Dense Passage Retrieval for Open-Domain Question Answering

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, Wen-tau Yih (EMNLP 2020)

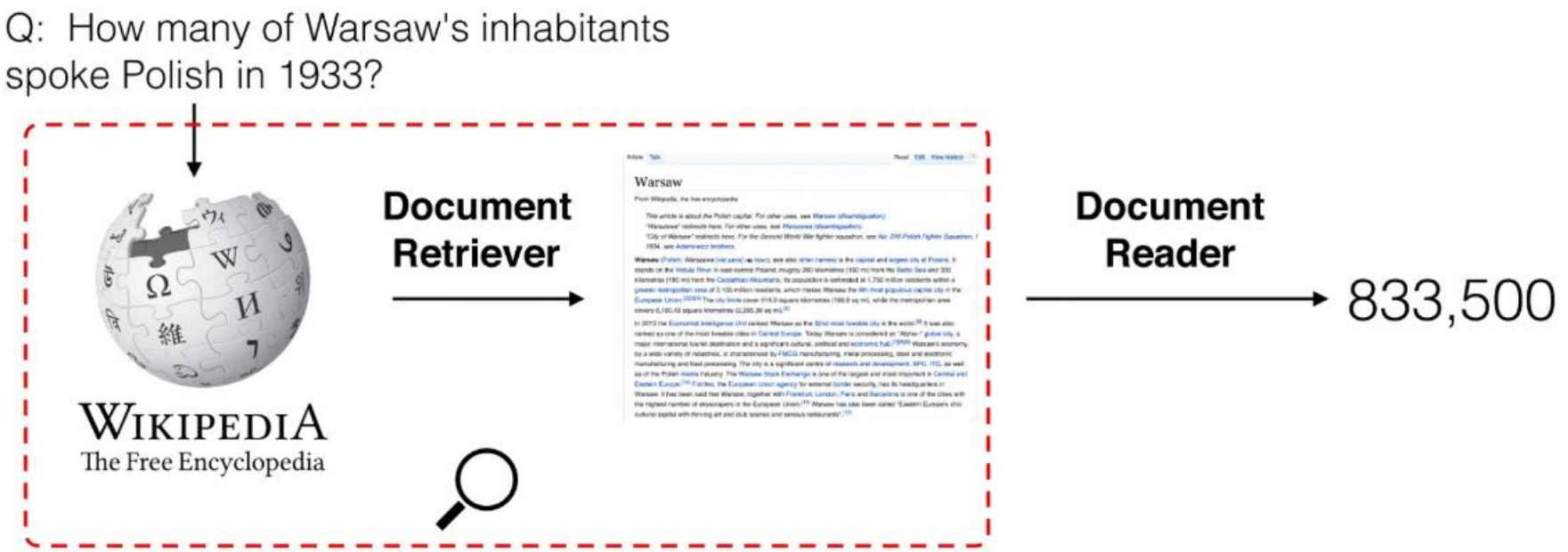
단국대학교 모바일시스템공학과 양윤성

Contents

1-(1). Background

What is Open-domain QA?

➢ Retrieve(Search) the open domain for hints to answer the question



1-(1). Background

Retriever and Reader in Open-domain QA

Input: a large collection of documents $\mathcal{D} = D_1, D_2, \ldots, D_N$ and Q

Output: an answer string A

Retriever: $f(\mathcal{D}, Q) \longrightarrow P_1, \ldots, P_K$ K is pre-defined (e.g., 100)

Reader: $g(Q, \{P_1, \ldots, P_K\}) \longrightarrow A$ A reading comprehension problem!

