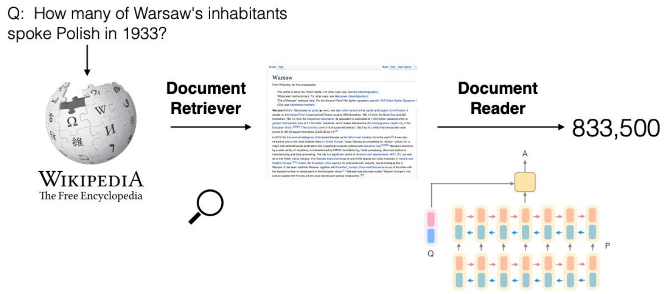
EMNLP 2022

Bridging the Training-Inference Gap for Dense Phrase Retrieval

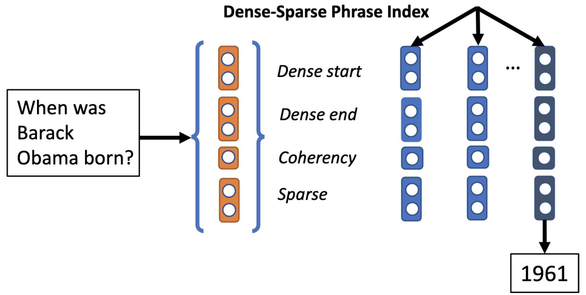
Gyuwan Kim   Jinhuk Lee   Barlas Oğuz   Wenhan Xiong
Yizhe Zhang   Yashar Mehdad   William Yang Wang

Open-Domain Question Answering (ODQA)

- Retriever-reader approach
  (e.g., DrQA, ORQA, REALM, DPR)



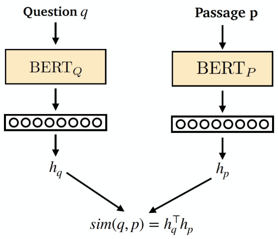
- Retrieval-only approach
  (e.g., DenSPI, Sparc, DensePhrases)



Reading Wikipedia to Answer Open-domain Questions (Chen et al., ACL 2017)
Real-Time Open-Domain Question Answering with Dense-Sparse Phrase Index (Seo et al., ACL 2019)

Problems in Dense Retrieval

- Training a dual encoder → constructing an index for efficient search
- Components of a dense retrieval system (training/validation/indexing/search) are **loosely connected each other**
    - e.g., model training does not directly optimize the retrieval performance from the full corpus
- **Building a large-scale index is expensive**, so even validating dense retrievers from different training objectives is challenging
    - More serious for dense phrase retrieval where the index size is on a billion scale



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

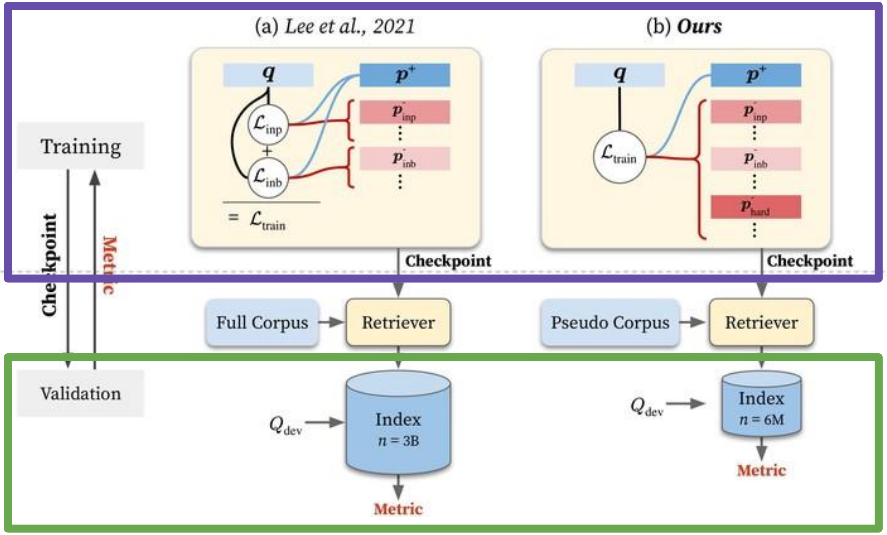
Dense Passage Retrieval for Open-Domain Question Answering (Karpukhin et al., EMNLP 2020)

Outline

- Goal: **minimize the training-inference gap** of dense retrievers to achieve better retrieval performance (focusing on **dense phrase retrieval**)

