Bridging the Training-Inference Gap for Dense Phrase Retrieval

Gyuwan Kim[1] Jinhuk Lee[2] Barlas Oğuz[3] Wenhan Xiong[3]
Yizhe Zhang[3] Yashar Mehdad[3] William Yang Wang[1]

[1]University of California, Santa Barbara [2]Korea University [3]Meta AI

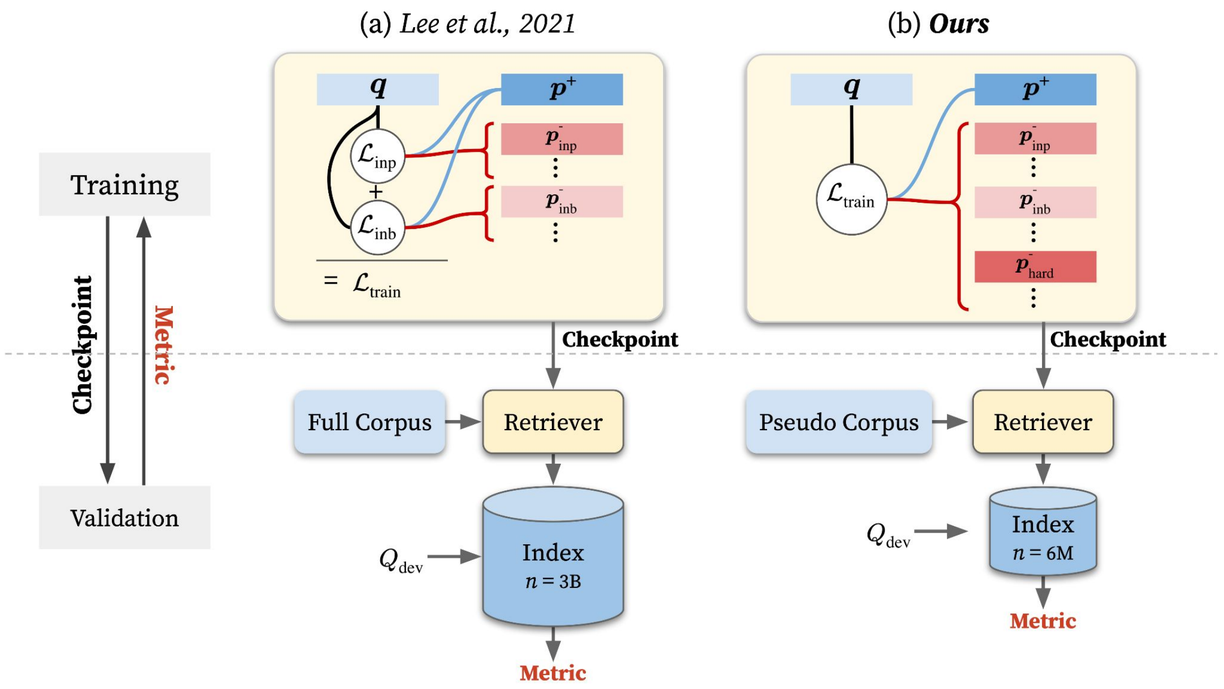
Motivation

- Components to build a dense retrieval system
  - ➢ Training a dual encoder
  - ➢ Selecting the best model with validation
  - ➢ Constructing an index for efficient search

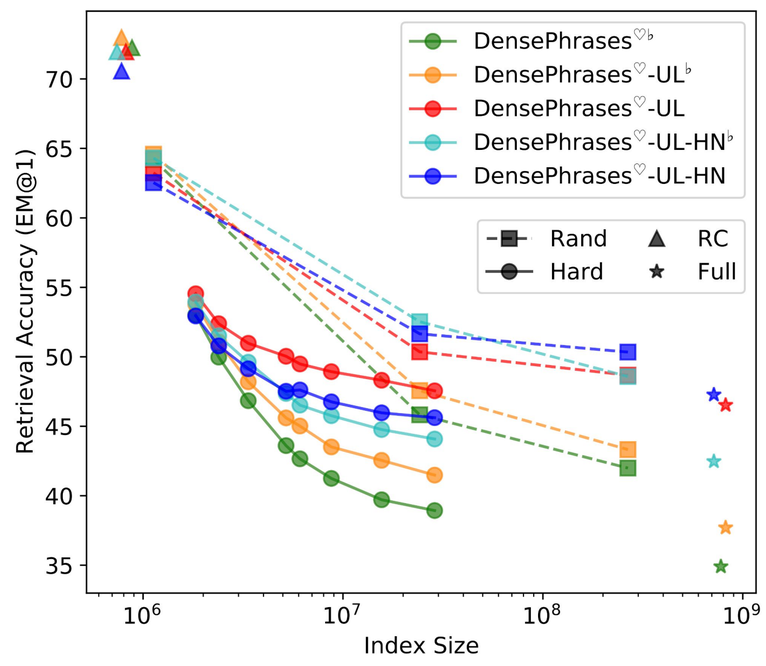
  are loosely connected each other
  e.g., model training does not directly optimize the retrieval performance from the full corpus
- Goal: **minimize the training-inference gap** of dense retrievers to achieve better retrieval performance (focusing on *phrase retrieval*)

