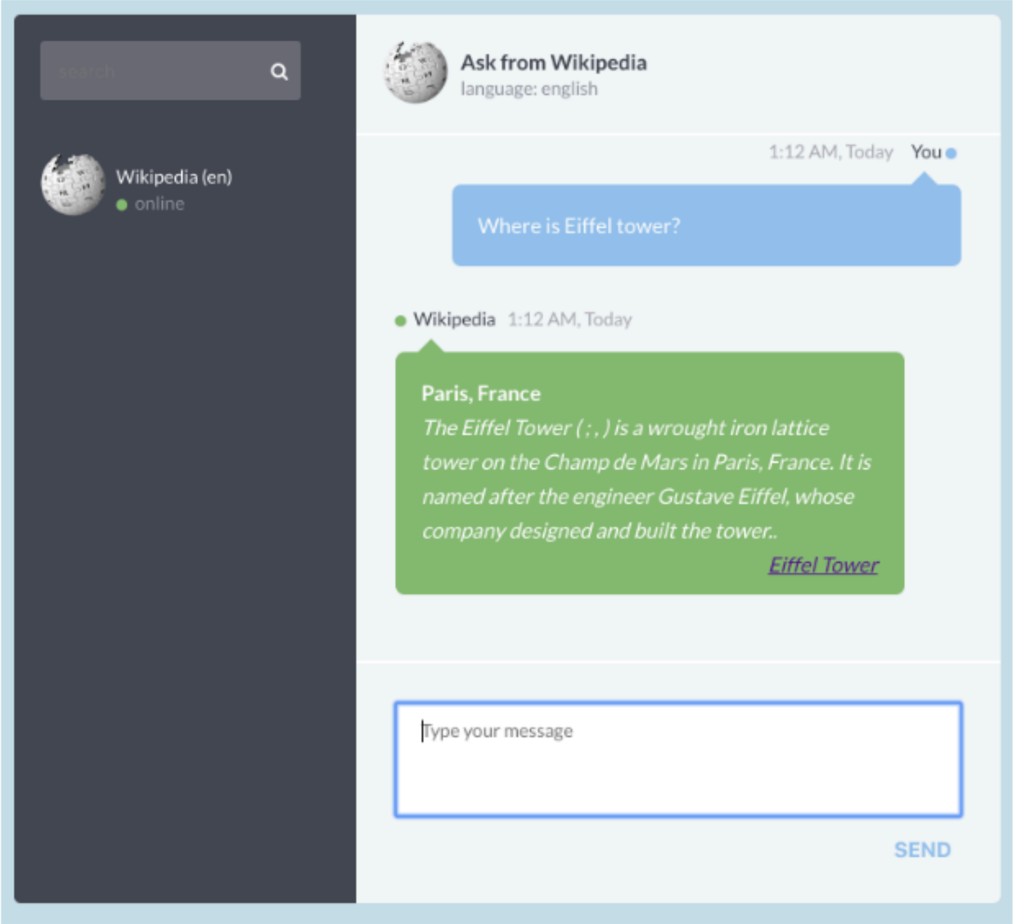
Open-domain Question Answering

Presented by Kun Lu and Chris Sciavolino

03/10/2020

Open-domain Question Answering



DrQA Web UI: https://github.com/zaghaghi/drqa-webui

A Brief History of Open-domain Question Answering

- **Simmons et al. (1964)** did first exploration of answering questions from an expository text based on matching dependency parses of a question and answer
- **Murax (Kupiec 1993)** aimed to answer questions over an online encyclopedia using IR and shallow linguistic processing
- **The NIST TREC QA track** begun in 1999 first rigorously investigated answering fact questions over a large collection of documents
- **IBM's Jeopardy! System (DeepQA, 2011)** brought attention to a version of the problem; it used an ensemble of many methods
- **DrQA (Chen et al. 2017)** uses IR followed by neural reading comprehension to bring deep learning to Open-domain QA

IBM's Watson and Jeopardy! Challenge

IBM Watson defeated two of Jeopardy's greatest champions in 2011

Sample questions:

Q: Even a broken one of these on your wall is right twice a day

A: clock. Watson got it correctly.

Q: Its largest airport is named for a World War II Hero; its second largest for a World War II Battle

A: Chicago. Watson didn't get it correctly.

vs Reading Comprehension

vs Reading Comprehension

1. Much Harder!

vs Reading Comprehension

1. Much Harder!

Combining challenges of both large-scale open-domain QA and of machine comprehension

vs Reading Comprehension

2. Very General!

vs Reading Comprehension

2. Very General!

the question can be any open-domain questions (instead of questions posed after reading the passage) and this meets people’s real information seeking

Overview

- (Chen et al, ACL’ 2017) **Reading Wikipedia to Answer Open-Domain Questions**
- (Lee et al, ACL’ 2019) **Latent Retrieval for Weakly Supervised Open Domain Question Answering**

Overview

- (Chen et al, ACL’ 2017) **Reading Wikipedia to Answer Open-Domain Questions**
- (Lee et al, ACL’ 2019) **Latent Retrieval for Weakly Supervised Open Domain Question Answering**

Reading Wikipedia to Answer Open-Domain Questions

Danqi Chen*
Computer Science
Stanford University
Stanford, CA 94305, USA
danqi@cs.stanford.edu

Adam Fisch, Jason Weston & Antoine Bordes
Facebook AI Research
770 Broadway
New York, NY 10003, USA
{afisch,jase,abordes}@fb.com

Agenda

1. Introduction of DrQA
2. Document Retriever
3. Document Reader
4. Data
5. Results

Agenda

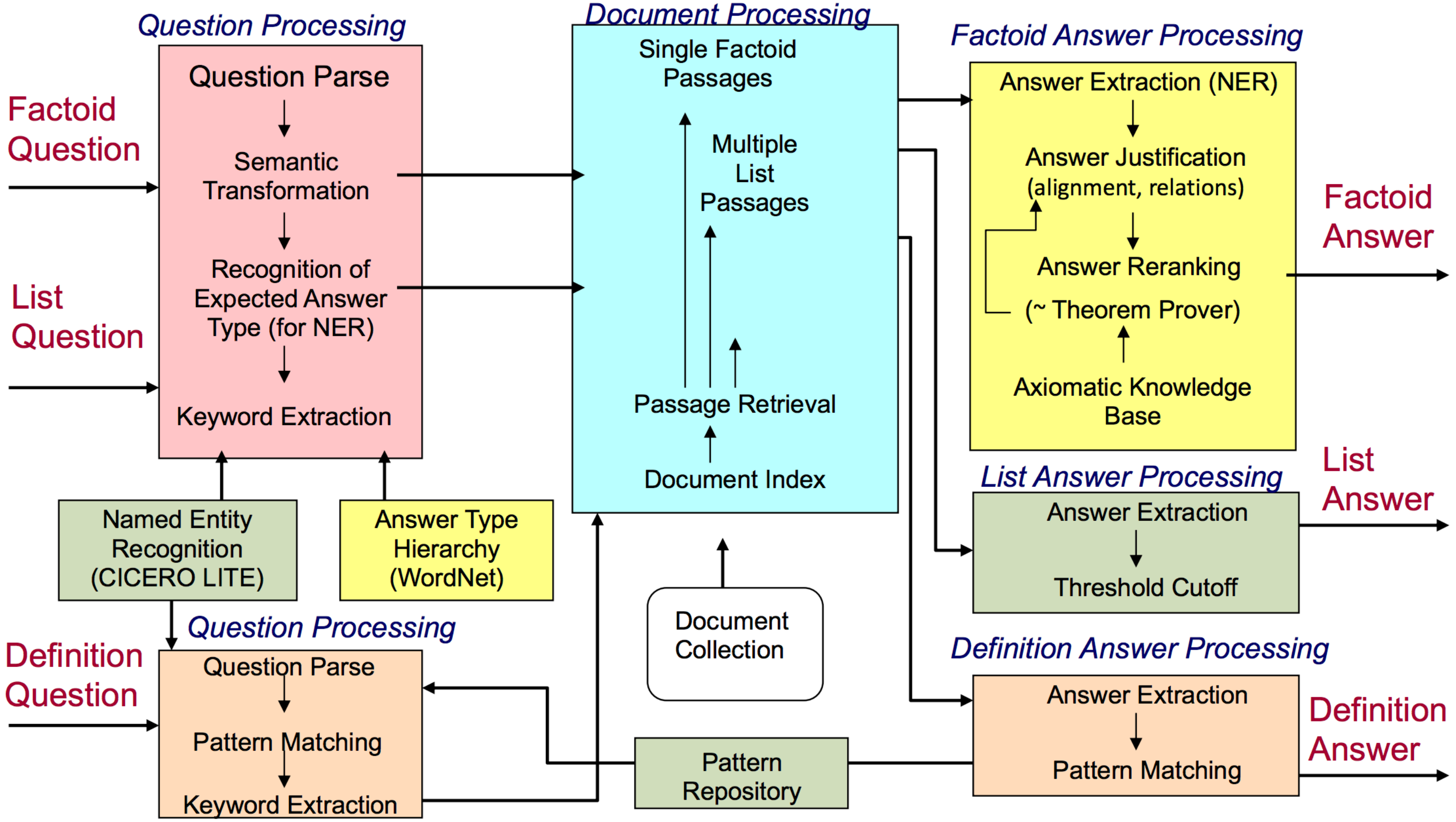
1. Introduction of DrQA
2. Document Retriever
3. Document Reader
4. Data
5. Results

Traditional QA System

Too Complicated…

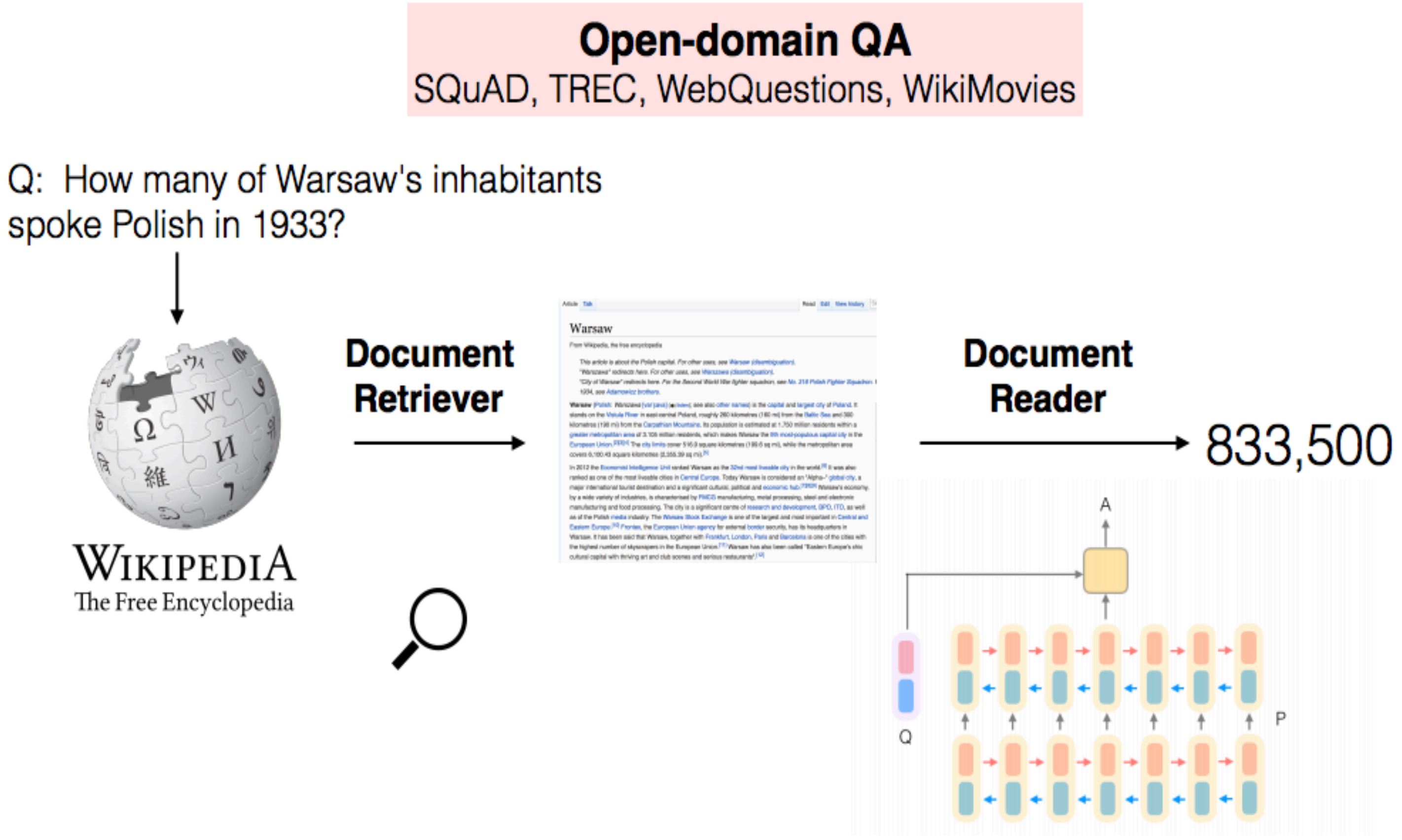
Turn-of-the Millennium Full NLP QA:

[architecture of LCC (Harabagiu/Moldovan) QA system, circa 2003]
Complex systems but they did work fairly well on "factoid" questions



System: DrQA

- Part 1. Document Retriever
  - ✓Finding relevant articles
- Part 2. Document Reader
  - ✓Extracting answers



Contributions

- DrQA was trying to ***reduce this complex problem into a simple two-stage retriever and reader problem***, by combining the challenges from IR and reading comprehension (and this was just a few months after SQuAD came out)

- DrQA: ***can be applied to any large collection of documents*** (e.g. the whole Web) but we chose to use the English Wikipedia as the knowledge source, which consists of 5M articles.

