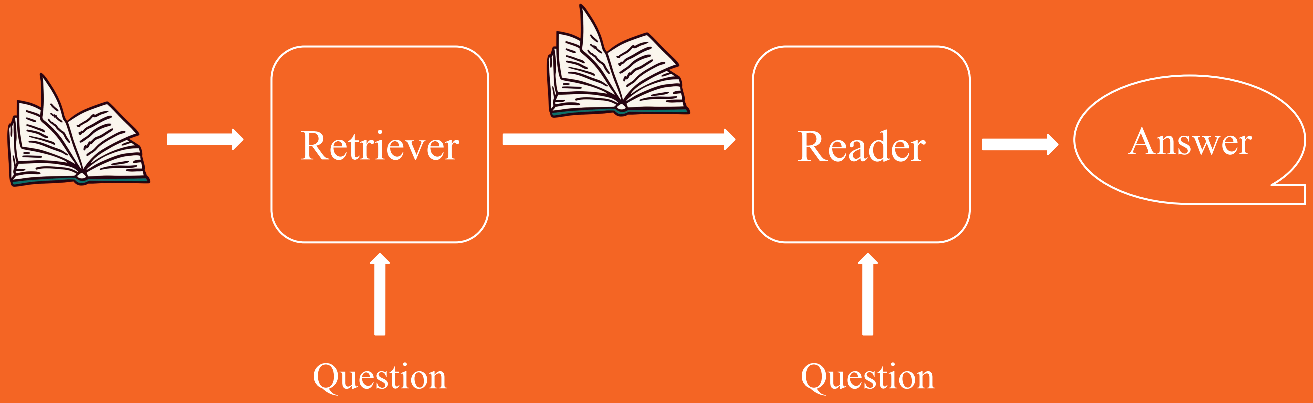
# Dense Passage Retrieval for Open-Domain Question Answering

University of Southern California

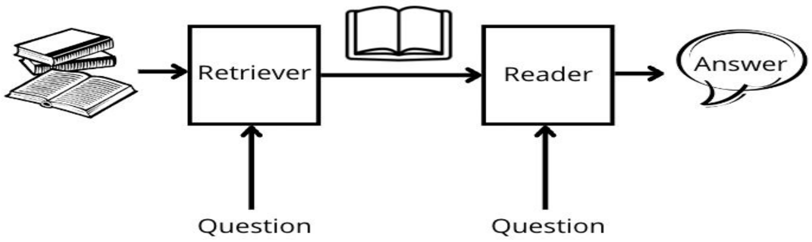
Team:
- Reuel
- Sayf
- Suraj
- Hetvi
- Kritika



# Objective, Motivation and Related Works:

What is ODQA?

- QA: build computer systems that automatically answer questions posed by humans in a natural language
- Open Domain: not limited to a specific domain
- Has 2 parts:
  - Information Retrieval (IR)
  - Question Answering (QA)




Generally, IR tasks are implemented with TF-IDF or BM-25

***BM25 (Best Match-25)***

- represents words using high dimensional and weighted sparse vectors (generated using TF-IDF)
- fails to map synonyms to be close to each other (due to sparse nature of representation)

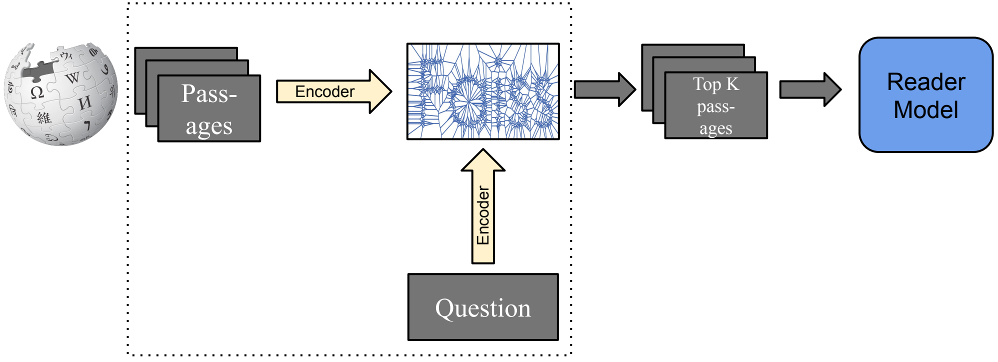
***Failed to capture syntactic meanings!***

