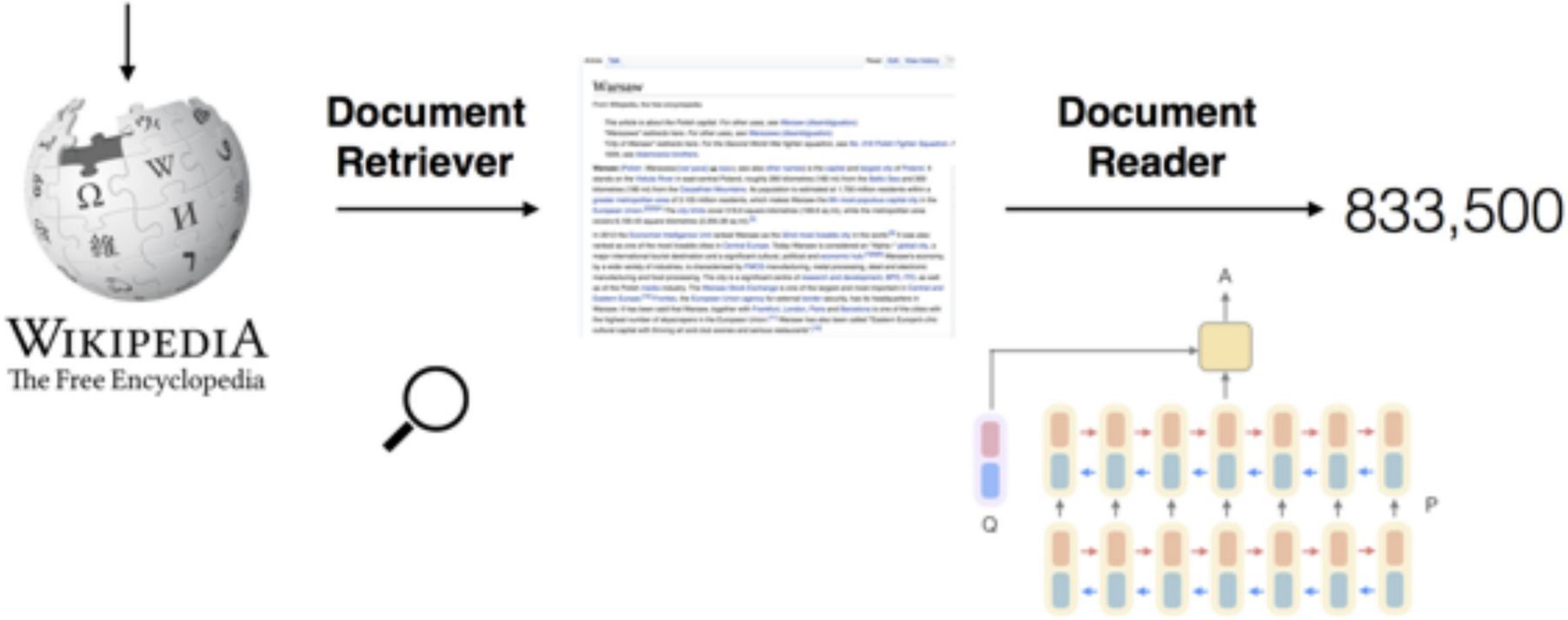
Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval

2022.Sep.24

김기범, Alookso

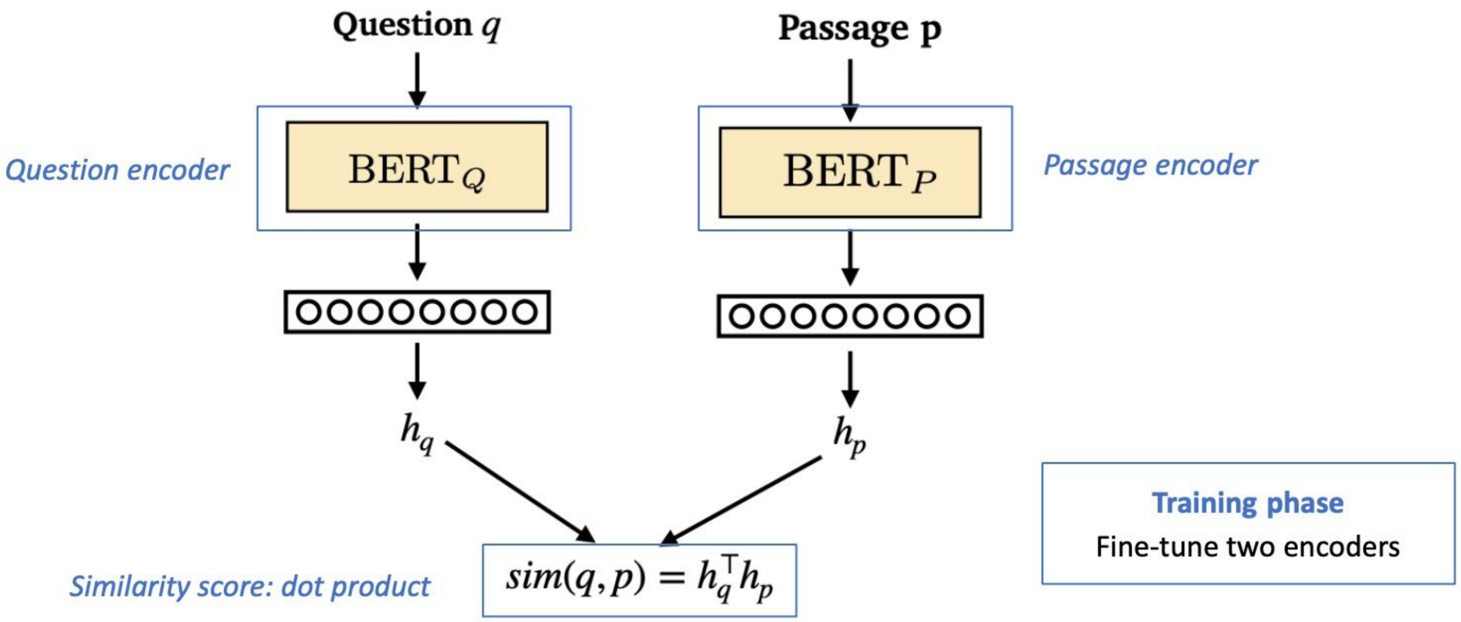
Open domain QA

Q: How many of Warsaw's inhabitants spoke Polish in 1933?



retrieval

- 주어진 문장에 대해 문장 pool/corpus에서 유사도가 높은 문장들을 찾아내는 것.
- Sparse vs Dense retrieval: embedding vector sparseness.

