In-Context 검색 증강형 한국어 언어 모델

이성민[1], 이정[2], 서대룡[3], 전동현[4], 강인호[5], 나승훈*[6]
전북대학교[1,2,*6], 네이버[3,4,5]
{cap1232, fhqlatm, nash}@jbnu.ac.kr, {daeryong.seo, donghyeon.jeon, once.ihkang}@navercorp.com

In-Context Retrieval-Augmented Korean Language Model

Sung-Min Lee[1], Joung Lee[2], Daeryong Seo[3], Donghyeon Jeon[4], Inho Kang[5], Seung-Hoon Na*[6]
Jeonbuk National University[1,2,*6], NAVER Corporation[3,4,5]

요 약

검색 증강형 언어 모델은 입력과 연관된 문서들을 검색하고 텍스트 생성 과정에 통합하여 언어 모델의 생성 능력을 강화한다. 본 논문에서는 사전 학습된 대규모 언어 모델의 추가적인 학습 없이 In-Context 검색 증강으로 한국어 언어 모델의 생성 능력을 강화하고 기존 언어 모델 대비 성능이 증가함을 보인다. 특히 다양한 크기의 사전 학습된 언어 모델을 활용하여 검색 증강 결과를 보여 모든 규모의 사전 학습 모델에서 Perplexity가 크게 개선된 결과를 확인하였다. 또한 오픈 도메인 질의응답(Open-Domain Question Answering) 과업에서도 EM-19, F1-27.8 향상된 결과를 보여 In-Context 검색 증강형 언어 모델의 성능을 입증한다.

주제어: Retrieval-augmented, RALM, In-Context, 검색 증강, 언어 모델

1. 서론

최근 자연어 처리 분야에서 비약적인 성장을 이루고 있는 대규모 언어 모델(Large Language Model, LLM)은 인간의 언어를 효과적으로 이해하기 위하여 방대한 양의 데이터와 매개변수(parameter)를 기반으로 학습된다. 대표적으로 최근 공개되어 큰 이목을 끈 ChatGPT, Bard 및 CLOVA-X와 같은 대화형 서비스의 기반 언어 모델을 예시로 들 수 있다. 이들은 각각 GPT-4[1], LaMDA[2], HYPERCLOVA[3]과 같은 초거대 규모의 생성형 언어 모델(Generative Language Model)을 기반으로 학습되었으며 기존 데이터로부터 학습하지 못한 과업에 대해서도 뛰어난 성능을 보여 대규모 언어 모델의 생성 능력을 입증한다.

이러한 초거대 언어 모델을 특정 과업에 맞추어 Fine-tuning (미세 조정)하는 것은 모델의 모든 매개변수를 저장하여 메모리 부담이 증대될 뿐만 아니라 학습에 필요한 GPU 자원 또한 크게 요구되는 단점이 존재한다. 이에 따라 추가적인 학습 없이 사전 학습된 모델을 범용적으로 활용하는 연구가 최근 연구 동향으로 자리 잡았다. 대표적으로 GPT-3[4]는 모든 사전 학습된 대규모 언어 모델이 입력 프롬프트를 통해 내재된 매개변수로부터 In-Context 학습을 수행할 수 있음을 보이며, Chain-of-Thought[5]는 여러 단계로 구분된 프롬프트를 모델에 입력하여 생성되는 연쇄적인 사고가 인간의 직관적인 사고 과정을 모방할 수 있음을 보인다.

하지만 여러 기법에도 불구하고 언어 모델은 학습하지 못한 과업에 대해 여전히 취약한 모습을 보인다. 모델이 학습된 이후에도 추가적인 인과 관계를 학습하지 못하므로 혼동된 생성 결과를 도출하는 문제도 존재한다. 이러한 문제를 해결하기 위해 RAG[6]나 REALM[7]과 같은 선행 연구는 검색 증강형 언어 모델 구조를 제안하였다. 검색 증강형 언어 모델은 입력과 연관된 문서를 검색하고 생성 과정에 통합하여 적절한 출력을 생성하도록 학습된다. 이를 통해 모델은 사전 학습된 매개변수 일부 혹은 전부를 새로이 학습하지 않고도 최신 정보에 기반한 출력을 생성할 수 있다. 또한 엄격하게 검수된 데이터베이스를 활용하여 모델이 부적절하거나 올바르지 못한 답변을 생성하지 않도록 간접적으로 제어할 수 있다는 장점을 보인다.

