확산 언어 모델을 활용한 한국어 다중 벡터 검색

전승재[1], 고동혁[1], 김장환[1], 양승원[2], 나승훈[3], 김태훈[1], 박천음[1]*

국립한밭대학교[1], 바이브온[2], 울산과학기술원[3]

{sjjeon, kohdh, kjh}@edu.hanbat.ac.kr, swyang@vibeon.ai, nash@unist.ac.kr, {thkim, parkce}@hanbat.ac.kr

Diffusion Language Model based Korean Multi-Vector Retrieval Model

Seungjae Jeon[1], Donghyeok Koh[1], Janghwan Kim[1],

Seungwon Yang[2], Seunghoon Na[3], Taehoon Kim[1], Cheoneum Park[1]*

Hanbat National Univ[1], VIBEON[2], UNIST[3]

요 약

ColBERT로 대표되는 Multi-Vector Retrieval(MVR)은 토큰 수준 임베딩을 통해 세부적인 의미 매칭을 이뤄낸다. 하지만 BERT 기반의 MVR은 거대 언어 모델(LLM)에 비해 낮은 capacity와 문맥 표현력을 보인다. 최근 LLM을 MVR의 임베딩에 적용하려는 시도들은 모델의 capacity를 확보했지만, autoregressive 모델의 causal masking의 영향으로 각 토큰의 이전 문맥만 참고할 수 있어 텍스트의 양방향 인코딩이 불가능하다. 본 논문은 높은 capacity와 양방향 특성을 가진 Diffusion Language Model(DLM)을 MVR의 임베딩 모델로서 사용하는 새로운 방법을 제안한다. LLM과 비슷한 크기로 학습된 DLM의 hidden_state를 임베딩 벡터로 사용하고 MaxSim으로 유사도를 계산하여 질의와 관련된 문서를 검색한다. 실험 결과, KorQuAD, MrTyDi-ko, MIRACL Korean, Ko-Strategy 등 네 개의 한국어 벤치마크에서 모두 state-of-the-art를 달성하였다.

I. 서론

정보 검색(Information Retrieval)분야에서 의미 유사도를 기반으로 검색을 수행하는 Dense Retrieval은 단일 벡터 검색(single-vector retrieval, SVR)과 다중 벡터 검색(multi-vector retrieval, MVR)으로 분류된다. SVR [1, 2]은 질의와 문서를 각각 하나의 벡터로 표현하고 유사도를 계산하여 질의와 관련된 문서를 검색하는 방법론으로, 구현이 간단하고 검색 속도가 빠르지만 임베딩 과정에서 세부 정보가 손실되는 한계가 있다. MVR [3, 4, 5]은 토큰 수준의 벡터 표현을 생성하고 질문-문서 간의 alignment score 기반 유사도를 계산하여 질의와 관련된 문서를 검색하는 방법론으로, SVR보다 임베딩 과정에서 손실되는 정보가 적어 토큰 단위의 정밀한 검색이 가능하다.

Dense Retrieval에서 BERT기반 임베딩의 낮은 capacity와 표현력을 보완하고자 거대언어모델(Large Language Model, LLM)을 임베딩 모델로서 사용하는 시도가 있다 [6]. 하지만 LLM의 causal masking의 단방향성 구조가 각 토큰의 후속 문맥을 반영하지 못하는 구조적 제약이 존재한다. LLM2Vec [7]은 causal mask를 제거하고 추가 학습을 통해 단방향성 문제를 완화하지만, SVR에 한정되어 있고 단방향성 편향을 보정한 방식에 그친다.

본 논문은 AR 모델의 causal masking 제약을 극복하고 토큰 간의 bidirection을 극대화하기 위해 Diffusion Language Model (DLM)을 기반으로 하는 멀티 벡터 검색 방법을 제안한다. 실험 결과, 제안 방법은 In-domain 벤치마크에서 nDCG@10 평균 0.831, out-of-domain 벤치마크에서 nDCG@10 0.735으로 모두 최고 성능을 달성하였다.

*Corresponding author

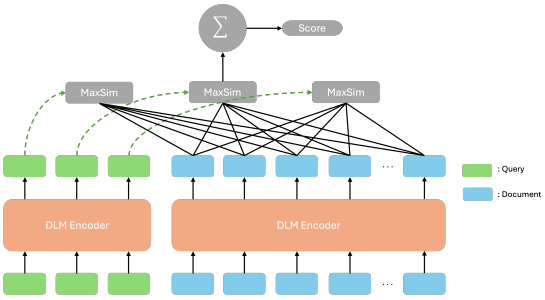

그림 1: 제안 방법의 Query-Document Score 도출 과정

II. 제안 방법

본 논문에서는 [그림 1]과 같이 DLM을 기반으로 다중벡터검색 방법을 제안하며, 모델 설명은 다음과 같다.