DrQA Demo

https://github.com/facebookresearch/DrQA

Hi!

Hello! Please ask a question.

What is question answering?

a computer science discipline within the fields of information retrieval and natural language processing

Who was the winning pitcher in the 1956 World Series?

Don Larsen

What is the answer to life, the universe, and everything?

42

Agenda

1. Introduction of DrQA
2. Document Retriever
3. Document Reader
4. Data
5. Results

Methods: two steps

1. TF-IDF bag-of-words vectors
2. Efficient bigram hashing (Weinberger et al., 2009)

Methods: two steps

- 1. TF-IDF bag-of-words vectors
- 2. Efficient bigram hashing (Weinberger et al., 2009)

TF-IDF bag-of-words vectors

- TF-IDF vectors:

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

- Improve unigram by considering local word order using n-gram
- Compare articles and questions to retrieve

TF-IDF bag-of-words vectors

- TF-IDF vectors:

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

High dimensional issue?

- Improve unigram by considering local word order using n-gram
- Compare articles and questions to retrieve

Methods: two steps

1. TF-IDF bag-of-words vectors
2. Efficient bigram hashing (Weinberger et al., 2009)

Efficient bigram hashing (Weinberger et al., 2009)

- Map the bigrams to 2^{24} bins with an unsighed ***murmur3*** hash
  - Preserving speed and memory efficiency (Weinberger et al., 2009)

- ***Murmur3:*** Map a word or string to a 32-bit or 128bit value
  - Online: http://murmurhash.shorelabs.com/

Feature Hashing

Sentence	Murmurhash3	Divide by	Reminder
john	3487894951	8	7
likes	1103617568	8	0
movies	3188341541	8	5

Index	Value
0	likes
1	
2	
3	
4	
5	movies
6	
7	john

Feature Hashing

Sentence	Murmurhash3	Divide by	Reminder
john	3487894951	8	7
likes	1103617568	8	0
movies	3188341541	8	5

Index	Value
0	likes
1	
2	
3	
4	
5	movies
6	
7	john

What if we have hash collisions?

Feature Hashing (Weinberger et al., 2009)

- Mathematical formula:

$$\phi_i^{(h,\xi)}(x) = \sum_{j:h(j)=i} \xi(j) x_j$$

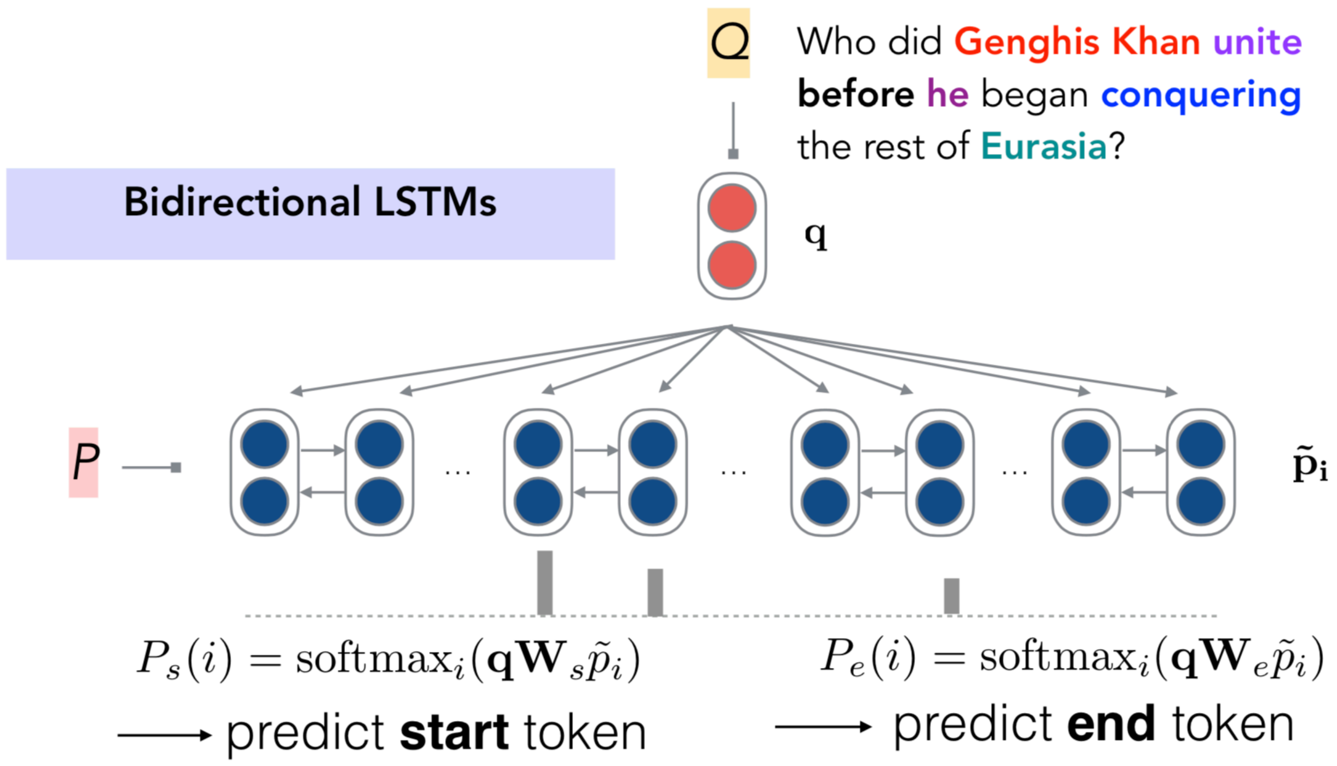
$$\textit{and } \langle x, x' \rangle_\phi := \left\langle \phi^{(h,\xi)}(x), \phi^{(h,\xi)}(x') \right\rangle .$$

- They proved exponential tail Bounds

Agenda

1. Introduction of DrQA
2. Document Retriever
3. Document Reader
4. Data
5. Results

Document Reader



Three steps:

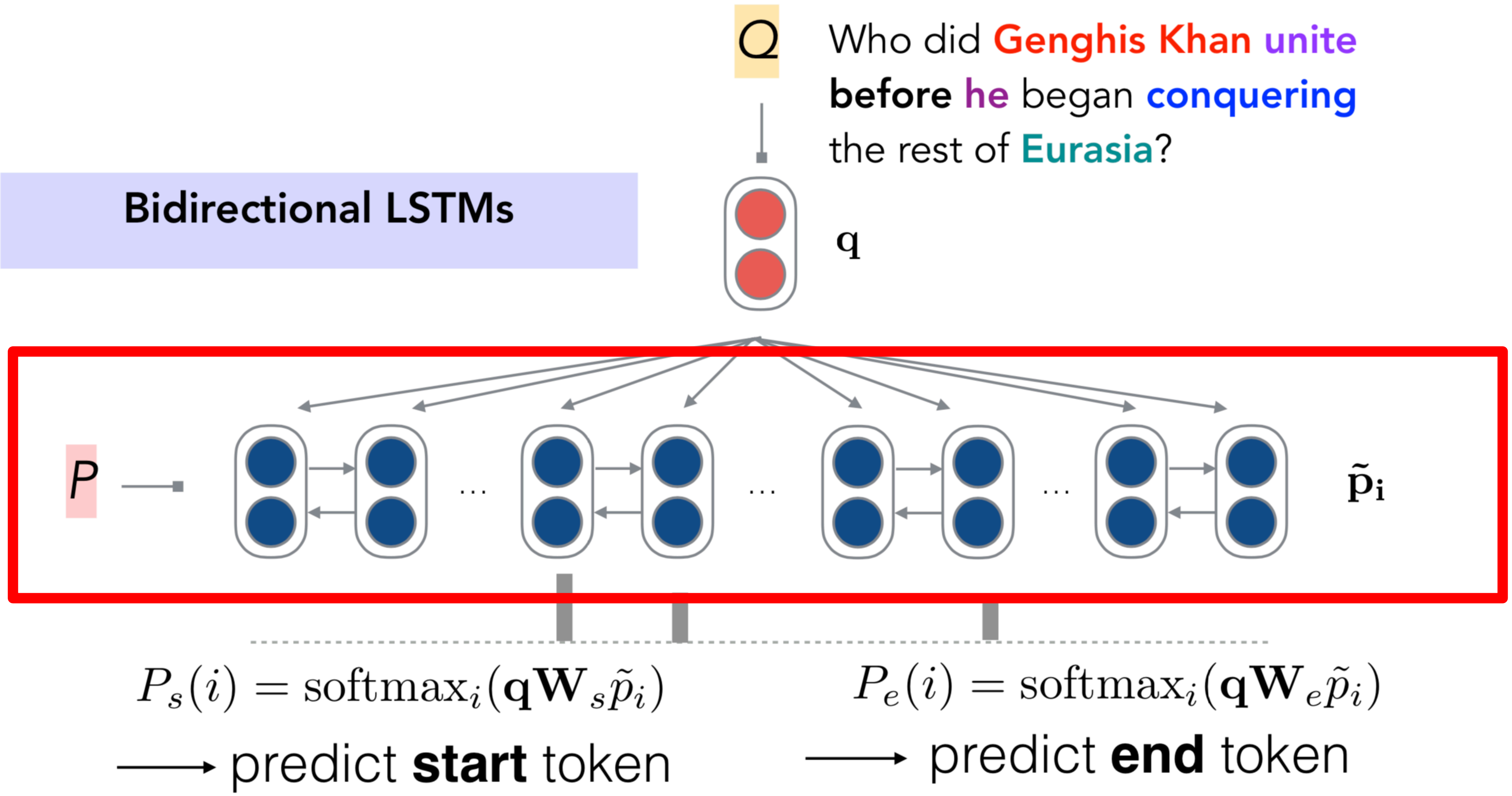
1. Paragraph encoding
2. Question encoding
3. Prediction

similar to AttentiveReader (Hermann et al, 2015; Chen et al, 2016)

Document Reader

Three steps:

1. Paragraph encoding
2. Question encoding
3. Prediction


Bidirectional LSTMs
Q
Who did Genghis Khan unite before he began conquering the rest of Eurasia?
q
P
p̃i
$P_s(i) = \mathrm{softmax}_i(\mathbf{q}\mathbf{W}_s\tilde{p}_i)$
→ predict **start** token
$P_e(i) = \mathrm{softmax}_i(\mathbf{q}\mathbf{W}_e\tilde{p}_i)$
→ predict **end** token

Paragraph encoding

- 1. Represent tokens p_i in a paragraph as a sequence of feature vectors $\tilde{p}_i \in \mathbb{R}^d$
  - Word embedding
  - Exact match
  - Token features
  - Aligned question embedding
- 2. Pass features $\tilde{p}_i$ as the input to a RNN (multi-layer Bidirectional LSTM):

$$\{\mathbf{p}_1, \ldots, \mathbf{p}_m\} = \text{RNN}(\{\tilde{\mathbf{p}}_1, \ldots, \tilde{\mathbf{p}}_m\})$$

Word Embeddings

- $f_{emb}(p_i) = \mathrm{E}(p_i)$
- 300-dimensional Glove word embeddings
- Keep most of the pre-trained word embeddings fixed and only fine tune the 1000 most frequent question key words: what, how, which… (crucial for QA system)

Exact match

- $f_{exact\ match}\ (p_i) = \ \mathbb{I}\ (p_i \in q)$
- Binary features indicating whether p_i can be exactly matched to one question word in q, either in original, lowercase, or lemma form

Token features:

- $f_{token}\ (p_i) = (POS(p_i), NER(p_i), TF(p_i))$
- Part of speech (POS)
- Entity recognition (NER)
- Normalized term frequency (TF)

