DMQA Open Seminar

Contrastive Learning for Retrieval Models

2025.01.24

고려대학교 산업경영공학과

Data Mining & Quality Analytics Lab.

이정민

KOREA UNIVERSITY

DMQA

발표자 소개

❖ 이정민(JungMin Lee)

- 고려대학교 산업경영공학과 대학원 재학
- Data Mining & Quality Analytics Lab.(김성범 교수님)
- 석박 통합 과정(2022.03~Present)

❖ Research Interest

- Uncertainty Quantification
- Label Noise Learning
- Large Language Models

❖ Contact

- jungmin9195@korea.ac.kr

Contents

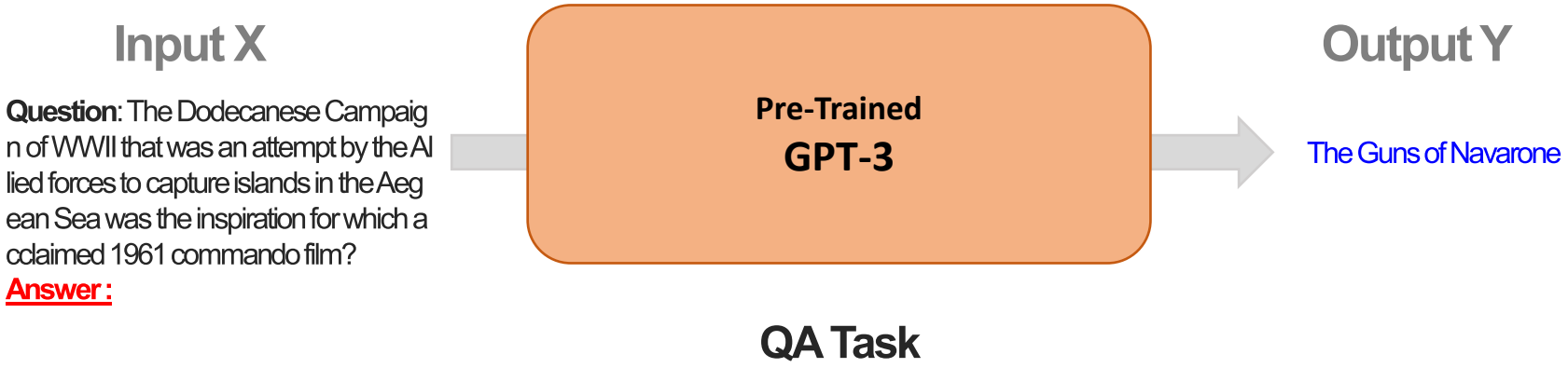
Introduction

Introduction

What is LLM?

❖ **Large Language Models (LLM)**

- 방대한 양의 텍스트로 사전 학습 되어 많은 지식을 축적한 언어 모델
- 대용량의 언어 모델을 통해 다양한 task를 수행할 수 있음



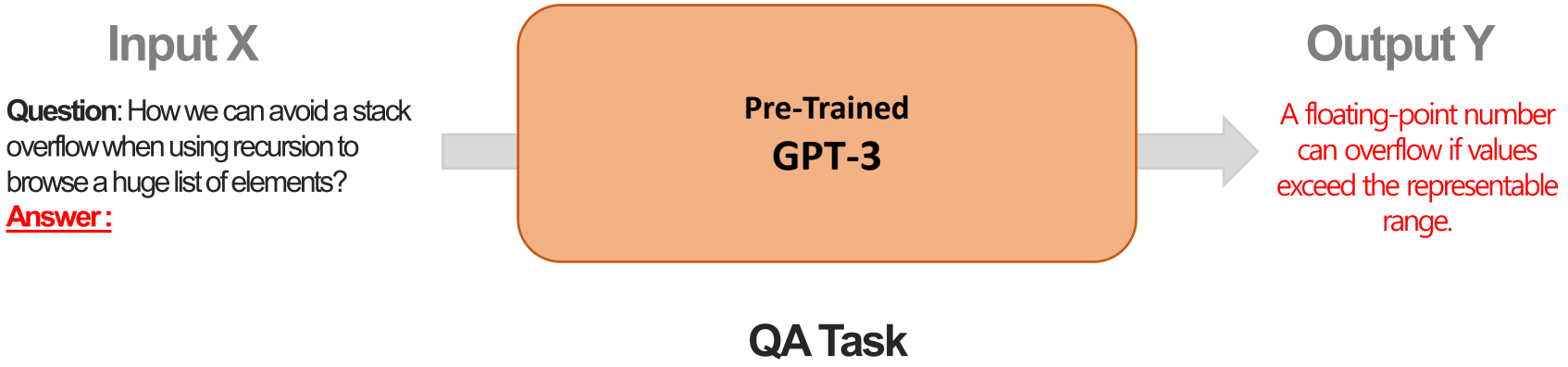
DMQA

Introduction

What is LLM?

❖ **Large Language Models (LLM)**

- 방대한 양의 텍스트로 사전 학습 되어 많은 지식을 축적한 언어 모델
- 대용량의 언어 모델을 통해 다양한 task를 수행할 수 있음
- 기존 LLM의 한계: 학습되지 않은 특정 도메인 텍스트에 대해서는 부적절한 답변 출력



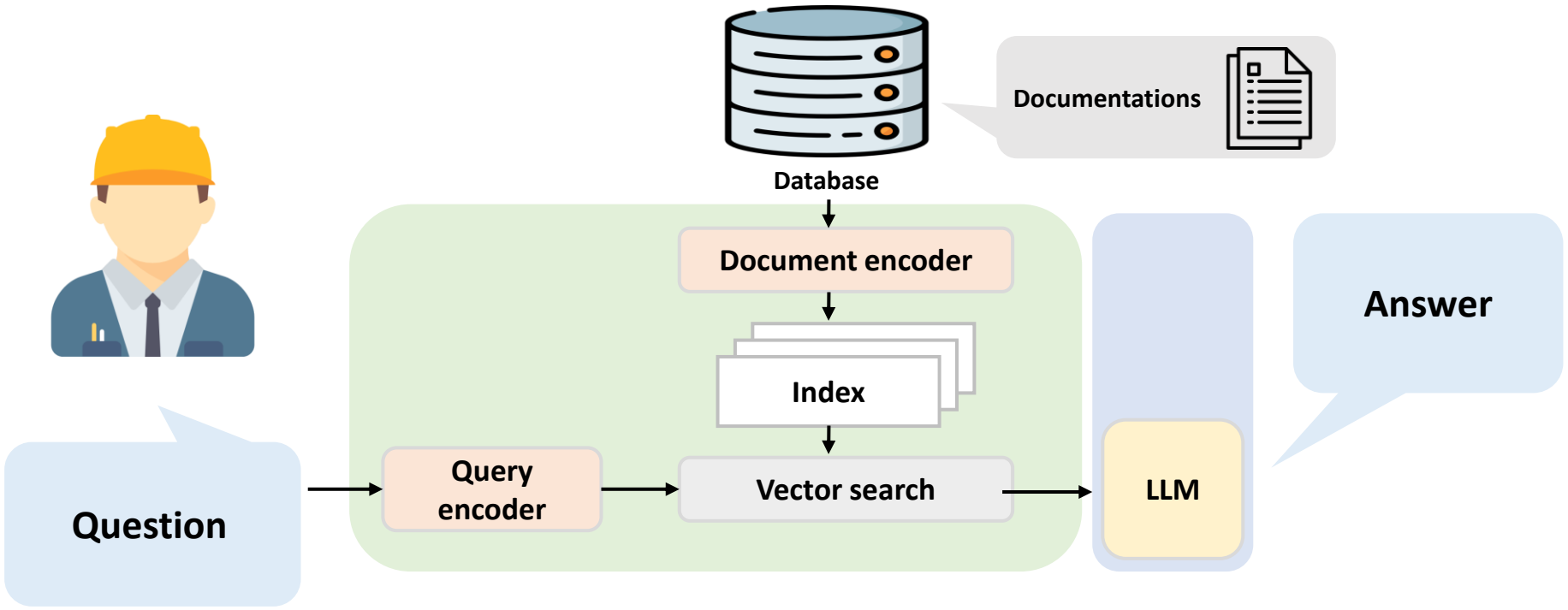
DMQA

Introduction

Retrieval-Augmented Generation

❖ RAG (Retrieval-Augmented Generation)

- 외부 Database의 정보를 활용하여 고품질의 답변을 생성하기 위한 프레임워크

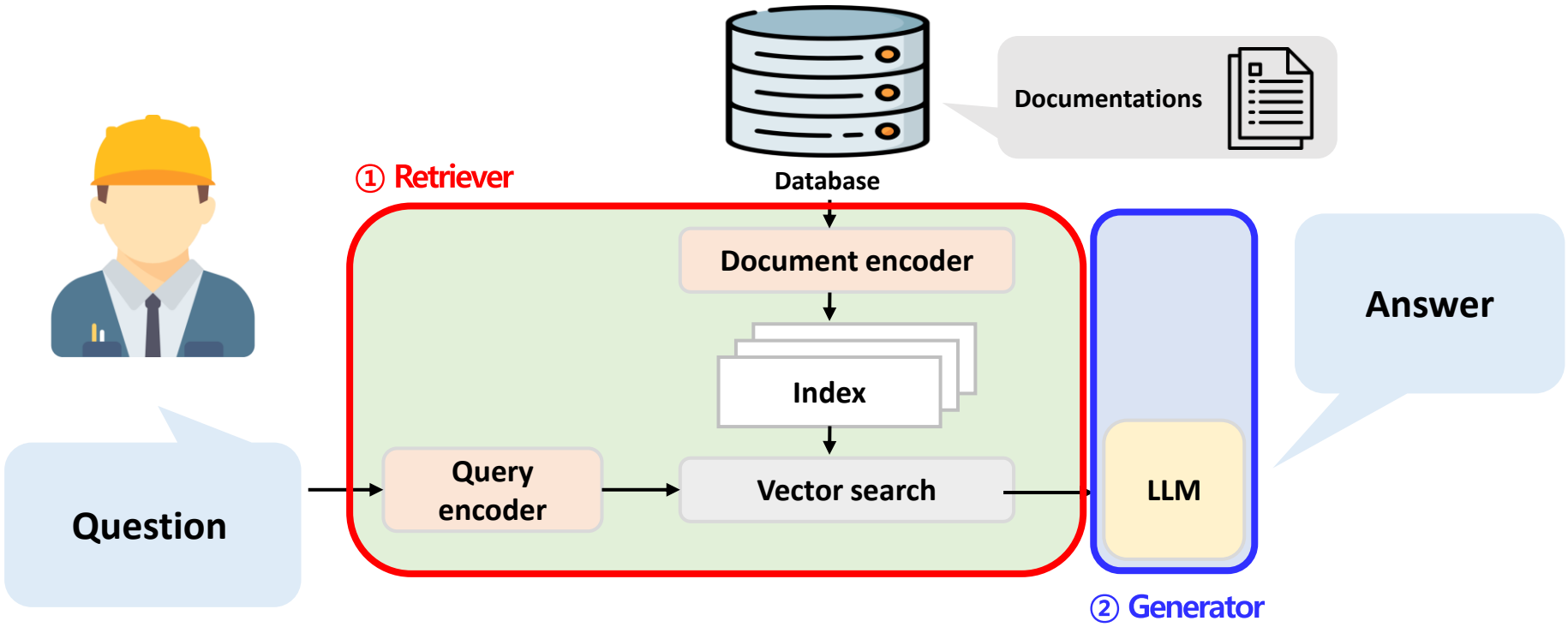


Introduction

Retrieval-Augmented Generation

❖ RAG (Retrieval-Augmented Generation)

1. Retriever가 질문과 관련된 정보를 데이터베이스에서 탐색하고 가장 연관된 문서를 가져옴
2. 가져온 정보를 질문과 함께 generator에 입력



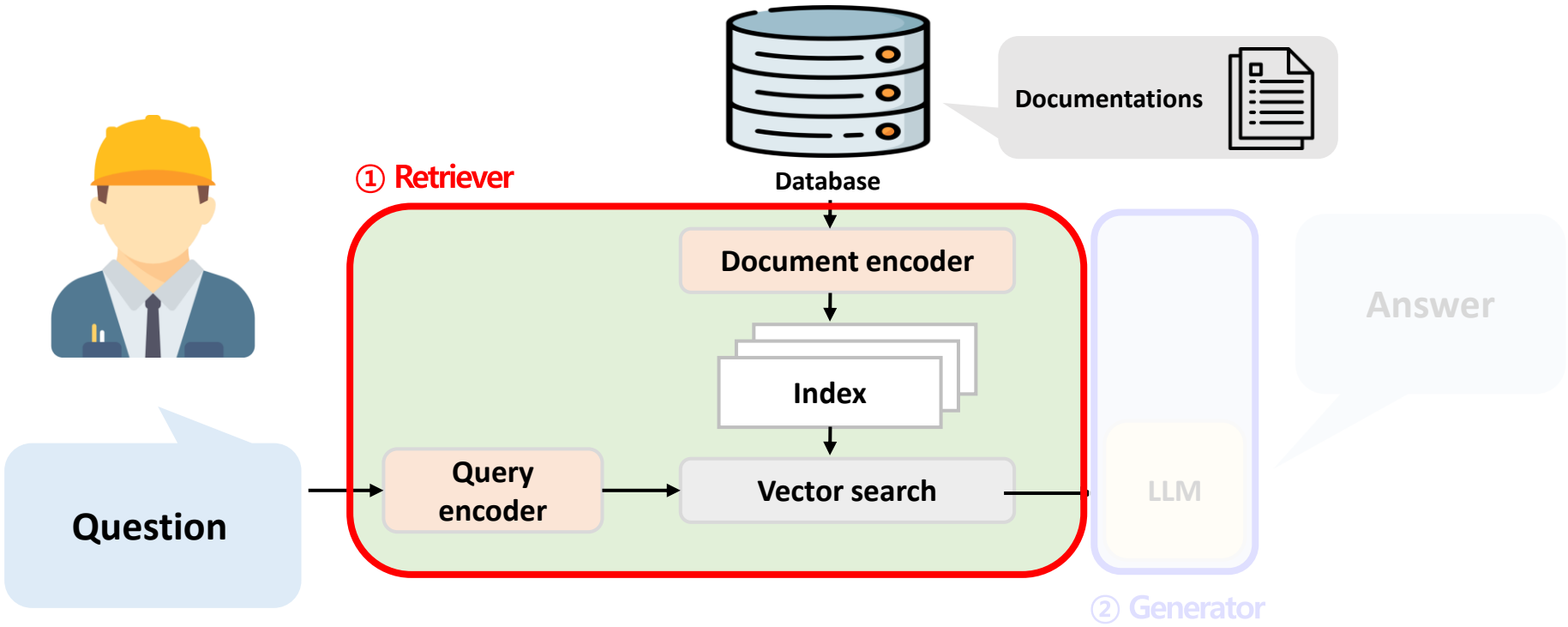
DMQA

Introduction

Retrieval-Augmented Generation

❖ RAG (Retrieval-Augmented Generation)