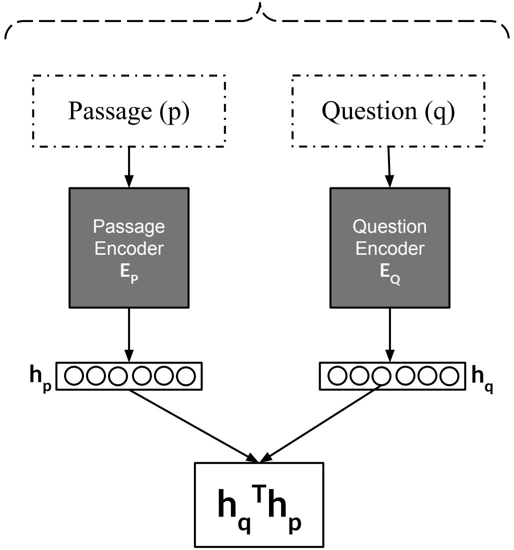
***ORQA (Open Retrieval Question Answering)***

- Joint learning on retriever and reader from question-answer string pairs
- A pseudo-task is defined using the sentence as the question and the context as the evidence
- Inverse Cloze Task is used for pretraining
- computationally expensive and sentences may not be an appropriate substitute for questions

*Can we train a better dense embedding model efficiently using only pairs of questions and passages and no pre-training?*

# Introducing: Dense Passage Retrieval (DPR)



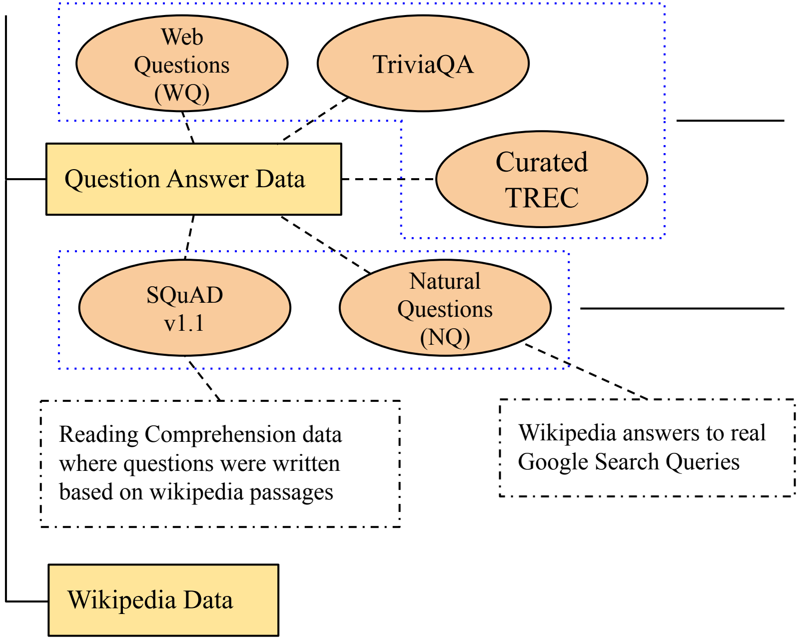

- Uses a dense **encoder ($E_P$)** to map a text passage to a *d* dimensional real-valued vector and builds an index for all the M passages that we will use for retrieval.
- At run-time, DPR applies a different **encoder ($E_Q$)** that maps the input question to a d-dimensional vector, and retrieves k passages of which vectors are the **closest** to the question vector.

$$\mathrm{sim}(q, p) = E_Q(q)^\intercal E_P(p).$$

Both encoders are **independent BERT based uncased** models with 768 hidden dimensions

Retriever Training mainly focuses on **fine tuning** these encoders to generate dense embeddings for the questions and passages

# Dataset and Preparation




These datasets do not contain any contexts. They have just question-answer pairs.

These datasets come with wikipedia passages for context

- First, the data is cleaned by removing semi-structured data, tables, lists etc
- Second, the data is split into passages with a maximum limit of 100 words

[TITLE][SEP][PASSAGE] ← Basic Retrieval Units

The goal here is to:
- Provide context for the datasets that do not have any contexts
- To do so, they take the top passage returned by BM-25 as the relevant context for the QA pair.

