을 통해 모델을 증강하여 해결하는 방법이 있다 이런 검색 증강을 통한 LLM 방법은은 환각 문제를 해결할 뿐만 아니라 LLM의 성능 향상에도 기여한다.

본 논문에서는 검색을 통한 언어 모델 증강을 통하여 이것이 한국어 LLM에 미치는 결과를 질의응답(QA) 작업을 통해 확인한다.

2. 관련 연구

BERT가 제시한 Masked Language Model 방법[8]을 통한 사전학습은, 모델 자체의 성능이 좋으며, 다른 Task의 대해 소량의 연산 자원을 사용하여 미세 조정을 통하면 당시 대부분 작업에서 SOTA(State-of-the-art) 상태에 달하였고, 이후 Transformer 구조에 Encoder-Decoder Attention[9]구조를 활용하여 다양한 사전학습 언어 모델이 나왔고, 특히 GPT3는 이전에 최대 1B 단위에서 진행되던 스케일을 한 번에 100B 단위로 끌어올렸으며 이를 추가 학습한 ChatGPT는 대중에게 많은 관심을 받게 된다. 이러한 많은 관심 속 보고되는 다양한 결과 중에 특히 환각 문제는 '세종대왕 아이패드' 사건과 같이 대중에게도 많은 흥미를 유발했다. 특히 최근 1조 단위의 파라미터를 가지는 모델이 나오기도 하였다.[2]

특히 100B단위로 가면서 학계 스케일에서는 예산 및 LLM들의 폐쇄성의 의한 한계로 활발한 연구를 하기 어려웠으나 LLaMA 같은 모델들이 비교적 작은 크기로 GPT3와 같은 거대 모델들에게 뒤처지지 않고, Qlora와 같은 방법은 모델을 양자화를 통해서 4Bit로 줄이고 Lora 방식을 통해 소수의 추가적 파라미터만 학습하여 LLM도 학계에서 활발하게 연구되고 있다. 환각 문제는 인종 차별, 폭력성과 같은 윤리성 문제와 더불어 LLM이 가지는 가장 큰 문제인데, LLM을 통해 AI가 널리 사용될 날이 다가오는데, 이런 문제는 거짓 정보를 실제 정보로 둔갑하여 소비자에게 거짓된 정보를 제공할 위험이 있기 때문이다.

