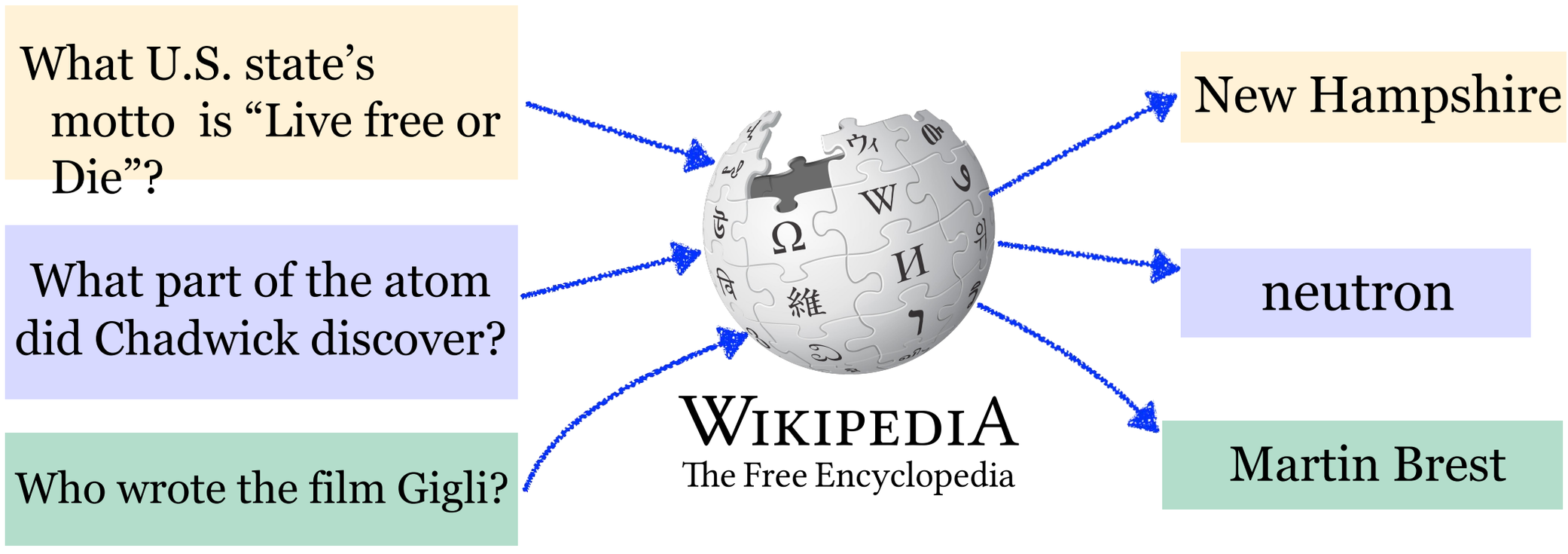
Reference for Open-domain QA, “Retrieval-augmented” models: https://github.com/danqi/acl2020-openqa-tutorial/tree/master/slides

Open-domain QA

- **Input**: question Q, D = English Wikipedia (~5 million documents)
- **Output**: answer A



"Machine Reading at Scale"

Chen et al., 2017. Reading Wikipedia to Answer Open-domain Questions

Dataset

Natural Questions [Kwiatkowski et al., 2019]

- Motivation
  - Large-scale end-to-end training data for QA
  - "Natural" questions from search engine query logs

Example 1

Question: what color was john wilkes booth's hair
Wikipedia Page: John_Wilkes_Booth
Long answer: Some critics called Booth "the handsomest man in America" and a "natural genius", and noted his having an "astonishing memory"; others were mixed in their estimation of his acting. He stood 5 feet 8 inches (1.73 m) tall, had jet-black hair , and was lean and athletic. Noted Civil War reporter George Alfred Townsend described him as a "muscular, perfect man" with "curling hair, like a Corinthian capital".