본 논문에서는 공개된 한국어 코퍼스로부터 새롭게 구축한 데이터베이스에 기반하여, In-Context 학습 방식과 검색 증강형 언어 모델을 결합한 In-Context 검색 증강형 한국어 언어 모델을 보인다. 이를 통해 사전 학습된 대규모 언어 모델 대비 향상된 모델의 생성 능력을 확인하고 오픈 도메인 질의응답(Open-Domain Question Answering) 과업에 적용하여 성능을 검증한 결과를 제시한다.

2. 관련 연구

검색 결과를 토대로 언어 모델의 생성 능력을 강화하는 시도는 이전부터 다양한 방식으로 연구되었다. 특히 오픈 도메인 질의응답과 같이 지식집약적인 과업에서 매개변수에 내재된 지식 외에 검색 결과에 기반한 외부 지식을 토대로 모델의 성능은

*교신저자 (Corresponding author)

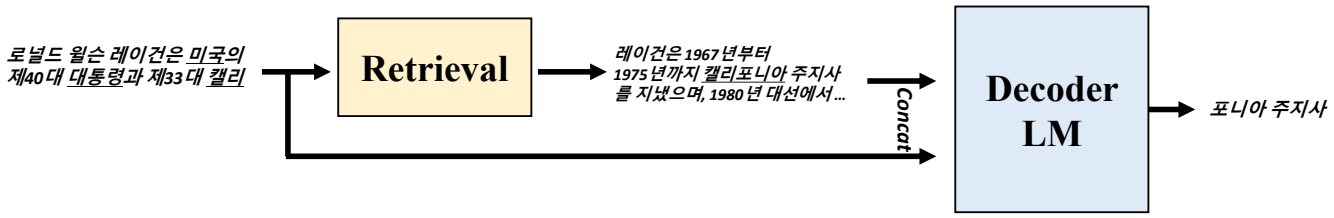

그림 1. In-Context 검색 증강형 한국어 언어 모델 구조도

나날이 발전되어왔다.

특히 검색 증강을 통한 언어 모델의 생성 능력을 강화하는 연구는 크게 두 가지로 나누어진다. 첫 번째는 kNN-LM[8]과 같이 Autoregressive(자기회귀) 언어 모델에서 다음 토큰을 생성하기 전 데이터베이스에서 검색된 결과에 기반하여 모델의 각 토큰별 생성 확률을 보정하는 방식이다. 두 번째는 본 연구와 같이 모델에 주어진 입력을 바탕으로 연관된 문서를 검색하고 이를 모델에 함께 입력하여 언어 모델의 생성 능력을 강화하는 방식이다.

2.1 검색 기반 생성형 언어 모델

RAG[6]은 문서 검색 결과를 Decoder의 입력으로 하여 텍스트를 생성하는 과정에서 검색 모듈(Retrieval)과 생성 모듈(Generator)을 통합한 End-to-End 방식의 모델을 제안한다. FiD[9]는 기존 Encoder-Decoder 구조에서 배치 단위의 인코딩 방식을 통해 길이가 긴 입력을 단일 과정으로 처리하여 오픈 도메인 질의응답 성능을 크게 향상시켰다. 더 나아가 검색된 문서를 사전 학습 단계에서 함께 활용하는 등 검색 증강된 학습 방식을 사전학습 단계에서부터 도입하여 모델의 검색 기반 생성 능력을 강화하는 방식도 연구되어 왔다 [10][7]. 또한 RETRO[11]은 수 조개의 토큰으로 구성된 데이터베이스를 구성하고, 문서의 임베딩(embedding) 공간에서 Nearest neighbor(최근접 이웃)을 탐색 후 이를 Chunked Cross Attention (CCA)을 통해 반영하는 확장된 모델 구조를 제안하였다.