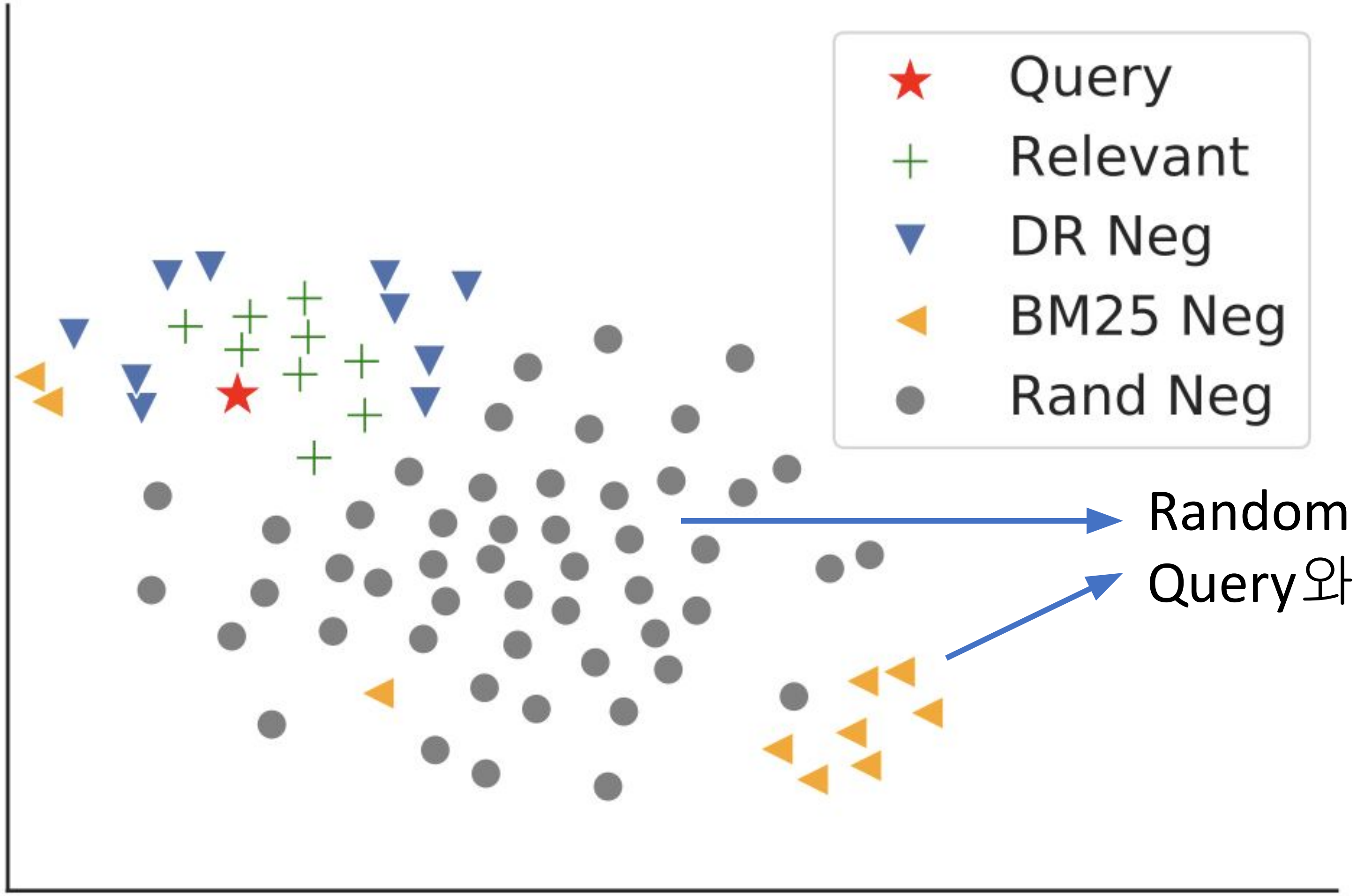


retrieval

- Dense retrieval을 활용한 연구들(1~3)에서는 neural networks를 이용하여 embedding을 기반으로 retrieval을 진행함.
- embedding의 dot product 기반 positive sample과 아래와 같은 negative sample들을 이용해 contrastive learning을 진행.
  - Random: 코퍼스 내의 random한 passage를 뽑는 방법
  - BM25: 코퍼스 내에서 BM25 기준으로 top-k
  - Gold: 학습셋 내의 다른 질의의 positive passage.

1. Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. **Latent retrieval for weakly supervised open domain question answering.** In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6086–6096, 2019.
2. ladimir Karpukhin, Barlas Og˘uz, Sewon Min, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. **Dense passage retrieval for open-domain question answering**. *arXiv preprint arXiv:2004.04906*, 2020.
3. Yi Luan, Jacob Eisenstein, Kristina Toutanove, and Michael Collins. **Sparse, dense, and attentional representations for text retrieval.** *arXiv preprint arXiv:2005.00181*, 2020.

negative sample



Random & BM25:
Query와 너무 멀다!

Negative sample representation의 t-SNE

negative sample

Diminishing Gradients of Uninformative Negatives

- negative samples과 query의 거리가 멀면 loss가 작아진다.
- zero loss를 만드는 negative samples는 gradients를 거의 0으로 만들고, model convergence에 미미한 영향을 미친다.

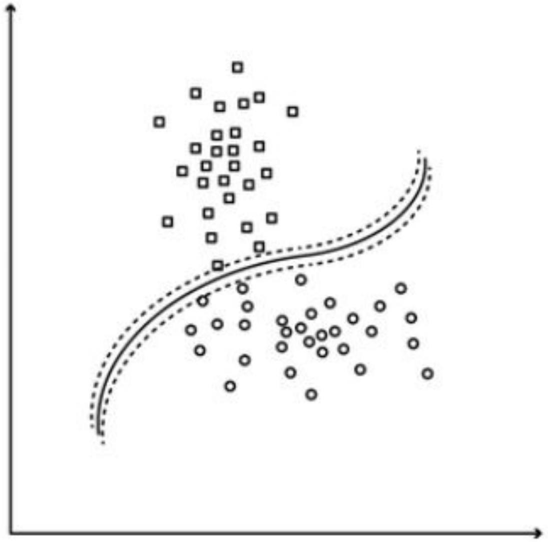
Inefficacy of Local In-Batch Negatives

- 전체 corpus 사이즈에 비해 batch size와 informative negative sample의 수가 적기 때문에, Local In-Batch에는 informative negative sample이 존재할 가능성이 적다.