Short answer: jet-black

Kwiatkowski et al., 2019. Natural Questions: A Benchmark for Question Answering Research

- Question source: Google search queries
- Answer source: Wikipedia page from top 5 search results
- Long answer: paragraph, table, list (HTML bounding box)
- Short answer: span(s), yes/no, NULL
- Annotation: a pool of ~50 annotators

MSMARCO

Q: will i qualify for osap if i'm new in canada

Candidate passages

Click passages to select or unselect them

Ontario.ca Francais. Français in order to apply online for funding consideration from The Ontario Student Assistance (PROGRAM), osap you must first register as a new user to this. website

Source: https://osap.gov.on.ca/OSAPSecurityWeb/public/agreement.xhtml

Visit the OSAP website for application deadlines. To get OSAP, you have to be eligible. You can apply using an online form, or you can print off the application forms. If you submit a paper application, you must pay an application fee. The online application is free.

Source: http://settlement.org/ontario/education/colleges-universities-and-institutes/financial-assistance-for-post-secondary-education/how-do-i-apply-for-the-ontario-student-assistance-program-osap/

Visit the OSAP website for application deadlines. To get OSAP, you have to be eligible. You can apply using an online form, or you can print off the application forms. If you submit a paper application, you must pay an application fee.

Source: http://settlement.org/ontario/education/colleges-universities-and-institutes/financial-assistance-for-post-secondary-education/how-do-i-apply-for-the-ontario-student-assistance-program-osap/

To be eligible to apply for financial assistance from the Ontario Student Assistance Program (OSAP), you must be a: 1 Canadian citizen; 2 Permanent resident; or. 3 Protected person/convention refugee with a Protected Persons Status Document (PPSD).

Source: http://settlement.org/ontario/education/colleges-universities-and-institutes/financial-assistance-for-post-secondary-education/who-is-eligible-for-the-ontario-student-assistance-program-osap/

You will not be eligible for a Canada-Ontario Integrated Student Loan, but can apply for a part-time loan through the Canada Student Loans Program. There are also grants, bursaries and scholarships available for both full-time and part-time students.

Source: http://www.campusaccess.com/financial-aid/osap.html

Selected passages

Visit the OSAP website for application deadlines. To get OSAP, you have to be eligible. You can apply using an online form, or you can print off the application forms. If you submit a paper application, you must pay an application fee. The online application is free.

Source: http://settlement.org/ontario/education/colleges-universities-and-institutes/financial-assistance-for-post-secondary-education/how-do-i-apply-for-the-ontario-student-assistance-program-osap/

To be eligible to apply for financial assistance from the Ontario Student Assistance Program (OSAP), you must be a: 1 Canadian citizen; 2 Permanent resident; or. 3 Protected person/convention refugee with a Protected Persons Status Document (PPSD).

Source: http://settlement.org/ontario/education/colleges-universities-and-institutes/financial-assistance-for-post-secondary-education/who-is-eligible-for-the-ontario-student-assistance-program-osap/

You will not be eligible for a Canada-Ontario Integrated Student Loan, but can apply for a part-time loan through the Canada Student Loans Program. There are also grants, bursaries and scholarships available for both full-time and part-time students.

Source: http://www.campusaccess.com/financial-aid/osap.html

Summarize the answer given by the selected passages:

No. You won't qualify.

Submit answer | Can't summarize

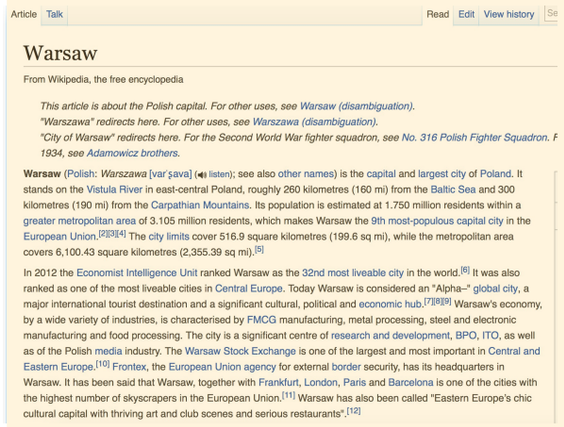
Field	Description
Query	A question query issued to Bing.
Passages	Top 10 passages from Web documents as retrieved by Bing. The passages are presented in ranked order to human editors. The passage that the editor uses to compose the answer is annotated as is_selected: 1.
Document URLs	URLs of the top ranked documents for the question from Bing. The passages are extracted from these documents.
Answer(s)	Answers composed by human editors for the question, automatically extracted passages and their corresponding documents.
Well Formed Answer(s)	Well-formed answer rewritten by human editors, and the original answer.
Segment	QA classification. E.g., tallest mountain in south america belongs to the ENTITY segment because the answer is an entity (Aconcagua).