2.2 In-Context 기반 생성 강화

최근 대규모 사전 학습 언어 모델이 공개됨에 따라 다양한 과업에서 뛰어난 Few-shot 성능이 입증되고 있다. 특히 해당 접근 방식은 대규모 언어 모델을 명시적으로 Fine-tuning 하지 않고 매개변수를 고정한 상태에서 활용하여 메모리를 효율적으로 절감할 수 있다. 하지만 과업에 따라서는 모델을 Fine-tuning 하는 방식에 비해 성능이 부족한 경우가 발생한다. 이에 따라 [12]는 사전 학습된 언어 모델을 활용하는 새로운 접근 방식으로 입력 의존적인 Prompt-tuning, Frozen Reader 및 재귀적 언어 모델을 제안하여 개선된 오픈 도메인 질의응답 성능을 보인다. 또한 In-Context RALM[13]에서는 기존 방식과 유사하나 대규모 언어 모델을 활용함에 있어 검색 결과에 기반한 In-Context 방식으로 모델의 생성 능력을 강화했다는 차이가 있다. 본 논문에서는 In-Context RALM을 한국어에 효과적으로 적용하여 In-Context 검색 결과에 기반한 언어 모델의 생성 능력을 강화할 수 있음을 보인다.

3. In-Context 검색 증강형 한국어 언어 모델

본 논문에서 적용하는 In-Context 검색 증강형 언어 모델 구조도는 그림 1과 같다.

Autoregressive 방식의 생성형 언어 모델 (e.g. GPT-2, GPT-3, OPT[14])에서 각 토큰 $x_1, ..., x_n$에 대한 전체 문장 생성 확률은 다음과 같다.

$$p(x_1, ..., x_n) = \prod_{i=1}^{n} p_\theta(x_i | x_{<i}), \tag{1}$$

수식 1의 $p(x_1, ..., x_n)$는 이전 토큰 $x_{<i}$가 주어질 때 다음 토큰 x_i을 예측하는 조건부 확률의 곱을 나타내며, θ는 모델 내의 매개변수를 의미한다.

검색 증강형 언어 모델은 입력 문장에 기반하여 검색된 문서를 활용하여 출력 문장을 생성하게 된다. 본 논문에서는 검색 모듈과 사전 학습된 언어 모델을 완전히 분리하여 검색을 수행하고 검색된 문서를 언어 모델의 입력 Prefix로 추가하여 이를 문맥삼아 다음 토큰들을 생성한다. 토큰 생성 과정은 수식 2와 같다.

$$p(x_1, ..., x_n) = \prod_{j=0}^{n_s-1} \prod_{i=1}^{s} p_\theta \left(x_{s \cdot j+i} | \left[\mathcal{R}(x_{<s \cdot j}); x_{<(s \cdot j+i)} \right] \right). \tag{2}$$

수식 2는 이전 토큰에 기반하여 데이터베이스 상의 연관 문서를 검색한 후 그 결과를 Prefix로 추가하여 토큰을 생성하는 과정을 나타낸다. $x_{<s \cdot j}$을 입력으로 하여 검색된 결과 $\mathcal{R}(x_{<s \cdot j})$를 기존 입력 $x_{<(s \cdot j+i)}$와 결합하여 토큰을 생성한다. [11]에서 언급한 바와 같이 매 토큰 생성 과정에서 검색을 수행했을 때 생성 품질이 가장 우수하지만, 메모리 효율성 및 추론 속도를 고려하여 Stride s마다 검색을 수행하여 문맥에 적합한 문서를

Model	Retrieval	Korean wikipedia	Modu news	Modu meeting
skt/kogpt2-base-v2 (125M)	-	60.145	26.528	29.039
	BM25	55.973	19.934	24.298
EleutherAI/polyglot-ko-1.3B	-	12.666	8.719	9.932
	BM25	12.268	7.104	8.805
EleutherAI/polyglot-ko-3.8B	-	10.590	7.443	8.811
	BM25	10.343	6.183	7.847
EleutherAI/polyglot-ko-5.8B	-	9.115	6.820	8.044
	BM25	8.902	5.711	7.240
EleutherAI/polyglot-ko-12.8B	-	8.494	6.474	7.747
	BM25	8.314	5.458	6.981

표 1. In-Context 검색 증강 적용 유무에 따른 언어 모델의 Perplexity

찾도록 구성된다. 이 때, n_s는 입력 문장의 길이 n을 Stride s로 나눈 값이며, s가 증가할수록 n_s가 작은 값으로 정의되어 검색 수행 빈도가 감소한다.