~ 0

$$-\log \frac{e^{\mathrm{sim}(q_i, p_i^+)}}{e^{\mathrm{sim}(q_i, p_i^+)} + \sum_{j=1}^{n} e^{\mathrm{sim}(q_i, p_{i,j}^-)}}.$$

negative sample



Classification problem에서 decision boundary 근처의 sample들의 정보량이 큰 것과 유사한듯?

negative sample: 정리

In-batch에서 뽑은 negative sample은 정보량이 적을 가능성이 높다.

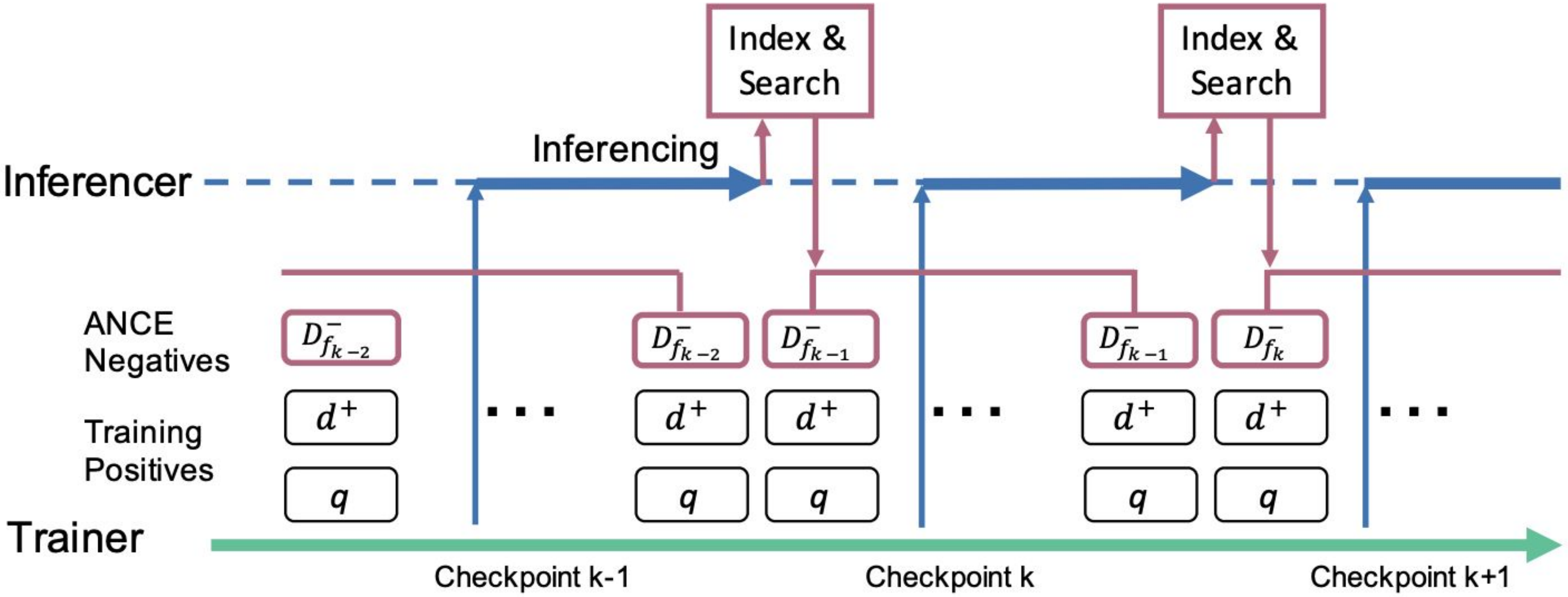
정보량이 적은 negative sample은 학습에 비효율적이다.

In-batch negative를 사용하는 것은 학습에 비효율적이다.

더 정보량이 많은, positive/query와 유사한(구분하기 어려운) negative sample을 사용해야 한다.