DrQA: a first neural open-domain QA system

[Chen et al., 2017]

Q: How many of Warsaw's inhabitants spoke Polish in 1933?

Document Retriever



Article Talk Read Edit View history

Warsaw

From Wikipedia, the free encyclopedia

This article is about the Polish capital. For other uses, see Warsaw (disambiguation).
"Warszawa" redirects here. For other uses, see Warszawa (disambiguation).
"City of Warsaw" redirects here. For the Second World War fighter squadron, see No. 316 Polish Fighter Squadron. F
1934, see Adamowicz brothers.

Warsaw (Polish: *Warszawa* [varˈʂava] (listen); see also other names) is the capital and largest city of Poland. It stands on the Vistula River in east-central Poland, roughly 260 kilometres (160 mi) from the Baltic Sea and 300 kilometres (190 mi) from the Carpathian Mountains. Its population is estimated at 1.750 million residents within a greater metropolitan area of 3.105 million residents, which makes Warsaw the 9th most-populous capital city in the European Union.[2][3][4] The city limits cover 516.9 square kilometres (199.6 sq mi), while the metropolitan area covers 6,100.43 square kilometres (2,355.39 sq mi).[5]

In 2012 the Economist Intelligence Unit ranked Warsaw as the 32nd most liveable city in the world.[6] It was also ranked as one of the most liveable cities in Central Europe. Today Warsaw is considered an "Alpha–" global city, a major international tourist destination and a significant cultural, political and economic hub.[7][8][9] Warsaw's economy, by a wide variety of industries, is characterised by FMCG manufacturing, metal processing, steel and electronic manufacturing and food processing. The city is a significant centre of research and development, BPO, ITO, as well as of the Polish media industry. The Warsaw Stock Exchange is one of the largest and most important in Central and Eastern Europe.[10] Frontex, the European Union agency for external border security, has its headquarters in Warsaw. It has been said that Warsaw, together with Frankfurt, London, Paris and Barcelona is one of the cities with the highest number of skyscrapers in the European Union.[11] Warsaw has also been called "Eastern Europe's chic cultural capital with thriving art and club scenes and serious restaurants".[12]

Document Reader

833,500

Information Retrieval

Reading Comprehension

Chen et al., 2017. Reading Wikipedia to Answer Open-domain Questions

Document Retriever

- A TF-IDF weighted term vector model (~Classic IR)

- This retriever is not *trainable*.
- Retriever at document level instead of paragraph level.

Chen et al., 2017. Reading Wikipedia to Answer Open-domain Questions

Document Reader

paragraph, document, arbitrary-length text blocks.

Cast as a *reading comprehension* problem:

- **Input** is a *passage P* and a question *Q*
- **Output** is an answer *A*

A restricted setting is that A needs to be a segment of text in P => "extractive question answering"

Stanford Question Answering Dataset [Rajpurkar et al., 2016]

Passage

Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl title. The game was played on February 7, 2016, at Levi's Stadium in the San Francisco Bay Area at Santa Clara, California.

