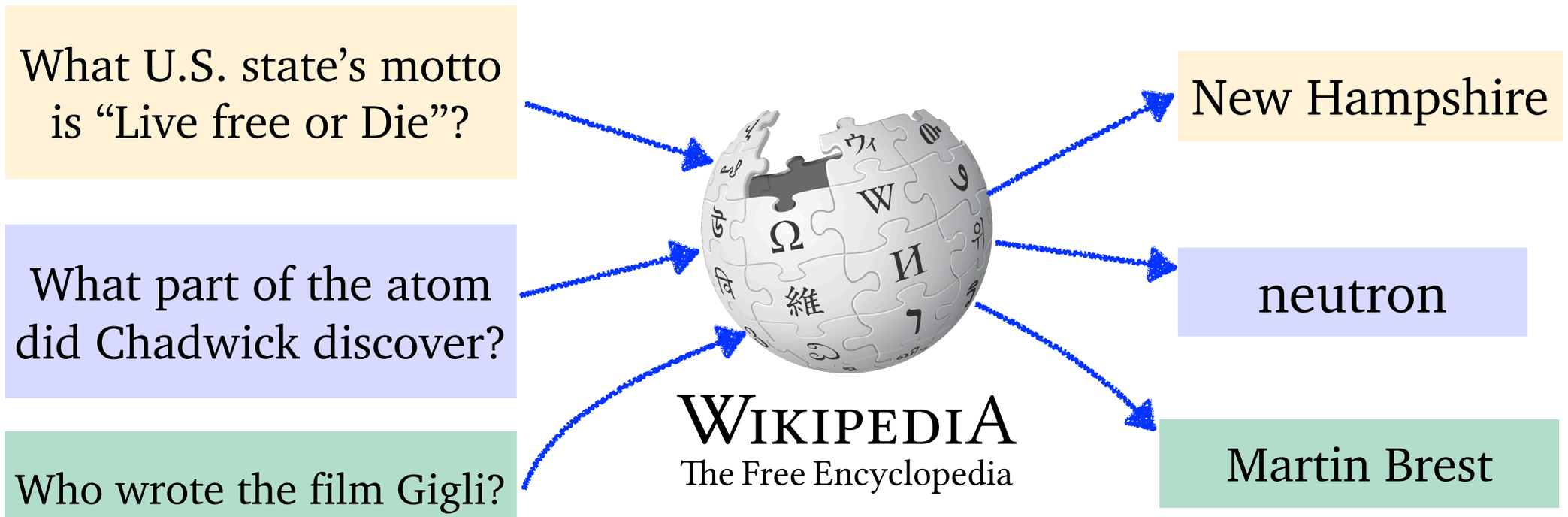
Part IV

Two-stage **retriever-reader** approaches

Problem setup

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



"Machine Reading at Scale"

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

Why is Wikipedia?

- It treats Wikipedia as a generic collection of articles and internal graph structured is not considered in this setting $\Longrightarrow$ easy to extend to any collection of documents.

- The search problem is challenging and realistic while its scale is still manageable (especially for academic research).

- Wikipedia contains a wealth of information of real-world facts. We don’t need to consider the *redundancy* problem too much here.

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

Document Reader

833,500

Information Retrieval

Reading Comprehension

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

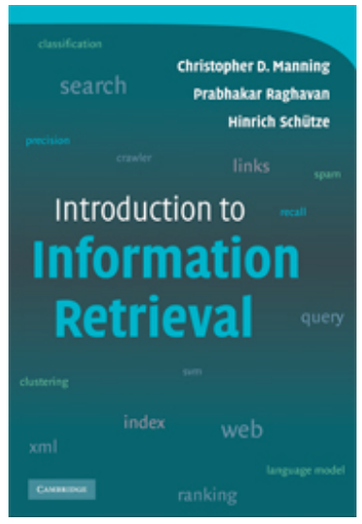
Document Retriever

- A TF-IDF weighted term vector model over unigrams/bigrams:

$$\text{tf-idf}(t, d, D) = \text{tf}(t, d) \times \text{idf}(t, D)$$

$$\text{tf}(t, d) = \log\left(1 + \text{freq}(t, d)\right)$$

$$\text{idf}(t, D) = \log\left(\frac{|D|}{|d \in D : t \in d|}\right)$$



tf =term frequency, idf = inverse document frequency
t: term, d: document (= one Wiki. article), D: corpus (= Wikipedia)

- Important - This retriever is not *trainable*.
- DrQA considers the retrieval problem at document level instead of paragraph level.

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

Document Reader

- It casts as a *reading comprehension* problem:
  - Input is a passage *P* and a question *Q*
  - Output is answer *A*. When the problem is restricted as a segment of text in the passage, the problem is also called as “extractive question answering”.