4. 실험

4.1 실험 설계

데이터셋	검색 문서 수	평가 데이터
한국어 위키피디아	600090	500
모두의 말뭉치 뉴스(2021)	580052	100
모두의 말뭉치 국회회의록	5295	100

표 2. 데이터셋 통계

본 연구에서 수행한 실험은 다양한 규모의 한국어 언어 모델(skt/kogpt2-base-v2[1], EleutherAI/polyglot-ko-{1.3, 3.8, 5.8, 12.8}B[2])을 활용하였다. 실험 및 평가를 위한 데이터셋으로는 한국어 위키피디아 데이터셋과, 모두의 말뭉치[3]의 뉴스(2021) 및 국회회의록 원시말뭉치를 활용하였다. 데이터셋 통계는 표 2와 같다. 또한 오픈 도메인 질의응답 과업에 대한 평가를 위해 KorQuAD 1.0[4] 데이터셋을 활용했다. 모든 실험은 NVIDIA RTX A6000 4기로 수행되었다.

4.2 실험 결과

In-Context 검색 증강형 한국어 언어 모델 성능 평가를 위해 표 2 제시한 데이터셋 중 일부를 샘플링하여 모델의 Perplexity를 측정하였다. Perplexity는 언어 모델을 생성 능력을 평가하기 위한 주된 지표 중 하나로, Perplexity가 낮을수록 언어 모델이 자연스럽고 정확한 결과를 생성한다고 평가된다. 모델 내의 검색 모듈은 pyserini[5]의 BM25 검색 모듈을 활용하였으며 실험 결과는 표 1과 같다.

In-Context 검색 증강을 적용한 모든 실험에서 개선된 Perplexity를 보인다. 특히 검색 증강을 적용한 polyglot-ko-3.8B 모델의 경우 모두의 말뭉치 뉴스 데이터셋에서 가장 큰 크기의 12.8B 모델 대비 개선된 Perplexity를 보이는 등 우수한 결과를 확인하였다. 모든 언어 모델은 정량적 평가를 위해 실험 시 Stride 값을 4로 설정하였으며, 검색을 위한 입력 문장의 길이를 32로 고정하여 진행하였다.

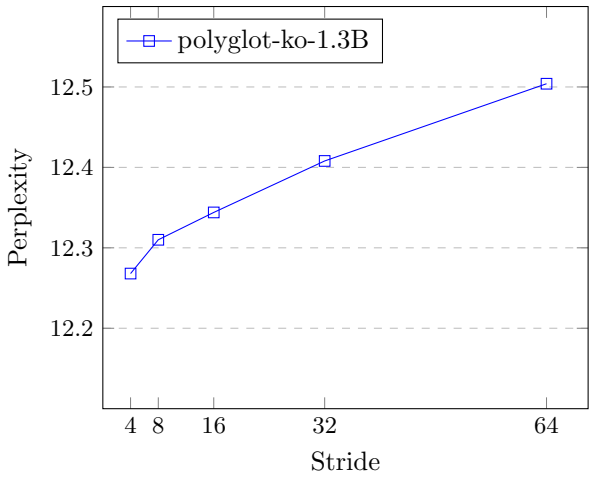

그림 2. Stride 길이에 따른 언어 모델의 Perplexity

그림 2는 Stride 길이에 따른 모델의 성능 차이를 확인하기 위해 각 Stride 별 Perplexity를 측정한 결과이다. 이 때, Stride가 4일 경우 가장 개선된 Perplexity를 나타내는 것으로 보아 검색 수행 빈도가 증가할수록 모델의 성능이 증가함을 보인다. 다만

[1] https://github.com/SKT-AI/KoGPT2
[2] https://github.com/EleutherAI/polyglot
[3] https://corpus.korean.go.kr/
[4] https://korquad.github.io
[5] https://github.com/castorini/pyserini

Stride가 작을수록 검색 수행 빈도가 급격히 증가하기 때문에 성능 향상에 비례하여 메모리 효율성 및 추론 속도에서 Trade-off가 발생하여 적절한 Stride를 설정하는 것이 중요하다.