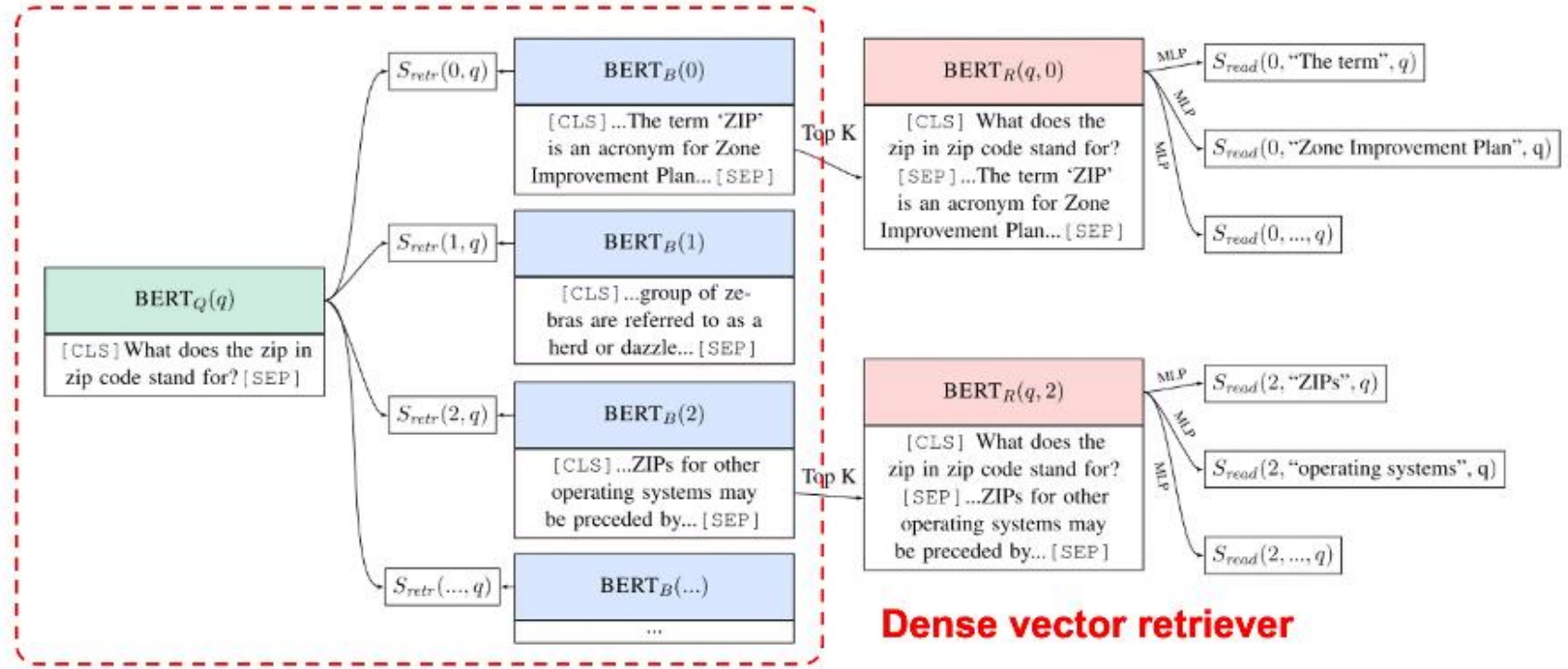
Sparse vector retriever

Who is the bad guy in lord of the rings?

Sala Baker is best known for portraying
the bad guy Sauron in the Lord of the Rings trilogy.

1-(1). Background

ORQA (Learnable Retriever)



Dense vector retriever

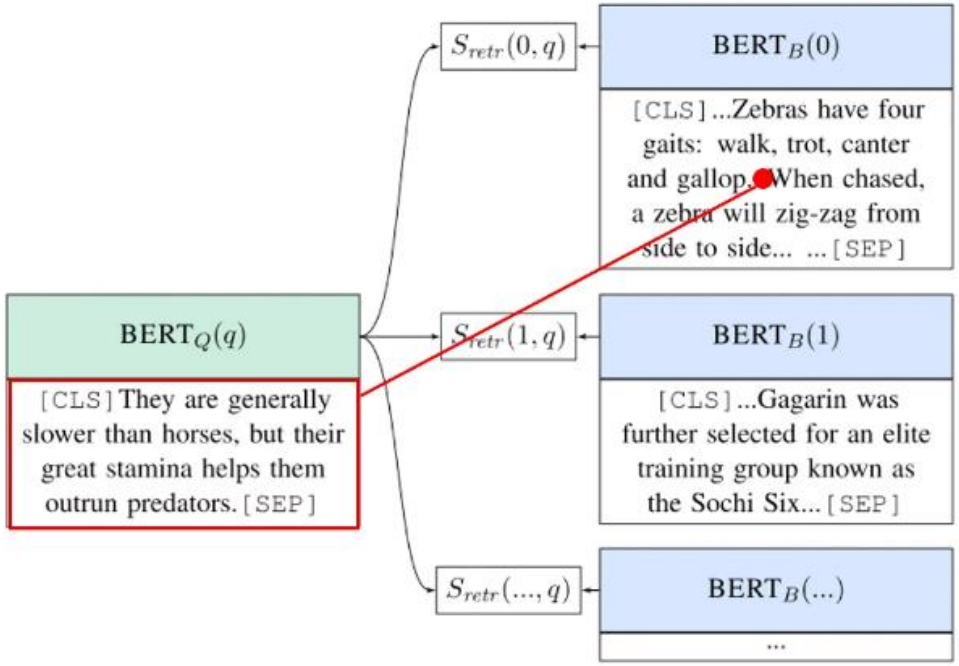
Who is the bad guy in lord of the rings?