(Rajpurkar et al, 2016):
Stanford Question Answering Dataset

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

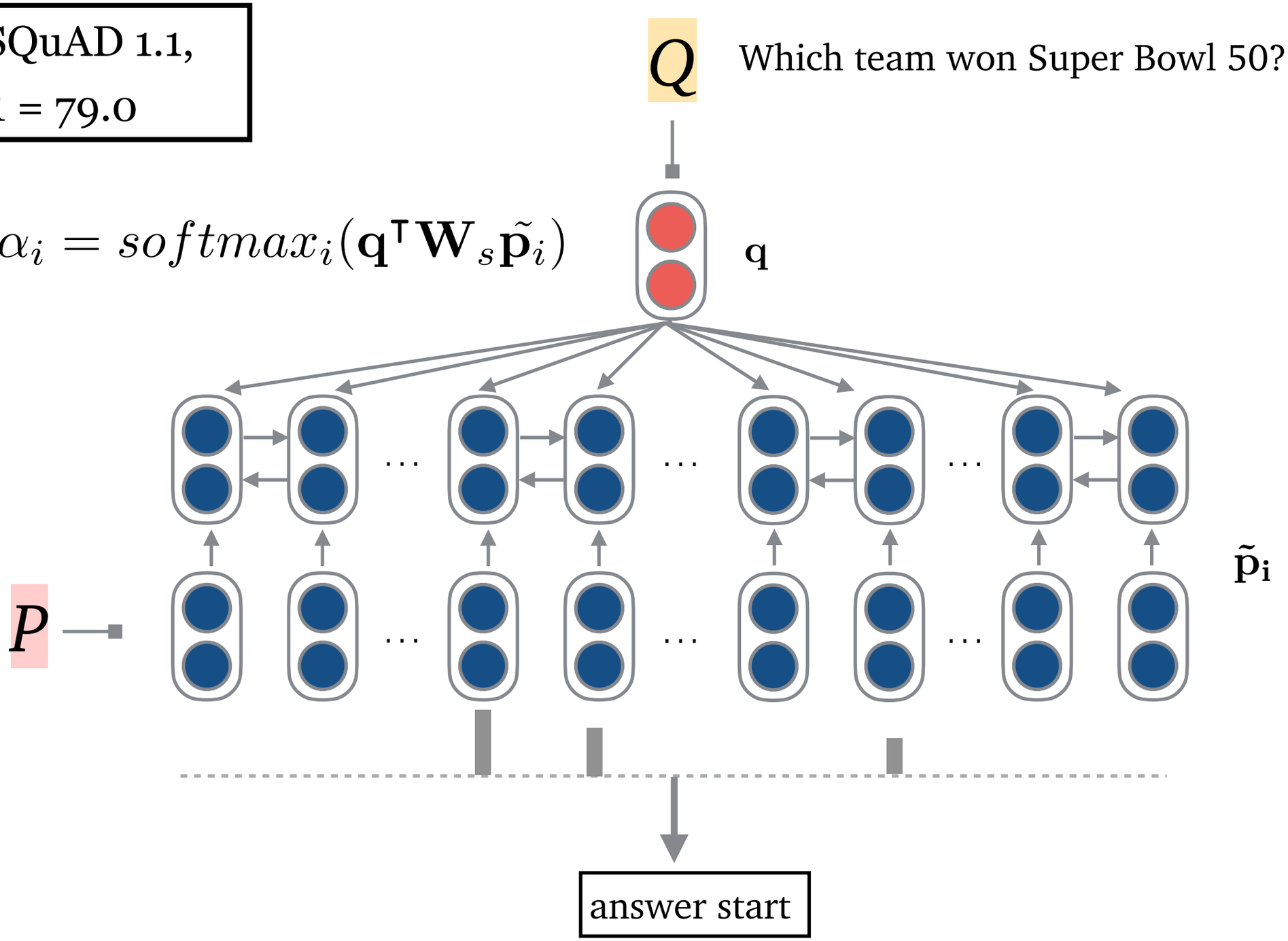
Chen et al., 2017. Reading Wikipedia to Answer Open-domain Questions