1. Retriever가 질문과 관련된 정보를 데이터베이스에서 탐색하고 가장 연관된 문서를 가져옴
2. 가져온 정보를 질문과 함께 generator에 입력



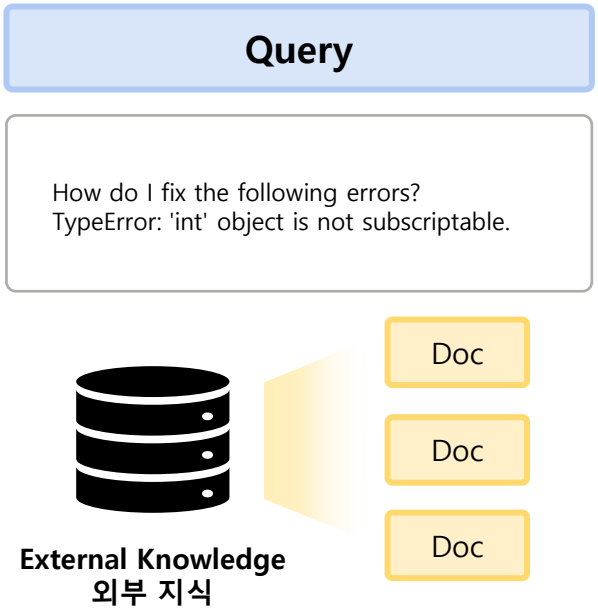
DMQA

Introduction

Retrieval Models

❖ Traditional Retrieval Models

- 딥러닝의 발전 이전에는, 통계 기반의 retrieval models 이 주로 사용 됨
  - ➢ TF-IDF / BM25

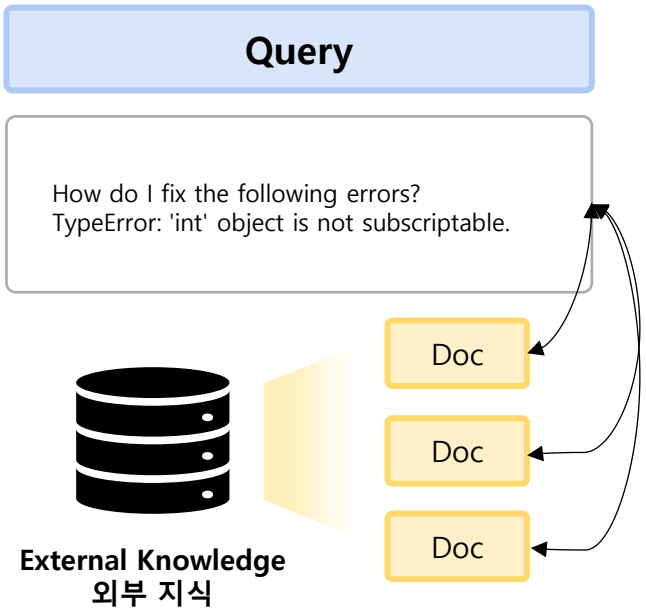


Introduction

Retrieval Models

❖ Traditional Retrieval Models

- 딥러닝의 발전 이전에는, 통계 기반의 retrieval models 이 주로 사용 됨
  - ➢ **TF-IDF** / BM25



토큰 t가 문서 d에서 나오는 회수

$$TFIDF = TF(t,d) \times IDF(t) = TF(t,d) \times \log\frac{N}{DF(t)}$$

토큰 t가 제공하는 정보의 양

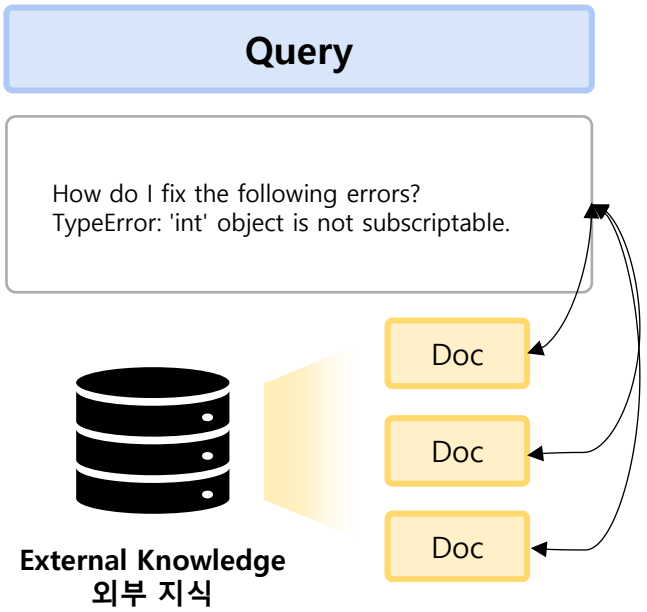
$$Score(D,Q) = \sum_{q \in Q} TFIDF(q,Q) \times TFIDF(q,D)$$

Introduction

Retrieval Models

❖ Traditional Retrieval Models

- 딥러닝의 발전 이전에는, 통계 기반의 retrieval models 이 주로 사용 됨
  - ➢ TF-IDF / **BM25**



토큰 t가 문서 d에서 나오는 회수

$$TFIDF = TF(t,d) \times IDF(t) = TF(t,d) \times \log\frac{N}{DF(t)}$$

토큰 t가 제공하는 정보의 양

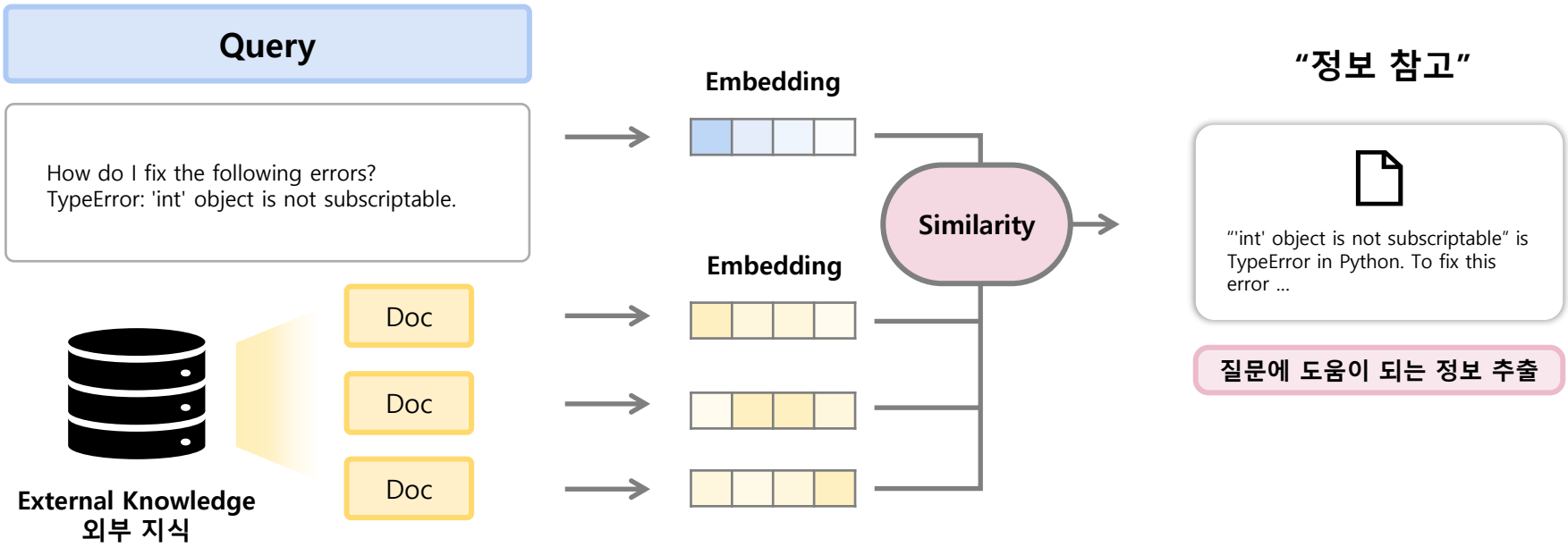
$$Score(D,Q) = \sum_{q \in Q} IDF(q,D) \times \frac{TFIDF(q,D) \times (k+1)}{TFIDF(q,D) + k \times (1 - b + b \times \frac{|D|}{d_{avg}})}$$

Introduction

Retrieval Models

❖ Dense Retrieval

- 임베딩 벡터 간 유사도 기반으로 쿼리와 관련된 정보 추출
- 관련 있는 정보를 얼마나 잘 참고하는지는 RAG 성능에 큰 영향을 줌

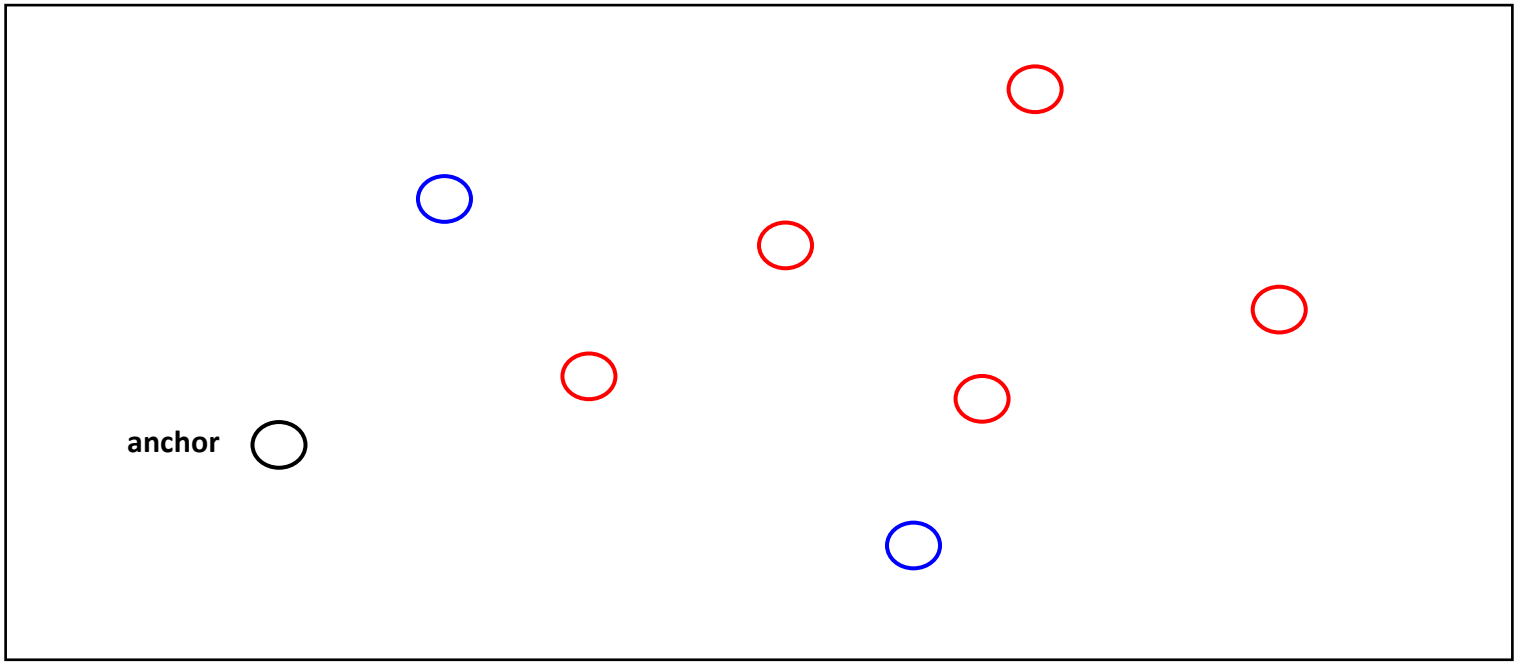


Introduction

Contrastive Learning

❖ Contrastive Learning

- Metric learning 방법론 중 하나로, 데이터 간 유사도 정보를 통해 거리 함수를 학습하는 방법론
- Main idea: Anchor를 기준으로 positive samples는 가깝도록, negative samples는 멀도록 학습



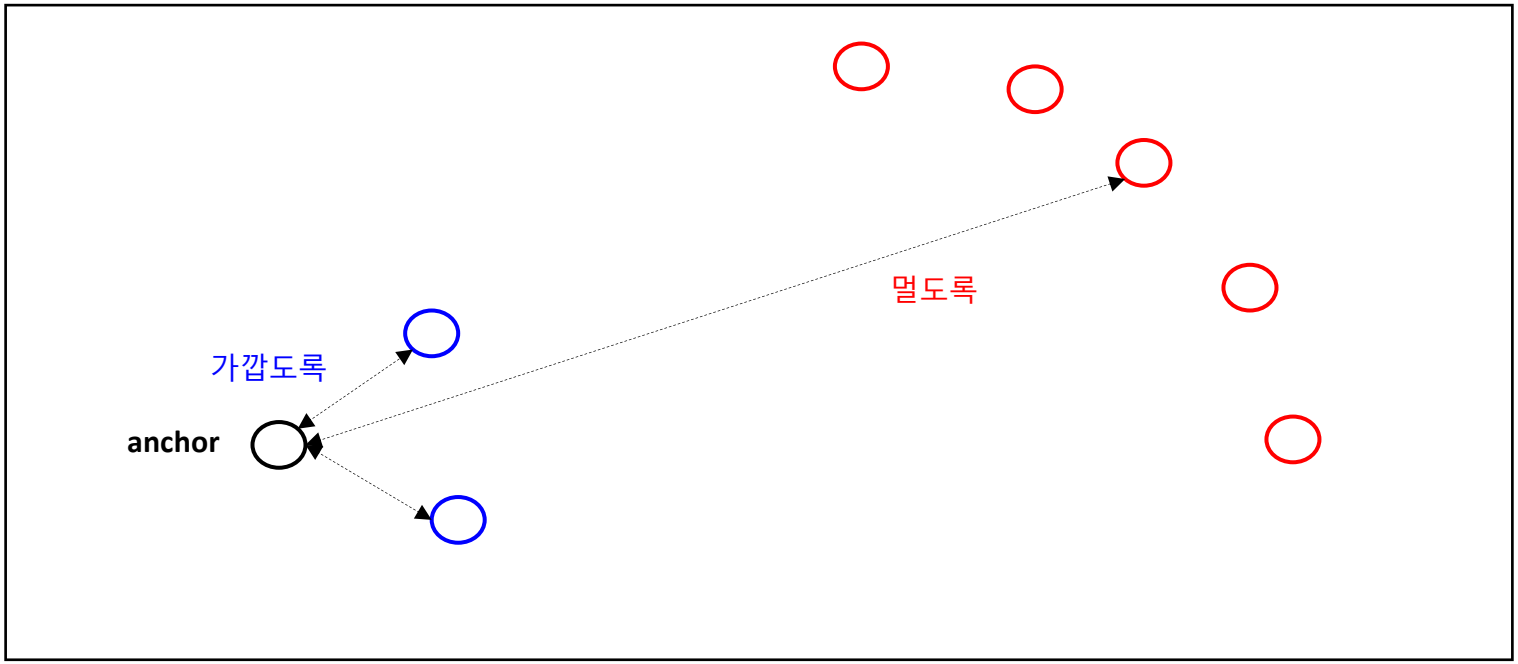
[Embedding space]