Sala Baker is best known for portraying the villain Sauron in the Lord of the Rings trilogy.

1-(1). Background

ORQA – Inverse Cloze Task (ICT)

- Proceed unsupervised pre-training
- Given a sentence (pseudo-question)
- Predict its context (pseudo-evidence)



1-(2). Problems

- ICT pre-training is computationally intensive
- Not completely clear that regular sentences are surrogates of questions
- The context encoder is not fine-tuned

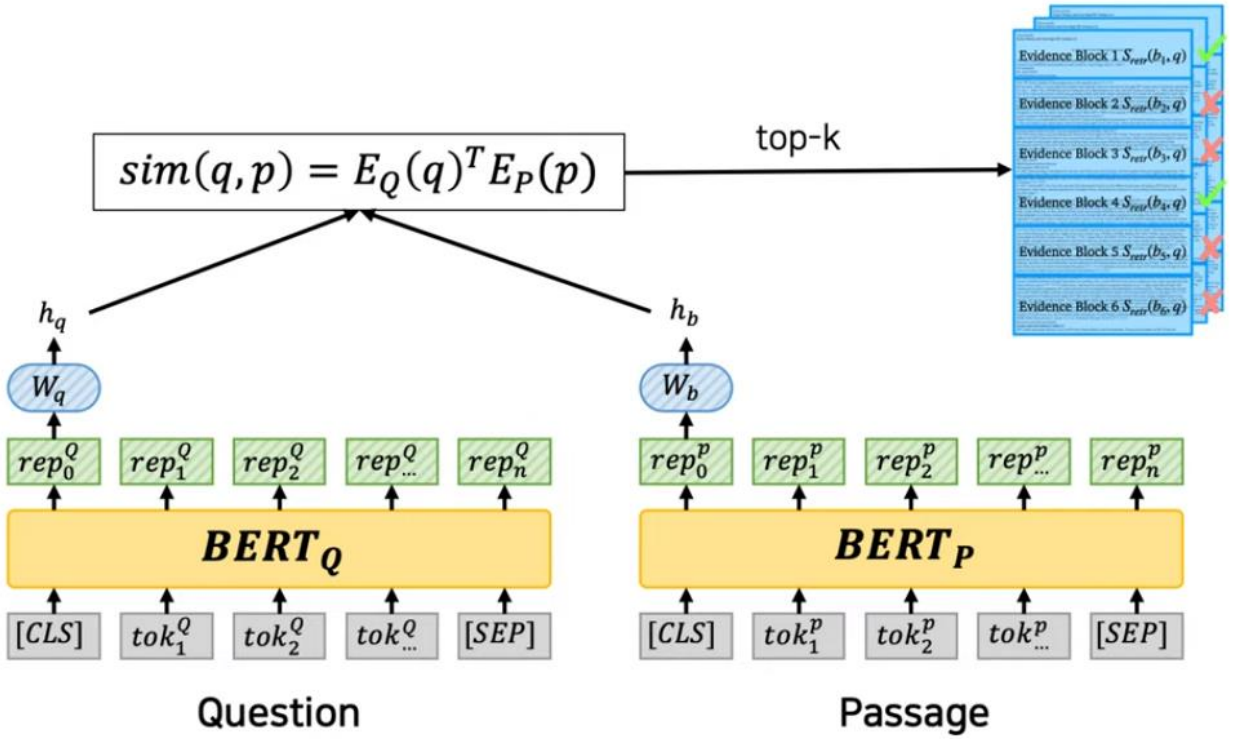
✓ Idea : Train using only pairs of questions and passages without additional pre-training