Document Reader

On SQuAD 1.1,

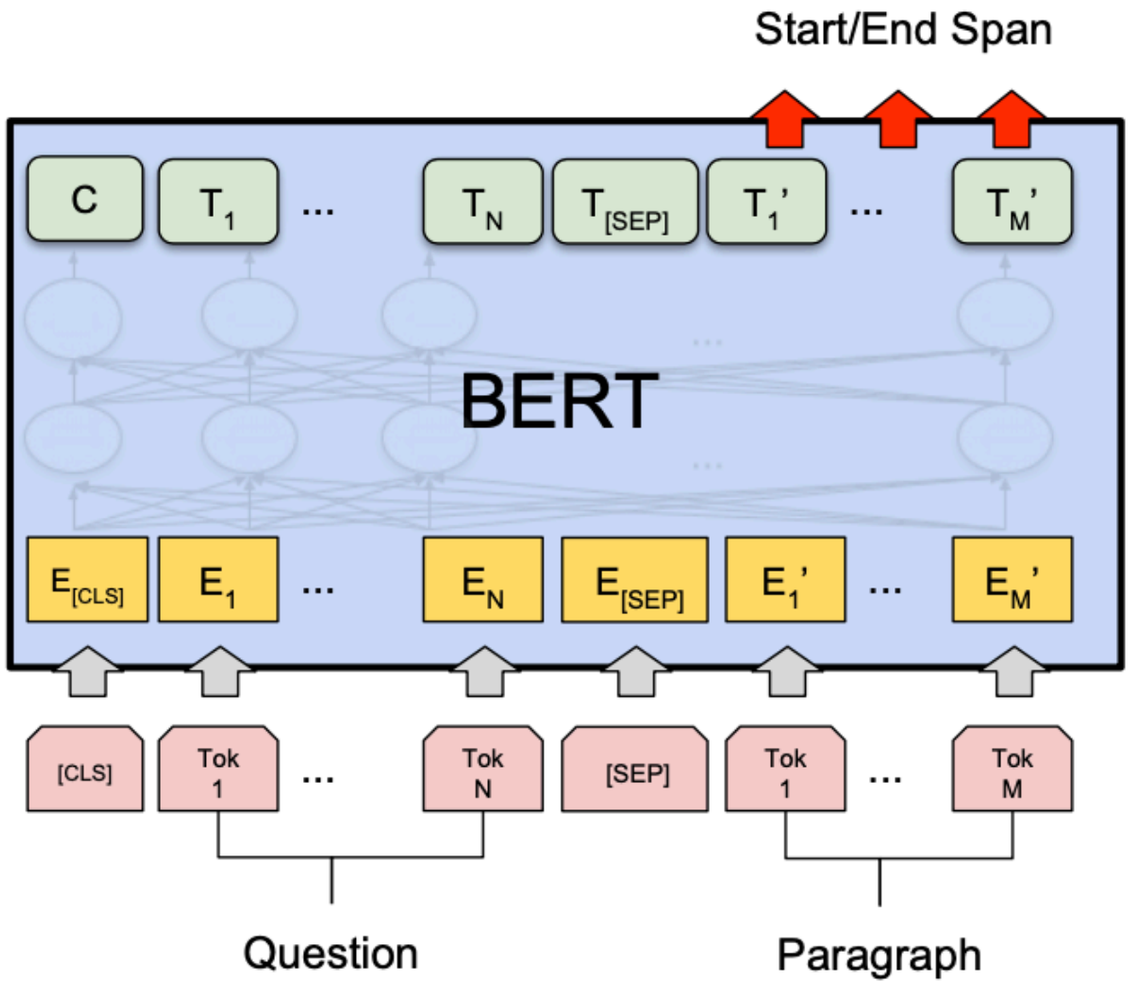
- F1 = 79.0

$$\alpha_i = softmax_i(\mathbf{q}^\intercal \mathbf{W}_s \tilde{\mathbf{p}}_i)$$



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,

- BiDAF +  : F1 = 85.8
-  : F1 = 90.9
- RoBERTa: F1 = 94.6

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

How to train the reader?

- Using an existing reading comprehension dataset (e.g., SQuAD)!

$$\mathcal{D}_{\text{rc}} = \{(P_i, Q_i, A_i)\}$$

Problem: very different distribution with real-world QA data.

- How about other QA datasets (e.g., WebQuestions, TREC-QA)?

$$\mathcal{D}_{\text{QA}} = \{(Q_i, A_i)\}$$

- Solution: create new training examples using our **retriever**!

(Q, A) ⟶ (P, Q, A)

if passage is retrieved and answer can be found in passage

Similar to distant supervision in information extraction (Mintz et al, 2009)

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

How to train the reader?

Question: What U.S. state's motto is "Live free or Die"?
Answer: New Hampshire

Passage

Live Free or Die

From Wikipedia, the free encyclopedia

"**Live Free or Die**" is the official motto of the U.S. state of **New Hampshire**, adopted by the state in 1945.[1] It is possibly the best-known of all state mottos, partly because it conveys an assertive independence historically found in American political philosophy and partly because of its contrast to the milder sentiments found in other state mottos.