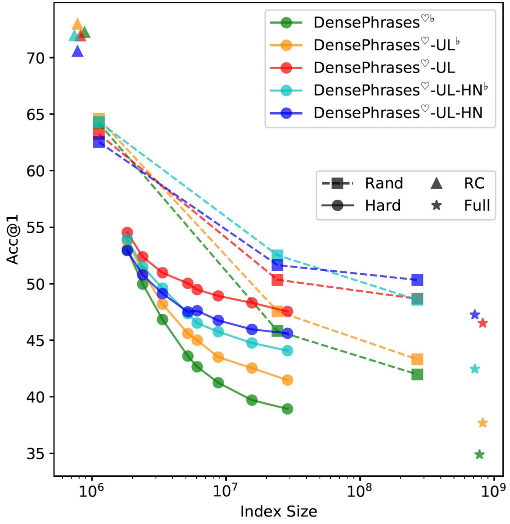
(2) Optimized Training

(1) Efficient Validation



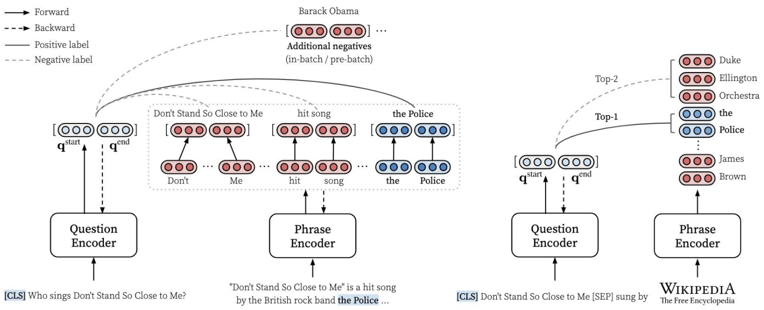
Efficient Validation of Dense Retrievers

- Measure retrieval accuracy on an index from a smaller subset of the full corpus (C)
  - Reading comprehension: corpus of only a single gold passage
  - 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
- The relative order of accuracy between models on hard subcorpus converges faster than random subcorpus



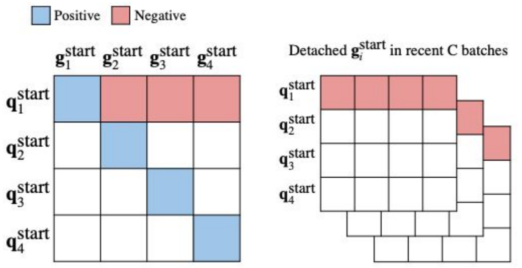
Background: Training of DensePhrases

- Contrastive learning with in-passage and in/pre-batch negatives
- Pre-training with generated question-answer pairs
- Knowledge distillation from a cross encoder
- Query-side fine-tuning



(a) Single-passage training w/ additional negatives

(b) Query-side fine-tuning & Inference



(a) In-batch Negatives ($B-1$)

(b) Pre-batch Negatives ($B \times C$)

$$\mathcal{L}_{\text{train}}^{\text{org}} = -\log \frac{e^{s(q,p^+;c)}}{e^{s(q,p^+;c)} + \sum_{(p^-;c)\in N_{\text{inp}}} e^{s(q,p^-;c)}} - \lambda \log \frac{e^{s(q,p^+;c)}}{e^{s(q,p^+;c)} + \sum_{(p^-;c')\in N_{\text{inb}}\cup N_{\text{prb}}} e^{s(q,p^-;c')}}$$

Learning Dense Representations of Phrases at Scale (Lee et al., ACL 2021)

Optimized Training of DensePhrases

- Unified loss (UL)
  - We should find an answer phrase among all possible candidates at once in the 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 HNs 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

$$\mathcal{L}_{\text{train}}^{\text{org}} = -\log \frac{e^{s(q,p^+;c)}}{e^{s(q,p^+;c)} + \sum_{(p^-;c)\in N_{\text{inp}}} e^{s(q,p^-;c)}} - \lambda \log \frac{e^{s(q,p^+;c)}}{e^{s(q,p^+;c)} + \sum_{(p^-;c')\in N_{\text{inb}}\cup N_{\text{prb}}} e^{s(q,p^-;c')}}$$

$$\mathcal{L}_{\text{train}} = -\log \frac{e^{s(q,p^+;c)}}{e^{s(q,p^+;c)} + \sum_{(p^-;c')\in N_{\text{inp}}\cup N_{\text{inb}}\cup N_{\text{prb}}\cup N_{\text{hard}}} \lambda(p^-) e^{s(q,p^-;c')}}$$

ODQA Experiment Results

- Both **unified loss** (UL) and **hard negatives** (HN) are shown to be effective
- Improving **passage retrieval** by 2~4 points in top-20 accuracy and **phrase retrieval** by 2~3 points in top-1 accuracy from the original DensePhrases

Phrase Retrieval

Model	NQ	TQA
	Acc@1	Acc@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**

Passage Retrieval

Model	Natural Questions					TriviaQA				
	Acc@1	Acc@5	Acc@20	MRR@20	P@20	Acc@1	Acc@5	Acc@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

Conclusion

- Develop an **efficient validation** metric measuring retrieval accuracy on a **subcorpus with hard passages** from a pre-trained dense retriever
- Optimize **training of dense phrase retrieval** using **unified loss** and **hard negatives** by bridging the training-inference gap
- Significantly improve phrase/passage retrieval accuracy from the original DensePhrases in open-domain question answering
- Encourage more works on dense phrase retrieval with an efficient development cycle