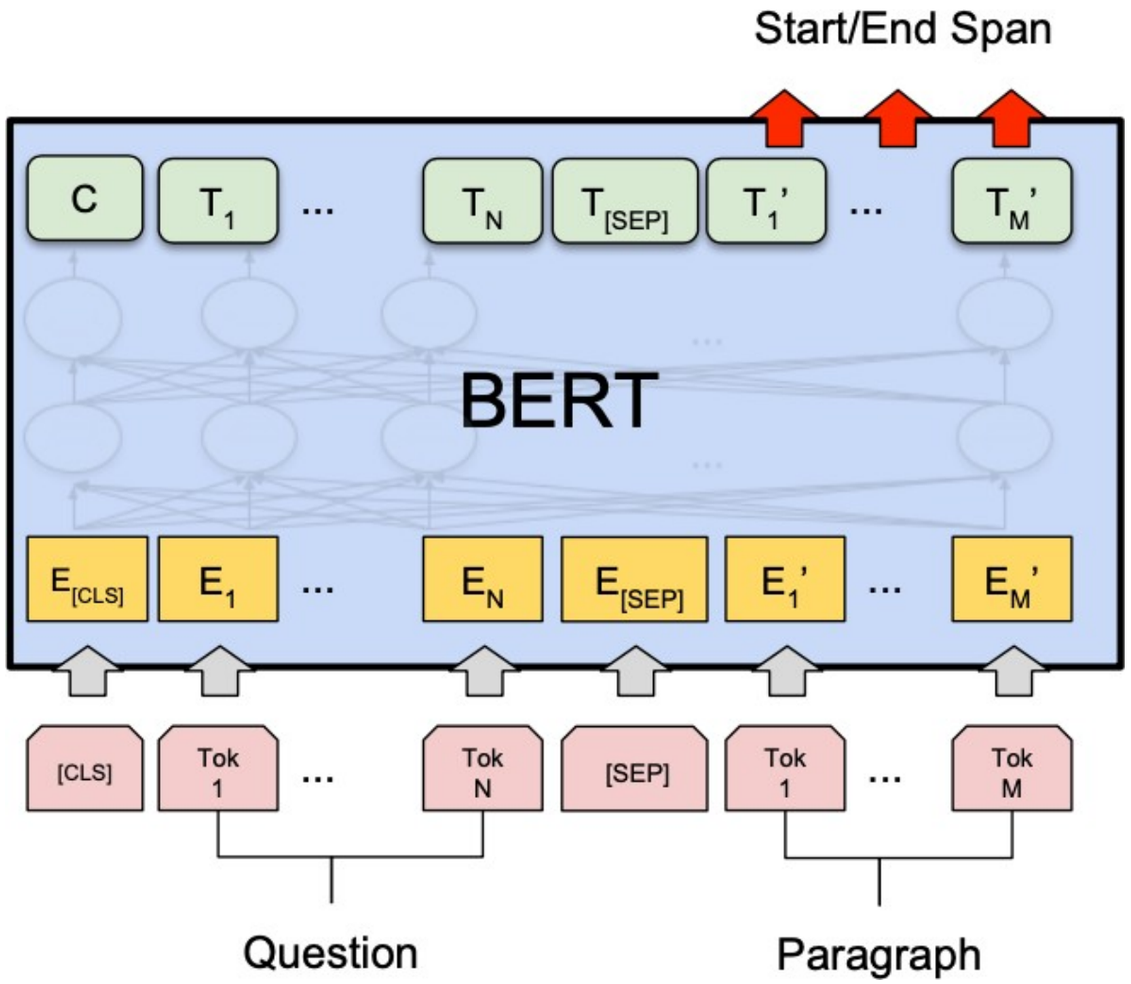
Question: Which NFL team won Super Bowl 50?
Answer: Denver Broncos