Introduction

Contrastive Learning

❖ Contrastive Learning

- Metric learning 방법론 중 하나로, 데이터 간 유사도 정보를 통해 거리 함수를 학습하는 방법론
- Main idea: Anchor를 기준으로 positive samples는 가깝도록, negative samples는 멀도록 학습

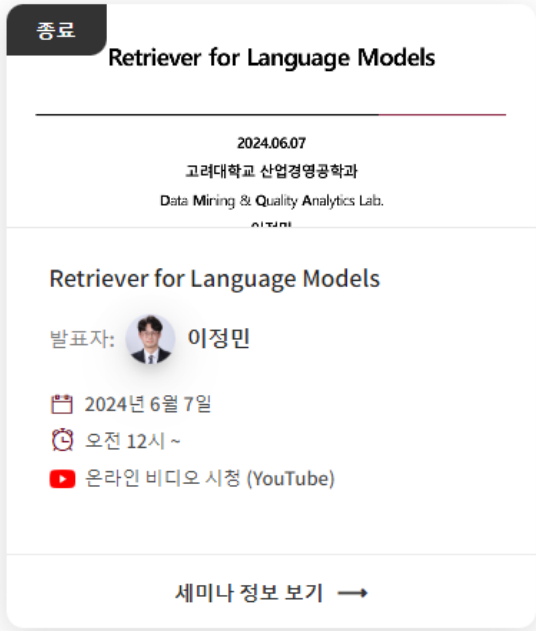


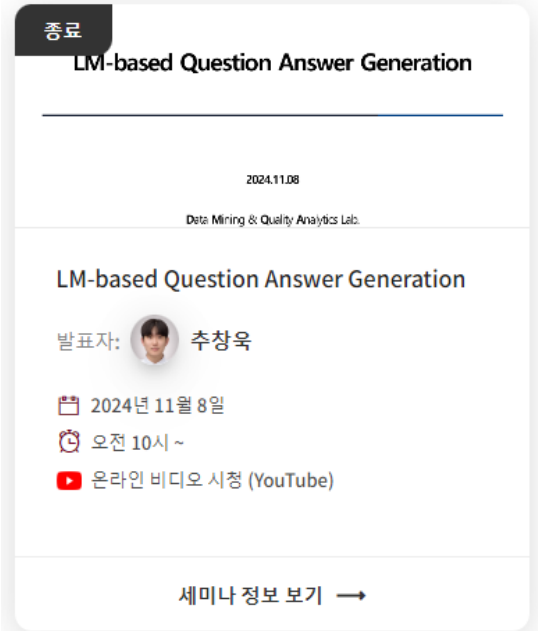
[Embedding space]

Introduction

Related Seminar

❖ Retrieval Models



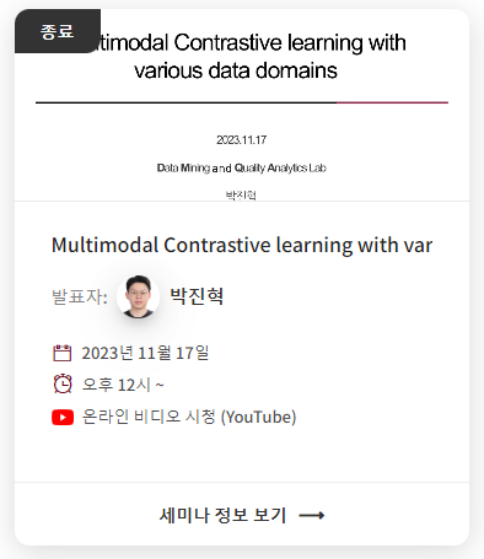


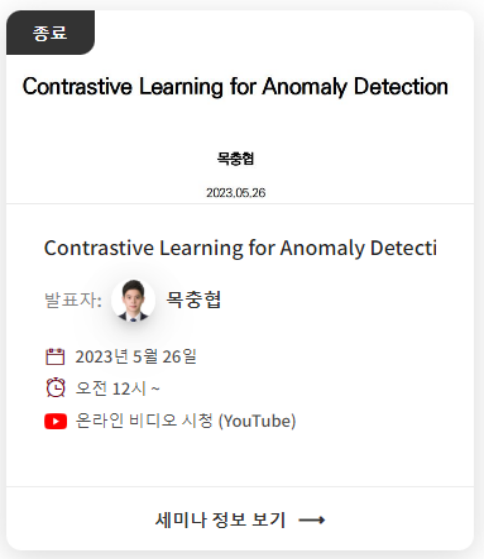
DMQA

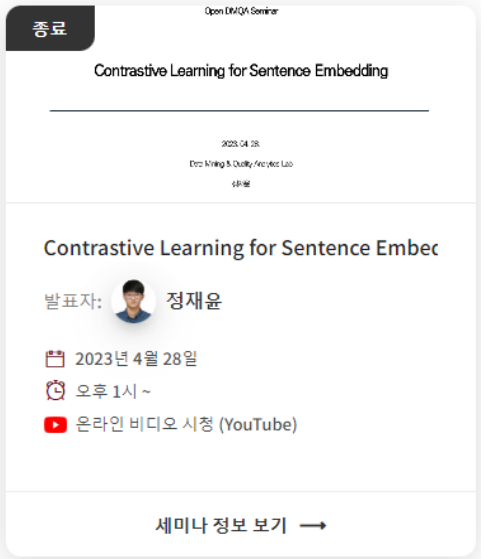
Introduction

Related Seminar

❖ Contrastive Learning







DMQA

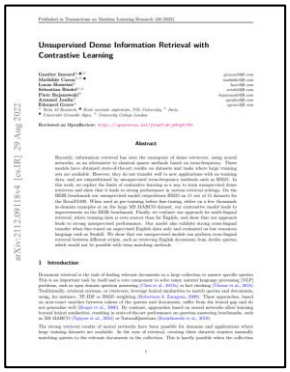
Introduction

Related Seminar

❖ Pre-Training Retrieval Models

Contrastive Learning

Contriever(2021)



PAIR(2021)

LaPraDoR(2022)

ReContriever(2023)

MvCR(2023)

MLM

SEED(2021)

RetroMAE(2022)

RetroMAE-2(2023)

Llama2Vec(2024)

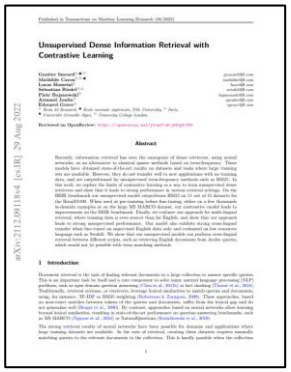
DMQA

Introduction

Related Seminar

❖ Pre-Training Retrieval Models

Contriever(2021) **PAIR(2021)** LaPraDoR(2022) **ReContriever(2023)** MvCR(2023)



Contrastive Learning

MLM

SEED(2021)

RetroMAE(2022)

RetroMAE-2(2023)

Llama2Vec(2024)

DMQA

Contrastive Learning for Retrieval Models

Contriever

Paper

❖ Unsupervised Dense Information Retrieval with Contrastive Learning (2021, arXiv)

- Contrastive learning을 통해 좋은 성능의 unsupervised retriever를 학습시킴

Unsupervised Dense Information Retrieval with Contrastive Learning

Gautier Izacard[◇,♣,♡] gizacard@fb.com
Mathilde Caron[◇,♡,♠] mathilde@fb.com
Lucas Hosseini[◇] hoss@fb.com
Sebastian Riedel[◇,△] sriedel@fb.com
Piotr Bojanowski[◇] bojanowski@fb.com
Armand Joulin[◇] ajoulin@fb.com
Edouard Grave[◇] egrave@fb.com
[◇] *Meta AI Research,* [♣] *Ecole normale supérieure, PSL University,* [♡] *Inria,*
[♠] *Université Grenoble Alpes,* [△] *University College London*

Reviewed on OpenReview: *https://openreview.net/forum?id=jKN1pXi7b0*

Abstract

Recently, information retrieval has seen the emergence of dense retrievers, using neural networks, as an alternative to classical sparse methods based on term-frequency. These models have obtained state-of-the-art results on datasets and tasks where large training sets are available. However, they do not transfer well to new applications with no training data, and are outperformed by unsupervised term-frequency methods such as BM25. In this work, we explore the limits of contrastive learning as a way to train unsupervised dense retrievers and show that it leads to strong performance in various retrieval settings. On the BEIR benchmark our unsupervised model outperforms BM25 on 11 out of 15 datasets for the Recall@100. When used as pre-training before fine-tuning, either on a few thousands in-domain examples or on the large MS MARCO dataset, our contrastive model leads to improvements on the BEIR benchmark. Finally, we evaluate our approach for multi-lingual retrieval, where training data is even scarcer than for English, and show that our approach leads to strong unsupervised performance. Our model also exhibits strong cross-lingual transfer when fine-tuned on supervised English data only and evaluated on low resources language such as Swahili. We show that our unsupervised models can perform cross-lingual retrieval between different scripts, such as retrieving English documents from Arabic queries, which would not be possible with term matching methods.

Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., & Grave, E. (2021). Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

DMQA

Contriever

Unsupervised Dense Information Retrieval with Contrastive Learning

❖ 연구 배경

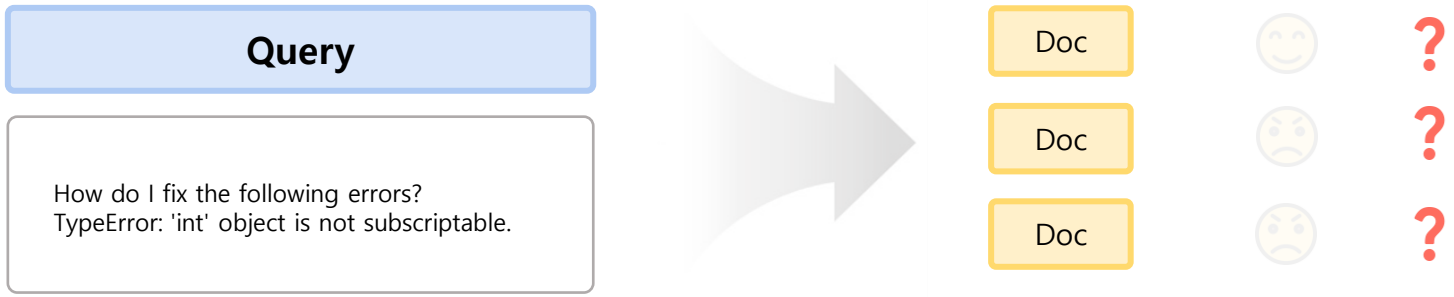
- No supervision 상황에서 어떻게 retrieval model을 잘 학습시킬까?

Contriever

Unsupervised Dense Information Retrieval with Contrastive Learning

❖ 연구 배경

- No supervision 상황에서 어떻게 retrieval model을 잘 학습시킬까?



Contriever

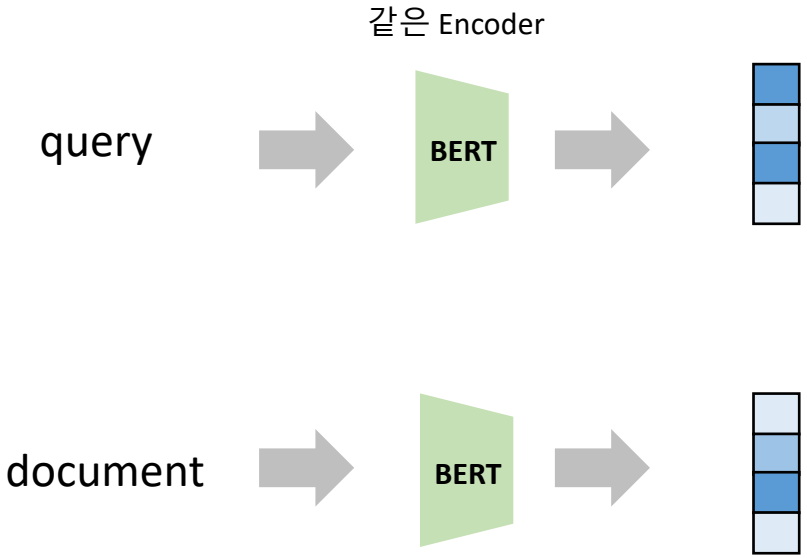
Unsupervised Dense Information Retrieval with Contrastive Learning

❖ Contrastive learning

- Positive documents와는 가깝도록, negative documents와는 멀도록
- Query와 documents는 같은 구조의 encoder 사용 → robustness

$$s(q,d) =< f_\theta(q), f_\theta(d) >$$

$$L(q,k_+) = -\frac{\exp(\frac{s(q,k_+)}{\gamma})}{\exp\left(\frac{s(q,k_+)}{\gamma}\right) + \sum_{i=1}^{K}\exp(\frac{s(q,k_i)}{\gamma})}$$