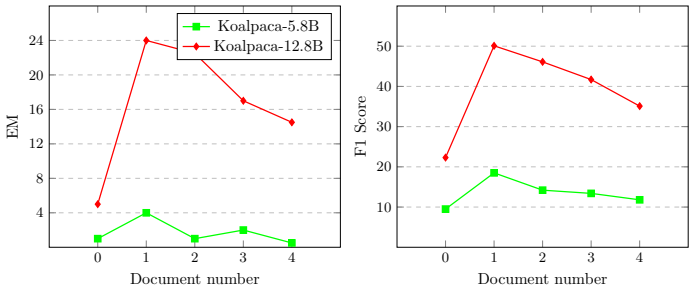

그림 3. 검색 문서 수에 따른 오픈 도메인 질의응답 성능

그림 3은 In-Context 검색 증강형 언어 모델의 Downstream 성능 평가를 위해 Korquad 1.0 데이터셋을 활용하여 오픈 도메인 질의응답 성능을 측정한 결과이다. 해당 실험은 Instruction-tuning이 적용된 Koalpaca[6] 모델을 사용하였으며 검색된 문서를 읽고 정답을 도출하도록 Instruction을 제시한 뒤 Zero-shot으로 성능을 평가하였다. Koalpaca-5.8B 및 12.8B 크기의 모델 모두 검색 증강을 수행하지 않은 실험 대비 큰 성능 향상을 보였으며 Koalpaca-12.8B 모델에서 최대 EM-19, F1-27.8 성능이 향상된 결과를 확인하였다. 특히 여러 문서를 검색하기보다 문서 하나를 검색하였을 때 가장 좋은 성능을 보여 모델이 생성 과정에서 정확한 문서 하나만을 활용하는 것이 성능 향상에 가장 효과적임을 보인다.

5. 결론

본 연구는 기존 한국어 언어 모델에 In-Context 검색 증강을 적용하여 개선된 생성 능력과 오픈 도메인 질의응답 성능 향상을 보인다. 특히 검색 증강을 적용한 3.8B 크기의 모델이 검색 증강을 적용하지 않은 12.8B 크기의 모델에 비해 개선된 Perplexity를 보여 우수한 성능을 입증한다. 다만 본 연구에서 제안한 모델은 검색 과정에서 BM25 검색 모듈만을 활용하기에, 향후 연구를 통해 신경망 학습 기반 Dense Retrieval 모델 적용 및 Re-ranking 기법을 도입하여 검색 모듈 개선을 통한 성능 향상을 기대하고자 한다. 또한 모델이 가장 적합한 문서 하나만을 활용하는 것이 효과적임을 확인하여 추후 검색 과정에서 활용되는 최적의 검색 프롬프트를 탐색하는 연구도 수행할 예정이다.

[6] https://github.com/Beomi/KoAlpaca

참고문헌

[1] OpenAI, "Gpt-4 technical report," 2023.

[2] R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du *et al.*, "Lamda: Language models for dialog applications," *arXiv preprint arXiv:2201.08239*, 2022.

[3] B. Kim, H. Kim, S.-W. Lee, G. Lee, D. Kwak, D. H. Jeon, S. Park, S. Kim, S. Kim, D. Seo *et al.*, "What changes can large-scale language models bring? intensive study on hyperclova: Billions-scale korean generative pretrained transformers," *arXiv preprint arXiv:2109.04650*, 2021.

[4] 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.

[5] 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.

[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] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "Retrieval augmented language model pre-training," *International conference on machine learning*, pp. 3929–3938, 2020.

[8] U. Khandelwal, O. Levy, D. Jurafsky, L. Zettlemoyer, and M. Lewis, "Generalization through memorization: Nearest neighbor language models," *arXiv preprint arXiv:1911.00172*, 2019.

[9] G. Izacard and E. Grave, "Leveraging passage retrieval with generative models for open domain question answering," *arXiv preprint arXiv:2007.01282*, 2020.

[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] S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. B. Van Den Driessche, J.-B. Lespiau, B. Damoc, A. Clark *et al.*, "Improving language models by retrieving from trillions of tokens," *International conference on machine learning*, pp. 2206–2240, 2022.

[12] Y. Levine, I. Dalmedigos, O. Ram, Y. Zeldes, D. Jannai, D. Muhlgay, Y. Osin, O. Lieber, B. Lenz, S. Shalev-Shwartz *et al.*, "Standing on the shoulders of giant frozen language models," *arXiv preprint arXiv:2204.10019*, 2022.

[13] O. Ram, Y. Levine, I. Dalmedigos, D. Muhlgay, A. Shashua, K. Leyton-Brown, and Y. Shoham, "In-context retrieval-augmented language models," *arXiv preprint arXiv:2302.00083*, 2023.

[14] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin *et al.*, "Opt: Open pre-trained transformer language models," *arXiv preprint arXiv:2205.01068*, 2022.