Aligned Question Embeddings

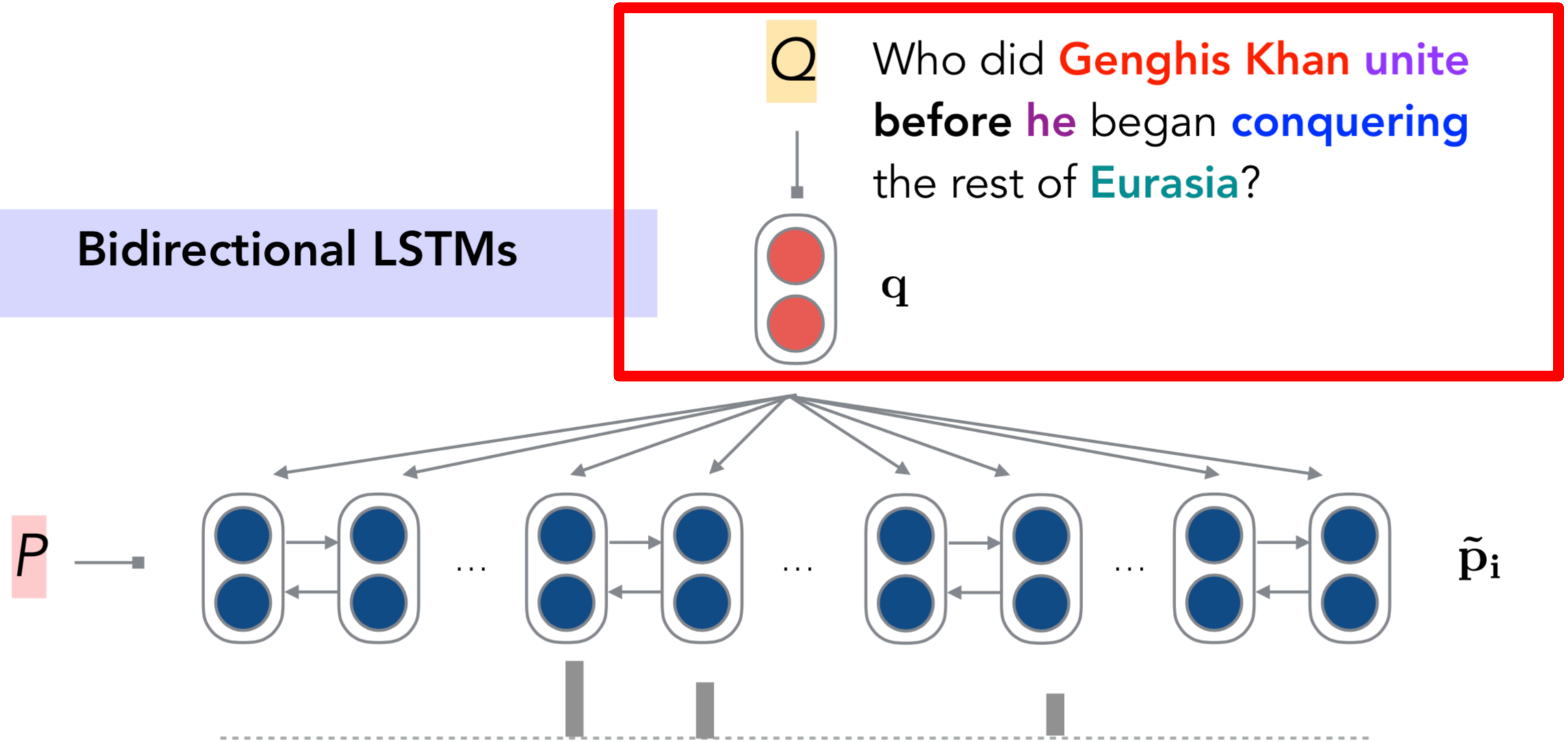
- $f_{align}(p_i) = \Sigma_j \, a_{i,j} \, \mathbb{E}\,(q_j)$
- Where
- $a_{i,j} = \dfrac{\exp(\alpha(\mathbb{E}(p_i) \cdot \alpha(\mathbb{E}(q_j))))}{\mathbb{E}_{j'} \exp(\alpha(\mathbb{E}(p_i) \cdot \alpha(\mathbb{E}(q_{j'}))))}$
- $a_{i,j}$ captures the similarity between p_i and q_j, and $\alpha(\cdot)$ is a single layer with ReLu nonlinearity

Document Reader

Three steps:

1. Paragraph encoding
2. Question encoding
3. Prediction

Bidirectional LSTMs



$$P_s(i) = \mathrm{softmax}_i(\mathbf{q}\mathbf{W}_s\tilde{p}_i)$$

→ predict **start** token

$$P_e(i) = \mathrm{softmax}_i(\mathbf{q}\mathbf{W}_e\tilde{p}_i)$$

→ predict **end** token

Question Encoding

- 1. Apply another RNN to top of word embeddings of q_i and get $\boldsymbol{q}_j$
- 2. Combining the resulting units into one single vector

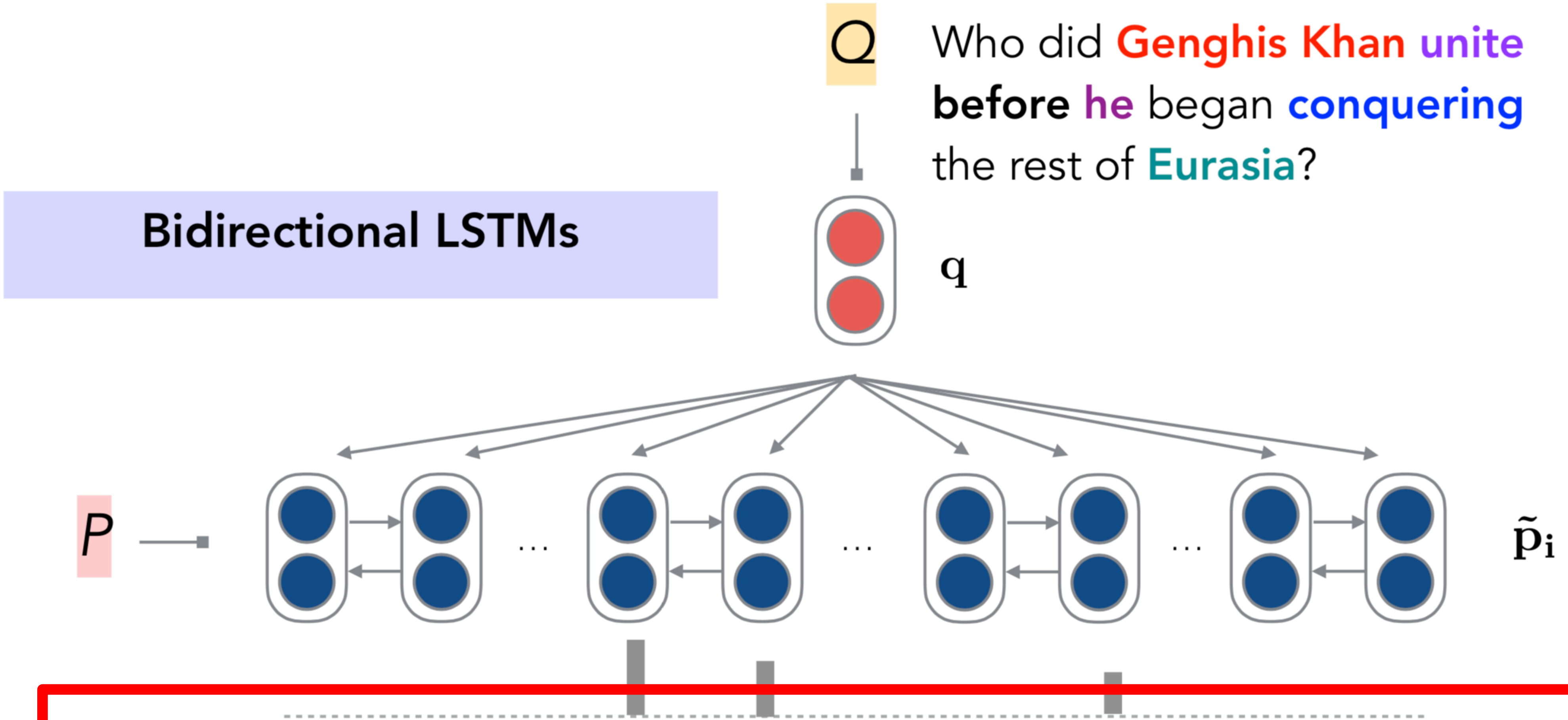
$$\boldsymbol{q} = \Sigma_{\mathrm{j}} b_j \boldsymbol{q}_j \ ,$$

Here $b_j = \frac{\exp(\boldsymbol{w} \cdot q_j)}{\Sigma_{j'} \exp(\boldsymbol{w} \cdot q_{j'})}$, and **w** is a weight vector to learn

Document Reader

Three steps:

1. Paragraph encoding
2. Question encoding
3. Prediction



$$P_s(i) = \mathrm{softmax}_i(\mathbf{q}\mathbf{W}_s\tilde{p}_i)$$

→ predict **start** token

$$P_e(i) = \mathrm{softmax}_i(\mathbf{q}\mathbf{W}_e\tilde{p}_i)$$

→ predict **end** token

Prediction

- Goal: predict the span of tokens that is most likely the correct answer
- Method: train two classifier independently for predicting two ends of span

$$\max_{i,j} P_{start}(i) \times P_{end}(j)$$

such that $i \leq j \leq i + 15$, where $P_{start}(i)$ and $P_{end}(j)$ is probability of each token being start and end:

$$P_{start}(i) \propto \exp(p_i W_s \boldsymbol{q})$$
$$P_{end}(i) \propto \exp(p_i W_e \boldsymbol{q})$$

Agenda

1. Introduction of DrQA
2. Document Retriever
3. Document Reader
4. Data
5. Results

Three types of data:

- Wikipedia: knowledge source for finding answers
- SQuAD: main source to train the Document Reader
- Three more QA datasets (CuratedTREC, WebQuestions, WikiMovies):

In addition to SQuAD, are used to test the DrQA

CuratedTREC

```
# 2.3
Q:What are titles of the group's releases?
Chocolate Starfish and Hot Dog Flavored Water
Significant Others
Three Dollar Bill, Y'all
Nookie
Break Stuff
```

```
# 4.4
Q:What movies did James Dean appear in?
East of Eden
Fixed Bayonet
Giant
Rebel Without a Cause
```

```
# 6.3
Q:Famous people who have been Rhodes scholars.
Maine Congressman Tom Allen
Australian Labor leader Kim Beazley
Alan Bersin
Newark Councilman, Cory Booker
Pres. candidate Bill Bradley
Wesley Clark
President Clinton
Peter Dawkins, Heisman Trophy Winner
Author and conceptual thinker Edward DiBono
Berkeley Law Prof., Judge William Fletcher
Environmentalist William Gronon
VA Sec. of Education Barbara Harmon
Kris Kristofferson
Alain Leroy Locke, 1st black Rhodes Scholar
Terrence Malick, Producer
Author Willie Morris
Mathew Polley, editor of "I Can't Believe It's Not the NYT"
Kurt Schork, correspondent killed
George Stephanapolis
Strobe Talbott
Lousiana Legislator David Vitter
Supreme Court Justice Byron White
Author John Edgar Wideman
Author Naomi Wolf
```

https://trec.nist.gov/data/qa.html

WebQuestions

1. Example 1:
   - Utterance: what is the name of justin bieber brother?
   - TargetValue: Jazmyn Bieber, Jaxon Bieber
   - Url: http://www.freebase.com/view/en/justin_bieber

2. Example 2:
   - Utterance: what character did natalie portman play in star wars?
   - TargetValue: Padm Amidala
   - Url: http://www.freebase.com/view/en/natalie_portman

https://nlp.stanford.edu/software/sempre/

WikiMovies

Doc: Wikipedia Article for Blade Runner (partially shown)
Blade Runner is a 1982 American neo-noir dystopian science fiction film directed by Ridley Scott and starring Harrison Ford, Rutger Hauer, Sean Young, and Edward James Olmos. The screenplay, written by Hampton Fancher and David Peoples, is a modified film adaptation of the 1968 novel "Do Androids Dream of Electric Sheep?" by Philip K. Dick. The film depicts a dystopian Los Angeles in November 2019 in which genetically engineered replicants, which are visually indistinguishable from adult humans, are manufactured by the powerful Tyrell Corporation as well as by other "mega-corporations" around the world. Their use on Earth is banned and replicants are exclusively used for dangerous, menial, or leisure work on off-world colonies. Replicants who defy the ban and return to Earth are hunted down and "retired" by special police operatives known as "Blade Runners". . . .
KB entries for Blade Runner (subset)
Blade Runner *directed_by* Ridley Scott Blade Runner *written_by* Philip K. Dick, Hampton Fancher Blade Runner *starred_actors* Harrison Ford, Sean Young, . . . Blade Runner *release_year* 1982 Blade Runner *has_tags* dystopian, noir, police, androids, . . .
IE entries for Blade Runner (subset)
Blade Runner, Ridley Scott *directed* dystopian, science fiction, film Hampton Fancher *written* Blade Runner Blade Runner *starred* Harrison Ford, Rutger Hauer, Sean Young. . . Blade Runner *labelled* 1982 neo noir special police, Blade *retired* Blade Runner Blade Runner, special police *known* Blade
Questions for Blade Runner (subset)
Ridley Scott directed which films? What year was the movie Blade Runner released? Who is the writer of the film Blade Runner? Which films can be described by dystopian? Which movies was Philip K. Dick the writer of? Can you describe movie Blade Runner in a few words?