ANCE Model

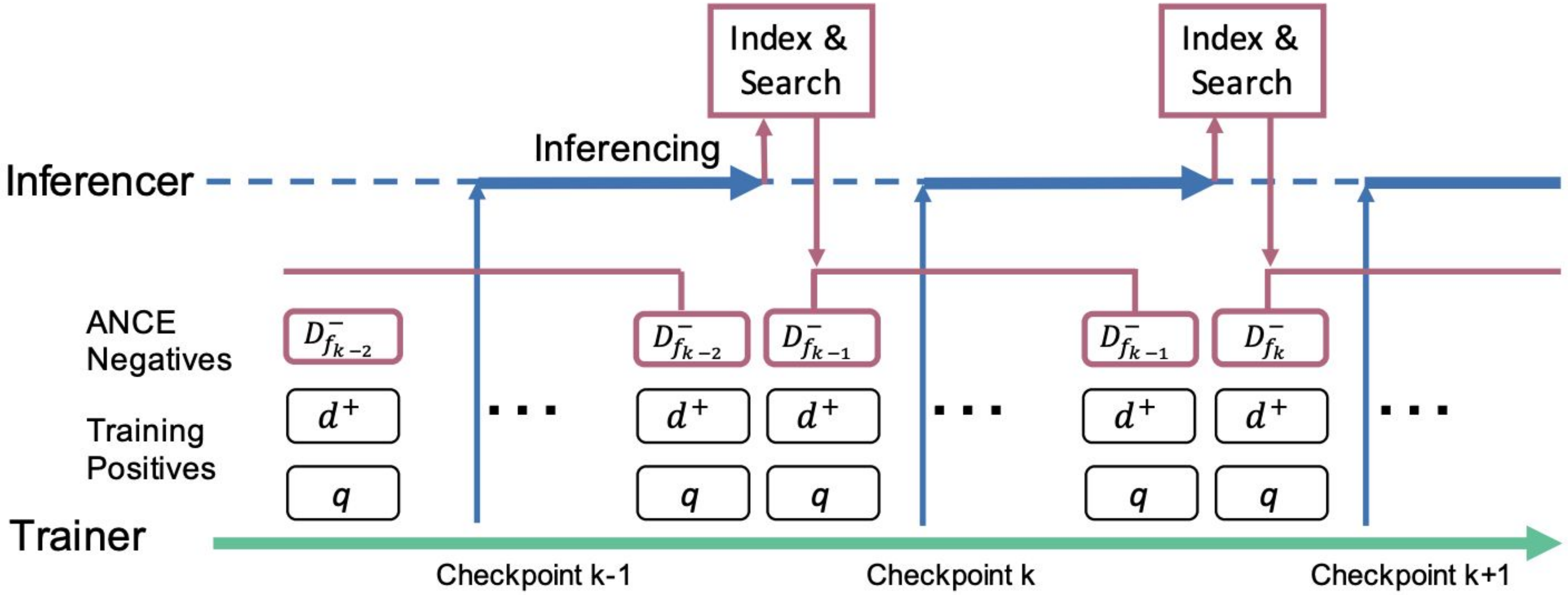
- Approximate nearest neighbor negative contrastive estimation
- Corpus 전체에 대한 ANN(Approximate Nearest Neighbor) index를 사용해 in-batch negative 대신 corpus에서 informative한 negative sample을 추출하는 방법.
- ANN index는 아래 두 단계로 이루어짐.
  - **Inference**: 전체 문서를 encoding
  - **Index**: ANN index를 계산.



ANCE Model

ANN index update와 학습을 비동기적으로 진행함.

- 학습중인 encoder을 이용해 representation 및 ANN Index 계산. (**m-batch마다**)
- 이와 동시에 Trainer는 ANN index로부터 얻은 negative sample을 이용해 retrieval model 학습.



ANCE Model

- ANN index 계산을 위해 벡터 유사도 검색을 GPU 가속으로 빠르게 사용할 수 있는 faiss를 사용.
  - Johnson, Jeff, Matthijs Douze, and Hervé Jégou. "Billion-scale similarity search with gpus." *IEEE Transactions on Big Data* 7.3 (2019): 535-547.