질의 토큰 $q = \{q_1, ..., q_n\}$ 와 문서 토큰 $d = \{d_1, ..., d_m\}$이 주어졌을 때, 다중 벡터 검색은 토큰 레벨 임베딩 벡터 $E_q \in \mathbb{R}^{N \times D}$ 와 $E_d \in \mathbb{R}^{M \times D}$ 을 변환하고 MaxSim 연산을 통해 유사도 스코어를 계산한다. 본 논문의 목표는 DLM을 통해 bidirectional context가 포함된 임베딩 벡터 E_q, E_d 를 생성하는 것이다.

본 논문은 마지막 Transformer 레이어의 hidden_state를 토큰별 임베딩 벡터로 사용한다. Dream-7B 모델을 사용할 경우, 입력 시퀀스에 대하여 hidden_state $H \in \mathbb{R}^{L \times 4096}$을 생성한다. 이때 denoising step $t = 0$ 으로 하여 masked 토큰이 포함되지 않은 input sequence가 모델에 입력된다.

마지막으로, 인코딩 된 질의 시퀀스의 토큰 벡터와 문단 시퀀스의

Type	Method	KorQuAD			MrTyDi-ko (low-resource)			MIRACL Korean (low-resource)			Average
		Hit@10	nDCG@10	MRR@10	Hit@10	nDCG@10	MRR@10	Hit@10	nDCG@10	MRR@10	nDCG@10
	BM25	0.631	0.521	0.463	0.438	0.326	0.285	0.477	0.391	0.338	0.413
SVR	mE5-large	0.869	0.799	0.762	0.748	0.644	0.604	0.765	0.662	0.638	0.703
	BGE-M3	0.897	0.785	0.744	0.738	0.640	0.590	0.810	0.700	0.660	0.708
	ReasonIR-8B	0.877	0.773	0.753	0.704	0.626	0.595	0.668	0.590	0.560	0.661
	LLaMA3-8B + LLM2Vec	0.865	0.768	0.720	0.695	0.600	0.562	0.652	0.553	0.519	0.640
	DIFFEMBED	0.887	0.799	0.772	0.710	0.636	0.601	0.708	0.620	0.588	0.685
MVR	ColBERTv2	0.738	0.645	0.612	0.501	0.423	0.388	0.470	0.390	0.364	0.487
	Dream-7B (Ours)	**0.967**	**0.929**	**0.920**	**0.899**	**0.790**	**0.763**	**0.867**	**0.774**	**0.762**	**0.831**
	LLaDA-8B (Ours)	0.849	0.760	0.745	0.824	0.748	0.721	0.837	0.732	0.720	0.747

표 1: In-domain 벤치마크 평가 결과

Type	Method	Hit@10	nDCG@10	MRR@10
	BM25	0.493	0.407	0.363
SVR	mE5-large	0.756	0.662	0.641
	BGE-M3	0.733	0.661	0.618
	ReasonIR-8B	0.764	0.669	0.643
	LLaMA3-8B + LLM2Vec	0.661	0.585	0.545
	DIFFEMBED	0.726	0.635	0.592
MVR	ColBERTv2	0.422	0.337	0.301
	Dream-7B (Ours)	**0.811**	**0.735**	**0.703**
	LLaDA-8B (Ours)	0.750	0.659	0.625

표 2: Out-of-domain 벤치마크(Ko-StrategyQA) 평가 결과

토큰 벡터 간의 유사도를 계산하기 위해 ColBERT [3]에서 제안된 MaxSim을 사용한다. MaxSim Scoring은 [식 1]과 같다.

$$S(q,d) = \sum_{i=1}^{N} \max_{j=1}^{M} (E_q^i \cdot E_d^{j^T}) \quad (1)$$

III. 실험

본 장에서는 제안 방법의 유효성을 검증하기 위해 기존 SVR, MVR 방법론과 비교 실험을 진행한다.

실험 환경. 실험에서 사용된 데이터셋은 KorQuAD-1.0[8], Ko-StrategyQA[9], MrTyDi-ko[10], MIRACL-ko[11]이며, 모두 한국어이다. KorQuAD, MrTyDi, MIRACL의 학습셋으로 모델 학습, 개발셋으로 평가를 수행한다. 이때 MrTyDi와 MIRACL는 학습 데이터의 양이 적은 low-resource 데이터셋이다. 학습셋이 존재하지 않은 Ko-StrategyQA는 Out-of-domain 평가를 수행한다. 모델 비교를 위하여, decoder-only LLM 기반의 SVR 모델인 mE5-large[6], ReasonIR[12], LLM2Vec[7]과, BERT 기반 모델인 ColBERT-v2[3], DLM 기반 모델인 DiffuEMBED[13]을 사용한다. 실험 평가 매저는 Hit, nDCG, MRR으로 각각 top-10결과를 평가한다.