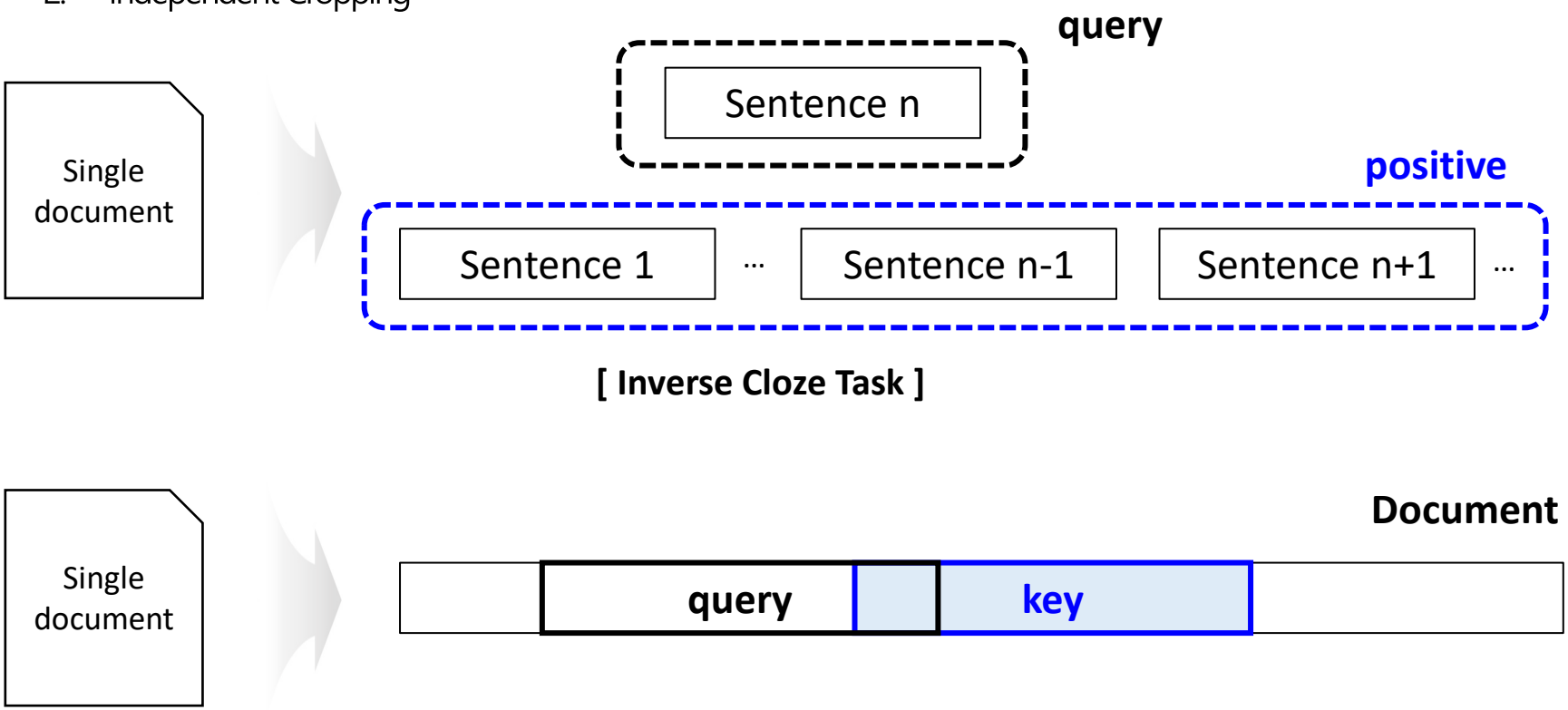


Contriever

Unsupervised Dense Information Retrieval with Contrastive Learning

❖ **Define positive pairs**

1. Inverse Cloze Task
2. Independent Cropping

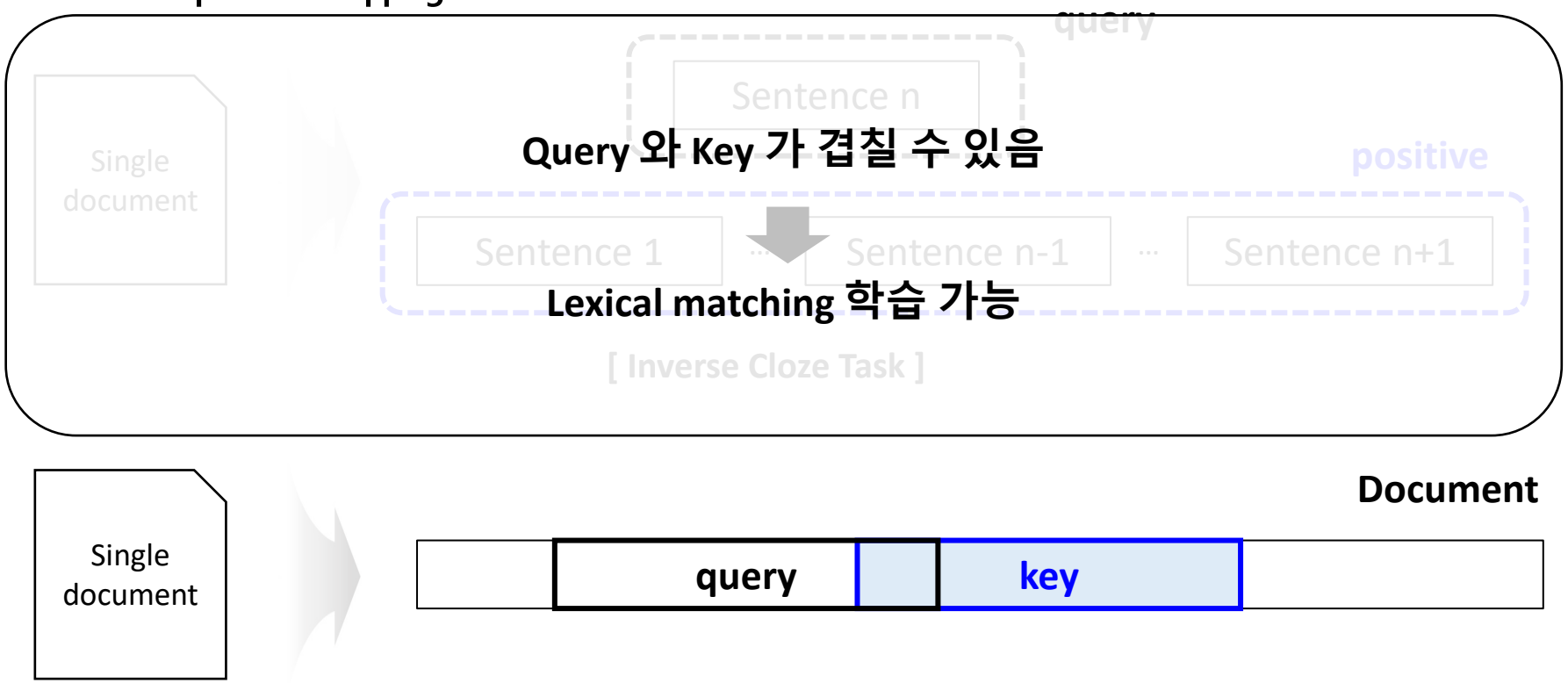


Contriever

Unsupervised Dense Information Retrieval with Contrastive Learning

❖ **Define positive pairs**

1. Inverse Cloze Task
2. **Independent Cropping**



[Independent Cropping]

DMQA

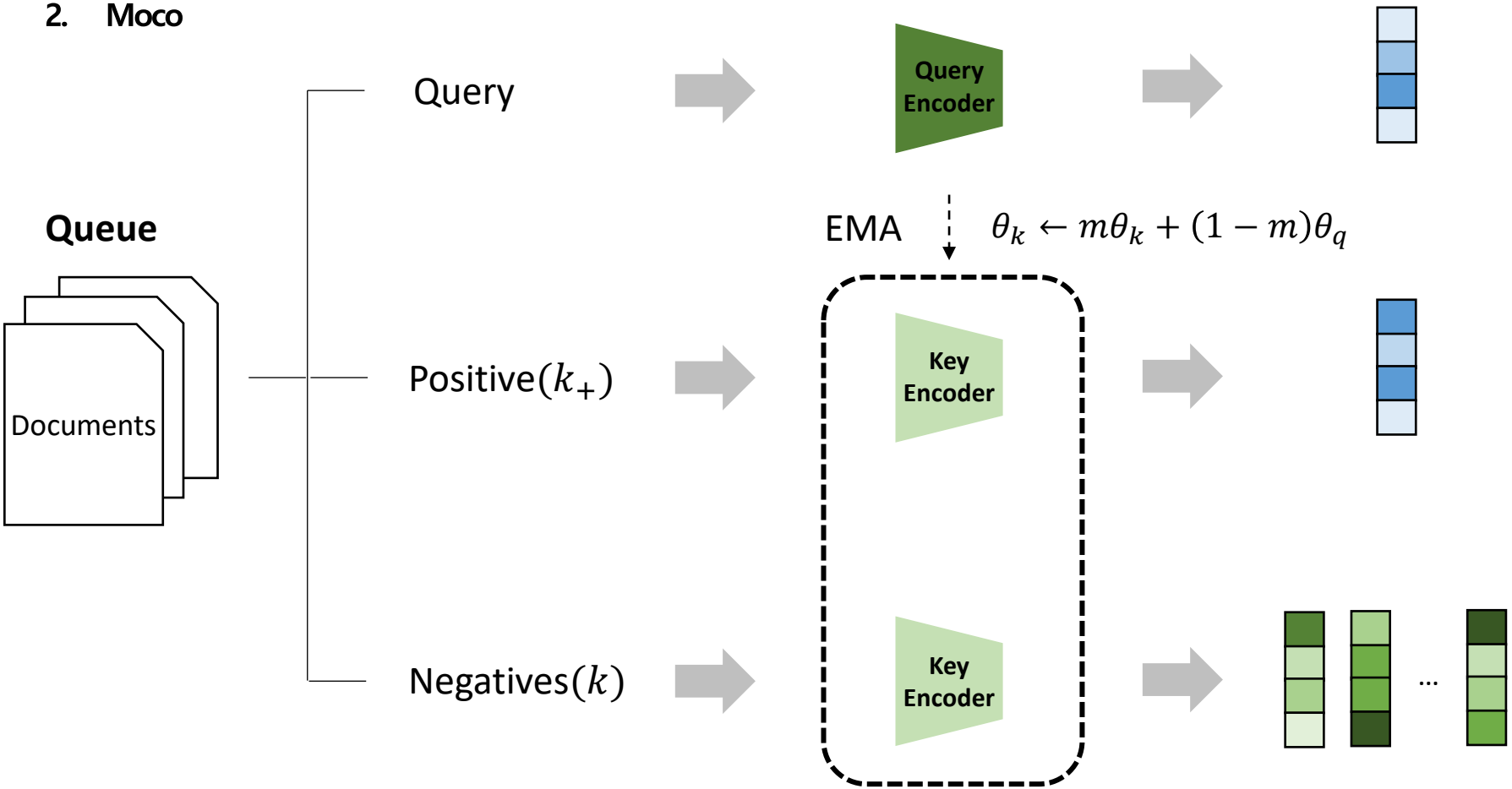
Contriever

Unsupervised Dense Information Retrieval with Contrastive Learning

$$L(q, k_+) = -\frac{\exp(\frac{s(q,k_+)}{\gamma})}{\exp\left(\frac{s(q,k_+)}{\gamma}\right) + \sum_{i=1}^{K} \exp(\frac{s(q,k_i)}{\gamma})}$$

❖ **Define negative pairs**

1. In-batch
2. **Moco**



DMQA

Contriever

Unsupervised Dense Information Retrieval with Contrastive Learning

❖ Experiments

- 미세 조정 없이 실험

	NaturalQuestions			TriviaQA		
	R@5	R@20	R@100	R@5	R@20	R@100
Inverse Cloze Task (Sachan et al., 2021)	32.3	50.9	66.8	40.2	57.5	73.6
Masked salient spans (Sachan et al., 2021)	41.7	59.8	74.9	53.3	68.2	79.4
BM25 (Ma et al., 2021)	-	62.9	78.3	-	**76.4**	**83.2**
Contriever	**47.8**	**67.8**	**82.1**	**59.4**	74.2	**83.2**
supervised model: DPR (Karpukhin et al., 2020)	-	78.4	85.4	-	79.4	85.0
supervised model: FiD-KD (Izacard & Grave, 2020a)	73.8	84.3	89.3	77.0	83.6	87.7

Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., & Grave, E. (2021). Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Contriever

Unsupervised Dense Information Retrieval with Contrastive Learning

❖ Experiments

- MS MARCO 데이터셋으로 미세 조정 후 BEIR 벤치마크에 대한 zero-shot 실험

Table 2: **BEIR Benchmark.** We report nDCG@10 on the test sets from the BEIR benchmark for bi-encoder methods without re-ranker. We also report the average and number of datasets where a method is the best ("Best on") over the entire BEIR benchmark (excluding three datasets because of their licence). Bold is the best overall. MS MARCO is excluded from the average. "CE" refers to cross-encoder.

	BM25	BM25+CE	DPR	ANCE	TAS-B	Gen-Q	ColBERT	Splade v2	Ours	Ours+CE
MS MARCO	22.8	41.3	17.7	38.8	40.8	40.8	40.1	43.3	40.7	**47.0**
Trec-COVID	65.6	**75.7**	33.2	65.4	48.1	61.9	67.7	71.0	59.6	70.1
NFCorpus	32.5	**35.0**	18.9	23.7	31.9	31.9	30.5	33.4	32.8	34.4
NQ	32.9	53.3	47.4	44.6	46.3	35.8	52.4	52.1	49.8	**57.7**
HotpotQA	60.3	70.7	39.1	45.6	58.4	53.4	59.3	68.4	63.8	**71.5**
FiQA	23.6	34.7	11.2	29.5	30.0	30.8	31.7	33.6	32.9	**36.7**
ArguAna	31.5	31.1	17.5	41.5	42.9	**49.3**	23.3	47.9	44.6	41.3
Touche-2020	**36.7**	27.1	13.1	24.0	16.2	18.2	20.2	36.4	23.0	29.8
CQADupStack	29.9	37.0.	15.3	29.6	31.4	34.7	35.0	-	34.5	**37.7**
Quora	78.9	82.5	24.8	85.2	83.5	83.0	85.4	83.8	**86.5**	82.4
DBPedia	31.3	40.9	26.3	28.1	38.4	32.8	39.2	43.5	41.3	**47.1**
Scidocs	15.8	16.6	7.7	12.2	14.9	14.3	14.5	15.8	16.5	**17.1**
FEVER	75.3	**81.9**	56.2	66.9	70.0	66.9	77.1	78.6	75.8	**81.9**
Climate-FEVER	21.3	25.3	14.8	19.8	22.8	17.5	18.4	23.5	23.7	**25.8**
Scifact	66.5	68.8	31.8	50.7	64.3	64.4	67.1	**69.3**	67.7	69.2
Avg. w/o CQA	44.0	49.5	26.3	41.3	43.7	43.1	45.1	50.6	47.5	51.2
Avg.	43.0	48.6	25.5	40.5	42.8	42.5	44.4	-	46.6	50.2
Best on	1	3	0	0	0	1	0	1	1	9

Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., & Grave, E. (2021). Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Contriever

Unsupervised Dense Information Retrieval with Contrastive Learning

❖ Experiments

- In-domain 데이터셋에 대한 few-shot 실험

Table 3: **Few-shot retrieval.** Test nDCG@10 after training on a small in-domain training set. We compare BERT and our model, with and without an intermediate fine-tuning step on MS MARCO. Note that our unsupervised pre-training alone outperforms BERT with intermediate MS MARCO fine-tuning.