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

Putting it together

Training time:

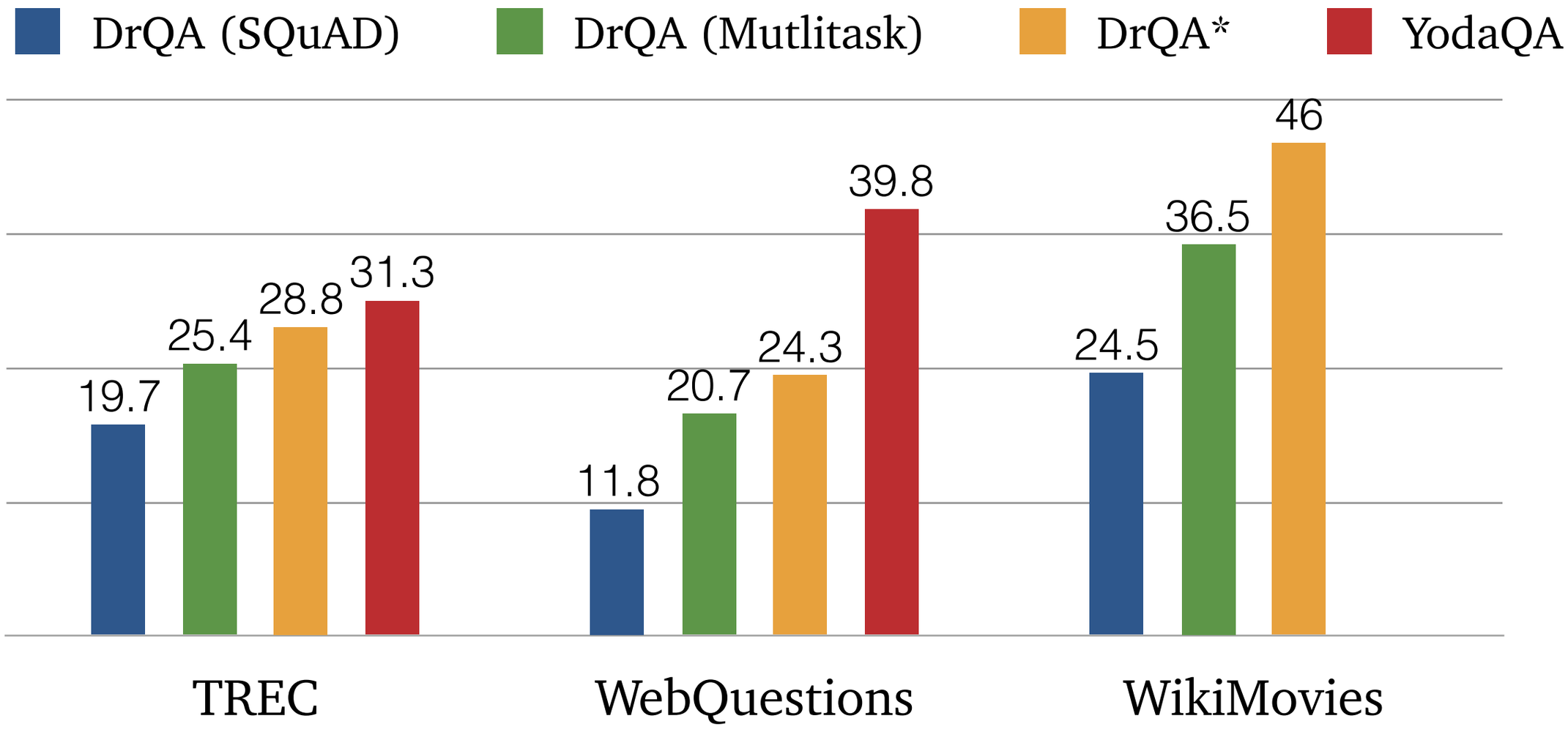
- Document retriever: not trained
- Document reader: a reading comprehension model trained on SQuAD + distantly-supervised data generated from QA datasets

Inference time:

- The document retriever returns *top 5 documents*
- The reader reads every (natural) passage in these 5 documents and predicts an answer and its span score.
- The system finally returns the answer with the highest (unnormalized) span score.

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

Experiments



% of questions answered correctly
(top-1 prediction, exact match)

DrQA*: see [Raison et al., 2018]
YodaQA: http://ailao.eu/yodaqa/

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

How can we do better?

- DrQA considers retrieval at document level.

  Does passage-level retriever work better?

- Answers in the retrieved passages might not be directly comparable at inference time.

  Does multi-passage training help?