Contents

2-(1). Research Objective

- Dense Passage Retriever (DPR)
- Focus on improving retriever performance
- Mapping information for each passage to a low-dimensional space
- Then extract top-k passages associated with question effectively

2-(2). Model Architecture



- Inner product(similarity) of question embedding and passage embedding
- Use two different BERT encoders

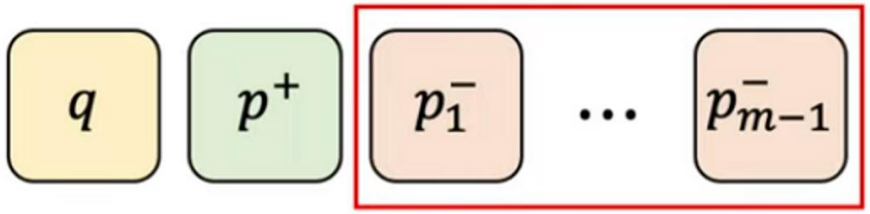
2-(3). DPR Training

$$D = \{< q_i, p_i^+, p_{i,1}^-, \cdots, p_{i,m}^- >\}$$

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

- Goal : To create vector space such that relevant pairs of questions and passages will have smaller distance than irrelevant ones

2-(4). Training : Positive/Negative Passages



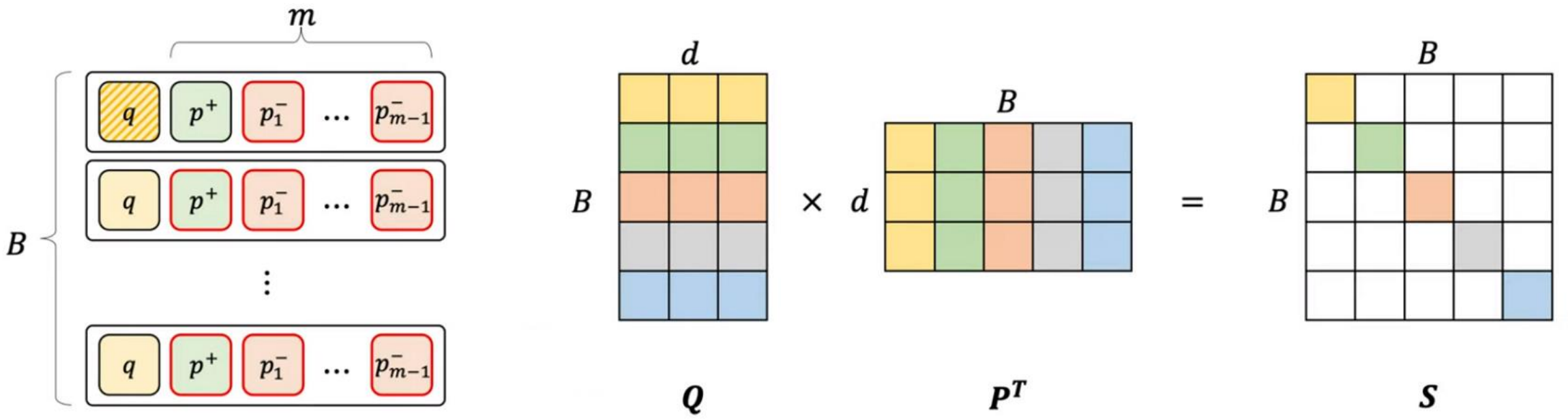
Random: any passages from the corpus

BM25: don't contain the answer,
but match most question tokens

Gold: positive passages paired with other questions

- Training method to choose negative examples more accurately

2-(5). Training : In-Batch Negatives



- Result : Represent as a similarity matrix
- Any (q_i, p_i) pair is a positive example when i = j, and negative otherwise

Contents

3-(1). Datasets

Natural Questions (Kwiatkowski et al., 2019): The questions were mined from real Google search queries and the answers were spans in Wikipedia articles identified by annotators

TriviaQA (Joshi et al., 2017): Contains a set of trivia questions with answers that were originally scraped from the Web