- ANCE는 dense retrieval model에 모두 적용할 수 있으며, 본 연구의 벤치마크를 위해 [Luan et al., 2020]논문 모델을 사용.
  - BERT Siamese/Dual Encoder, dot projuct similarity, negative log likelihood loss

Experiment - Dataset

- Retrieval
  - TREC 2019 Deep Learning Track
    - large scale **retrieval** dataset
    - Bing 검색엔진의 쿼리-관련 문서 레이블링
- OpenQA
  - Natural Question
  - TriviaQA

Experiment: Retrieval

Table 1: Results in TREC 2019 Deep Learning Track. Results not available are marked as “n.a.”, not applicable are marked as “–”. Best results in each category are marked bold.

	MARCO Dev Passage Retrieval		TREC DL Passage NDCG@10		TREC DL Document NDCG@10	
	MRR@10	**Recall@1k**	**Rerank**	**Retrieval**	**Rerank**	**Retrieval**
Sparse & Cascade IR						
BM25	0.240	0.814	–	0.506	–	0.519
Best DeepCT	0.243	n.a.	–	n.a.	–	0.554
Best TREC Trad Retrieval	0.240	n.a.	–	0.554	–	0.549
BERT Reranker	–	–	**0.742**	–	0.646	–
Dense Retrieval						
Rand Neg	0.261	0.949	0.605	0.552	0.615	0.543
NCE Neg	0.256	0.943	0.602	0.539	0.618	0.542
BM25 Neg	0.299	0.928	0.664	0.591	0.626	0.529
DPR (BM25 + Rand Neg)	0.311	0.952	0.653	0.600	0.629	0.557
BM25 $\rightarrow$ Rand	0.280	0.948	0.609	0.576	0.637	0.566
BM25 $\rightarrow$ NCE Neg	0.279	0.942	0.608	0.571	0.638	0.564
BM25 $\rightarrow$ BM25 + Rand	0.306	0.939	0.648	0.591	0.626	0.540
ANCE (FirstP)	**0.330**	**0.959**	0.677	**0.648**	0.641	0.615
ANCE (MaxP)	–	–	–	–	**0.671**	**0.628**

random sampling in batch

random sampling from BM25 top 100

BM25 Warm Up

Experiment: OpenQA

Table 2: Retrieval results (Answer Coverage at Top-20/100) on Natural Questions (NQ) and Trivial QA (TQA) in the setting from Karpukhin et al. (2020).

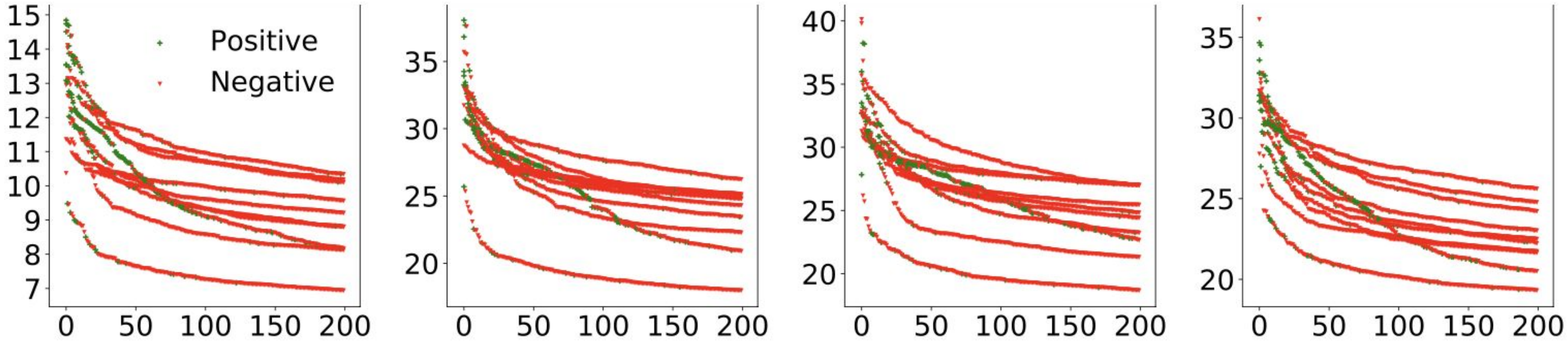
	Single Task		**Multi Task**	
	NQ	**TQA**	**NQ**	**TQA**
Retriever	Top-20/100	Top-20/100	Top-20/100	Top-20/100
BM25	59.1/73.7	66.9/76.7	–/–	–/–
DPR	78.4/85.4	79.4/85.0	79.4/86.0	78.8/84.7
BM25+DPR	76.6/83.8	79.8/84.5	78.0/83.9	79.9/84.4
ANCE	**81.9/87.5**	**80.3/85.3**	**82.1/87.9**	**80.3/85.2**

Experiment: Efficiency

Table 5: Efficiency of ANCE Search and Training.