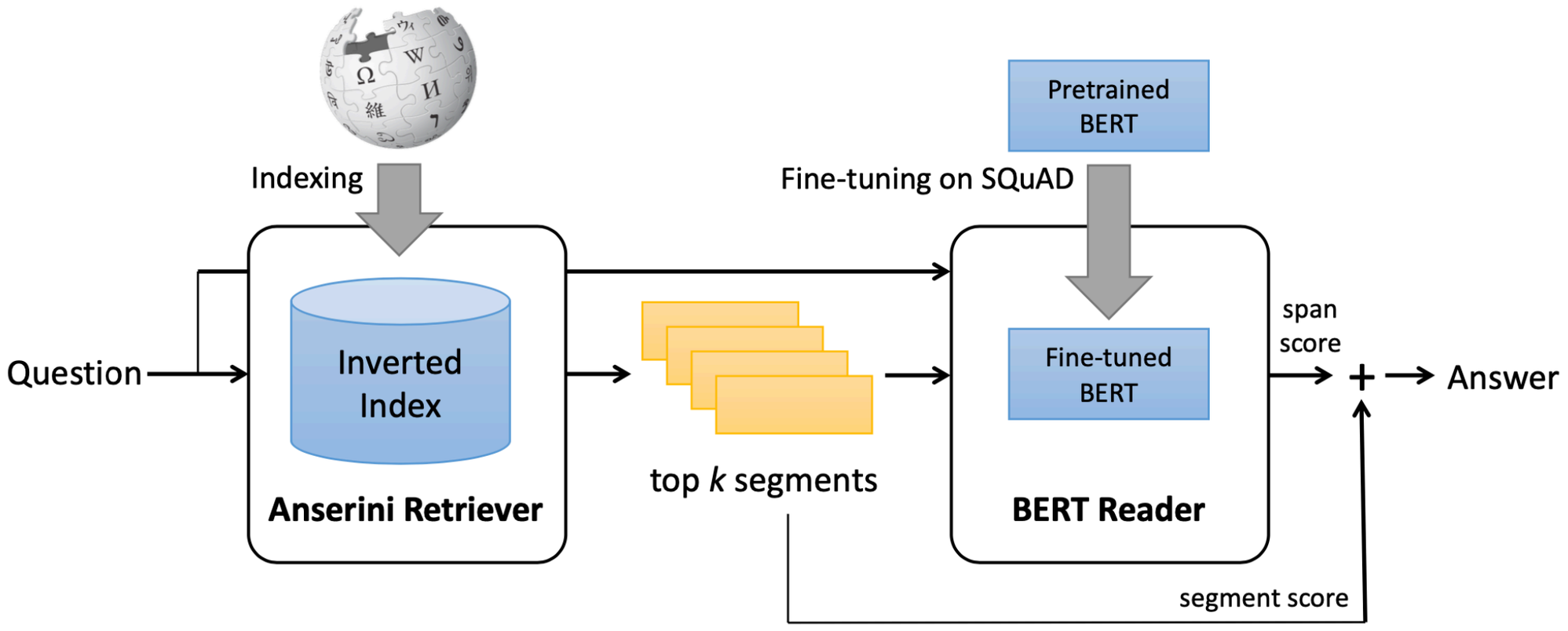
- The importance of each passage has been omitted.

  Can we train a ranker on the retrieved passages?

- The retriever is not trained!

  The focus of the next part.

BERTserini [Yang et al., 2019]



Anserini Retriever:
Lucene with BM25, operated on 29.5M paragraphs

BERT Reader:
Trained on SQuAD

Scoring from both retriever and reader:

$$S = (1 - \mu) \cdot S_{\text{Anserini}} + \mu \cdot S_{\text{BERT}}$$

This system is only evaluated on SQuAD though:

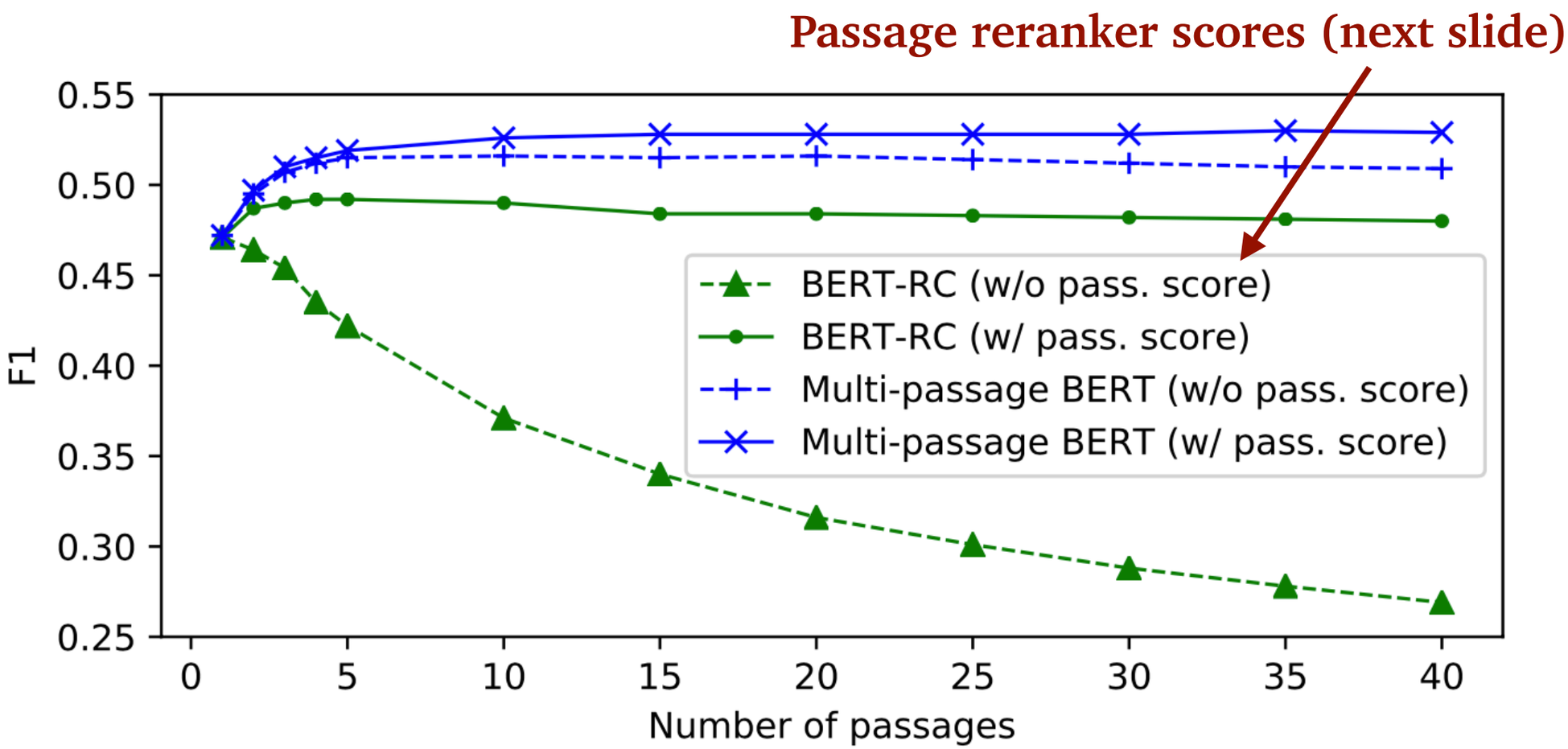
27.1 (DrQA, SQuAD only) $\longrightarrow$ **38.6**