WebQuestions (Berant et al., 2013): Consists of questions selected using Google Suggest API, where the answers are entities in Freebase

CuratedTREC (Baudiš and Šedivy, 2015): Sources questions from TREC QA tracks as well as various Web sources and is intended for open-domain QA from unstructured corpora

SQuAD v1.1 (Rajpurkar et al., 2016): Annotators were presented with a Wikipedia paragraph, and asked to write questions that could be answered from the given text

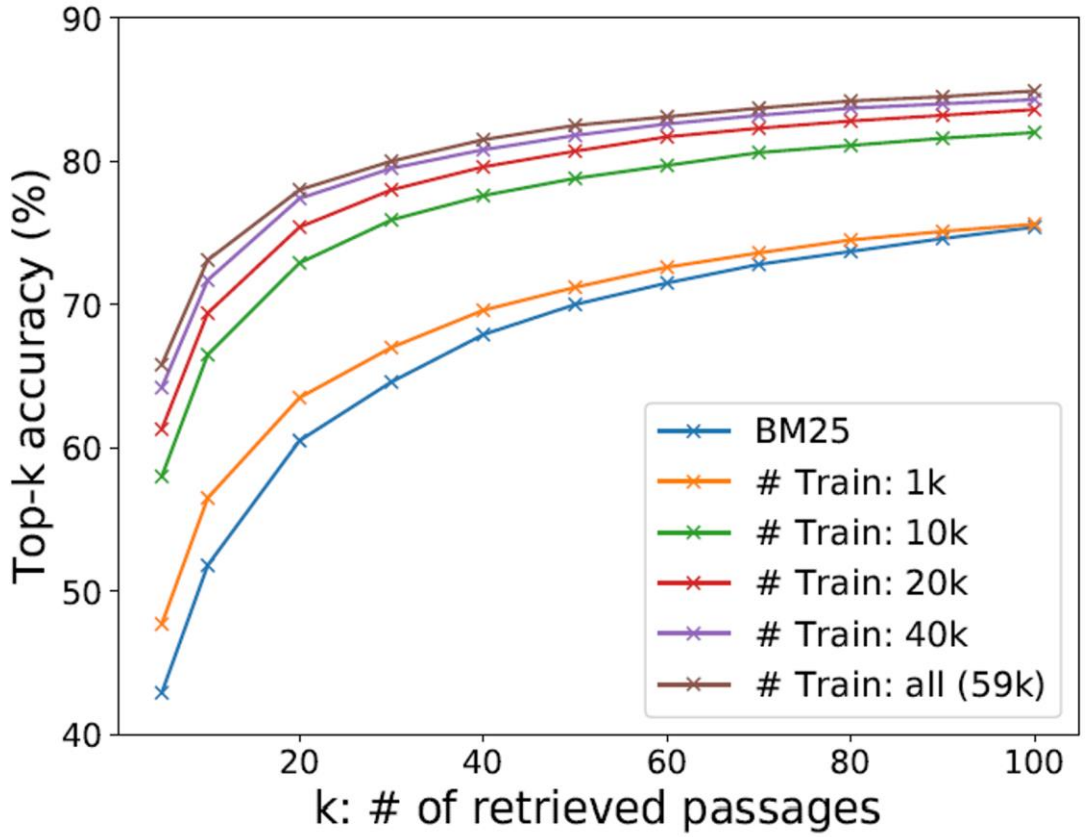
3-(2). Retrieval Accuracy

Training	Retriever	Top-20					Top-100				
		NQ	TriviaQA	WQ	TREC	SQuAD	NQ	TriviaQA	WQ	TREC	SQuAD
None	BM25	59.1	66.9	55.0	70.9	68.8	73.7	76.7	71.1	84.1	80.0
Single	DPR	78.4	79.4	73.2	79.8	63.2	85.4	**85.0**	81.4	89.1	77.2
	BM25 + DPR	76.6	79.8	71.0	85.2	**71.5**	83.8	84.5	80.5	92.7	**81.3**
Multi	DPR	**79.4**	78.8	**75.0**	**89.1**	51.6	**86.0**	84.7	**82.9**	93.9	67.6
	BM25 + DPR	78.0	**79.9**	74.7	88.5	66.2	83.9	84.4	82.3	**94.1**	78.6

$$\text{BM25}(q,p) + \lambda \cdot \text{sim}(q,p)$$

- Probability of getting answer when retriever get passages (top-20, 100)
- As a result, better performance with DPR