	Additional data	SciFact	NFCorpus	FiQA
# queries		729	2,590	5,500
BM25	-	66.5	32.5	23.6
BERT	-	75.2	29.9	26.1
Contriever	-	84.0	33.6	36.4
BERT	MS MARCO	80.9	33.2	30.9
Contriever	MS MARCO	**84.8**	**35.8**	**38.1**

Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., & Grave, E. (2021). Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Contriever

Unsupervised Dense Information Retrieval with Contrastive Learning

❖ Experiments – Ablation studies

- Moco vs. in batch: 성능 차이가 거의 없기 때문에 큰 batch size를 요구하지 않는 Moco 선택

Table 6: **MoCo vs. in-batch negatives.** In this table, we report nDCG@10 on the BEIR benchmark for in-batch negatives and MoCo, without fine-tuning on the MS MARCO dataset.

	NFCorpus	NQ	FiQA	ArguAna	Quora	DBPedia	SciDocs	FEVER	AVG
MoCo	26.2	13.1	13.7	33.0	69.5	20.0	11.9	57.6	30.1
In-batch negatives	24.2	21.6	13.0	33.7	74.9	17.9	13.6	56.1	31.9

Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., & Grave, E. (2021). Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Contriever

Unsupervised Dense Information Retrieval with Contrastive Learning

❖ Experiments – Ablation studies

- 대체적으로 negatives 의 수가 클수록 성능 향상이 이루어짐

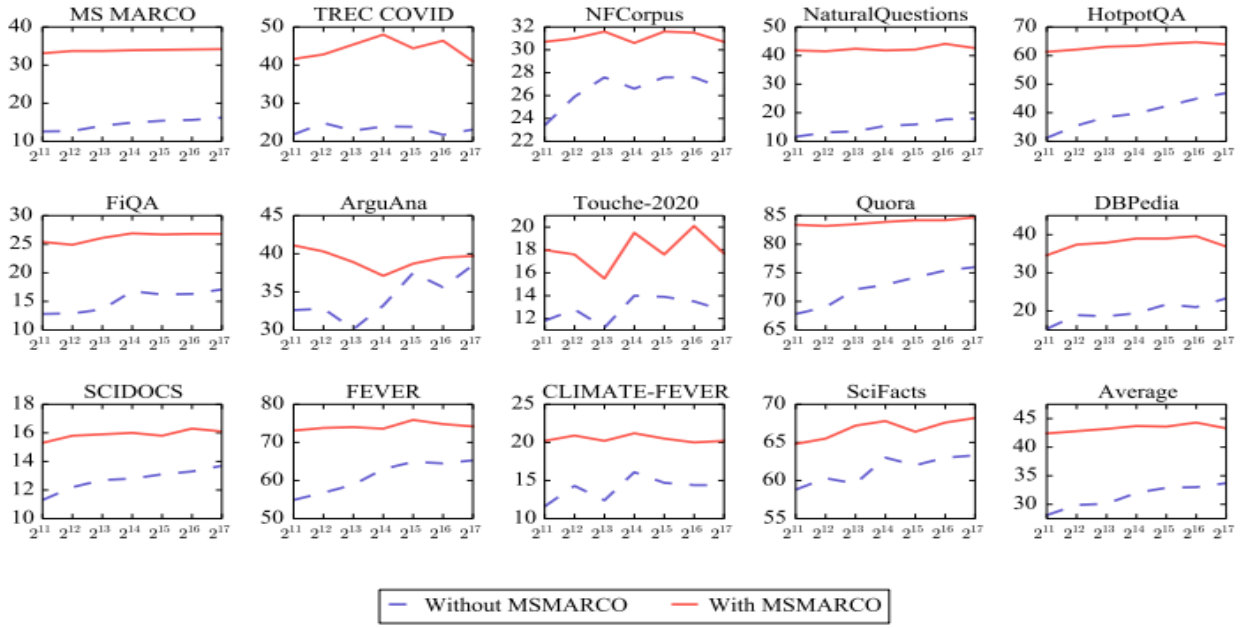


Figure 2: **Impact of the number of negatives.** We report nDCG@10 as a function of the queue size, with and without fine-tuning on MS MARCO. We report numbers using the MoCo framework where the keys for the negatives are computed with the momentum encoder and stored in a queue.

Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., & Grave, E. (2021). Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

PAIR

Paper

❖ PAIR: Leveraging Passage-Centric Similarity Relation for Improving Dense Passage Retrieval (2022, ACL)

- Query-centric & Passage-centric을 모두 고려

PAIR: Leveraging Passage-Centric Similarity Relation for Improving Dense Passage Retrieval

Ruiyang Ren[1,3*†], Shangwen Lv[2*], Yingqi Qu[2], Jing Liu[2‡], Wayne Xin Zhao[3,4‡]
Qiaoqiao She[2], Hua Wu[2], Haifeng Wang[2] and Ji-Rong Wen[3,4]
[1]School of Information, Renmin University of China; [2]Baidu Inc.
[3]Beijing Key Laboratory of Big Data Management and Analysis Methods
[4]Gaoling School of Artificial Intelligence, Renmin University of China
{reyon.ren, jrwen}@ruc.edu.cn, batmanfly@gmail.com
{lvshangwen, quyingqi, liujing46, sheqiaoqiao, wu_hua, wanghaifeng}@baidu.com

Abstract

Recently, dense passage retrieval has become a mainstream approach to finding relevant information in various natural language processing tasks. A number of studies have been devoted to improving the widely adopted dual-encoder architecture. However, most of the previous studies only consider query-centric similarity relation when learning the dual-encoder retriever. In order to capture more comprehensive similarity relations, we propose a novel approach that leverages both query-centric and **PA**ssage-centric s**I**milarity **R**elations (called **PAIR**) for dense passage retrieval. To implement our approach, we make three major technical contributions by introducing formal formulations of the two kinds of similarity relations, generating high-quality pseudo labeled data via knowledge distillation, and designing an effective two-stage training procedure that incorporates passage-centric similarity relation constraint. Extensive experiments show that our approach significantly outperforms previous state-of-the-art models on both MSMARCO and Natural Questions datasets[1].

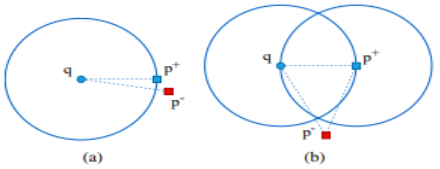


Figure 1: An illustrative case of a query q, its positive passage p^+ and negative passage p^-: (a) Query-centric similarity relation enforces $s(q, p^+) > s(q, p^-)$; (b) Passage-centric similarity relation further enforces $s(p^+, q) > s(p^+, p^-)$, where $s(p^+, q) = s(q, p^+)$. We use the distance (*i.e.*, dissimilarity) for visualization: the longer the distance is, the less similar it is.

ing question answering (Lee et al., 2019; Xiong et al., 2020b), information retrieval (Luan et al., 2021; Khattab and Zaharia, 2020), dialogue (Ji et al., 2014; Henderson et al., 2017) and entity linking (Gillick et al., 2019; Wu et al., 2020).

Typically, the dual-encoder architecture is used

Ren, R., Lv, S., Qu, Y., Liu, J., Zhao, W. X., She, Q., ... & Wen, J. R. (2021). PAIR: Leveraging passage-centric similarity relation for improving dense passage retrieval. *arXiv preprint arXiv:2108.06027*.

DMQA

PAIR

PAIR: Leveraging Passage-Centric Similarity Relation for Improving Dense Passage Retrieval

❖ 연구 배경

- Query-centric 정보 만 사용할 경우 → positive passage와 negative passage를 구별하기 어려워짐

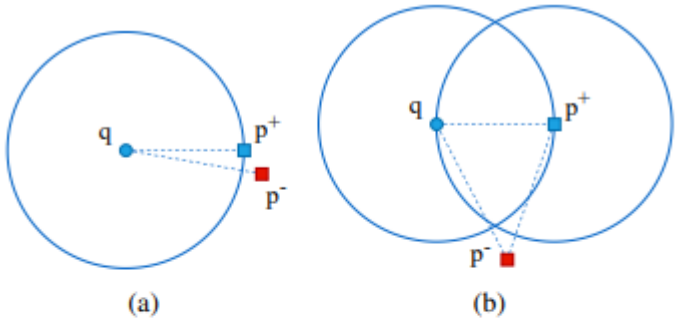


Figure 1: An illustrative case of a query q, its positive passage p^+ and negative passage p^-: (a) Query-centric similarity relation enforces $s(q, p^+) > s(q, p^-)$; (b) Passage-centric similarity relation further enforces $s(p^+, q) > s(p^+, p^-)$, where $s(p^+, q) = s(q, p^+)$. We use the distance (*i.e.*, dissimilarity) for visualization: the longer the distance is, the less similar it is.

$$s(q,p) = E_Q(q)^T \cdot E_P(p)$$

$$s^Q(q,p^+) > s^Q(q,p^-)$$

$$s^P(p^+,q) > s^P(p^+,p^-)$$

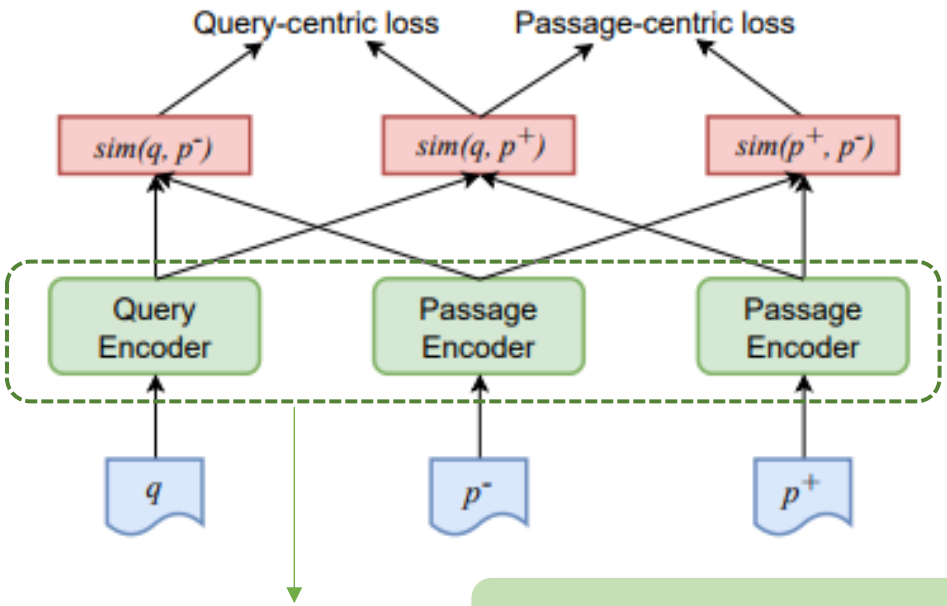
Ren, R., Lv, S., Qu, Y., Liu, J., Zhao, W. X., She, Q., ... & Wen, J. R. (2021). PAIR: Leveraging passage-centric similarity relation for improving dense passage retrieval. *arXiv preprint arXiv:2108.06027*.

PAIR

PAIR: Leveraging Passage-Centric Similarity Relation for Improving Dense Passage Retrieval

❖ PAssage-centric sImilarity Relations (PAIR)

- Query-centric & Passage-centric 정보를 같이 사용하여 Contrastive learning 수행



$$L_Q = -\frac{1}{N}\sum_{<q,p+>} \log \frac{e^{s^Q(q,p^+)}}{e^{s^Q(q,p^+)} + \sum_{p^-} e^{s^Q(q,p^-)}}$$

$$L_P = -\frac{1}{N}\sum_{<q,p+>} \log \frac{e^{P(p^+,q)}}{e^{P(p^+,q)} + \sum_{p^-} e^{s^P(p^+,p^-)}}$$

$$\boldsymbol{L = (1-\alpha) \times L_Q + \alpha \times L_P}$$

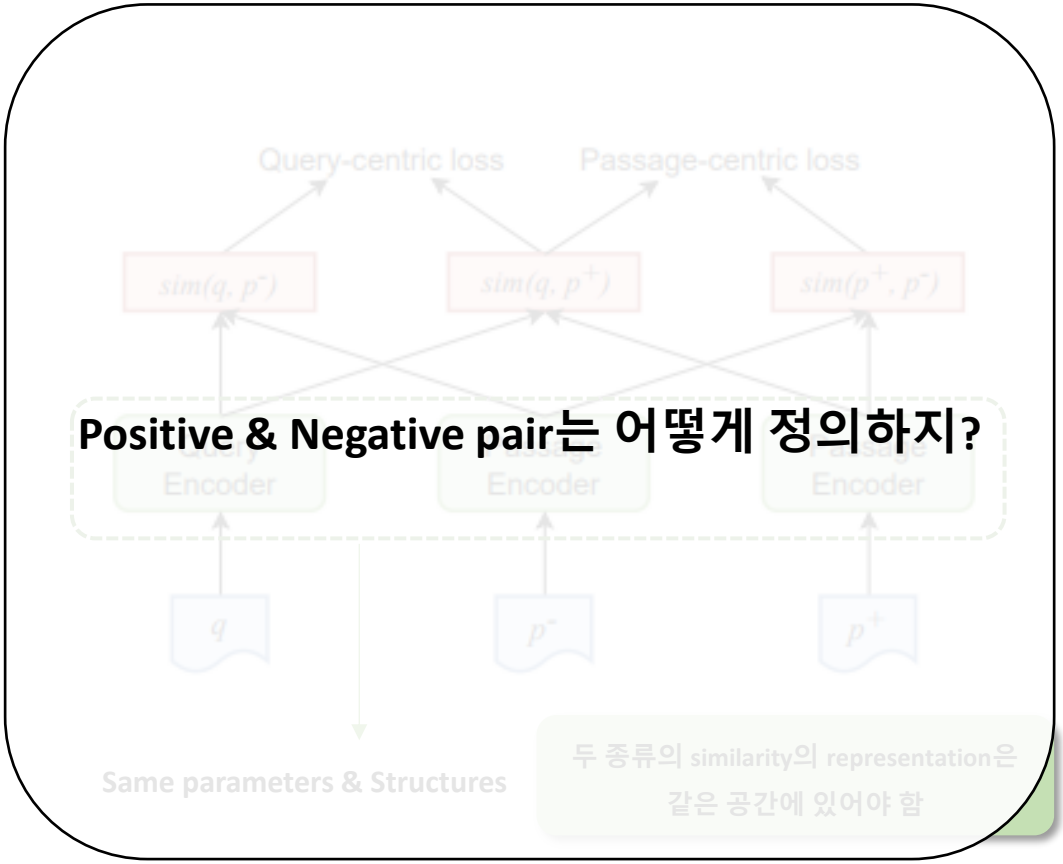
Ren, R., Lv, S., Qu, Y., Liu, J., Zhao, W. X., She, Q., ... & Wen, J. R. (2021). PAIR: Leveraging passage-centric similarity relation for improving dense passage retrieval. *arXiv preprint arXiv:2108.06027*.