Yang et al., 2019. End-to-End Open-Domain Question Answering with BERTserini

Multi-passage training

[Clark and Gardner, 2018; Wang et al., 2019]

- **Shared normalization**: process paragraphs independently, but compute the span probability across spans in all paragraphs in each mini-batch



Clark and Gardner, 2018. Simple and Effective Multi-Paragraph Reading Comprehension
Wang et al., 2019. Multi-passage BERT: A Globally Normalized BERT Model for Open-domain Question Answering

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.
- Let the retriever return the top N passages and we can train a ranker component to select the best passage.

$$
\begin{aligned}
\mathbf{u}_i &= \text{MaxPooling}(\mathbf{H}_i^{\text{Rank}}), \\
\mathbf{C} &= \text{Tanh}\left(\mathbf{W}^c[\mathbf{u}_1; \mathbf{u}_2; ...; \mathbf{u}_N] + \mathbf{b}^c \otimes \mathbf{e}_N\right), \\
\gamma &= \text{Softmax}(\mathbf{w}^c\mathbf{C}),
\end{aligned}
$$

$\mathbf{u}_i$ is a representation of (P_i, Q)

- This reranker can be easily trained using distant supervision: whether the passage contains the answer or not.
- A better solution is to use training signal from the reader (next slide).

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 Answering

Reinforced ranker-reader

Algorithm 1 Reinforced Ranker-Reader (R^3)

1: **Input:** $\mathbf{a}^g$, $\mathbf{q}$, passages from IR
2: **Output:** Θ
3: **Initialize:** $\Theta \leftarrow$ pre-trained Θ with a baseline method[6]
4: **for** each $\mathbf{q}$ in dataset **do**
5: For question $\mathbf{q}$, sample K passages from the top N passages retrieved by IR model for training. [7]
6: Randomly sample a positive passage $\tau \sim \pi(\tau|\mathbf{q})$
7: Extract the answer $\mathbf{a}^{rc}$ through RC model
8: Get reward r according to $R(\mathbf{a}^g, \mathbf{a}^{rc}|\tau)$.
9: Updating Ranker (ranking model) through policy gradient $r\frac{\partial}{\partial\Theta}\log(\pi(\tau|\mathbf{q}))$
10: Updating Reader (RC model) through supervised gradient $\frac{\partial}{\partial\Theta}L(\mathbf{a}^g|\tau,\mathbf{q})$
11: **end for**

$$R(\mathbf{a}^g, \mathbf{a}^{rc}|\tau) = \begin{cases} 2, & \text{if } \mathbf{a}^g == \mathbf{a}^{rc} \\ f1(\mathbf{a}^g, \mathbf{a}^{rc}), & \text{else if } \mathbf{a}^g \cap \mathbf{a}^{rc} != \emptyset \\ -1, & \text{else} \end{cases}$$

Wang et al., 2018. R³: Reinforced Ranker-Reader for Open-Domain Question Answering