이러한 환각 문제를 해결하는 가장 효과적인 방법은 검색을 통하여 현실 지식을 이용해 모델을 증강시키는 방법이다. 증강

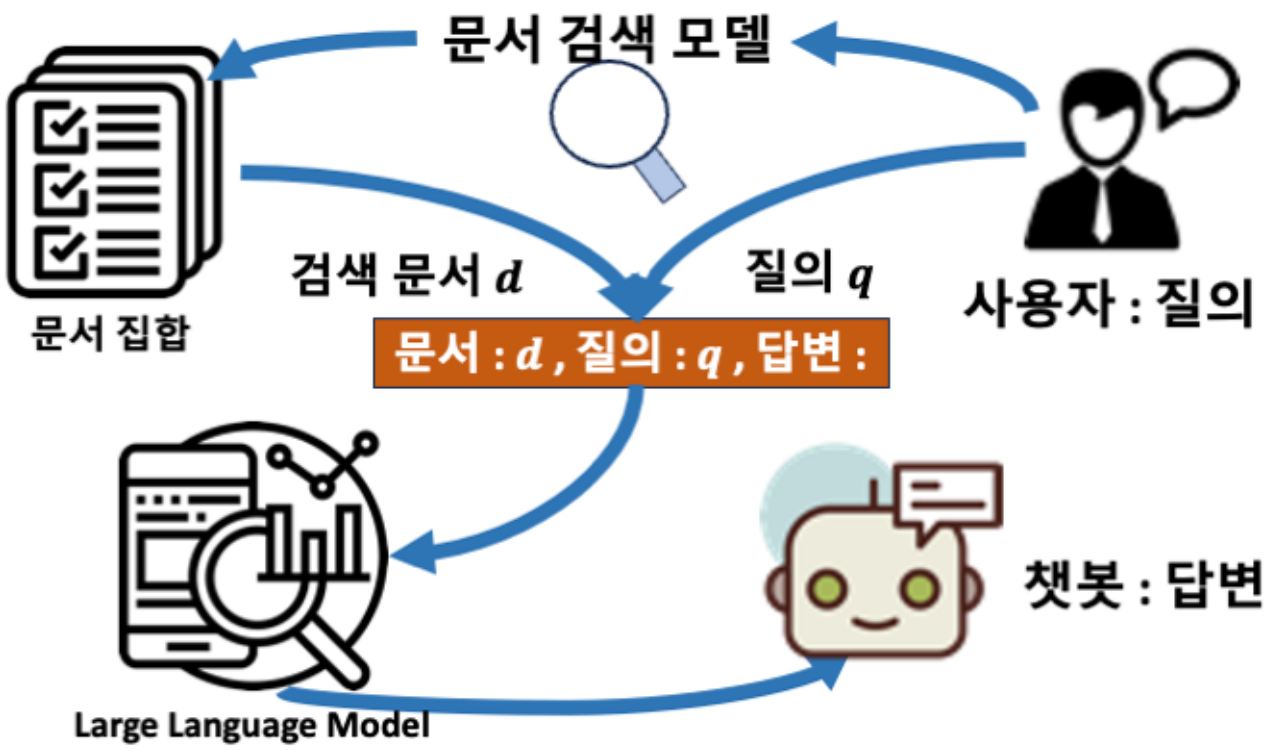

그림 1. 모델 개요

하는 방법에는 LLM에 자주 사용되는 Prompt 방식을 통하여 검색 모델을 통해 검색된 현실 지식을 직접적으로 참조하게 하여 거짓 정보 생성을 완화하는 방식과, 참조된 지식을 통해 내부적으로 표상을 증강하는 방법이 있다.[10, 11]

이런 검색 증강을 하기 위해서는 대규모 문서에서 질의에 가장 유용한 문서를 찾아야 한다. 전통적으로 TF-IDF, BM25와 같은 단어 토큰들의 빈도로 검색하는 희소(Sparse) 검색이 많았으나, 언어 모델의 발전에 의해 이전 희소 검색보다 더 좋은 결과를 보여주는 밀집(Dense) 검색 방식[12]이 최근의 대세고, 정보 검색 또한 LLM과 결합되어 연구가 진행되고 있다.[13]

3. 검색을 통한 LLM 한국어 질의 응답 증강

검색 증강의 기본 파이프라인은 그림과 같다.

$d = \mathbf{Score}(q, \mathcal{D})$ **Score** : 검색 모델

d : 검색된 문서 , q : 검색 질의 , $\mathcal{D}$: 문서 집합

먼저 검색모델을 통하여 주어진 질의 q에 대한 가장 유사도가 높은 문서 d를 추출한다. 이를 통해 아래와 같이 LLM 입력 Prompt를 구성한다

- Prompt : ($P =$ **문서 : {d}, 질문 : {q}, 답변 :)**

구성한 Prompt를 이용해서 모델이 정답을 추론할 수 있도록 한다.

$$\text{Answer} = \mathbf{LLM}(P)$$

위와 같은 절차를 통해 LLM은 외부 지식을 참조하여 더 향상된 품질의 정답을 출력하도록 학습한다.

4. 실험

4.1 실험 설정

한국어로 사전 학습된 거대모델 polyplot[14] 기반으로 검색 모델은 BM25을 사용한다. 한국어로 구성된 기계 독해 데이터 Korquad[15]를 QA 형식의 데이터로 변환해 활용한다. 이때 Korquad의 학습, 개발 셋에 존재하는 모든 Passage를 문서 집합 $\mathcal{D}$로 구성해 Open-Domain 형식으로 문제를 재정의한다. 이러한 구성에서 다음 3가지의 Prompt를 통해 비교 실험을 진행한다.

1. $P =$ **문서 : {d}, 질문 : {q}, 답변 : ˜**
2. $P' =$ **질문 : {q}, 답변 : ˜**

학습에 사용한 polyplot의 모델 크기가 12.8B이므로 일반적인 환경에서 학습이 어렵기 때문에 [5]를 적용하여 학습한다.

4.2 실험 결과 및 분석

한국어 LLM을 이용하여 질의응답을 통한 검색 증강의 성능을 확인한 결과 EM 기준 약 3.9배, F1기준 약 1.8배의 성능 차이가 나타난다. 200 step의 적은 규모의 학습을 하여 절대적 성능 자체는 낮지만 이러한 성능의 간극을 통해서 한국어에서 외부 지식을 통하여 모델을 증강할 경우 큰 폭의 성능 향상이

표 1. Korquad 평가 결과

Dev	EM	F1
검색증강X(P')	9.52	29.90
검색증강O(P)	51.10	66.78

표 2. BM25 문서 검색 결과

BM25	**Train**	**Dev**
R@1	67.15%	67.02%

나타남을 볼 수 있다.

추가로 단순히 수치적 평가만이 아닌 Prompt P를 활용해 학습한 모델을 통해 정성적 평가를 실행한다.