PAIR

PAIR: Leveraging Passage-Centric Similarity Relation for Improving Dense Passage Retrieval

❖ **PAssage-centric sImilarity Relations (PAIR)**

- Query-centric & Passage-centric 정보를 같이 사용하여 Contrastive learning 수행



$$L_Q = -\frac{1}{N}\sum_{<q,p+>} \log \frac{e^{s^Q(q,p^+)}}{e^{s^Q(q,p^+)} + \sum_{p^-} e^{s^Q(q,p^-)}}$$

$$L_P = -\frac{1}{N}\sum_{<q,p+>} \log \frac{e^{P(p^+,q)}}{e^{P(p^+,q)} + \sum_{p^-} e^{s^P(p^+,p^-)}}$$

$$\boldsymbol{L = (1-\alpha) \times L_Q + \alpha \times L_P}$$

Ren, R., Lv, S., Qu, Y., Liu, J., Zhao, W. X., She, Q., ... & Wen, J. R. (2021). PAIR: Leveraging passage-centric similarity relation for improving dense passage retrieval. *arXiv preprint arXiv:2108.06027*.

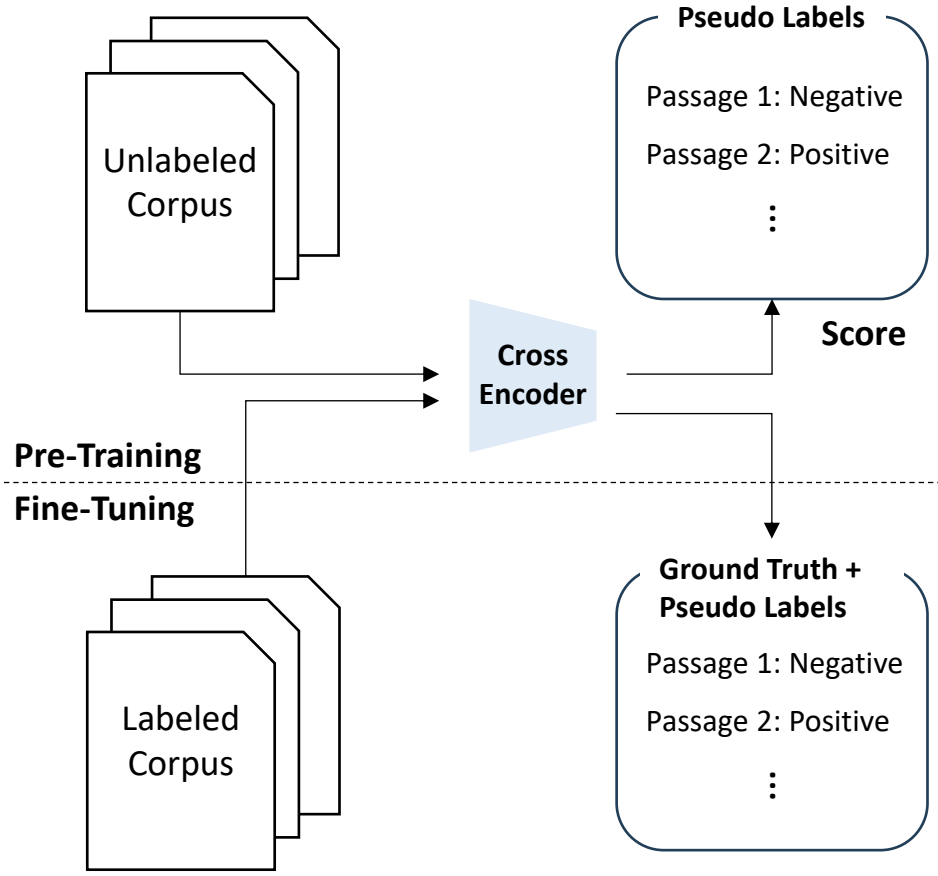
DMQA

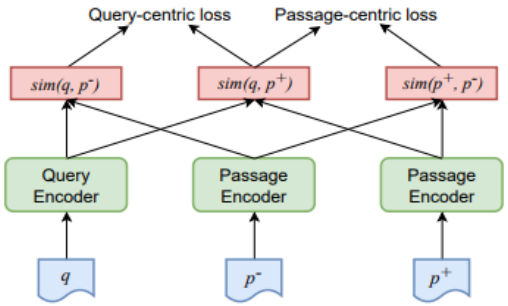
PAIR

PAIR: Leveraging Passage-Centric Similarity Relation for Improving Dense Passage Retrieval

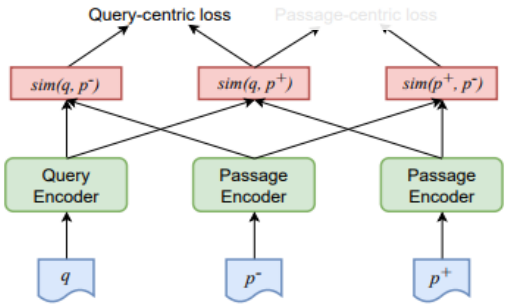
❖ Two-stage Training Procedure

$$L = (1 - \alpha) \times L_Q + \alpha \times L_P$$





L_q

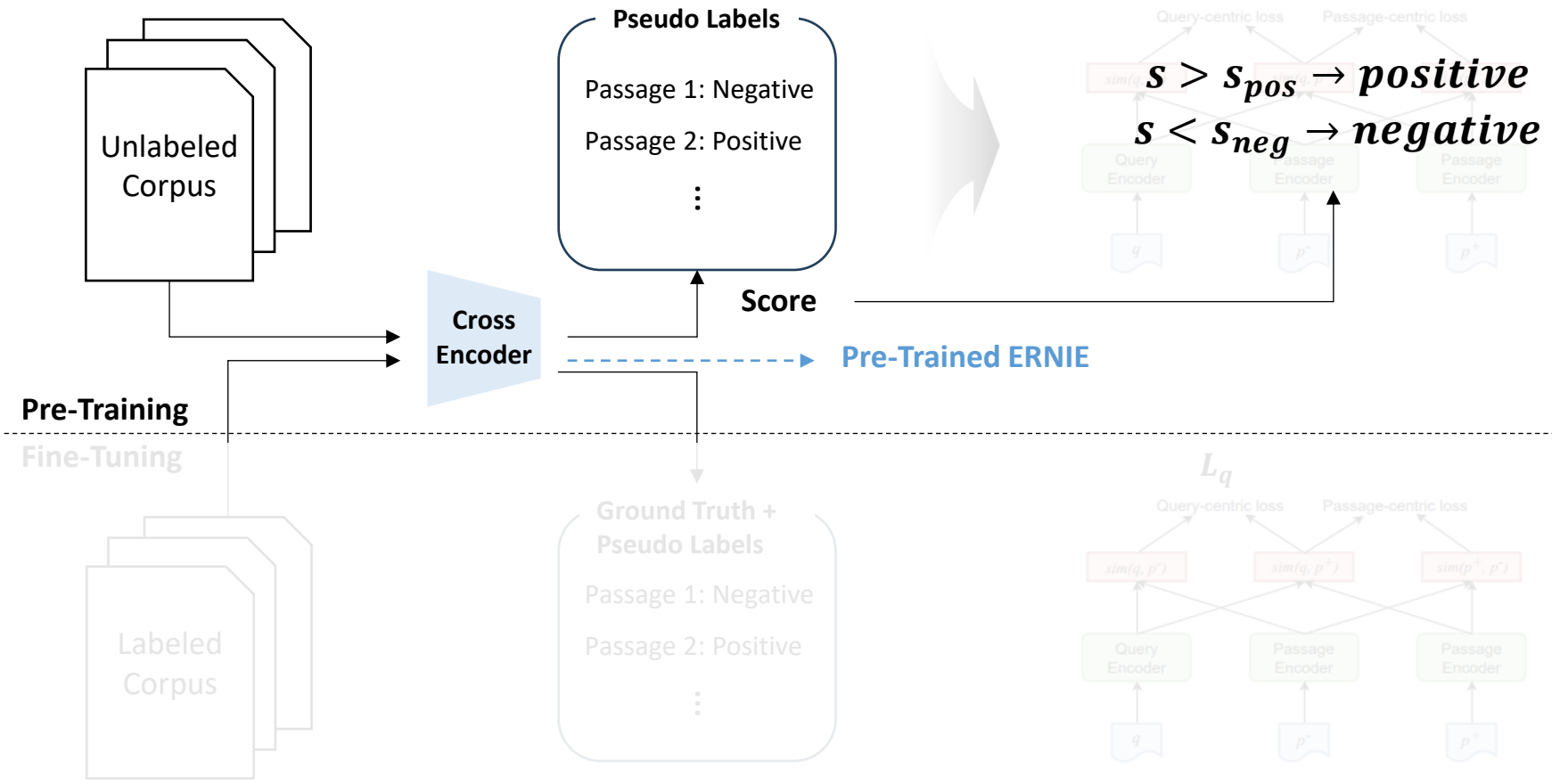


DMQA

PAIR

PAIR: Leveraging Passage-Centric Similarity Relation for Improving Dense Passage Retrieval

❖ Two-stage Training Procedure



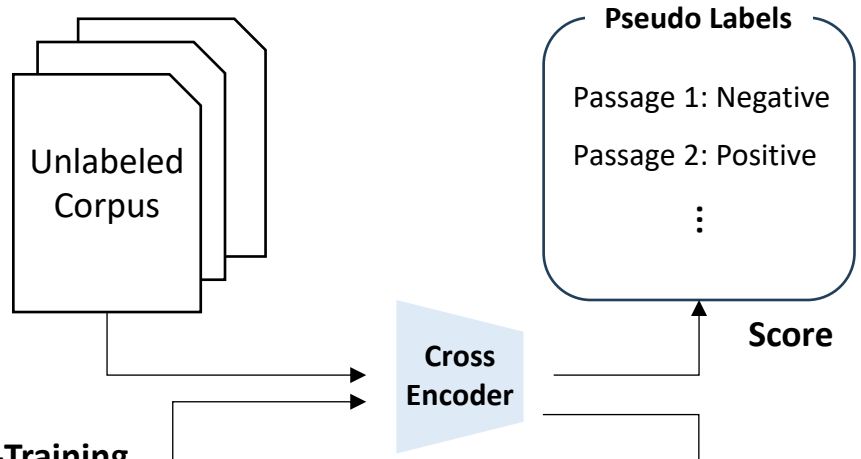
DMQA

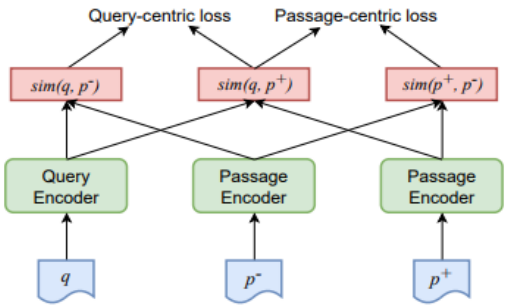
PAIR

PAIR: Leveraging Passage-Centric Similarity Relation for Improving Dense Passage Retrieval

❖ Two-stage Training Procedure

$$L = (1 - \alpha) \times L_Q + \alpha \times L_P$$



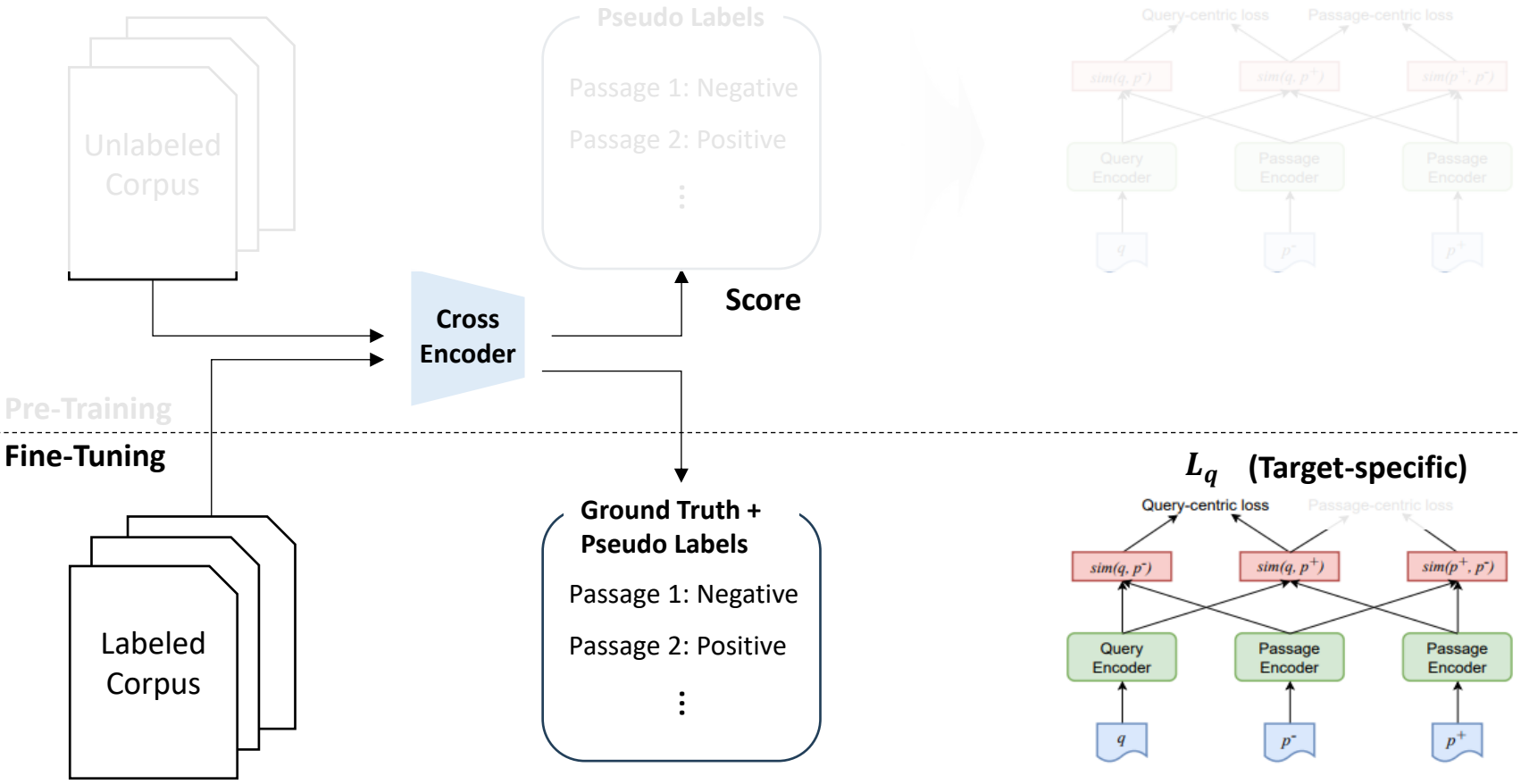


DMQA

PAIR

PAIR: Leveraging Passage-Centric Similarity Relation for Improving Dense Passage Retrieval