Question: What does AFC stand for?
Answer: American Football Conference

Question: What year was Super Bowl 50?
Answer: 2016

Chen et al., 2017. Reading Wikipedia to Answer Open-domain Questions

Document Reader



Question = Segment A

Passage = Segment B

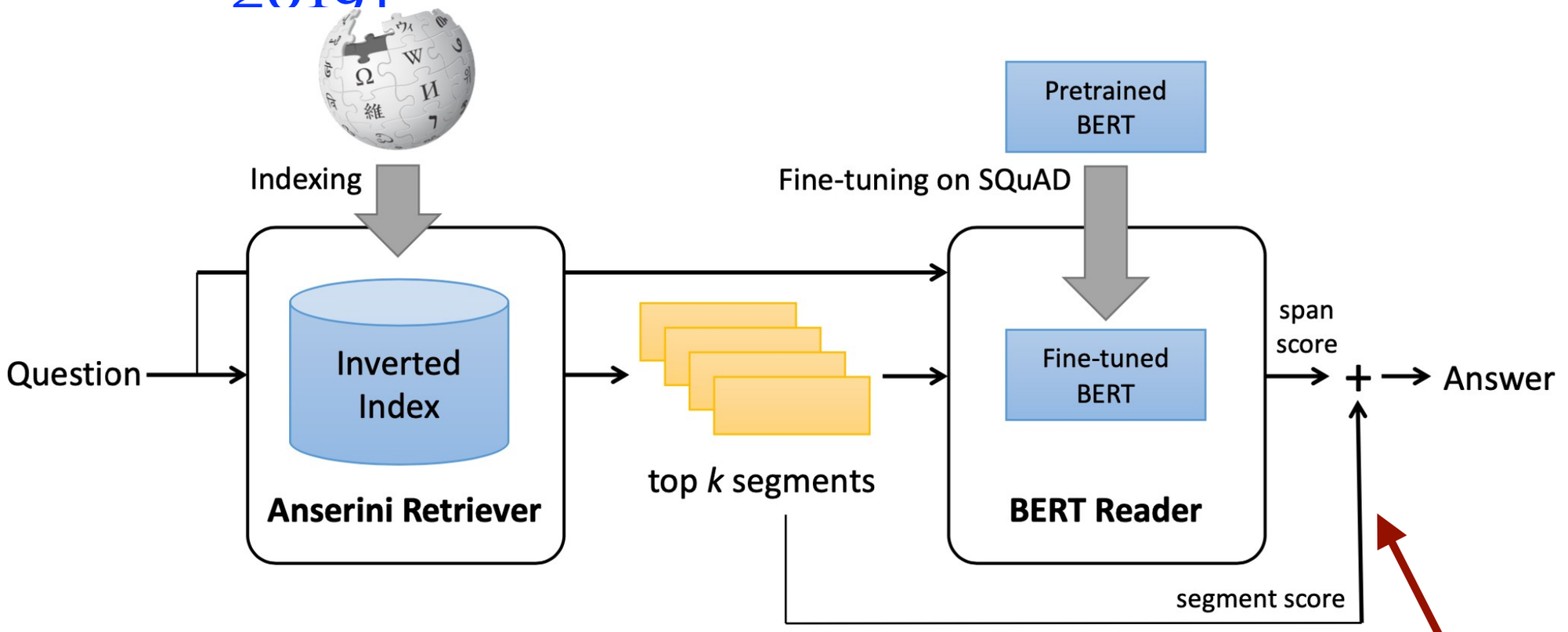
Answer = predicting two endpoints in segment B

On SQuAD 1.1,

-  : F1 = 90.9
- RoBERTa: F1 = 94.6

Devlin et al., 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

BERTserini [Yang et al., 2019]



Anserini Retriever
[Yang et al. 2017]:
Lucene with BM25, operated on 29.5M paragraphs