Table 1: WIKIMOVIES: Questions, Doc, KB and IE sources.

https://research.fb.com/downloads/babi/

Number of Questions

Dataset	Train		Test
	Plain	DS	
SQuAD	87,599	71,231	10,570[†]
CuratedTREC	1,486*	3,464	694
WebQuestions	3,778*	4,602	2,032
WikiMovies	96,185*	36,301	9,952

Table 2: Number of questions for each dataset used in this paper. DS: distantly supervised training data. *: These training sets are not used as is because no paragraph is associated with each question. [†]: Corresponds to SQuAD development set.

Not used, since no paragraph is associated with each question

Distantly Supervised Data

(Q, A) ⟶ (P, Q, A) if P is retrieved and A can be found in P

Q: What part of the atom did Chadwick discover?

WebQuestions

A: neutron

Atom

From Wikipedia, the free encyclopedia

The atomic mass of these isotopes varied by integer amounts, called the whole number rule.[23] The explanation for these different isotopes awaited the discovery of the **neutron**, an uncharged particle with a mass similar to the proton, by the physicist **James Chadwick** in 1932. Isotopes were then explained as elements with the same number of protons, but different numbers of neutrons within the nucleus.

Example training data

Dataset	Example	Article / Paragraph
SQuAD	**Q**: How many provinces did the Ottoman empire contain in the 17th century? **A**: 32	**Article**: Ottoman Empire **Paragraph**: ... At the beginning of the 17th century the empire contained 32 provinces and numerous vassal states. Some of these were later absorbed into the Ottoman Empire, while others were granted various types of autonomy during the course of centuries.
CuratedTREC	**Q**: What U.S. state's motto is "Live free or Die"? **A**: New Hampshire	**Article**: Live Free or Die **Paragraph**: "Live Free or Die" is the official motto of the U.S. state of New Hampshire, adopted by the state in 1945. 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.
WebQuestions	**Q**: What part of the atom did Chadwick discover?[†] **A**: neutron	**Article**: Atom **Paragraph**: ... The atomic mass of these isotopes varied by integer amounts, called the whole number rule. The explanation for these different isotopes awaited the discovery of the neutron, an uncharged particle with a mass similar to the proton, by the physicist James Chadwick in 1932. ...
WikiMovies	**Q**: Who wrote the film Gigli? **A**: Martin Brest	**Article**: Gigli **Paragraph**: Gigli is a 2003 American romantic comedy film written and directed by Martin Brest and starring Ben Affleck, Jennifer Lopez, Justin Bartha, Al Pacino, Christopher Walken, and Lainie Kazan.

Agenda

1. Introduction of DrQA
2. Document Retriever
3. Document Reader
4. Data
5. Results

Three Parts of Evaluation

1. Document Retriever Evaluation
2. Document Reader Evaluation
3. Full Wikipedia Question Answering

Three Parts of Evaluation

1. Document Retriever Evaluation
2. Document Reader Evaluation
3. Full Wikipedia Question Answering

Document retrieval results

All beat built-in Wikipedia Search API

Dataset	Wiki Search	Doc. Retriever plain	Doc. Retriever +bigrams
SQuAD	62.7	76.1	**77.8**
CuratedTREC	81.0	85.2	**86.0**
WebQuestions	73.7	**75.5**	74.4
WikiMovies	61.7	54.4	**70.3**

Table 3: Document retrieval results. % of questions for which the answer segment appears in one of the top 5 pages returned by the method.

Three Parts of Evaluation

1. Document Retriever Evaluation
2. Document Reader Evaluation
3. Full Wikipedia Question Answering

On SQuAD

Method	Dev		Test	
	EM	F1	EM	F1
Dynamic Coattention Networks (Xiong et al., 2016)	65.4	75.6	66.2	75.9
Multi-Perspective Matching (Wang et al., 2016)†	66.1	75.8	65.5	75.1
BiDAF (Seo et al., 2016)	67.7	77.3	68.0	77.3
R-net†	n/a	n/a	71.3	79.7
DrQA (Our model, Document Reader Only)	**69.5**	**78.8**	$\boxed{70.0}$	$\boxed{79.0}$

Table 4: Evaluation results on the SQuAD dataset (single model only). †: Test results reflect the SQuAD leaderboard (https://stanford-qa.com) as of Feb 6, 2017.

Surpass all the published results and can match the top performance on the SQuAD leaderboard at the time of writing

Three Parts of Evaluation

1. Document Retriever Evaluation
2. Document Reader Evaluation
3. Full Wikipedia Question Answering

Full Wikipedia Question Answer

Three versions of DrQA:

- SQuAD: A single Document Reader model is trained on the SQuAD training set only and used on all evaluation sets

- Fine-tune (DS): A Document Reader model is pre-trained on SQuAD and then fine-tuned for each dataset independently using its distant supervision (DS) training set

- Multitask (DS): A single Document Reader model is jointly trained on the SQuAD training set and all the DS sources

Full Wikipedia Results

Multitask performs the best, reasonable performance across four datasets

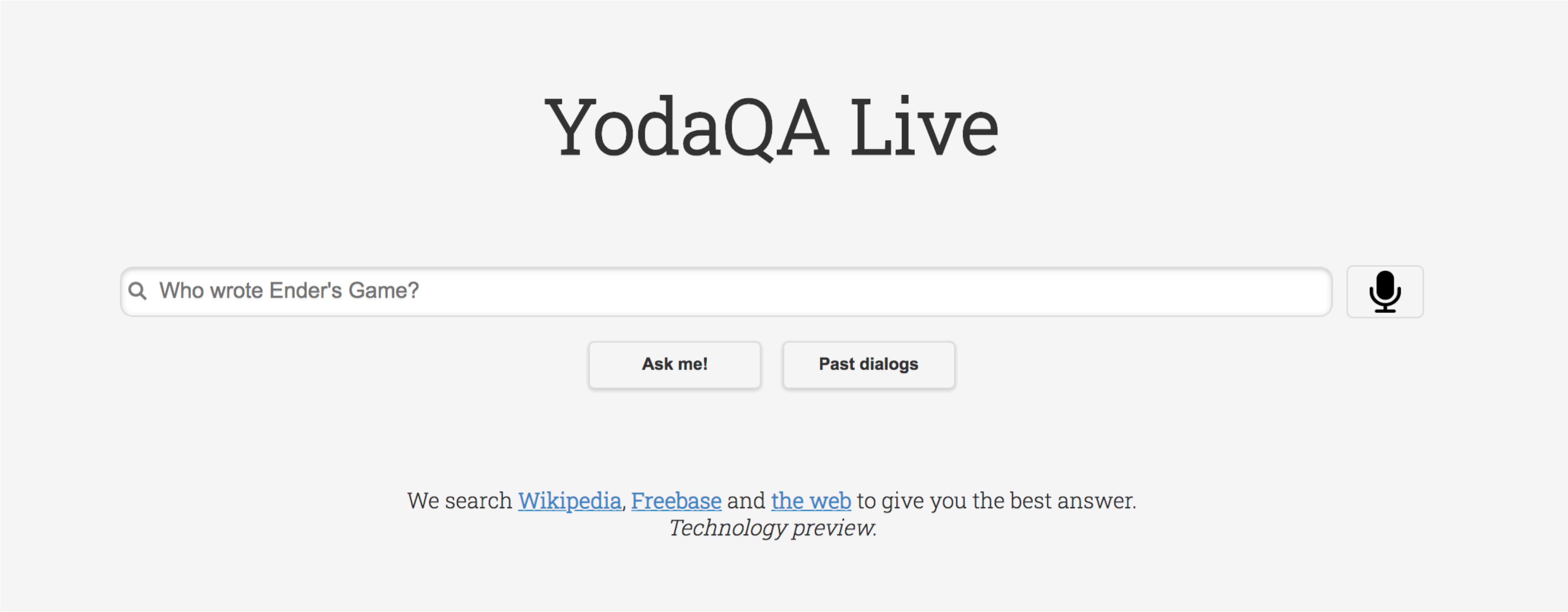
Dataset	YodaQA	DrQA		
		SQuAD	+Fine-tune (DS)	+Multitask (DS)
SQuAD *(All Wikipedia)*	n/a	27.1	28.4	29.8
CuratedTREC	31.3	19.7	25.7	25.4
WebQuestions	39.8	11.8	19.5	20.7
WikiMovies	n/a	24.5	34.3	36.5

Table 6: Full Wikipedia results. Top-1 exact-match accuracy (in %, using SQuAD eval script). +Fine-tune (DS): Document Reader models trained on SQuAD and fine-tuned on each DS training set independently. +Multitask (DS): Document Reader single model trained on SQuAD and all the distant supervision (DS) training sets jointly. YodaQA results are extracted from https://github.com/brmson/yodaqa/wiki/Benchmarks and use additional resources such as Freebase and DBpedia, see Section 2.

Seems to be better in these two tasks, anything wrong?

What is YodaQA?

YodaQA is an open source system modeled after IBM's DeepQA (Watson) system, which is a hybrid system which answers questions based on different types of data, including unstructured text, websites, databases etc.



Nothing wrong!

- It is not a direct comparison between YodaQA and DrQA as YodaQA relies on additional resources such as Freebase, while DrQA is more challenging by using single source

- WebQuestions is a dataset which is designed to answer questions over Freebase

Main Take-Aways

- DrQA was the first attempt to scale up reading comprehension to open-domain question answering, by combining IR techniques and neural reading comprehension models.

- Although we achieved good accuracy on SQuAD in 2017 (EM = 70.. vs state-of-the-art EM = 90 in 2020), the final QA accuracy still remains low: 20.7 - 36.5.

- Distant supervision + multi-task learning helps!

Latent Retrieval for Weakly Supervised Open Domain Question Answering

Kenton Lee **Ming-Wei Chang** **Kristina Toutanova**
Google Research
Seattle, WA
{kentonl,mingweichang,kristout}@google.com

ORQA

1. Introduction & Motivation
2. Model
3. Evaluation
4. Analysis

ORQA

1. **Introduction & Motivation**
2. Model
3. Evaluation
4. Analysis

Limitations of Current Models

DrQA Model

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

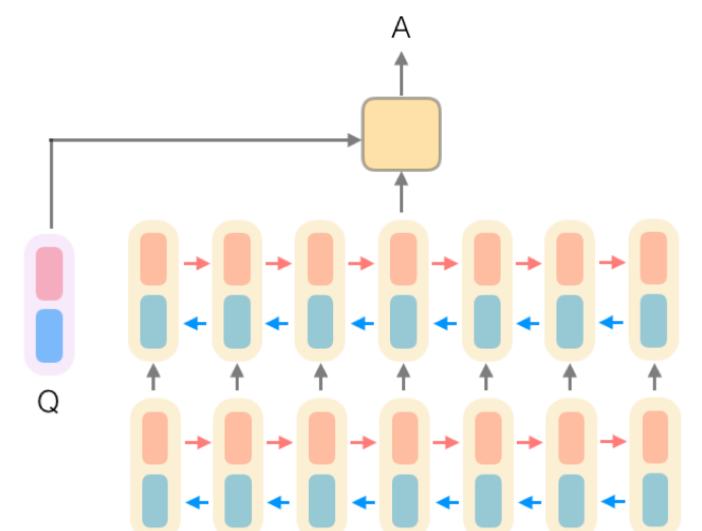
This article is about the Polish capital. For other uses, see Warsaw (disambiguation).
"Warszawa" redirects here. For other uses, see Warszawa (disambiguation).
"City of Warsaw" redirects here. For the Second World War fighter squadron, see No. 316 Polish Fighter Squadron. F 1934, see Adamowicz brothers.

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



What if we *learned* the information retrieval section?

Problem Space

Typically in open-domain question answering

Input: question string q

Output: answer string a

Where do you get the evidence to go from input to output?
- In reading comprehension, it’s given to you
- Here, it can come from *anywhere*

It's a modeling choice here!

Discussion Question

Discuss the differences between unsupervised QA, strongly supervised QA, weakly supervised QA settings in open-domain question answering.

Discussion Question

Discuss the differences between unsupervised QA, strongly supervised QA, weakly supervised QA settings in open-domain question answering.

