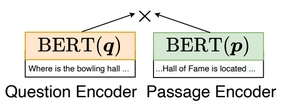
Questions Are All You Need to Train a Dense Passage Retriever

Devendra Singh Sachan[1,2], Mike Lewis[3], Dani Yogatama[4], Luke Zettlemoyer[3,5], Joelle Pineau[1,2,3], Manzil Zaheer[6]
[1]McGill University, [2]Mila-Quebec AI Institute, [3]Meta AI, [4]USC, [5]University of Washington, [6]Google DeepMind
sachande@mila.quebec

Introduction

Dense Passage Retriever (DPR)



- Embeds question and passages in the same embedding space.
- Retrieved passages are very useful for tasks like *question answering.*

Our Approach to train DPR

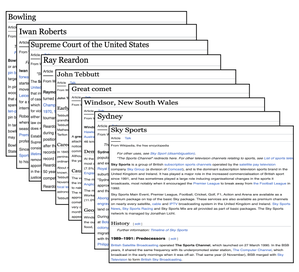
Training Data =

When was the last time the detroit lions won a championship?

Who played general chang in star trek 6?

Where did the rule of 72 come from?

Collection of questions



Database of documents (evidence)

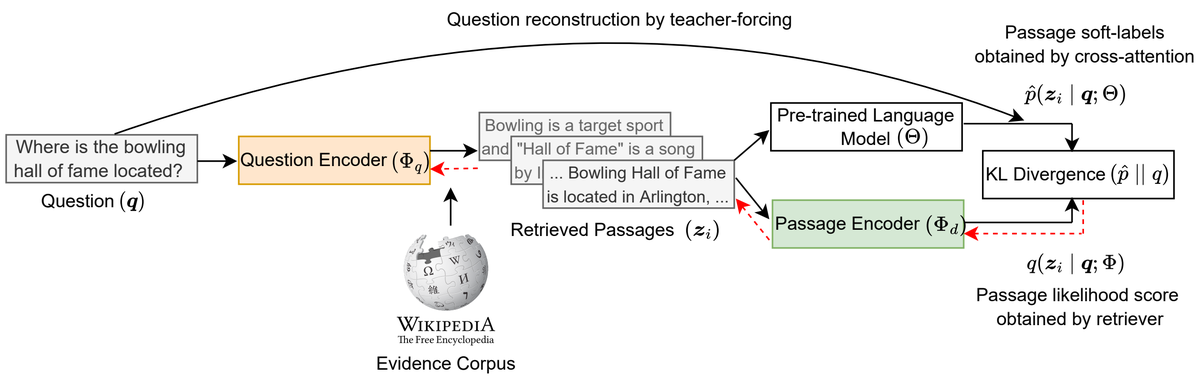
Ex: English Wikipedia ~ 21M passages

Key Idea

Leverage knowledge contained in **Large Language Models** to score candidate passages.

Method

ART: Autoencoding-based Retriever Training



Step 1: K-NN Search over Evidence

- Compute question similarity with all evidence passages
- Select top-K (such as 32) passages with highest scores $\mathcal{Z} = \{z_1, \ldots, z_K\}$

Step 2: Retriever Likelihood Calculation

- Calculate scores using "*current*" passage encoder weights

$$s(q, z_i; \Phi) = f_q(q; \Phi_q)^\top f_d(z_i; \Phi_d)$$

- Define retriever distribution

$$q(z_i \mid q, \mathcal{Z}; \Phi) = \text{softmax}\ s(q, z_i), \forall z_i$$

Step 3: Zero-Shot Relevance Score Estimation

- Use a **LLM** to score question tokens conditioned on each passage (teacher-forcing)

$$p(z_i \mid q; \Theta) \propto \frac{1}{|q|} \sum_t \log p(q_t \mid q_{<t}, z_i; \Theta)$$

- Obtain soft relevance scores *i.e.*, teacher distribution as

$$\hat{p}(z_i \mid q, \mathcal{Z}) = \text{softmax}\ p(z_i \mid q; \Theta), \forall z_i$$

Step 4: Loss Calculation and Backpropagation

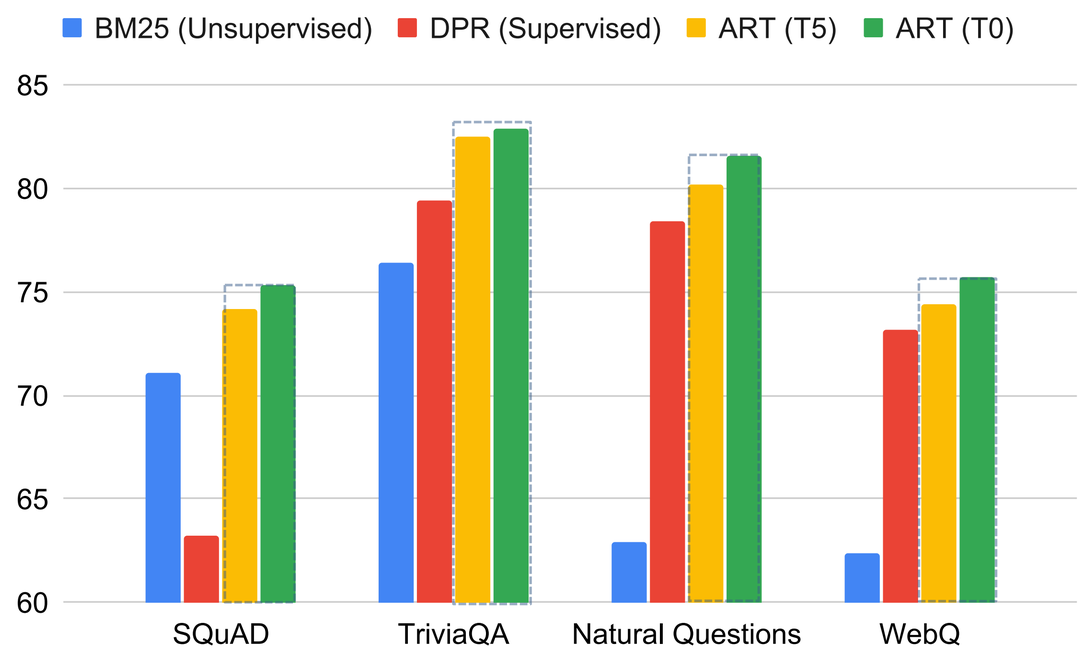
- Align *retriever distribution* with *relevance score distribution*