Passage re-ranker: results

	Quasar-T		$\text{SQuAD}_{\text{OPEN}}$		WikiMovies		CuratedTREC		WebQuestions	
	F1	**EM**	**F1**	**EM**	**F1**	**EM**	**F1**	**EM**	**F1**	**EM**
GA (Dhingra et al. 2017)	26.4	26.4	-	-	-	-	-	-	-	-
BiDAF (Seo et al. 2017)	28.5	25.9	-	-	-	-	-	-	-	-
DrQA (Chen et al. 2017a)	-	-	-	28.4	-	34.3	-	25.7	-	**19.5**
Single Reader (SR)	$38.5^{.2}$	$31.5^{.2}$	$35.4^{.2}$	$26.9^{.2}$	$38.8^{.1}$	$37.7^{.1}$	$33.6^{.6}$	$27.4^{.4}$	$22.0^{.2}$	$15.2^{.3}$
Simple Ranker-Reader (SR^2)	$38.8^{.2}$	$31.9^{.2}$	$35.8^{.2}$	$27.2^{.2}$	$39.3^{.1}$	$38.1^{.1}$	$33.4^{.6}$	$27.7^{.5}$	$22.5^{.3}$	$15.6^{.4}$
Reinforced Ranker-Reader (R^3)	$\mathbf{40.9}^{.3}$	$\mathbf{34.2}^{.3}$	$\mathbf{37.5}^{.2}$	$\mathbf{29.1}^{.2}$	$\mathbf{39.9}^{.1}$	$\mathbf{38.8}^{.1}$	$\mathbf{34.3}^{.6}$	$\mathbf{28.4}^{.6}$	$\mathbf{24.6}^{.3}$	$17.1^{.3}$
DrQA-MTL (Chen et al. 2017a)	-	-	-	29.8	-	36.5	-	25.4	-	20.7
YodaQA (Baudiš and Šedivỳ 2015)	-	-	-	-	-	-	-	31.3	-	39.8

Single Reader < Simple Ranker-Reader < Reinforced Ranker-Reader

The improvement is relatively limited though.

Wang et al., 2018. R^3: Reinforced Ranker-Reader for Open-Domain Question Answering

Hard EM Learning [Min et al., 2019]

When a retrieved passage contains multiple mentions of the answer, we don't know which span is the correct one.

Given

Q: Which composer did pianist Clara Wieck marry in 1840?
A: Robert Schumann

Retrieved passage

Robert Schumann was a German composer and influential music critics of the Romantic era. (…) Robert Schumann himself refers to it as "an affliction of the whole hand" (…) Robert Schumann is mentioned in a 1991 episode of Seinfeld "The Jacket" (…) Clara Schumann was a German musician and composer. Her husband was the composer Robert Schumann. (…) Brahms met Joachim in Hanover, made a very favorable impression on him, and got from him a letter of introduction to Robert Schumann.

Min et al., 2019. A Discrete Hard EM Approach for Weakly Supervised Question Answering