❖ Two-stage Training Procedure



PAIR

PAIR: Leveraging Passage-Centric Similarity Relation for Improving Dense Passage Retrieval

❖ Experiments

Methods	PLM	MSMARCO Dev			Natural Questions Test		
		MRR@10	R@50	R@1000	R@5	R@20	R@100
BM25 (anserini) (Yang et al., 2017)	-	18.7	59.2	85.7	-	59.1	73.7
doc2query (Nogueira et al., 2019b)	-	21.5	64.4	89.1	-	-	-
DeepCT (Dai and Callan, 2019)	-	24.3	69.0	91.0	-	-	-
docTTTTTquery (Nogueira et al., 2019a)	-	27.7	75.6	94.7	-	-	-
GAR (Mao et al., 2020)	-	-	-	-	-	74.4	85.3
DPR (single) (Karpukhin et al., 2020)	$BERT_{base}$	-	-	-	-	78.4	85.4
DPR-E	$ERNIE_{base}$	32.5	82.2	97.3	68.4	80.7	87.3
ANCE (single) (Xiong et al., 2020a)	$RoBERTa_{base}$	33.0	-	95.9	-	81.9	87.5
ME-BERT (Luan et al., 2021)	$BERT_{large}$	34.3	-	-	-	-	-
NPRINC (Lu et al., 2020)	$BERT_{base}$	31.1	-	97.7	73.3	82.8	88.4
ColBERT (Khattab and Zaharia, 2020)	$BERT_{base}$	36.0	82.9	96.8	-	-	-
RocketQA (Qu et al., 2020)	$ERNIE_{base}$	37.0	85.5	97.9	74.0	82.7	88.5
PAIR (Ours)	$ERNIE_{base}$	**37.9**	**86.4**	**98.2**	**74.9**	**83.5**	**89.1**

Ren, R., Lv, S., Qu, Y., Liu, J., Zhao, W. X., She, Q., ... & Wen, J. R. (2021). PAIR: Leveraging passage-centric similarity relation for improving dense passage retrieval. *arXiv preprint arXiv:2108.06027*.

PAIR

PAIR: Leveraging Passage-Centric Similarity Relation for Improving Dense Passage Retrieval

❖ Experiments – Ablation study

- 실험 결과적으로는 knowledge distillation (pre-training)의 효과가 가장 크다는 것을 확인
  - ➢ w/o PSR: passage-centric relation 제외 (pre-training 시)
  - ➢ w/o KD: knowledge distillation 제외 (labeled data만 사용)
  - ➢ w/ PSR FT: fine-tuning 시 passage-centric relation 추가 사용
  - ➢ w/o SP: separate encoders
  - ➢ w/o PT: pre-training 제외

Methods	R@5	R@20	R@100
Complete (PAIR)	**74.9**	**83.5**	**89.1**
w/o PSR	73.6	83.3	88.8
w/o KD	70.9	82.7	88.1
w/ PSR FT	74.6	83.4	89.0
w/o SP	74.0	83.4	88.9
w/o PT	73.0	82.8	88.5

Ren, R., Lv, S., Qu, Y., Liu, J., Zhao, W. X., She, Q., ... & Wen, J. R. (2021). PAIR: Leveraging passage-centric similarity relation for improving dense passage retrieval. *arXiv preprint arXiv:2108.06027*.

PAIR

PAIR: Leveraging Passage-Centric Similarity Relation for Improving Dense Passage Retrieval

❖ Experiments – 정성적 평가

- Passage-centric relationship을 고려 하지 않을 때 부적절한 passage를 회수

Query	Top 1 passage retrieved by PAIR (correct)	Top 1 passage retrieved by PAIR$_{\neg PSR}$ (incorrect)
Which animal is the carrier of the **H1N1** virus ?	**H1N1** strains caused a small percentage of all human flu infections in 2004–2005. Other strains of **H1N1** are endemic *in pigs* (swine influenza) and in birds (avian influenza) …	**H5N1** is a subtype virus which can cause illness in humans and many other animal species. A bird-adapted strain of **H5N1**, called HPAIA (**H5N1**) for …
Where is **gall bladder** situated in human body?	The **gall bladder** is a small hollow organ where bile is stored …In humans, the pear-shaped **gall bladder** lies *beneath the liver*, although the structure and position …	The **urinary bladder** is a hollow muscular organ in humans and some other animals that collects and stores urine from the kidneys before disposal by urination …

Ren, R., Lv, S., Qu, Y., Liu, J., Zhao, W. X., She, Q., ... & Wen, J. R. (2021). PAIR: Leveraging passage-centric similarity relation for improving dense passage retrieval. *arXiv preprint arXiv:2108.06027*.

ReContriever

Paper

❖ Unsupervised Dense Retrieval with Relevance-Aware Contrastive Pre-Training (2023, ACL)

- Pseudo-positive examples들이 부정확할 수 있음을 지적

Unsupervised Dense Retrieval with Relevance-Aware Contrastive Pre-Training

Yibin Lei[1]*, Liang Ding[2]†, Yu Cao[3], Chantong Zan[4], Andrew Yates[1], Dacheng Tao[2]
[1]University of Amsterdam [2]JD Explore Academy
[3]Tencent IEG [4]China University of Petroleum (East China)
{y.lei, a.c.yates}@uva.nl, {liangding.liam, dacheng.tao}@gmail.com
rainyucao@tencent.com, zanct@s.upc.edu.cn

Abstract

Dense retrievers have achieved impressive performance, but their demand for abundant training data limits their application scenarios. Contrastive pre-training, which constructs pseudo-positive examples from unlabeled data, has shown great potential to solve this problem. However, the pseudo-positive examples crafted by data augmentations can be irrelevant. To this end, we propose relevance-aware contrastive learning. It takes the intermediate-trained model itself as an imperfect oracle to estimate the relevance of positive pairs and adaptively weighs the contrastive loss of different pairs according to the estimated relevance. Our method consistently improves the SOTA unsupervised Contriever model (Izacard et al., 2022) on the BEIR and open-domain QA retrieval benchmarks. Further exploration shows that our method can not only beat BM25 after further pre-training on the target corpus but also serves as a good few-shot learner. Our code is publicly available at https://github.com/Yibin-Lei/ReContriever.

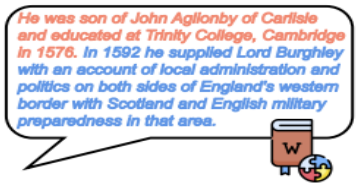


Figure 1: A text snippet from Wikipedia, where two nearby sentences are quite irrelevant. Random cropping may lead to a false positive query-passage pair.

Meanwhile, collecting human-annotated data for new domains is always hard and expensive. Thus improving dense retrievers with limited annotated data becomes essential, considering the significant domain variations of practical retrieval tasks.

Contrastive pre-training, which first generates pseudo-positive examples from a universal corpus and then utilizes them to contrastively pre-train retrievers, has shown impressive performance with-

Lei, Y., Ding, L., Cao, Y., Zan, C., Yates, A., & Tao, D. (2023). Unsupervised Dense Retrieval with Relevance-Aware Contrastive Pre-Training. *arXiv preprint arXiv:2306.03166*.

DMQA

ReContriever

Unsupervised Dense Retrieval with Relevance-Aware Contrastive Pre-Training

❖ 연구 배경

- 과연 pseudo-positive labels은 정확할까? (false positive)

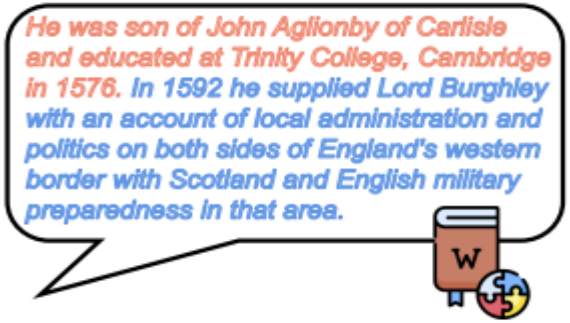
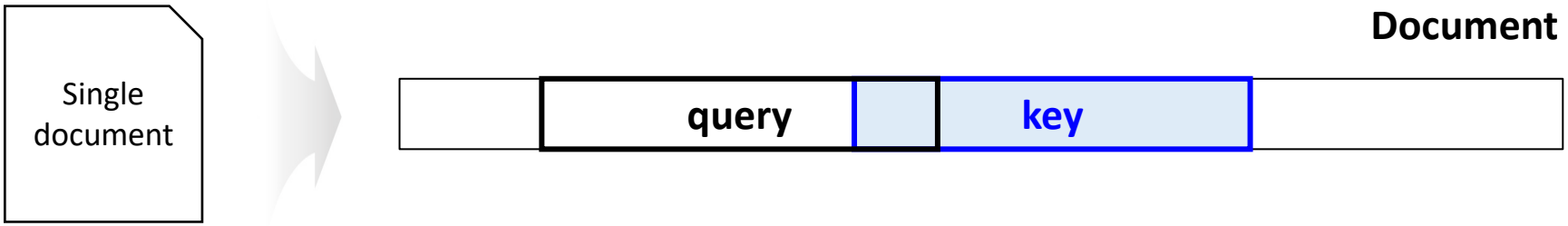


Figure 1: A text snippet from Wikipedia, where two nearby sentences are quite irrelevant. Random cropping may lead to a false positive query-passage pair.



[Independent Cropping in Contriever]

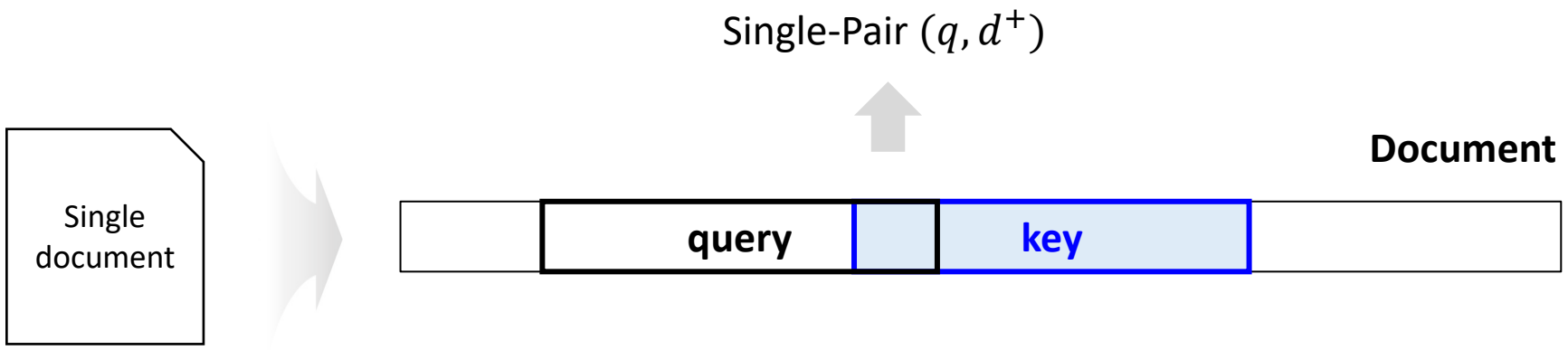
Lei, Y., Ding, L., Cao, Y., Zan, C., Yates, A., & Tao, D. (2023). Unsupervised Dense Retrieval with Relevance-Aware Contrastive Pre-Training. *arXiv preprint arXiv:2306.03166*.