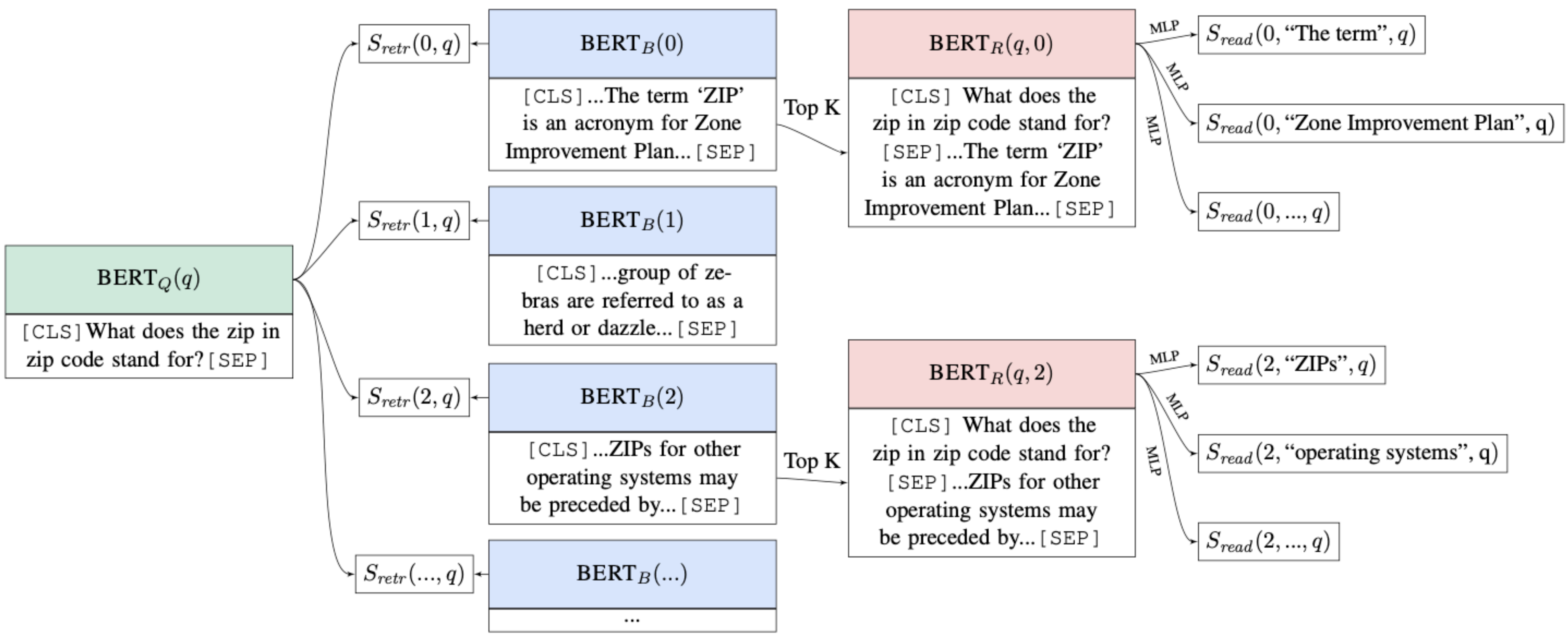
Unsupervised: No training data or question-answer pairs
Strongly supervised: Assumes reading comprehension dataset *with gold evidence* and question-answer pairs
Weakly supervised: Only have access to question-answer pairs *without* any gold evidence

Task	Training		Evaluation		Example
	Evidence	Answer	Evidence	Answer	
Reading Comprehension	given	span	given	string	SQuAD (Rajpurkar et al., 2016)
Open-domain QA					
Unsupervised QA	none	none	none	string	GPT-2 (Radford et al., 2019)
Strongly Supervised QA	given	span	heuristic	string	DrQA (Chen et al., 2017)
Weakly Supervised QA					
Closed Retrieval QA	heuristic	string	heuristic	string	TriviaQA (Joshi et al., 2017)
Open Retrieval QA	**learned**	string	**learned**	string	**ORQA (this work)**

ORQA

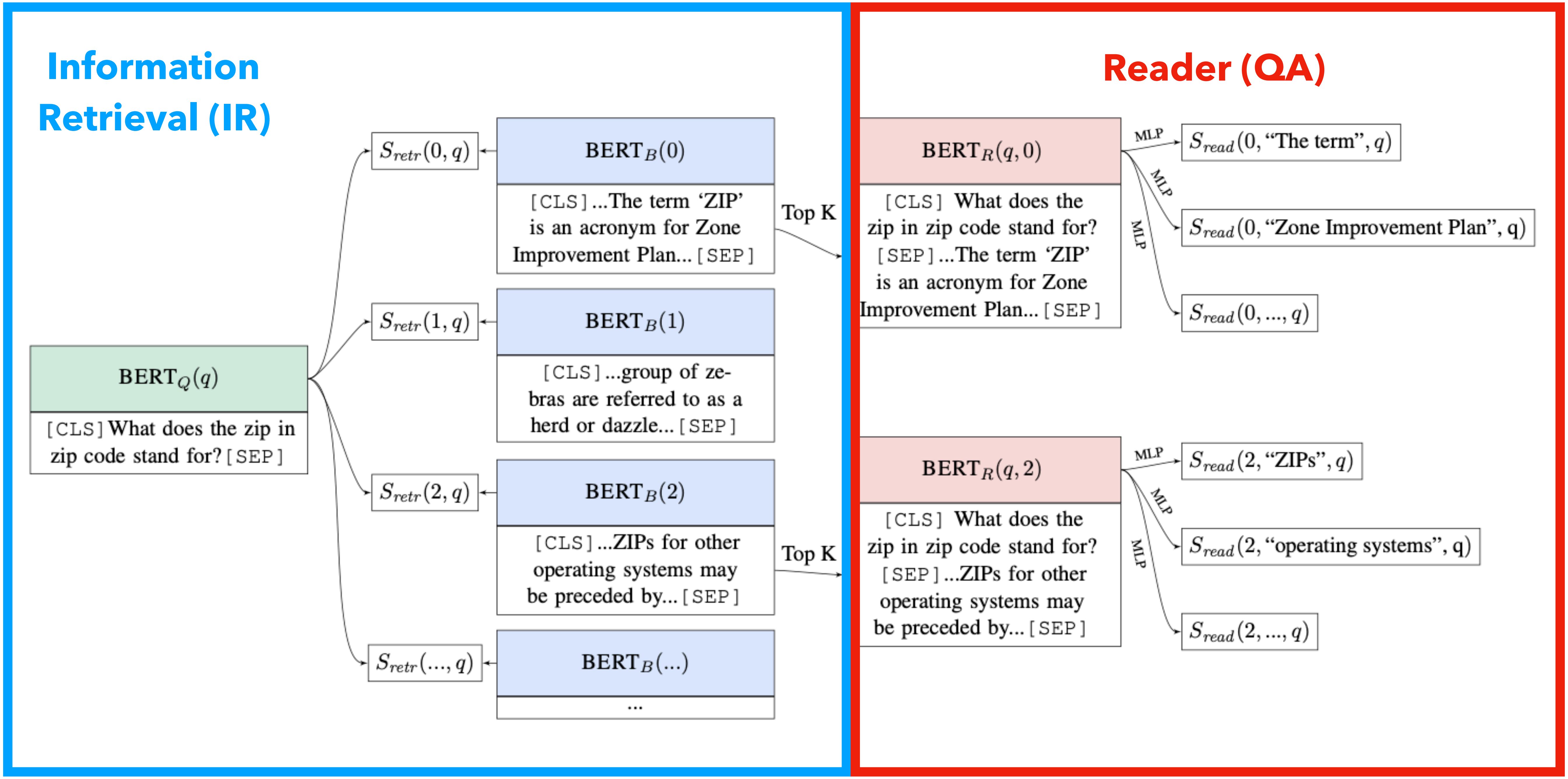
1. Introduction & Motivation
2. **Model**
3. Evaluation
4. Analysis

The Model!



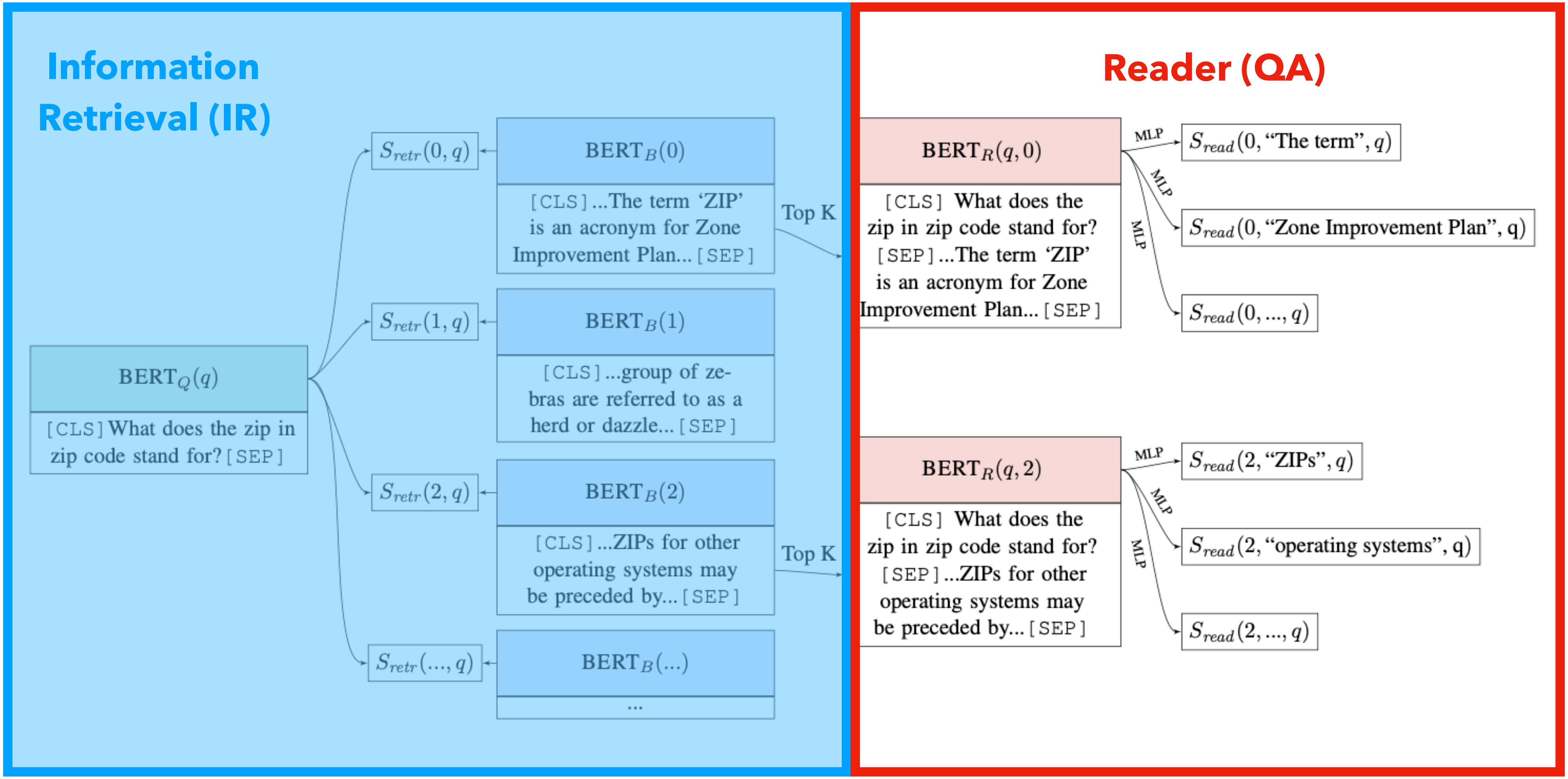
This is from 2019, it's time for BERT models!

The Model!



Information Retrieval (IR)
Reader (QA)
BERT_Q(q)
[CLS] What does the zip in zip code stand for? [SEP]
S_retr(0, q)
S_retr(1, q)
S_retr(2, q)
S_retr(..., q)
BERT_B(0)
[CLS] ...The term 'ZIP' is an acronym for Zone Improvement Plan... [SEP]
BERT_B(1)
[CLS] ...group of zebras are referred to as a herd or dazzle... [SEP]
BERT_B(2)
[CLS] ...ZIPs for other operating systems may be preceded by... [SEP]
BERT_B(...)
...
Top K
Top K
BERT_R(q, 0)
[CLS] What does the zip in zip code stand for? [SEP] ...The term 'ZIP' is an acronym for Zone Improvement Plan... [SEP]
BERT_R(q, 2)
[CLS] What does the zip in zip code stand for? [SEP] ...ZIPs for other operating systems may be preceded by... [SEP]
MLP
S_read(0, "The term", q)
S_read(0, "Zone Improvement Plan", q)
S_read(0, ..., q)
S_read(2, "ZIPs", q)
S_read(2, "operating systems", q)
S_read(2, ..., q)

The Model!