$$\mathcal{L}(\Phi) = \mathbb{KL}(\hat{p}(z_i \mid q, \mathcal{Z}) \mid\mid \underbrace{q(z_i \mid q, \mathcal{Z}; \Phi)}_{\text{retriever training}})$$

Step 5: Periodically Update Evidence Embeddings

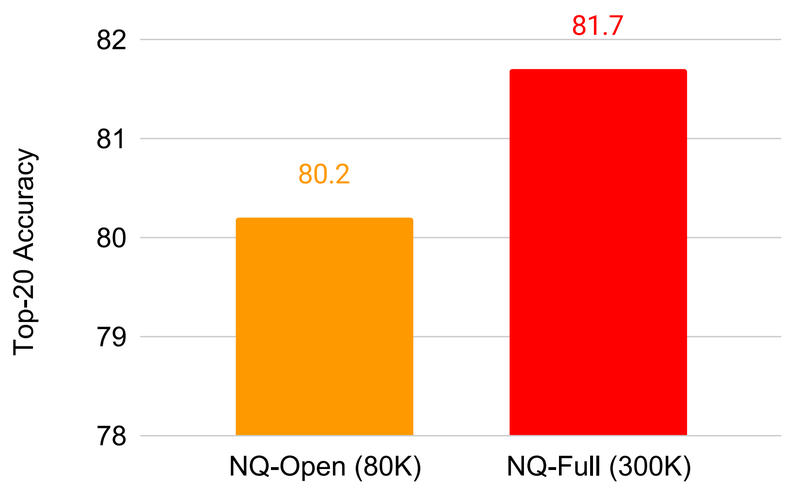
Results

Retrieval Accuracy (Top-20)



- ART outperforms unsupervised retrievers such as BM25.
- It matches or exceeds the performance of supervised models.

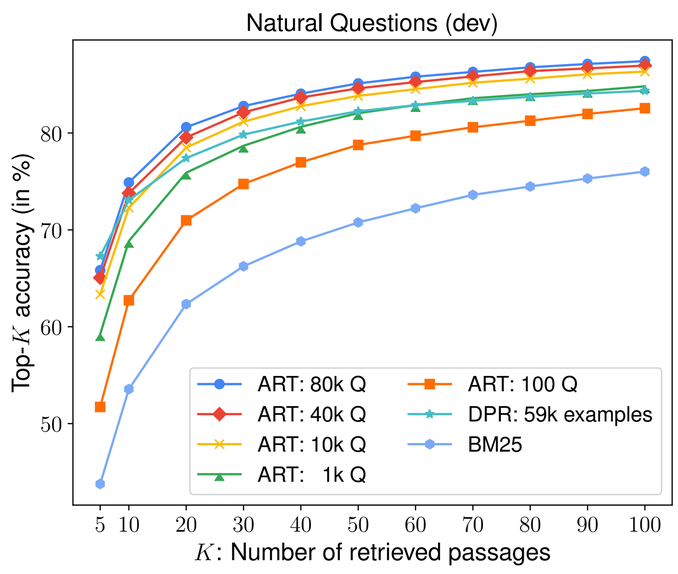
Robustness to Question Answerability



- ART is robust to natural questions distribution
- It can be trained on a mix of answerable and unanswerable questions

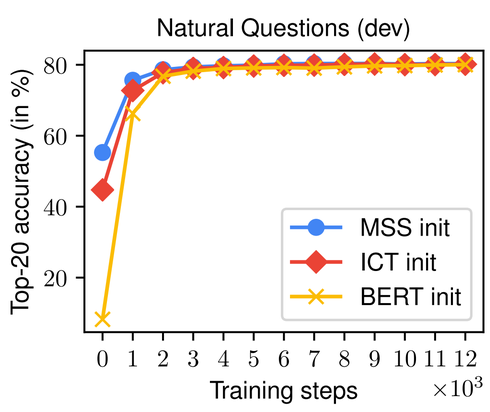
Analysis

Sample Efficiency



- Outperforms BM25 with just 100 examples

Retriever Initialization



- Different initializations converge to the same final state
- Useful for bootstrapping retriever from BERT

Conclusion

- **ART**: an approach to train dense retriever using unaligned pairs of questions and passages.
- Custom hard negative mining approaches are **not required**.
- Uses off-the-shelf large language models as a black-box (w/o finetuning).

Related Work

1. *Improving Passage Retrieval with Zero-Shot Question Generation*, Devendra Sachan et al., EMNLP 2022
2. *RocketQAv2: A Joint Training Method for Dense Passage Retrieval and Passage Re-ranking,* Ruiyang Ren et al., EMNLP 2021

Code: *https://github.com/DevSinghSachan/art*