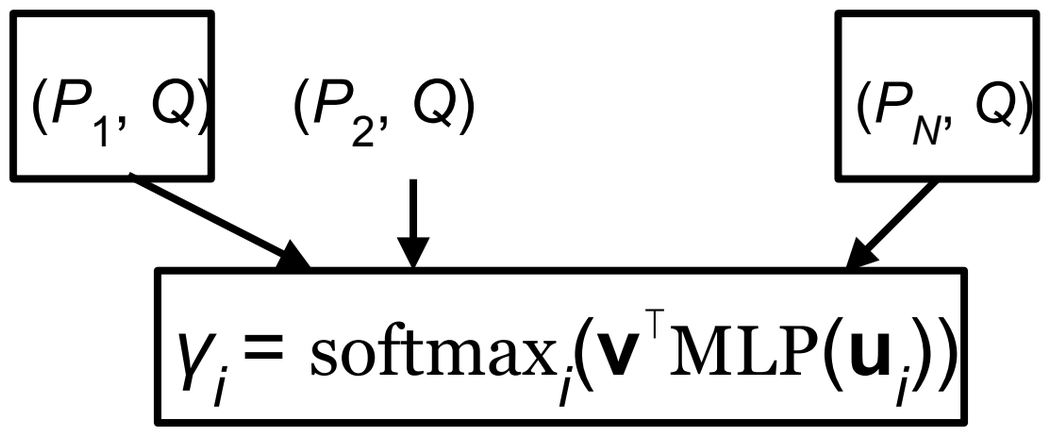
Classic IR on Passages

BERT Reader:
Trained on SQuAD

Scoring from both retriever and reader:
$S = (1 - \mu) \cdot S_{\text{Anserini}} + \mu \cdot S_{\text{BERT}}$

Training a passage re-ranker [Wang et al., 2018]

- Training a "deep" re-ranker model on retrieved passages can help further identify the relevance of the passages.



- This reranker can be easily trained using **distant supervision**: whether the passage contains the answer or not.

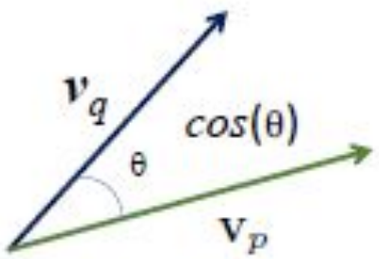
- Expensive == Hence used for RERANKING

Wang et al., 2018. R³: Reinforced Ranker-Reader for Open-Domain Question Answering
Lin et al., 2018. Denoising Distantly Supervised Open-Domain Question Answering
Wang et al., 2019. Multi-passage BERT: A Globally Normalized BERT Model for Open-domain Question

Dense IR replacing classic IR?

Classic:
Fast (supported by inverted index)
Exact entity matching

Dense:
Expensive
Match “you know who”



$d_1 >> d_2$

sparse repr: [0… 1 …. 1 …0..1] $\in \mathbb{R}^{d_1}$

dense repr: [1.03, -5.72, 6.42, .., 9.91] $\in \mathbb{R}^{d_2}$

sparse

“How many provinces did the Ottoman empire contain in the 17th century?”

“What part of the atom did Chadwick discover?”

dense

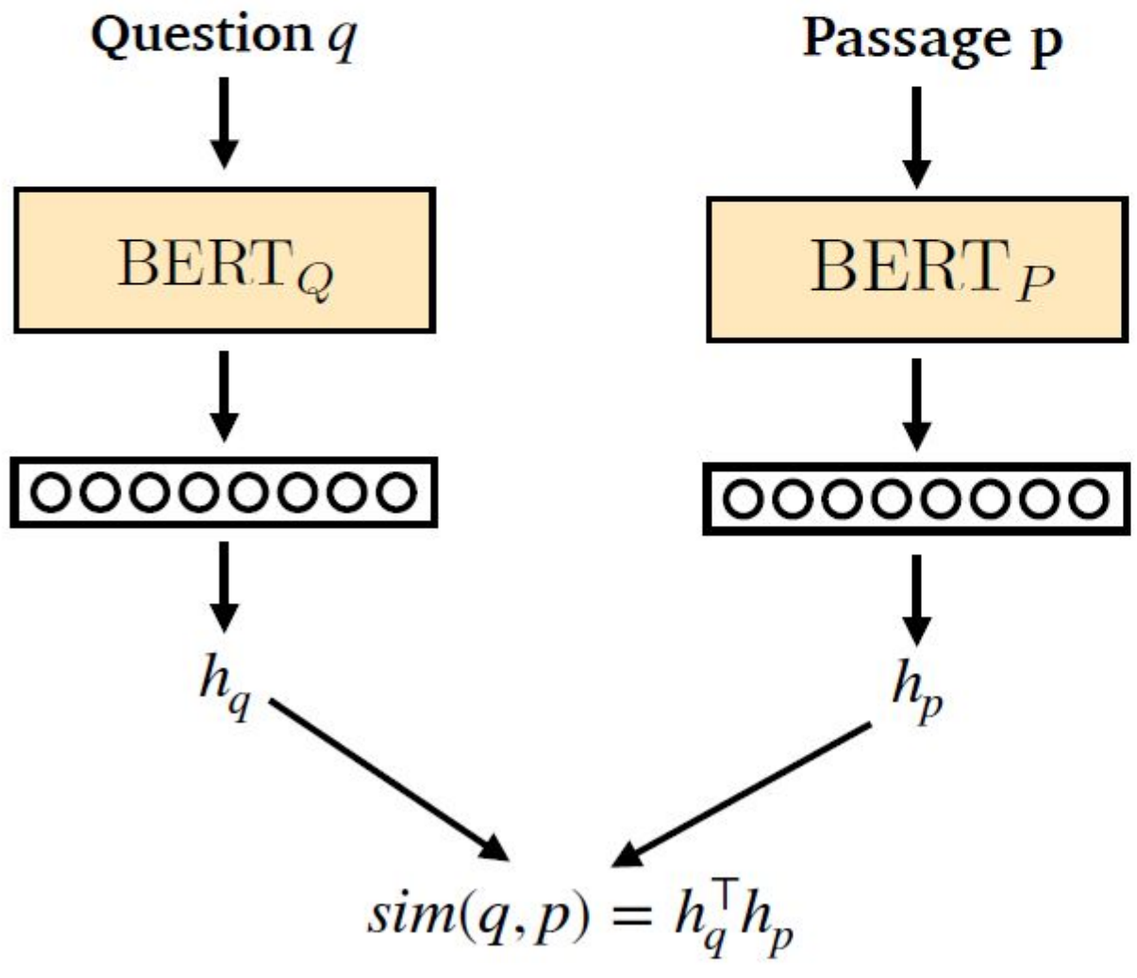
“Who is the bad guy in lord of the rings?”