3-(3). Sample Efficiency



- Importance of DPR-training → 1000 examples already outperforms BM25

3-(4). In-Batch Negative Training

Type	#N	IB	Top-5	Top-20	Top-100
Random	7	✗	47.0	64.3	77.8
BM25	7	✗	50.0	63.3	74.8
Gold	7	✗	42.6	63.1	78.3
Gold	7	✓	51.1	69.1	80.8
Gold	31	✓	52.1	70.8	82.1
Gold	127	✓	55.8	73.0	83.1
G.+BM25$^{(1)}$	31+32	✓	65.0	77.3	84.4
G.+BM25$^{(2)}$	31+64	✓	64.5	76.4	84.0
G.+BM25$^{(1)}$	127+128	✓	**65.8**	**78.0**	**84.9**

Top Block Analysis

- Low k → BM25
- High k → Gold
- But negative passage type has no significant impact

3-(4). In-Batch Negative Training

Type	#N	IB	Top-5	Top-20	Top-100
Random	7	✗	47.0	64.3	77.8
BM25	7	✗	50.0	63.3	74.8
Gold	7	✗	42.6	63.1	78.3
Gold	7	✓	51.1	69.1	80.8
Gold	31	✓	52.1	70.8	82.1
Gold	127	✓	55.8	73.0	83.1
G.+BM25$^{(1)}$	31+32	✓	65.0	77.3	84.4
G.+BM25$^{(2)}$	31+64	✓	64.5	76.4	84.0
G.+BM25$^{(1)}$	127+128	✓	**65.8**	**78.0**	**84.9**

Middle Block Analysis

- Compare in-batch and no in-batch
- Large batch size → High performance

3-(4). In-Batch Negative Training

Type	#N	IB	Top-5	Top-20	Top-100
Random	7	✗	47.0	64.3	77.8
BM25	7	✗	50.0	63.3	74.8
Gold	7	✗	42.6	63.1	78.3
Gold	7	✓	51.1	69.1	80.8
Gold	31	✓	52.1	70.8	82.1
Gold	127	✓	55.8	73.0	83.1
G.+BM25$^{(1)}$	31+32	✓	65.0	77.3	84.4
G.+BM25$^{(2)}$	31+64	✓	64.5	76.4	84.0
G.+BM25$^{(1)}$	127+128	✓	**65.8**	**78.0**	**84.9**

Bottom Block Analysis

- Add hard negative passage

  (based on BM25)

- 1 hard p- is better than 2 hard p-

- Best performance

  : Gold + 1 hard negative passage

3-(5). Question Answering

Training	Model	NQ	TriviaQA	WQ	TREC	SQuAD
Single	BM25+BERT (Lee et al., 2019)	26.5	47.1	17.7	21.3	33.2
Single	ORQA (Lee et al., 2019)	33.3	45.0	36.4	30.1	20.2
Single	HardEM (Min et al., 2019a)	28.1	50.9	-	-	-
Single	GraphRetriever (Min et al., 2019b)	34.5	56.0	36.4	-	-
Single	PathRetriever (Asai et al., 2020)	32.6	-	-	-	**56.5**
Single	$\text{REALM}_{\text{Wiki}}$ (Guu et al., 2020)	39.2	-	40.2	46.8	-
Single	$\text{REALM}_{\text{News}}$ (Guu et al., 2020)	40.4	-	40.7	42.9	-
Single	BM25	32.6	52.4	29.9	24.9	38.1
	DPR	**41.5**	56.8	34.6	25.9	29.8
	BM25+DPR	39.0	57.0	35.2	28.0	36.7
Multi	DPR	**41.5**	56.8	**42.4**	49.4	24.1
	BM25+DPR	38.8	**57.9**	41.1	**50.6**	35.8

Contents

4. Conclusion

- DPR can outperform and potentially replace traditional method in open-domain question answering
- Dense vector retriever > Sparse vector retriever
- Improved retrieval performance → Obtained new SOAT results on multiple question-answering benchmarks

Thank You!

Dense Passage Retrieval for Open-Domain Question Answering