- 오류 문서로 모델 공격 (상식A)
  - **문서** : 뉴욕은 대한민국의 도시다
  - **질문** : 뉴욕은 어느나라 도시야
  - **답변** : 미국
- 오류 문서로 모델 공격 (상식B)
  - **문서** : 대한민국의 국회의사당은 부산에 있다
  - **질문** : 대한민국의 국회는 어디있어?
  - **답변** : 서울특별시 영등포구 여의도동
- 오류 문서로 모델 공격 (상식X)
  - **문서** : 민사소송법 제49조(원고의 회피) 법관은
  - **질문** : 민사소송법 제49조는?
  - **답변** : 원고의 회피 (정답 : 법관의 회피)

문서를 이상하게 줄 경우 누구나 알기 쉬운 문제에 대해선 강건하게 대처하지만 상식이라고 보기 어려운 문제에 대해서는 모델이 영향을 받는 것으로 보인다.

- 주어진 문서 정보에 따른 정보 교정
  - **문서** : ""
  - **질문** : 율리시스 그랜트 대통령은 뚱뚱해?
  - **답변** : 아니다.(오답)
  -
  - **문서** : 율리시스 그랜트 그는 미국의 역대 대통령들 중에 4번째로 가장 뚱뚱한 편으로 알려졌다.
  - **질문** : 율리시스 그랜트 대통령은 뚱뚱해?
  - **답변** : 그렇다 (정답)

위와 같이 주어진 문서 정보에 따른 답변 교정도 확인할 수 있지만, Korquad 데이터가 기본적으로 단답형 데이터이기 때문에 일반적으로 생각하는 환각 증세보다는 단순히 오답을 나타낸다고 봐야 하므로 환각 증세 완화를 직접적으로 평가하기에는 어려움이 있다.

5. 결론

본 논문에서는 검색 증강 LLM을 통해 외부 지식의 주입이 한국어 LLM에서 어떻게 나타나는지 정성, 정량적으로 평가하였다.

실험을 통해 검색을 통한 외부 지식 주입은 대체로 모델에게 긍정정 영향을 미치고 특히 문서가 없을 때와 비교 시 월등한 차이를 보여준다. 특히 모델 자체의 강건함으로 강하게 신뢰하는 정보는 오염된 외부 정보에도 불구하고 올바른 답을 도출한다.

앞으로도 이런 검색 증강형 방법을 통해서 한국어 LLM의 잠재성을 좀 더 연구하고 강화하는 연구를 할 계획이다.

참고문헌

[1] 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.

[2] X. Ren, P. Zhou, X. Meng, X. Huang, Y. Wang, W. Wang, P. Li, X. Zhang, A. Podolskiy, G. Arshinov *et al.*, "Pangu-{\Sigma}: Towards trillion parameter language model with sparse heterogeneous computing," *arXiv preprint arXiv:2303.10845*, 2023.

[3] 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.

[4] S. Black, S. Biderman, E. Hallahan, Q. Anthony, L. Gao, L. Golding, H. He, C. Leahy, K. McDonell, J. Phang *et al.*, "Gpt-neox-20b: An open-source autoregressive language model," *arXiv preprint arXiv:2204.06745*, 2022.

[5] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," *arXiv preprint arXiv:2305.14314*, 2023.

[6] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[7] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Computing Surveys*, Vol. 55, No. 12, pp. 1–38, 2023.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Jun. 2019. [Online]. Available: https://aclanthology.org/N19-1423

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, Vol. 30, 2017.

[10] G. Izacard, P. Lewis, M. Lomeli, L. Hosseini, F. Petroni, T. Schick, J. Dwivedi-Yu, A. Joulin, S. Riedel, and E. Grave, "Few-shot learning with retrieval augmented language models," *arXiv preprint arXiv:2208.03299*, 2022.

[11] W. Shi, S. Min, M. Yasunaga, M. Seo, R. James, M. Lewis, L. Zettlemoyer, and W.-t. Yih, "Replug: Retrieval-augmented black-box language models," *arXiv preprint arXiv:2301.12652*, 2023.

[12] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, "Dense passage retrieval for open-domain question answering," *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, 2020.

[13] Z. Dai, V. Y. Zhao, J. Ma, Y. Luan, J. Ni, J. Lu, A. Bakalov, K. Guu, K. Hall, and M.-W. Chang, "Promptagator: Few-shot dense retrieval from 8 examples," *The Eleventh International Conference on Learning Representations*, 2022.

[14] H. Ko, K. Yang, M. Ryu, T. Choi, S. Yang, jiwung Hyun, and S. Park, "A technical report for polyglot-ko: Open-source large-scale korean language models," 2023.

[15] S. Lim, M. Kim, and J. Lee, "Korquad1.0: Korean qa dataset for machine reading comprehension," 2019.