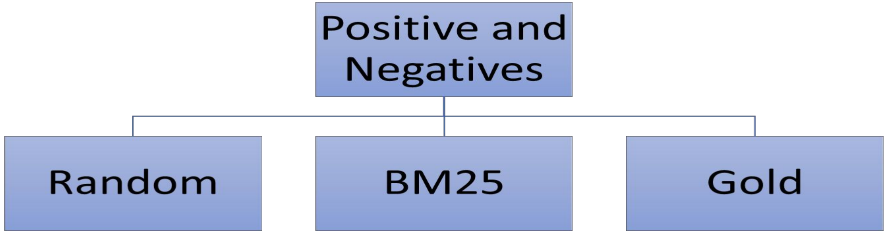
# Methodology

- The goal is to create a vector space such that relevant pairs of questions and passages will have smaller distance (i.e., higher similarity than the irrelevant ones

$$\mathcal{D} = \{\langle q_i, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^- \rangle\}_{i=1}^m$$

## Positive and Negative Passages

- Positive examples are available explicitly
- Negative Passage or Irrelevant Passage are those which needs to be selected from a pool




- Random
  - Any random passage from the corpus
- BM25
  - Top passages returned by BM25 which don't contain the answer but match most question tokens
- Gold
  - Positive passages paired with other questions which appear in the training set

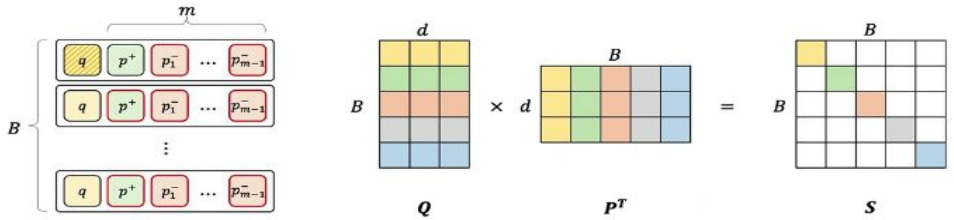
## Best Model uses

**Gold Passages + one BM25 negative passage**

## Loss Function

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

# In Batch Negatives




- B set of questions (denoted by Q) and passages (denoted by P) represented with d dimension
- Dot product between P and Q generates a similarity matrix S
- Diagonal values represent the positive passages

# Experimental Setup

- Batch Size: 128
- Epochs: 40 (for larger dataset) or 100 (for smaller dataset)
- Learning Rate: $10^{-5}$
- Optimizer: Adam
- Dropout Rate: 0.1
- Linear Scheduler

# Training

Training was done with:

- **Single** dataset (Encoder is trained specifically with one dataset)
- **Multi-dataset** (Encoder trained with combined dataset - except for SQuAD)

Results were tabulated for the DPR Model, as well as the traditional BM-25 model and a combination of the two.

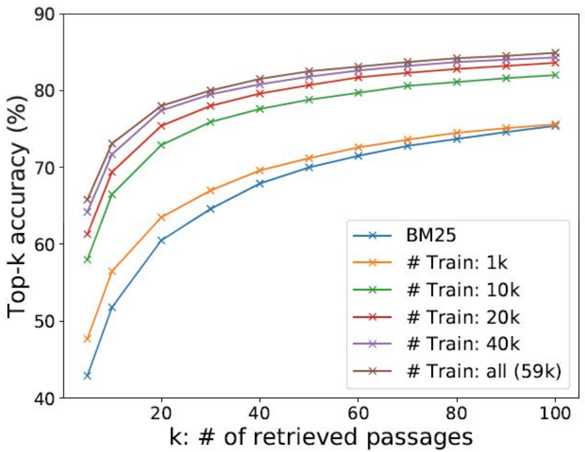
**Accuracy Metric used:**

***Top-K accuracy:*** effectiveness of model to identify correct passage out of the top k passages that have been retrieved

# DPR Results

| Training | Retriever | Top-20 NQ | Top-20 TriviaQA | Top-20 WQ | Top-20 TREC | Top-20 SQuAD | Top-100 NQ | Top-100 TriviaQA | Top-100 WQ | Top-100 TREC | Top-100 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 |

Graphical comparison between DPR and BM-25:




# Inferences

- DPR outperformed BM-25 in almost all datasets
- DPR+BM25 provides best results
- From observation, SQuAD performs rather poorly. This can be attributed to the fact that annotators wrote the question after seeing the passage - giving BM-25 an advantage

# Reader (QA System)

- Uses representation for the passages learnt from pre-trained BERT base uncased model (from CLS token)
- Makes use of cross attention that experimentally has shown to perform better for small set of top passages