Sala Baker is an actor and stuntman from New Zealand. He is best known for portraying the ***villain*** *Sauron in the Lord of the Rings trilogy by Peter Jackson.*

Dense Passage Retrieval (DPR)

[Karpukhin et al., 2020]

Key contribution: you can train a dense retrieval only from a small number of Q/A pairs, without any pre-training!



How to get positives and negatives?

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

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

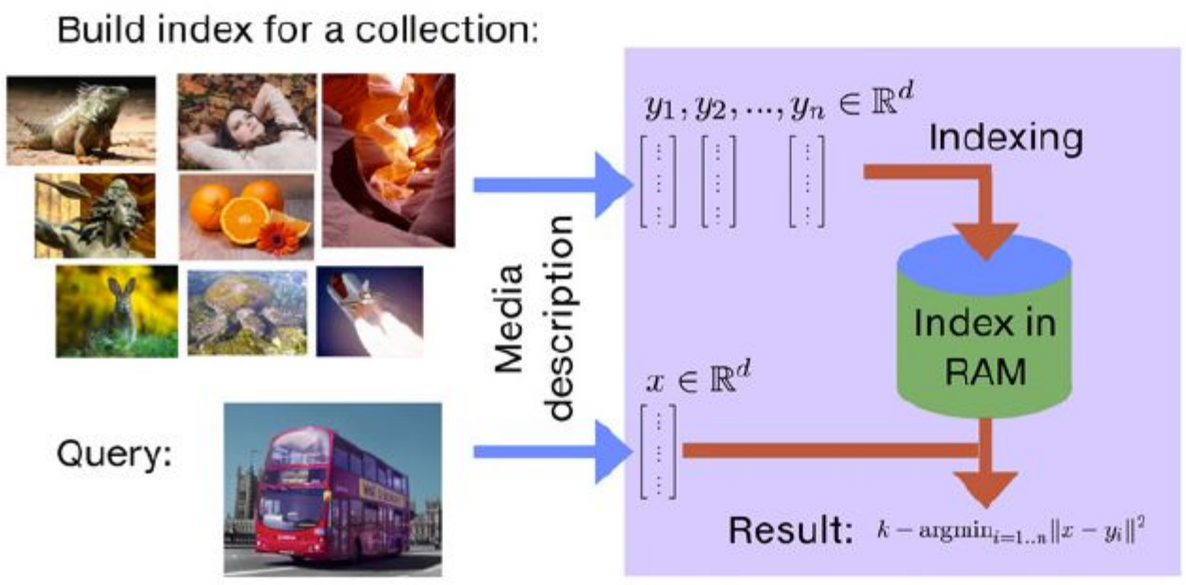
Karpukhin et al., 2020. Dense Passage Retrieval for Open-Domain Question Answering