Operation	Offline	Online
BM25 Index Build	3h	–
BM25 Retrieval	–	37ms
BERT Rerank	–	1.15s
Sparse IR Total (BM25 + BERT)	–	**1.42s**
ANCE Inference		
Encoding of Corpus/Per doc	10h/4.5ms	–
Query Encoding	–	2.6ms
ANN Retrieval (batched q)	–	9ms
Dense Retrieval Total	–	**11.6ms**
ANCE Training		
Encoding of Corpus/Per doc	10h/4.5ms	–
ANN Index Build	10s	–
Neg Construction Per Batch	72ms	–
Back Propagation Per Batch	19ms	–

Experiment



(a) ANCE FirstP (100%) (b) NCE Neg (0%) (c) Rand Neg (0%) (d) BM25+Rand (7%)

Figure 3: The top DR scores for 10 random TREC DL testing queries. The x-axes are their ranking order. The y-axes are their retrieval scores minus corpus average. All models are warmed up by BM25 Neg. The percentages are the overlaps between the testing and training negatives near convergence.

- TREC DL task에 대한 랜덤 쿼리 10개의 retrieval score plot.
- X축: DR score로 정렬된 sample index, y축: (DR score – mean DR score)
- 괄호 안 %는 top 100 highest scored negative sample이 해당 query 결과에 얼마나 포함되어있는지에 대한 비율.

Experiment

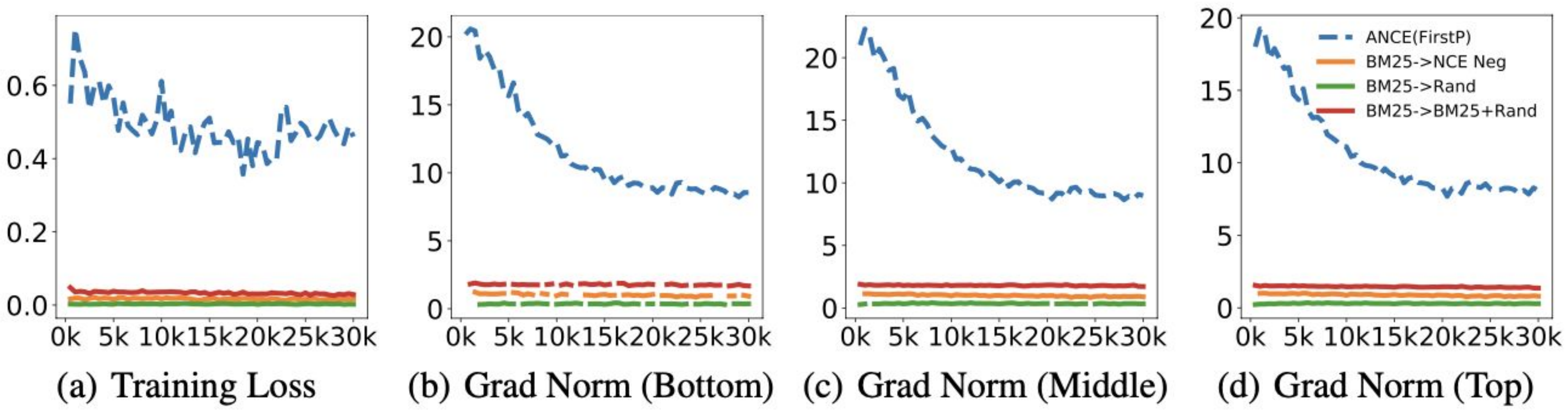


Figure 4: The loss and gradient norms during DR training (after BM25 warm up). The gradient norms are on the bottom (1-4), middle (5-8), and top (9-12) BERT layers. The x-axes are training steps.

- zero loss를 만드는 negative samples는 gradients를 거의 0으로 만들고, model convergence에 미미한 영향을 미친다.