Information Retrieval (IR)
Reader (QA)
$\mathrm{BERT}_Q(q)$
[CLS] What does the zip in zip code stand for? [SEP]
$S_{retr}(0, q)$
$S_{retr}(1, q)$
$S_{retr}(2, q)$
$S_{retr}(..., q)$
$\mathrm{BERT}_B(0)$
[CLS] ...The term 'ZIP' is an acronym for Zone Improvement Plan... [SEP]
$\mathrm{BERT}_B(1)$
[CLS] ...group of zebras are referred to as a herd or dazzle... [SEP]
$\mathrm{BERT}_B(2)$
[CLS] ...ZIPs for other operating systems may be preceded by... [SEP]
$\mathrm{BERT}_B(...)$
...
Top K
Top K
$\mathrm{BERT}_R(q, 0)$
[CLS] What does the zip in zip code stand for? [SEP] ...The term 'ZIP' is an acronym for Zone Improvement Plan... [SEP]
MLP
MLP
MLP
$S_{read}(0, \text{"The term"}, q)$
$S_{read}(0, \text{"Zone Improvement Plan"}, q)$
$S_{read}(0, ..., q)$
$\mathrm{BERT}_R(q, 2)$
[CLS] What does the zip in zip code stand for? [SEP] ...ZIPs for other operating systems may be preceded by... [SEP]
MLP
MLP
MLP
$S_{read}(2, \text{"ZIPs"}, q)$
$S_{read}(2, \text{"operating systems"}, q)$
$S_{read}(2, ..., q)$

Information Retrieval (IR)

Question q

What does the zip
in zip code stand for?

All of Wikipedia

History[edit]
Early history and five-digit ZIP Codes[edit]

A 1963 U.S. Post Office sign featuring Mr. ZIP.

The early history and context of postal codes began with postal district/zone numbers. The United States Post Office Department (USPOD) implemented postal zones for many large cities in 1943.[3]
For example:
Mr. John Smith
3256 Epiphenomenal Avenue
Minneapolis *16*, Minnesota
The "16" is the number of the postal zone in the specific city.[citation needed]
By the early 1960s, a more organized system was needed, and non-mandatory five-digit ZIP Codes were introduced nationwide on July 1, 1963. The USPOD issued its *Publication 59: Abbreviations for Use with ZIP Code* on October 1, 1963, with the list of two-letter state abbreviations which are generally written with both letters capitalized.[4] An earlier list, publicized in June 1963, had proposed capitalized abbreviations ranging from two to five letters.[4] According to *Publication 59*, the two-letter standard was "based on a maximum 23-position line, because this has been found to be the most universally acceptable line capacity basis for major addressing systems",[4] which would be exceeded by a long city name combined with a multi-letter state abbreviation, such as "Sacramento, Calif." along with the ZIP Code. The abbreviations have remained unchanged, with the exception of Nebraska, which was changed from NB to NE in 1969 at the request of the Canadian postal administration, to avoid confusion with the Canadian province of New Brunswick.[4]
Robert Moon is considered the father of the ZIP Code; he submitted his proposal in 1944 while working as a postal inspector.[5][6] The post office only credits Moon with the first three digits of the ZIP Code, which describe the sectional center facility (SCF) or "sec center". An SCF is a central mail processing facility with those three digits. The fourth and fifth digits, which give a more precise locale within the SCF, were proposed by Henry Bentley Hahn Sr.[7]
The SCF sorts mail to all post offices with those first three digits in their ZIP Codes. The mail is sorted according to the final two digits of the ZIP Code and sent to the corresponding post offices in the early morning. Sectional centers do not deliver mail and are not open to the public (although the building may include a post office that is open to the public), and most of their employees work the night shift. Items of mail picked up at post offices are sent to their own SCFs in the afternoon, where the mail is sorted overnight. In the case of large cities, the last two digits as assigned generally coincided with the older postal zone number.[citation needed]
For example:
Mr. John Smith
3256 Epiphenomenal Avenue
Minneapolis, MN 554*16*
In 1967, these became mandatory for second- and third-class bulk mailers, and the system was soon adopted generally. The United States Post Office used a cartoon character, which it called Mr. ZIP, to promote the use of the ZIP Code. He was often depicted with a legend such as "USE ZIP CODE" in the selvage of panes of postage stamps or on the covers of booklet panes of stamps.[citation needed]
In 1971, Elmira (NY) *Star-Gazette* reporter Dick Baumbach found out the White House was not using a ZIP Code on its envelopes. Herb Klein, special assistant to President Nixon, responded by saying the next printing of envelopes would include the ZIP Code.[8]
ZIP+4[edit]
In 1983, the U.S. Postal Service introduced an expanded ZIP Code system that it called *ZIP+4*, often called "plus-four codes", "add-on codes", or "add-ons". A ZIP+4 Code uses the basic five-digit code plus four additional digits to identify a geographic segment within the five-digit delivery area, such as a city block, a group of apartments, an individual high-volume receiver of mail, a post office box, or any other unit that could use an extra identifier to aid in efficient mail sorting and delivery. However, initial attempts to promote universal use of the new format met with public resistance and today the plus-four code is not required.[9] In general, mail is read by a multiline optical character reader (MLOCR) that almost instantly determines the correct ZIP+4 Code from the address—along with the even more specific delivery point—and sprays an Intelligent Mail barcode (IM) on the face of the mail piece that corresponds to 11 digits—nine for the ZIP+4 Code and two for the delivery point.
For Post Office Boxes, the general (but not invariable) rule is that each box has its own ZIP+4 code. The add-on code is often one of the following: the last four digits of the box number (e.g. PO Box 107050, Albany, NY 12201-7050), zero plus the last three digits of the box number (e.g., PO Box 17727, Eagle River, AK 99577-0727), or, if the box number consists of fewer than four digits, enough zeros are attached to the front of the box number to produce a four-digit number (e.g., PO Box 77, Juneau, AK 99750-0077). However, there is no uniform rule, so the ZIP+4 Code must be looked up individually for each box[citation needed] (e.g. using the USPS's official ZIP Code Lookup tool, and being sure to enter just city and state, not the 5-digit ZIP[10]).
Postal bar code[edit]
The ZIP Code is often translated into an Intelligent Mail barcode that is printed on the mailpiece to make it easier for automated machines to sort. A barcode can be printed by the sender (some word-processing programs such as WordPerfect[11] include the feature), but this is not recommended, as the address-to-ZIP lookup tables can be significantly out of date. It is better to let the post office put one on when it processes the piece.[citation needed] In general, the post office uses OCR technology, though in some cases a human might have to read and enter the address.[citation needed]
Customers who send bulk mail can get a discount on postage if they have printed the barcode themselves and have presorted the mail. This requires more than just a simple font; mailing lists must be standardized with up-to-date Coding Accuracy Support System (CASS)-certified software that adds and verifies a full, correct ZIP+4 Code and an additional two digits representing the exact delivery point.[citation needed] Furthermore, mail must be sorted in a specific manner to an 11-digit code with at least 150 mailpieces for each qualifying ZIP Code and must be accompanied by documentation confirming this. These steps are usually done with PAVE-certified software that also prints the barcoded address labels and the barcoded sack or tray tags.[citation needed]
The assignment of delivery point digits (the 10th and 11th digits) is intended to ensure that every single mailable point in the country has its own 12-digit number. The delivery-point digits are calculated based on the primary or secondary number of the address. The USPS publishes the rules for calculating the delivery point in a document called the CASS Technical Guide.[12] However, when confronted with two addresses like 18 and 18C, often CASS will assign the same 12-digit number to two distinct mail delivery points. The last digit is always a check digit, which is obtained by summing all 5, 9 or 11 digits, taking the residue modulo 10 of this sum (i.e., the remainder after dividing by 10) and finally subtracting this from 10. (Thus, the check digit for 10001-0001 00 would be 7, since 1+1+1=3, 3≡3(mod 10) and 10−3=7.)[citation needed]
Structure and allocation[edit]
Scope and international mail[edit]
ZIP Codes designate delivery points within the United States (and its territories). There are generally no ZIP Codes for deliveries to other countries, except for the independent countries of the Federated States of Micronesia, the Republic of the Marshall Islands, and the Republic of Palau, each of which is integrated into the U.S. postal system under a Compact of Free Association. Another exception are ZIP Codes used for overseas stations of USA armed forces.[13]
Mail to U.S. diplomatic missions overseas is addressed as if it were addressed to a street address in Washington, D.C. The four-digit diplomatic pouch number is used as a building number, while the city in which the embassy or consulate is located is combined with the word "Place" to form a street name. Each mission uses a ZIP+4 Code consisting of **20521** and the diplomatic pouch number. For example, the mailing address of the U.S. Embassy in New Delhi, India would be:
Embassy of the United States of America
9000 New Delhi Place
Washington, DC 20521-9000[14]
However, individuals posted at diplomatic missions overseas are now assigned a Diplomatic Post Office address and unique box number. The Zip Code identifies the diplomatic mission destination and is different from the diplomatic pouch number in the example above. While delivered through the pouch system, mail to such addresses are not considered "Diplomatic Pouch" materials, and as such must adhere to the mailing regulations of the host country. An example address is:
JOHN ADAMS
UNIT 8400 BOX 0000
DPO AE 09498-0048[15]
By type and use[edit]
There are four types of ZIP Codes:
- Unique: assigned to a single high-volume address
- Post Office Box-only: used only for PO Boxes at a given facility, not for any other type of delivery
- Military: used to route mail for the U.S. military
- Standard: all other ZIP Codes.

Unique ZIP Codes are used for governmental agencies, universities, businesses, or buildings that receive such extremely high volumes of mail that they need their own ZIP Codes. Government examples include 20505 for the Central Intelligence Agency in Washington, D.C.; 81009 for the Federal Citizen Information Center of

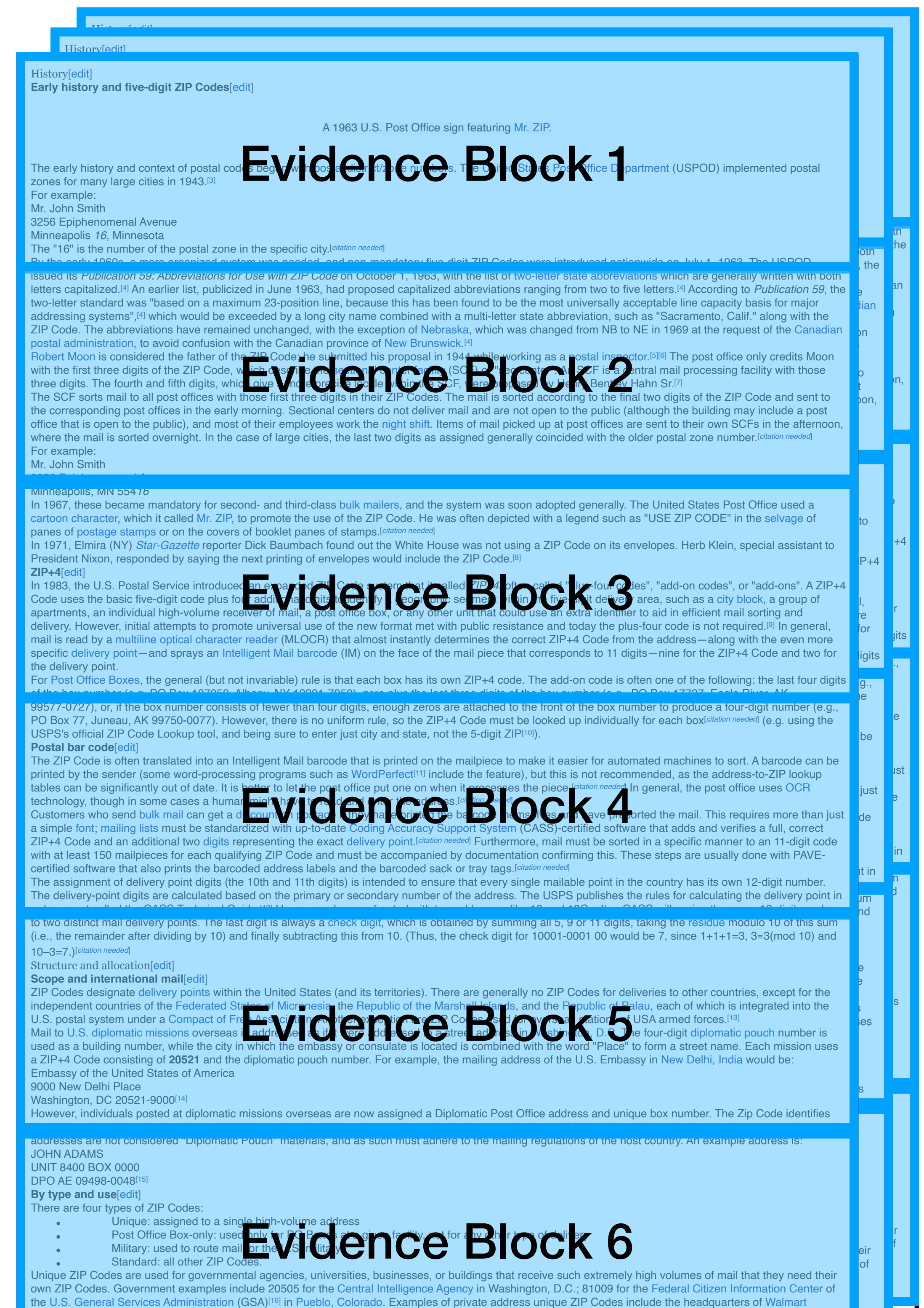
Information Retrieval (IR)

1. Segment all document into B evidence blocks

Question q

What does the zip in zip code stand for?