Indexing

Typically expensive in high-D
Lots of inner-products

MIPS=Indexing for kNN (such as inverted index for tf-idf)

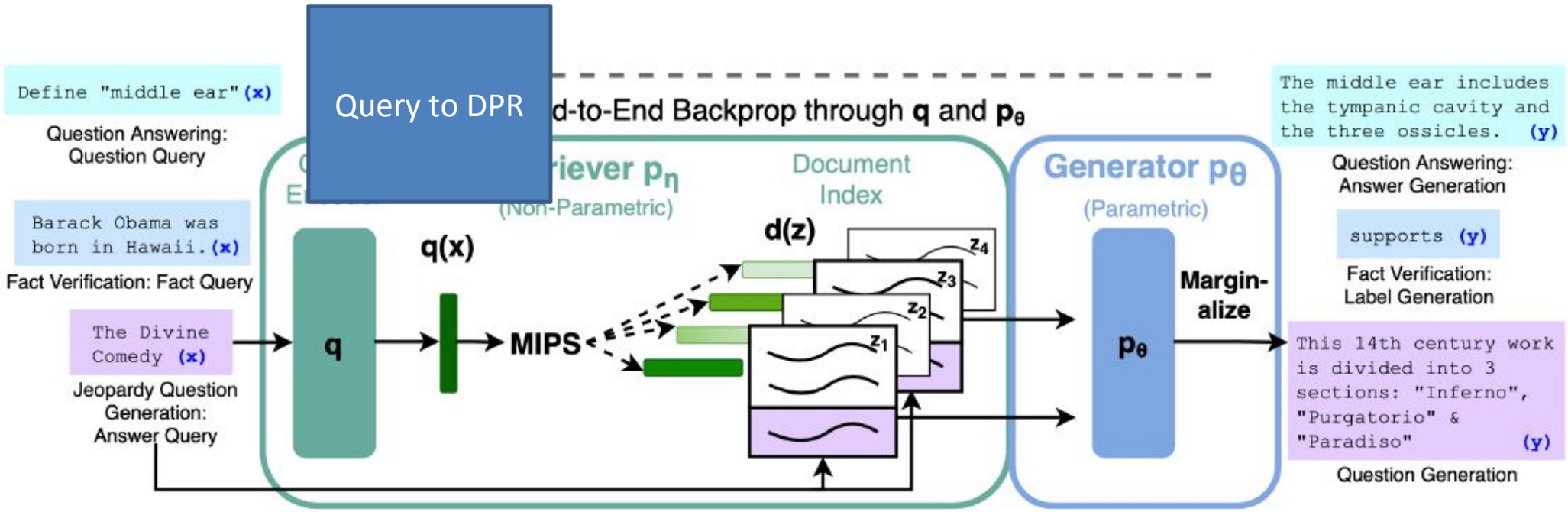
- Now we have much better techniques and tools to support fast maximum inner product search (MIPS):
- In-memory data structure and indexing schemes



e.g., FAISS
[Johnson et al., 2017]

Two Google papers for "joint" optimization

"Retrieval Augmented Generation"



Answer generat
Not extracted

$$p_{\text{RAG-Sequence}}(y|x) = \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) \prod_i^N p_\theta(y_i|x, z, y_{1:i-1})$$

retrieval model

seq2seq model

Lewis et al., 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

“Retrieval Augmented LM” == Expensive

REALM: **Re**trieval-**a**ugmented **L**anguage **M**odel [Guu et al., 2020]

Pre-training on both retriever and reader!

$$p(y \mid x) = \sum_{z \in \mathcal{Z}} p(y \mid z, x)\, p(z \mid x)$$

$\mathcal{Z}$ = Top(K) passages