실험 결과. [표 1]은 제안 방법과 기존 검색 모델들의 성능을 비교한 결과이다. 제안한 Dream-7B 기반 MVR은 KorQuAD에서 nDCG@10 0.929를 달성하여 기존 ColBERT-v2 대비 28.4%p, LLM 기반 SVR 모델인 mE5-large 대비 13.0%p 향상된다.

특히 MrTyDi-ko(1,295개)와 MIRACL-ko(868개) 같은 학습 데이터가 제한된 데이터셋에서도 Dream-7B(Ours)는 ColBERTv2 대비 각각 37.7%p, 38.4%p 상승한다. 이는 DLM의 any-order modeling과 bidirectional attention를 통해 제한된 데이터에서도 효과적인 representation 학습이 가능함을 보인다[14].

[표 2]의 Out-of-domain 평가의 경우, Dream-7B는 Ko-StrategyQA에서 nDCG@10 0.735를 달성하여 ColBERT-v2 (0.337) 대비 39.8%p 성능 향상을 보인다. Reasoning이 필요한 복잡한 질의에서 DLM의 bidirectional context가 효과적임을 알 수 있다. 제안 방법과 DLM 기반 SVR 모델인 DIFFEMBED (0.635)와 비교할 경우, Dream-7B 기반의 제안 방법이 10.0%p 높은 성능을 보이는데, 이는 토큰 수준의 정밀한 매칭이 검색 성능 향상에 기여한 것으로 사료된다.

IV. 결론

본 논문에서는 검색 성능 고도화를 위해 DLM 기반 MVR 방법을 제안하였으며, Dream-7B 백본을 기준으로 4개 한국어 벤치마크에서 모두 최고 성능을 달성하였다. 실험 결과, 제안 방법은 4개 벤치마크에서 모든 평가 방법을 포함하여, SVR 대비 약 6-14%p, MVR (ColBERT) 대비 약 23-40%p 가량 향상되었다. 향후 연구로는 영어 벤치마크에서 실험을 수행하여, 언어별 강건한 방법임을 확인할 예정이다.

참 고 문 헌

[1] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, "Dense passage retrieval for open-domain QA," in *Proc. EMNLP 2020*.

[2] T. Formal, B. Piwowarski, and S. Clinchant, "SPLADE: Sparse lexical and expansion model for first stage ranking," in *Proc. SIGIR 2021*.

[3] K. Santhanam, O. Khattab, J. Saad-Falcon, C. Potts, and M. Zaharia, "ColBERTv2: Effective and efficient retrieval via lightweight late interaction," in *Proc. NAACL-HLT 2022*.

[4] J. Lee, Z. Dai, S. M. K. Duddu, T. Lei, I. Naim, M.-W. Chang, and V. Y. Zhao, "Rethinking the role of token retrieval in multi-vector retrieval," in *Proc. NeurIPS 2023*).

[5] C. Park, S. Jeong, M. Kim, K. Lim, and Y.-H. Lee, "SCV: Light and effective multi-vector retrieval with sequence compressive vectors," in *Proc. COLING 2025*.

[6] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, "Improving text embeddings with large language models," in *Proc. ACL 2024*, pp. 11897–11916.

[7] P. BehnamGhader, V. Adlakha, M. Mosbach, D. Bahdanau, N. Chapados, and S. Reddy, "LLM2Vec: Large language models are secretly powerful text encoders," in *Proc. COLM 2024*. arXiv:2404.05961.

[8] S. Lim, M. Kim, and J. Lee, "KorQuAD1.0: Korean QA dataset for machine reading comprehension," 2019.

[9] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers, "MTEB: Massive text embedding benchmark," in *Proc. EACL 2023*.

[10] X. Zhang, X. Ma, P. Shi, and J. Lin, "Mr. TyDi: A multi-lingual benchmark for dense retrieval," in *Proc. 1st Workshop on Multilingual Representation Learning* (*MRL@EMNLP 2021*), pp. 127–137, 2021.

[11] X. Zhang, N. Thakur, O. Ogundepo, E. Kamalloo, D. Alfonso-Hermelo, X. Li, Q. Liu, M. Rezagholizadeh, and J. Lin, "MIRACL: A multilingual retrieval dataset covering 18 diverse languages," *TACL*, vol. 11, pp. 1114–1131, 2023.

[12] R. Shao, R. Qiao, V. Kishore, N. Muennighoff, X. V. Lin, D. Rus, B. K. H. Low, S. Min, W.-t. Yih, P. W. Koh, and L. Zettlemoyer, "ReasonIR: Training retrievers for reasoning tasks," 2025.

[13] S. Zhang, Y. Zhao, L. Geng, A. Cohan, A. T. Luu, and C. Zhao, "Diffusion vs. autoregressive language models: A text embedding perspective," in *Proc. EMNLP 2025*.

[14] J. Ni, Q. Liu, L. Dou, C. Du, Z. Wang, H. Yan, T. Pang, and M. Q. Shieh, "Diffusion language models are super data learners," 2025.