All of Wikipedia



Information Retrieval (IR)

2. Run the question and each evidence block through BERT encoders

Question q

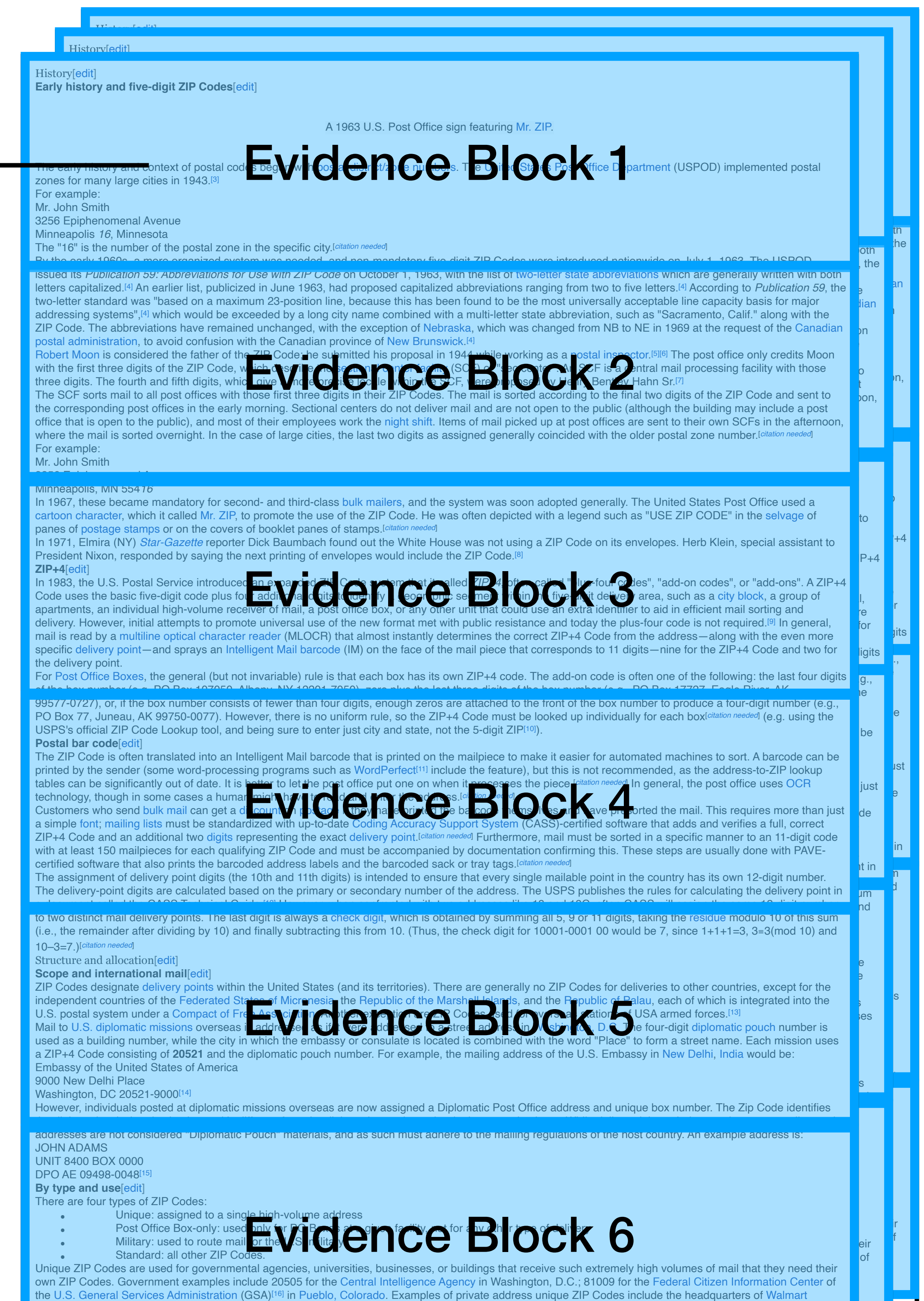
What does the zip in zip code stand for?

$BERT_Q$

Each evidence block b

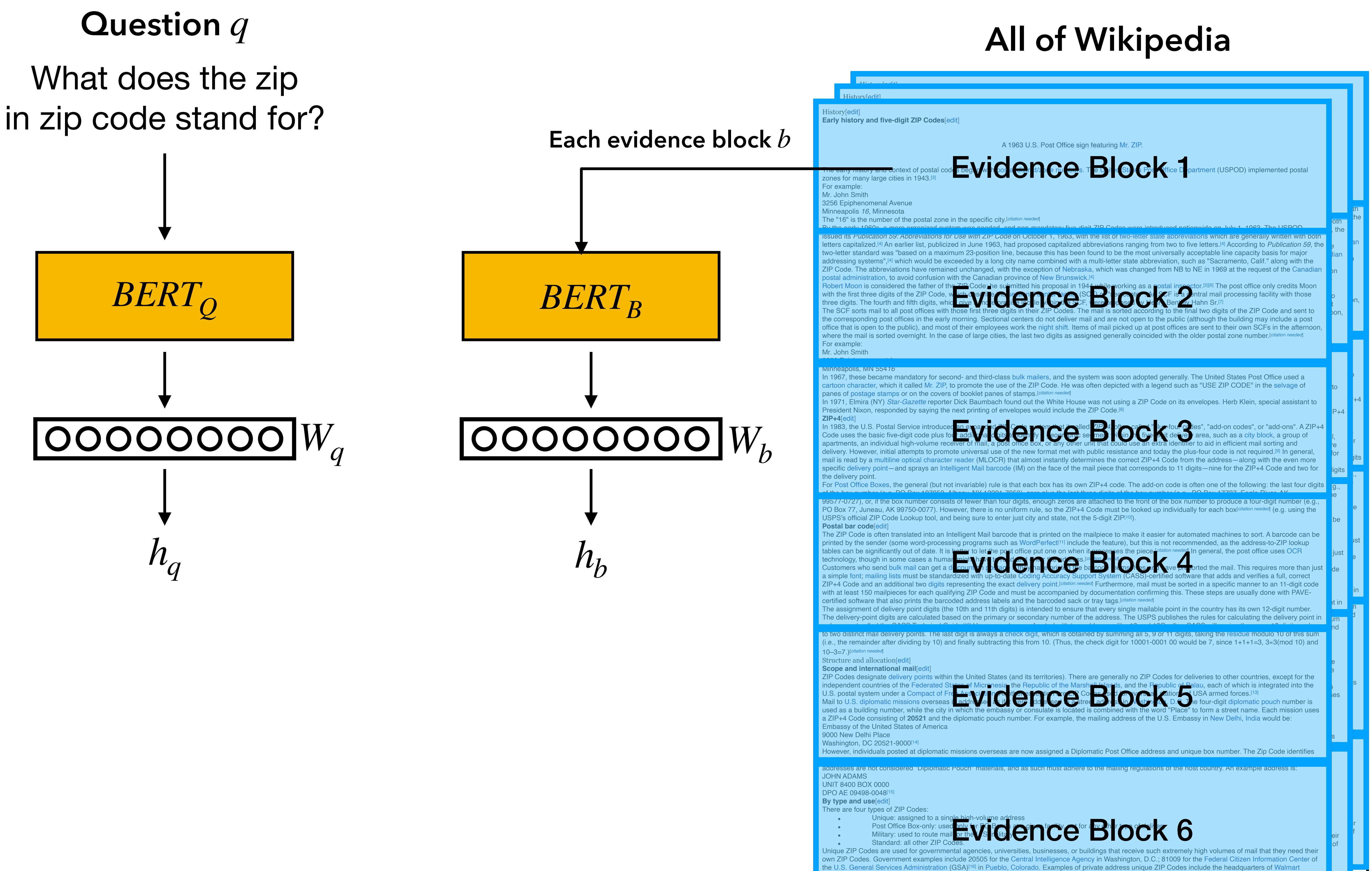
$BERT_B$

All of Wikipedia



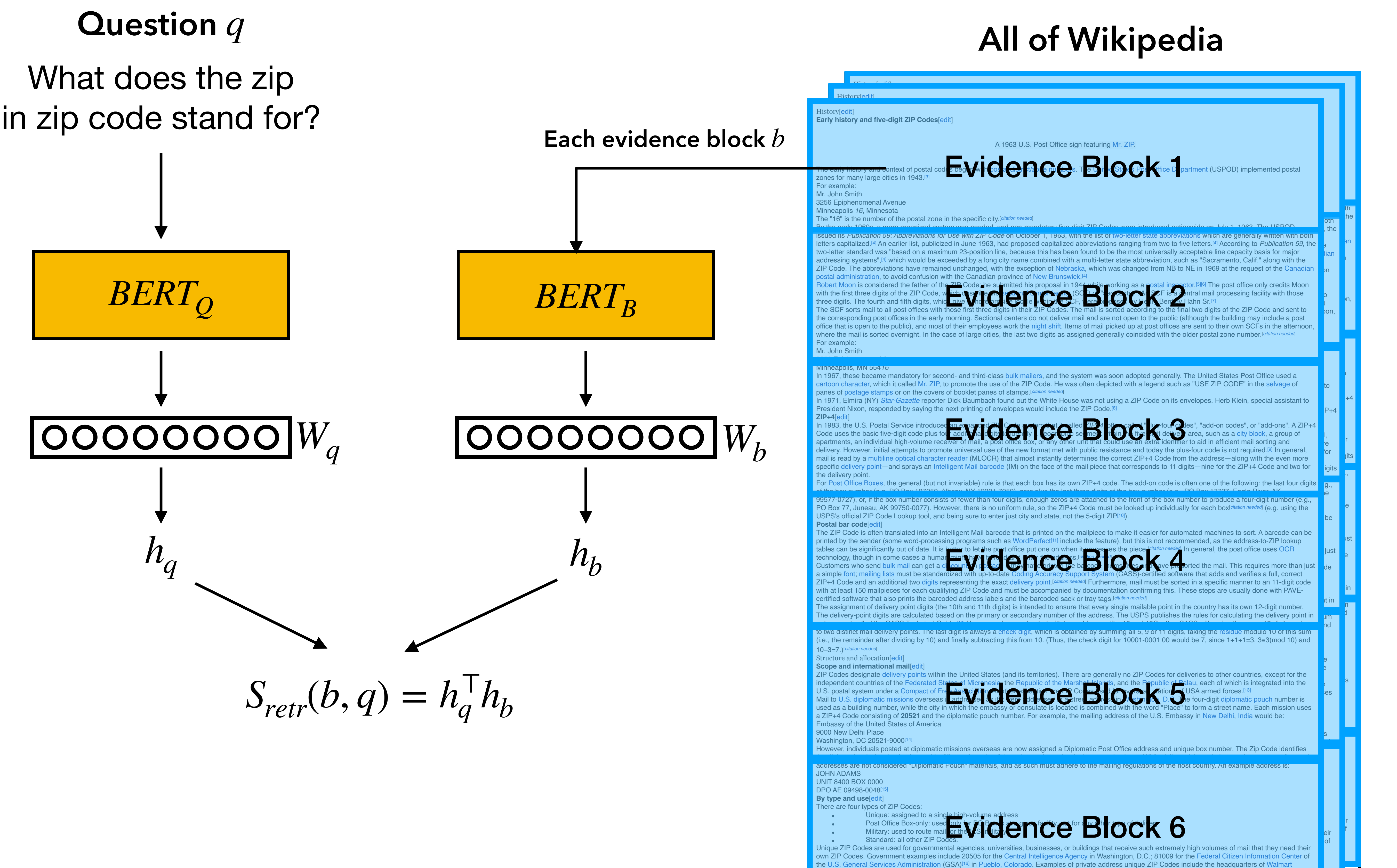
Information Retrieval (IR)