ReContriever

Unsupervised Dense Retrieval with Relevance-Aware Contrastive Pre-Training

❖ One-Document-Multi-Pair

- 다양한 pairs에서 relevance score를 확보하기 위해 multi-pair 생성



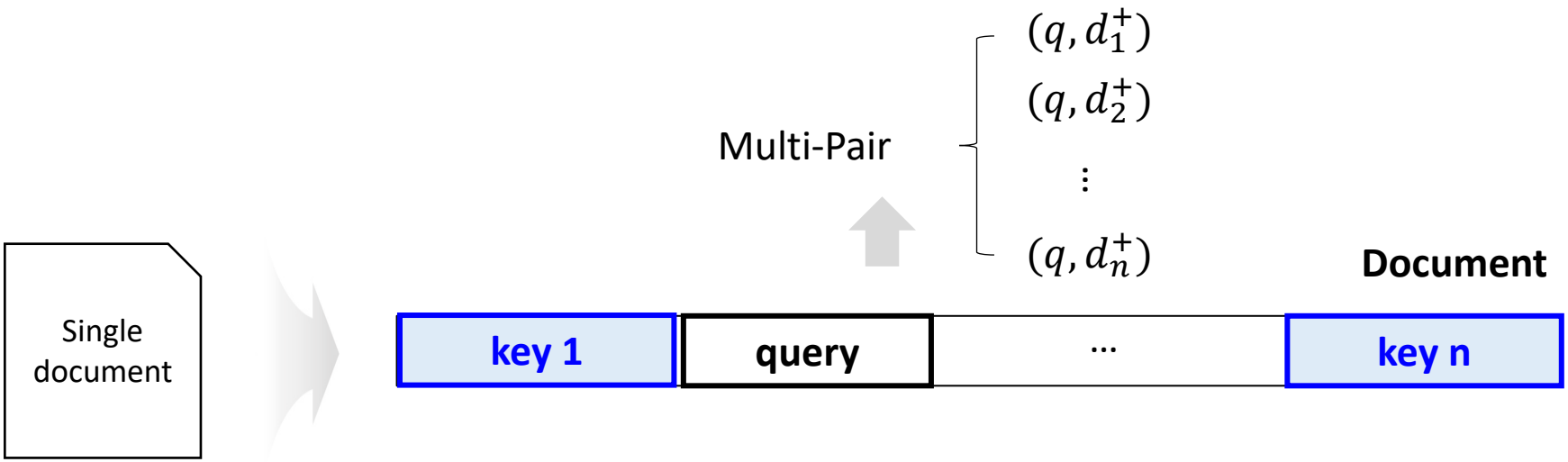
[Single-Pair in Contriever]

ReContriever

Unsupervised Dense Retrieval with Relevance-Aware Contrastive Pre-Training

❖ One-Document-Multi-Pair

- 다양한 pairs에서 relevance score를 확보하기 위해 multi-pair 생성



[Multi-Pair in ReContriever]

ReContriever

Unsupervised Dense Retrieval with Relevance-Aware Contrastive Pre-Training

❖ Relevance-Aware Contrastive Loss

- Relevance 정보를 contrastive loss의 가중치로 사용

$$q \left\{ \begin{matrix} d_1^+ \\ d_2^+ \\ \vdots \\ d_n^+ \end{matrix} \right. \quad \frac{s_\theta(q, d_j^+)}{\sum_{k=1}^{n} s_\theta(q, d_k^+)}$$

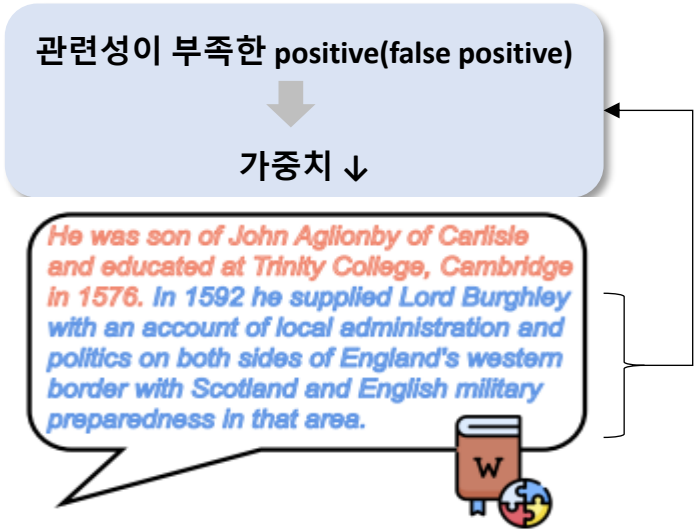


Figure 1: A text snippet from Wikipedia, where two nearby sentences are quite irrelevant. Random cropping may lead to a false positive query-passage pair.

Lei, Y., Ding, L., Cao, Y., Zan, C., Yates, A., & Tao, D. (2023). Unsupervised Dense Retrieval with Relevance-Aware Contrastive Pre-Training. *arXiv preprint arXiv:2306.03166*.

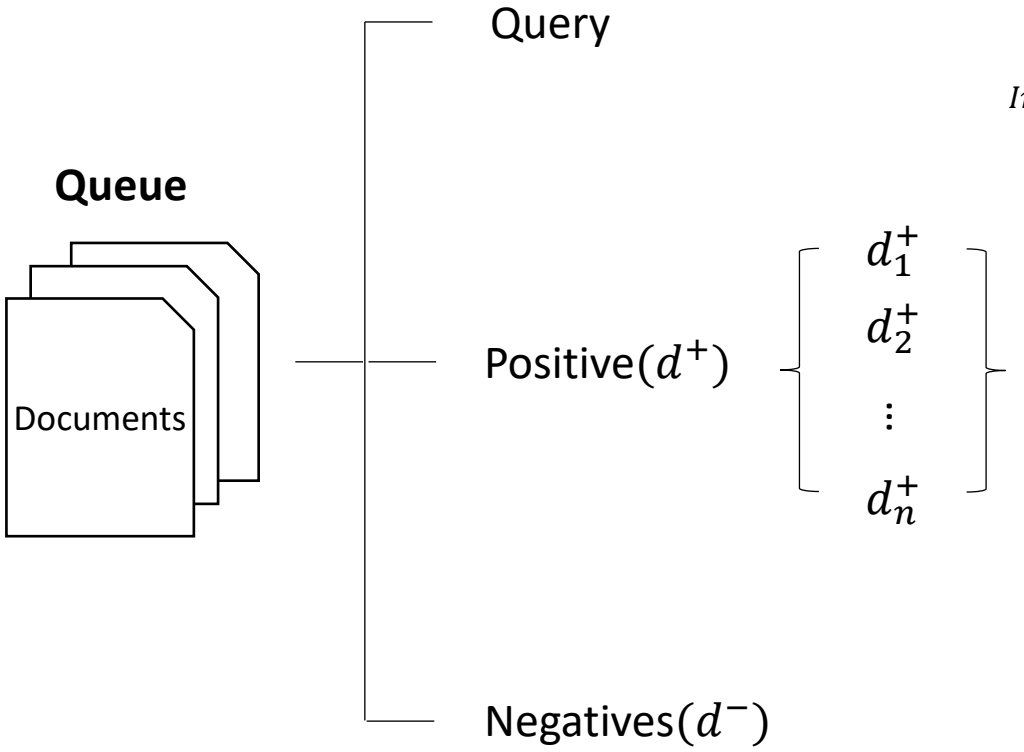
ReContriever

Unsupervised Dense Retrieval with Relevance-Aware Contrastive Pre-Training

❖ Relevance-Aware Contrastive Loss

- Relevance 정보를 contrastive loss의 가중치로 사용

$$Relevance\ Score = \frac{s_\theta(q, d_j^+)}{\sum_{k=1}^{n} s_\theta(q, d_k^+)}$$



$$InfoNCE(q, d) = -\log \frac{\exp\left(\frac{s(q, d^+)}{\gamma}\right)}{\exp\left(\frac{s(q, d^+)}{\gamma}\right) + \sum_{i=1}^{D} \exp\left(\frac{s(q, d_i^-)}{\gamma}\right)}$$

$$L_{relevance} = \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{s_\theta(q_i, d_{ij}^+)}{\sum_{k=1}^{n} s_\theta(q_i, d_{ik}^+)} InfoNCE(q_i, d_{ij}^+)$$

ReContriever

Unsupervised Dense Retrieval with Relevance-Aware Contrastive Pre-Training

❖ Experiments

- BEIR 벤치마크 실험
- BM25 와의 gap을 많이 줄임

DATASET	BM25	BERT	SimCSE	RetroMAE	coCondenser	Contriever	Contriever (reproduced)	ReContriever
MS MARCO	**22.8**	0.6	8.8	4.5	7.7	20.6	21.1	21.8†
Trec-COVID	**65.6**	16.6	38.6	20.4	17.3	27.4	42.0	40.5
NFCorpus	**32.5**	2.5	14.0	15.3	14.4	31.7	30.0	31.9†
NQ	**32.9**	2.7	12.6	3.4	3.9	25.4	29.5	31.0†
HotpotQA	**60.3**	4.9	23.3	25.0	24.4	48.1	44.1	50.1†
FiQA-2018	23.6	1.4	14.8	9.3	5.2	24.5	**26.2**	**26.2**
ArguAna	31.5	23.1	**45.6**	37.6	34.5	37.9	43.4	39.8
Touche-2020	**36.7**	3.4	11.6	1.9	3.0	16.7	16.7	16.6
CQADupStack	**29.9**	2.5	20.2	17.0	9.8	28.4	28.4	28.7†
Quora	78.9	3.9	81.5	69.0	66.7	83.5	83.6	**84.3**†
DBPedia	**31.3**	3.9	13.7	4.6	15.1	29.2	27.6	29.3†
SCIDOCS	**15.8**	2.7	7.4	7.4	1.9	14.9	15.0	15.6†
FEVER	**75.3**	4.9	20.1	7.1	25.3	68.2	66.9	68.9†
Climate-fever	**21.3**	4.1	17.6	4.4	9.8	15.5	15.6	15.6
SciFact	**66.5**	9.8	38.5	53.1	48.1	64.9	65	66.4
Avg	**41.7**	8.7	24.6	18.7	8.9	35.8	37.0	37.8
Avg Rank	**1.9**	7.9	4.9	6.1	6.3	3.4	2.7	2.2

Table 1: **NDCG@10 of BEIR Benchmark**. All models are **unsupervised trained without any human-annotated data**. **Bold** indicates the best result. The average and rank across the entire benchmark are included. Four datasets are excluded because of their licenses. "†" means ReContriever performs significantly better than our reproduced Contriever, as determined by a t-test with p-value 0.05 as threshold.

Lei, Y., Ding, L., Cao, Y., Zan, C., Yates, A., & Tao, D. (2023). Unsupervised Dense Retrieval with Relevance-Aware Contrastive Pre-Training. *arXiv preprint arXiv:2306.03166*.

ReContriever

Unsupervised Dense Retrieval with Relevance-Aware Contrastive Pre-Training

❖ Experiments

- Open-domain 벤치마크 실험

Model	NQ			TriviaQA			WQ		
	Top-5	Top-20	Top-100	Top-5	Top-20	Top-100	Top-5	Top-20	Top-100
Supervised Model									
DPR	-	78.4	85.4	-	79.4	85.0	-	73.2	81.4
Unsupervised Models									
BM25	43.8	62.9	78.3	**66.3**	**76.4**	83.2	41.8	62.4	75.5
RetroMAE	23.0	40.1	58.8	47.0	61.4	74.2	25.8	43.8	62.3
SimCSE	5.4	11.5	23.0	3.7	7.6	17.0	3.3	8.7	19.4
coCondenser	28.9	46.8	63.5	7.5	13.8	24.3	30.2	50.7	68.7
Spider	49.6	68.3	81.2	63.6	75.8	83.5	46.8	65.9	79.7
Contriever	47.3	67.8	80.6	59.5	73.9	82.9	43.5	65.7	80.1
Contriever (reproduced)	48.9	68.3	81.4	61.2	74.6	83.4	47.0	67.0	80.5
ReContriever	**50.3**†	**69.4**†	**82.6**†	63.4†	75.9†	**84.1**†	**48.3**	**68.0**	**81.1**

Table 2: **Recall of open-domain retrieval benchmarks**. **Bold**: the best results across unsupervised models. "†" means ReContriever performs significantly better than our reproduced Contriever, as determined by a t-test with p-value 0.05 as threshold.

Lei, Y., Ding, L., Cao, Y., Zan, C., Yates, A., & Tao, D. (2023). Unsupervised Dense Retrieval with Relevance-Aware Contrastive Pre-Training. *arXiv preprint arXiv:2306.03166*.

ReContriever

Unsupervised Dense Retrieval with Relevance-Aware Contrastive Pre-Training

❖ Experiments

Model	SciF	SCID	Arg	CQA	Avg.
BM25	66.5	15.8	31.5	29.9	35.9
Contriever	64.9	14.9	43.4	28.4	37.9
+ corpus pretrain	66.3	**17.1**	52.4	30.6	41.6$^{\Uparrow+3.7}$
ReContriever	66.4	15.6	39.8	28.4	37.6
+ corpus pretrain	**67.1**	16.6	**54.6**†	**30.7**	**42.3**$^{\Uparrow+4.7}$

Table 3: **NDCG@10 after further pre-training on the target domain corpus**. "⇑" denotes the gains of further pre-training. "†" means ReContriever performs significantly better than our reproduced Contriever.

Model	NQ		
	Top-5	Top-20	Top-100
	Reference		
DPR	-	78.4	85.4
BM25	43.8	62.9	78.3
	8 examples		
Spider	49.7	68.3	81.4
Contriever	51.7	70.6	83.1
ReContriever	52.9	71.6	84.2†
	32 examples		
Spider	50.2	69.4	81.7
Contriever	52.6	70.9	83.1
ReContriever	53.5	71.9†	84.7†
	128 examples		
Spider	57.0	74.3	85.3
Contriever	55.1	72.4	83.7
ReContriever	55.9	74.1†	85.1†

Table 4: **Few-shot Retrieval on NQ**. Results are report with Recall. "†" means ReContriever performs significantly better than our reproduced Contriever.

Lei, Y., Ding, L., Cao, Y., Zan, C., Yates, A., & Tao, D. (2023). Unsupervised Dense Retrieval with Relevance-Aware Contrastive Pre-Training. *arXiv preprint arXiv:2306.03166*.

ReContriever

Unsupervised Dense Retrieval with Relevance-Aware Contrastive Pre-Training

❖ Experiments – Ablation study

- Relevance-aware loss의 효과를 극명히 보여줌

Model	MS MARCO	NFCoprus	NQ	Hotpot	FiQA	Touche	Quora	SCIDOCS	Avg.
Contriever	19.1	25.1	26.7	43.2	23.2	18.6	82.3	14.6	31.6
+ relevance-aware loss	0.2	2.5	0.0	0.2	0.5	0.6	57.6	14.5	9.5
+ one-document-multiple-pair	19.9	29.5	27.5	44.2	21.9	15.9	82.8	14.5	32.0
ReContriever	20.8	28.1	29.6	49.9	23.4	18.2	83.3	14.7	33.5

Table 5: **Ablation Study.** Results are reported with NDCG@10.

Lei, Y., Ding, L., Cao, Y., Zan, C., Yates, A., & Tao, D. (2023). Unsupervised Dense Retrieval with Relevance-Aware Contrastive Pre-Training. *arXiv preprint arXiv:2306.03166*.

Conclusion

Conclusion

❖ Introduction

- Retrieval models는 크게 **contrastive learning** / MLM 방식으로 사전 학습이 이루어짐

❖ Contriever

- Contrastive learning 기반으로 Independent Cropping, MoCo 방식을 사용하여 positive & negative pairs를 정의

❖ PAIR

- Query-centric relationship 뿐만 아니라 passage-centric relationship을 같이 고려한 2 stage 학습
- Fine-tuning 시에는 target-specific query-centric relationship만을 고려

❖ ReContriever

- 기존 Contriever의 false positive 한계점을 극복하고자 relevance aware loss 제안
- Multi-Positive pairs의 relevance score를 통해 Contriever 모델 개선

References

References

- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.
- Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., & Grave, E. (2021). Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.
- Ren, R., Lv, S., Qu, Y., Liu, J., Zhao, W. X., She, Q., ... & Wen, J. R. (2021). PAIR: Leveraging passage-centric similarity relation for improving dense passage retrieval. *arXiv preprint arXiv:2108.06027*.
- Lei, Y., Ding, L., Cao, Y., Zan, C., Yates, A., & Tao, D. (2023). Unsupervised Dense Retrieval with Relevance-Aware Contrastive Pre-Training. *arXiv preprint arXiv:2306.03166*.

고맙습니다