Efficient Validation

- To expedite modeling innovation correctly, we measure retrieval accuracy on an index from **a smaller subset of the full corpus** (C)
- C_0: gold passages from the development set (minimal set ensuring to contain answers)
- *Random* Subcorpus (R_r): C_0 + random passages, $|R_r| = r|C|$
- *Hard* Subcorpus (H_k): C_0 + all context passages from top-k retrieval results using a pre-trained dense retriever



Comparison of (a) original and (b) proposed procedure of DensePhrases training (top) and validation (bottom)



Validation results with different size of random (r ∈ {0, 1/100, 1/10}) and hard (k ∈ {1, 2, 4, 8, 10, 16, 32, 64}) subcorpora, ♭ : before query-side fine-tuning

Optimized Training of DensePhrases

- *Unified loss* (UL)
  - ➢ We should find an answer phrase among **all** possible candidates **at once** in test time
  - ➢ Put all negatives together into contrastive targets with different λ coefficients
  - ➢ Use all tokens in context passages
    - # of negatives: in-passage (L-1), in-batch (B-1 → B*L-1), pre-batch (B*T → B*T*L)
- *Hard negatives* (HN)
  - ➢ Fix mistakes from the first round model
  - ➢ Mining: extract **model-based** hard negatives from top-k retrieval results for questions in the training set
  - ➢ Training: **fine-tune a dual encoder** by appending sampled hard negatives as negative targets for each training step

Experiments

- The relative order of accuracy between models on hard subcorpus converges quickly
- Both UL and HN are shown to be effective
- We improves **phrase retrieval** by 2-3% in top-1 accuracy and **passage retrieval** by 2-4% in top-20 accuracy from DensePhrases

Model	NaturalQ	TriviaQA
	EM@1	EM@1
DPR♢ + BERT reader	41.5	56.8
DPR♠ + BERT reader	41.5	56.8
RePAQ♢ (retrieval-only)	41.2	38.8
RePAQ♠ (retrieval-only)	41.7	41.3
DensePhrases♡	40.9	50.7
DensePhrases♠	41.3	53.5
DensePhrases♡-UL	43.5	51.3
DensePhrases♡-UL-HN	**44.0**	47.0
DensePhrases♠-UL	42.4	**55.5**

Model	NaturalQ					TriviaQA				
	Top-1	Top-5	Top-20	MRR@20	P@20	Top-1	Top-5	Top-20	MRR@20	P@20
DPR♢	46.0	68.1	79.8	55.7	16.5	54.4	-	79.4	-	-
DPR♠	44.2	66.8	79.2	54.2	17.7	54.6	70.8	79.5	61.7	30.3
DensePhrases♡	50.1	69.5	79.8	58.7	20.5	-	-	-	-	-
DensePhrases♠	51.1	69.9	78.7	59.3	22.7	62.7	75.0	80.9	68.2	38.4
DensePhrases♡-UL	57.1	75.7	83.7	65.2	22.0	62.0	74.6	80.6	67.6	33.3
DensePhrases♡-UL-HN	**58.6**	75.7	83.4	**66.1**	21.9	60.3	73.3	79.6	66.1	32.3
DensePhrases♠-UL	56.7	**75.9**	**83.8**	65.2	**23.7**	**65.0**	**76.6**	**82.7**	**70.2**	**39.0**

♢: trained on each dataset independently, ♠: trained on multiple datasets, ♡: trained on Natural Questions datasets