$P_i$: BERT representation of the i^th passage

$\hat{P}$ : vector of BERT representations for each passage

$W$: learnable weight vectors

Assign passage selection score to each of the top K retrieved passages

*calculated as*

$$P_{\text{selected}}(i) = \text{softmax}\left(\hat{\mathbf{P}}^{\intercal}\mathbf{w}_{\text{selected}}\right)_i$$

Extract an answer span from each passage and assigns a span score

*calculated as*

$$P_{\text{start},i}(s) \times P_{\text{end},i}(t)$$

where,

$$P_{\text{start},i}(s) = \text{softmax}\left(\mathbf{P}_i\mathbf{w}_{\text{start}}\right)_s$$

$$P_{\text{end},i}(t) = \text{softmax}\left(\mathbf{P}_i\mathbf{w}_{\text{end}}\right)_t$$

Final answer selected as the best span from the passage with highest selection score

# Reader Results

| 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 | REALM$_{\text{Wiki}}$ (Guu et al., 2020) | 39.2 | - | 40.2 | 46.8 | - |
| Single | 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 |

*Dataset Size (in number of questions):*

| Dataset | Train | | Dev | Test |
|---|---|---|---|---|
| Natural Questions | 79,168 | 58,880 | 8,757 | 3,610 |
| TriviaQA | 78,785 | 60,413 | 8,837 | 11,313 |
| WebQuestions | 3,417 | 2,474 | 361 | 2,032 |
| CuratedTREC | 1,353 | 1,125 | 133 | 694 |
| SQuAD | 78,713 | 70,096 | 8,886 | 10,570 |

**Accuracy Metric used:**

***Exact Match:*** strict metric used to measure the exactness of the retrieved answer with respect to the ground truth

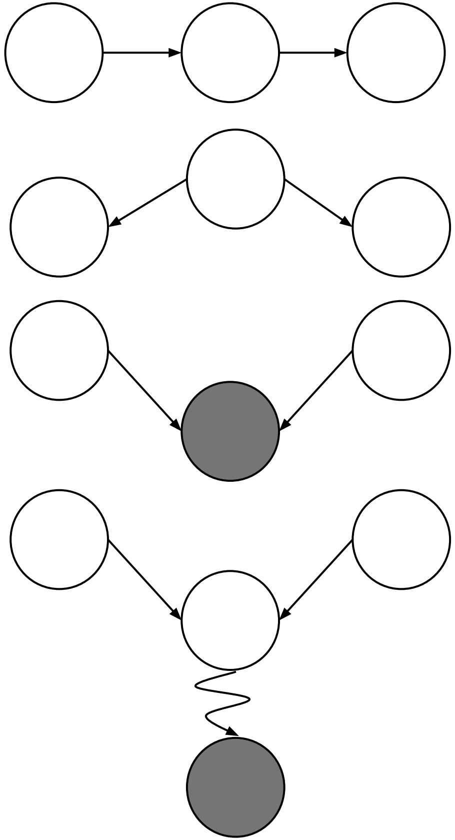
# Inferences

- Comparable results for large datasets (NQ and TriviaQA) between multi and single dataset training
- Clear advantage for multi-dataset training for small datasets
- DPR outperforms SOTA models in at least 4 out of 5 datasets

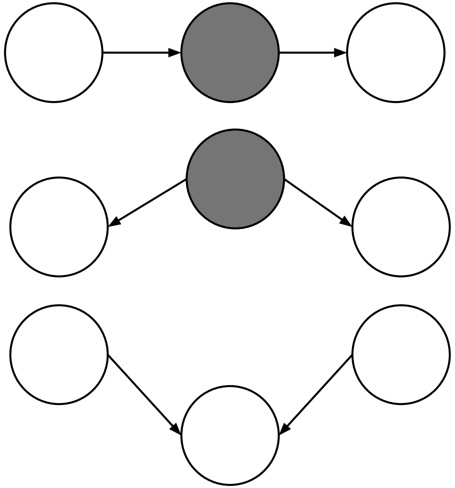
# Conclusions:

- Answered the question posed at the start of the presentation - ***provided an efficient dense representation for question passage pairs without pretraining***
- ***Outperformed*** and potentially could replace the traditional sparse retrieval component in open-domain question answering
- As a result of improved retrieval performance, they obtained ***new state-of-the-art results*** on multiple open-domain question answering benchmarks.

## Active



## Inactive



- If any one triple is inactive in a path, whole path is inactive
- All paths must be inactive to say d-separated/independent