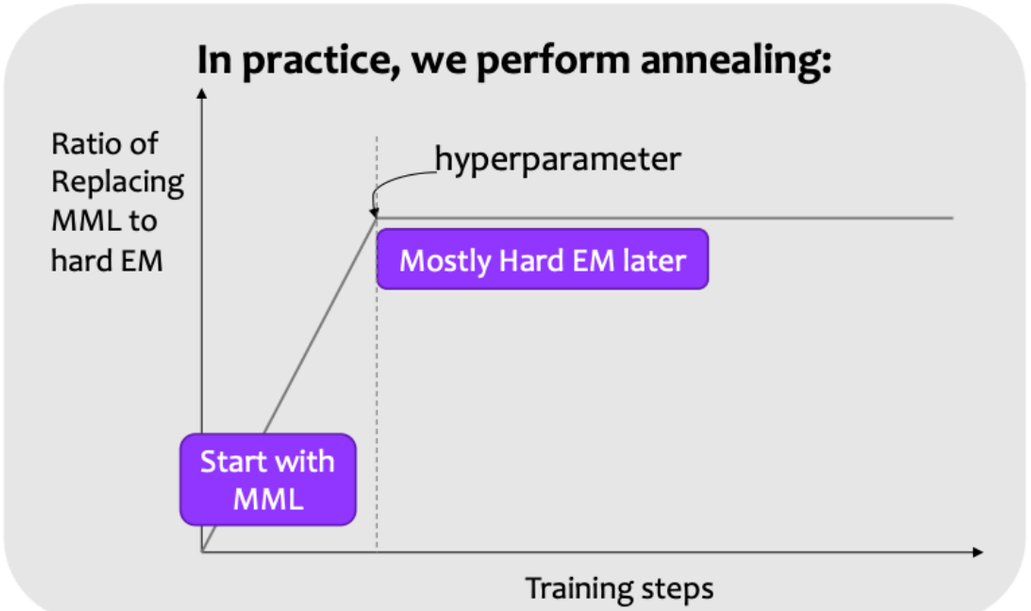
Hard EM Learning

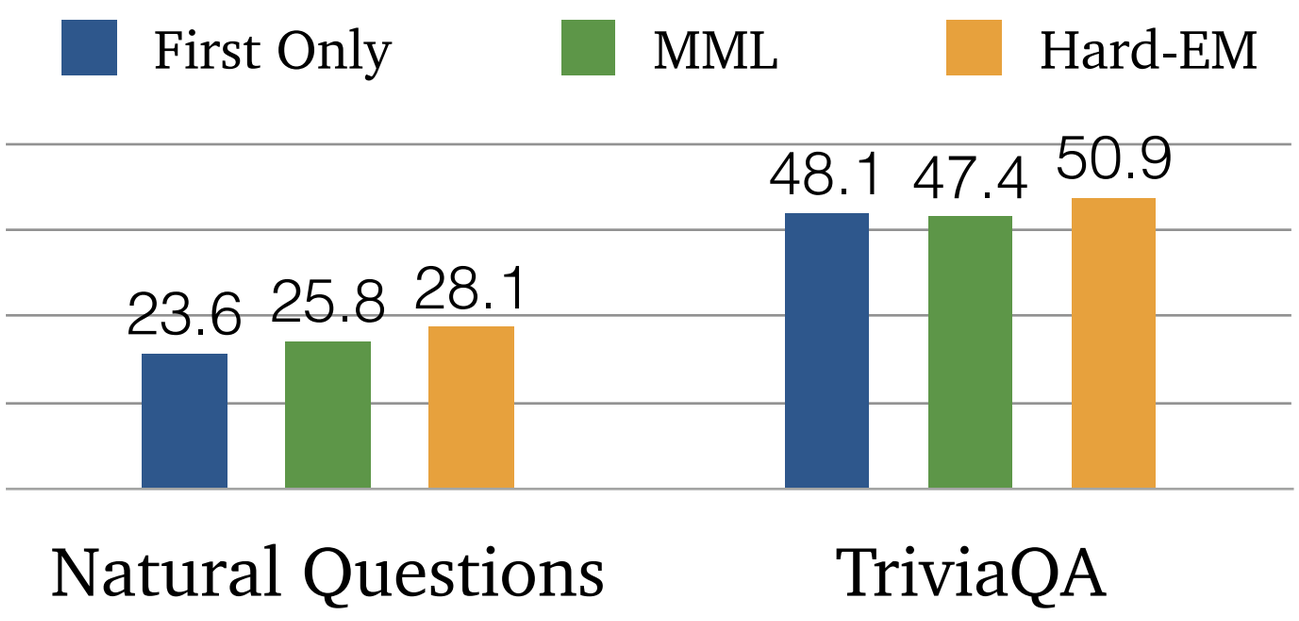
Hard-EM approach

First Only: $J(\theta) = -\log \mathbb{P}(z_1|x;\theta)$, where z_1 appears first in the given document among all $z_i \in Z$.

MML: $J(\theta) = -\log \Sigma_{i=1}^{n} \mathbb{P}(z_i|x;\theta)$.

Ours: $J(\theta) = -\log \max_{1 \le i \le n} \mathbb{P}(z_i|x;\theta)$.





Hard-EM > First Only, MML

Min et al., 2019. A Discrete Hard EM Approach for Weakly Supervised Question Answering

Training an answer re-ranker [Wang et al., 2018]

If every passage returns a candidate answer, can we re-rank the answer candidates based on all their evidence?

Question2: Which physicist , mathematician and astronomer discovered the first 4 moons of Jupiter

A1: Isaac Newton

P1: **Sir Isaac Newton** was an English physicist , mathematician , astronomer , natural philosopher , alchemist and theologian …

P2: **Sir Isaac Newton** was an English mathematician, astronomer, and physicist who is widely recognized as one of the most influential scientists …

Question2: Which physicist , mathematician and astronomer discovered the first 4 moons of Jupiter

A2: Galileo Galilei

P1: **Galileo Galilei** was an Italian physicist , mathematician , astronomer , and philosopher who played a major role in the Scientific Revolution .

P2: **Galileo Galilei** is credited with discovering the first four moons of Jupiter .

Wang et al., 2018. Evidence Aggregation for Answer Re-Ranking in Open-Domain Question Answering

Training an answer re-ranker

If every passage returns a candidate answer, can we re-rank the answer candidates based on all their evidence?

- ***Strength***-based re-ranker: if an answer candidate is supported by multiple pieces of evidence (with high confidence), the answer is more likely to be correct.
- ***Coverage***-based re-ranker: one answer candidate is more likely to be answer if the union of its evidence covers most information in the question.

This works for Quasar-T, SearchQA, TriviaQA when there is enough redundancy but is not evaluated on the Wikipedia setting yet.

Wang et al., 2018. Evidence Aggregation for Answer Re-Ranking in Open-Domain Question Answering

Summary

- Document/passage retriever + neural reading comprehension largely simplified the open-domain QA pipeline

- Several ways to further improve performance:
  - Globally normalized multi-passage training
  - Passage re-ranker
  - Answer re-ranker
  - Improved training methods