3. Pass outputs through a linear layer



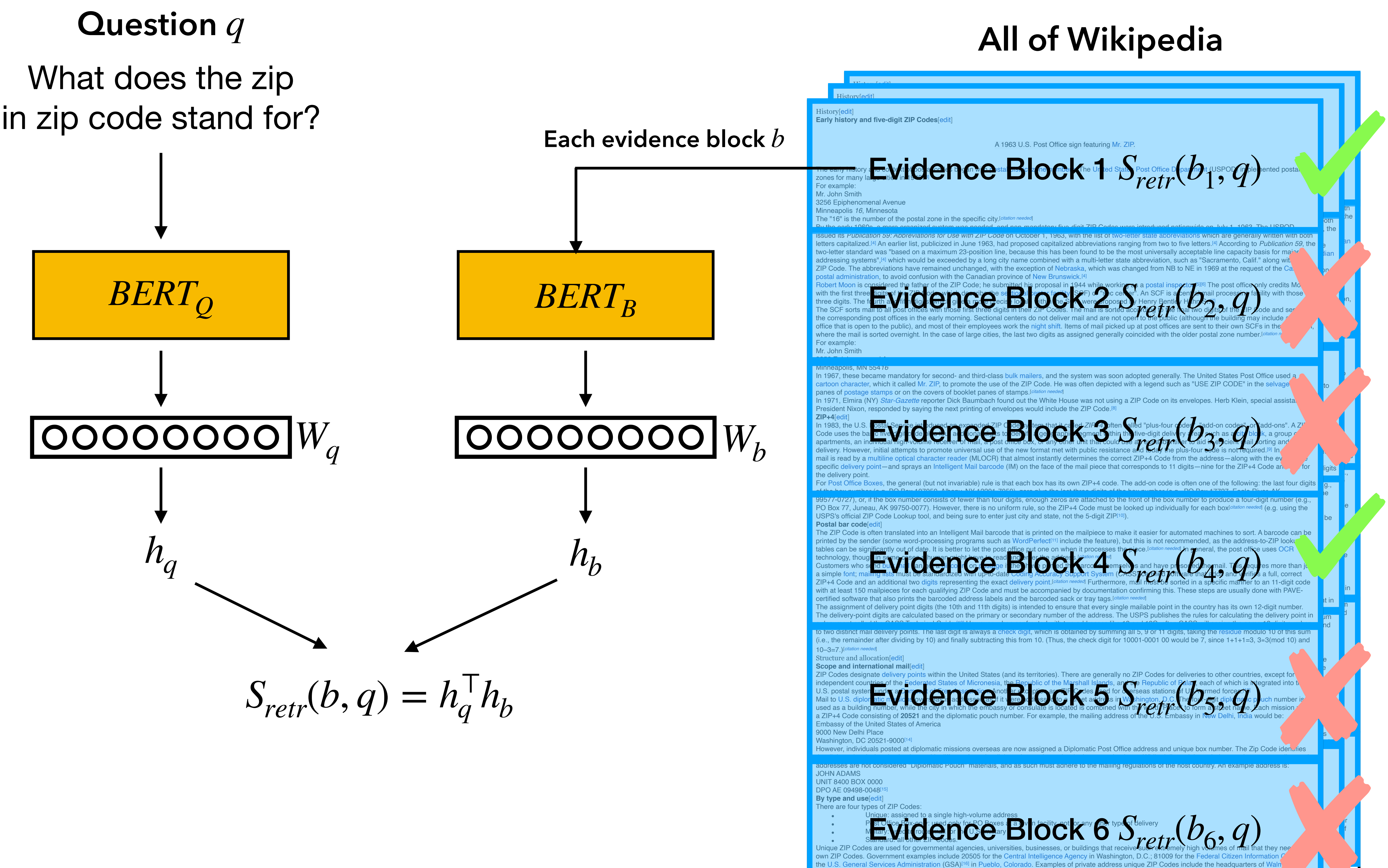
Information Retrieval (IR)

4. Score each evidence block as the inner product between h_q and h_b

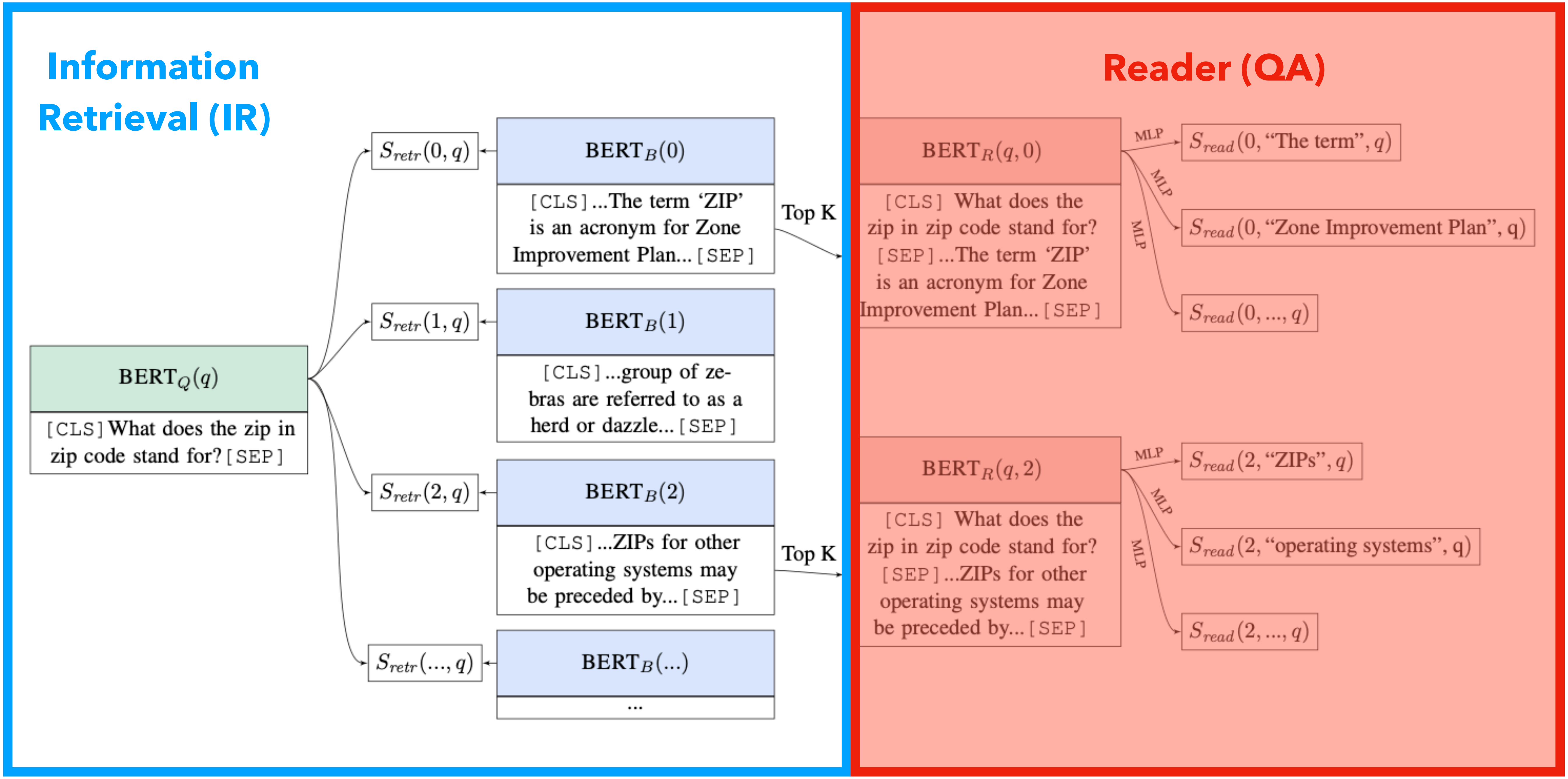


Information Retrieval (IR)

5. Pick and output the top k scoring evidence blocks.



The Model!

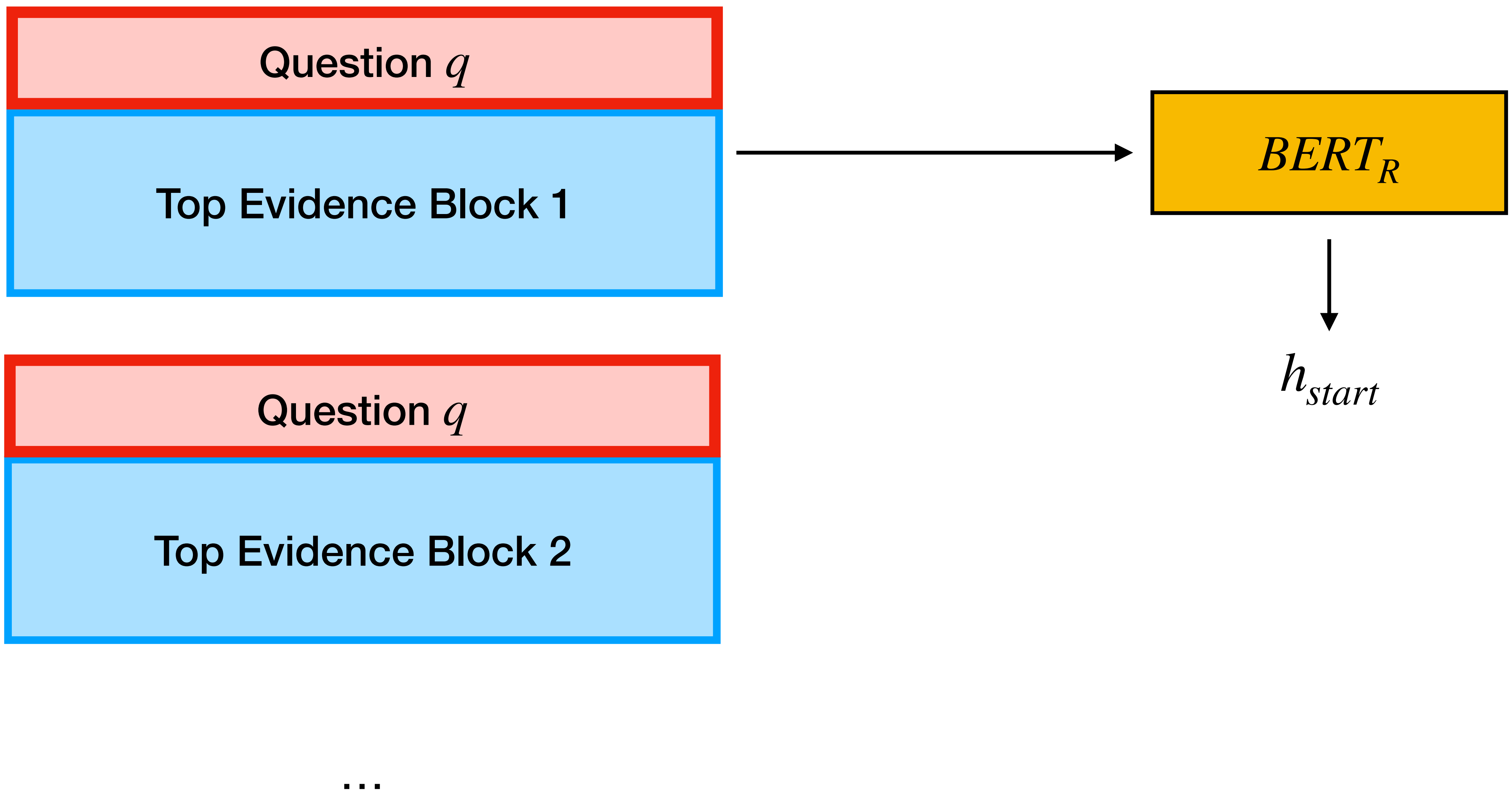


Information Retrieval (IR)
Reader (QA)
$\text{BERT}_Q(q)$
[CLS] What does the zip in zip code stand for? [SEP]
$S_{retr}(0, q)$
$S_{retr}(1, q)$
$S_{retr}(2, q)$
$S_{retr}(..., q)$
$\text{BERT}_B(0)$
[CLS] ...The term 'ZIP' is an acronym for Zone Improvement Plan... [SEP]
$\text{BERT}_B(1)$
[CLS] ...group of zebras are referred to as a herd or dazzle... [SEP]
$\text{BERT}_B(2)$
[CLS] ...ZIPs for other operating systems may be preceded by... [SEP]
$\text{BERT}_B(...)$
...
Top K
Top K
$\text{BERT}_R(q, 0)$
[CLS] What does the zip in zip code stand for? [SEP] ...The term 'ZIP' is an acronym for Zone Improvement Plan... [SEP]
MLP
$S_{read}(0, \text{"The term"}, q)$
$S_{read}(0, \text{"Zone Improvement Plan"}, q)$
$S_{read}(0, ..., q)$
$\text{BERT}_R(q, 2)$
[CLS] What does the zip in zip code stand for? [SEP] ...ZIPs for other operating systems may be preceded by... [SEP]
$S_{read}(2, \text{"ZIPs"}, q)$
$S_{read}(2, \text{"operating systems"}, q)$
$S_{read}(2, ..., q)$

Reader (QA)

Top-k Evidence Blocks
Concat. with Question

1. Pass each (question, block) pair from the retriever through the $BERT_R$ model to find the <u>start</u> token hidden state

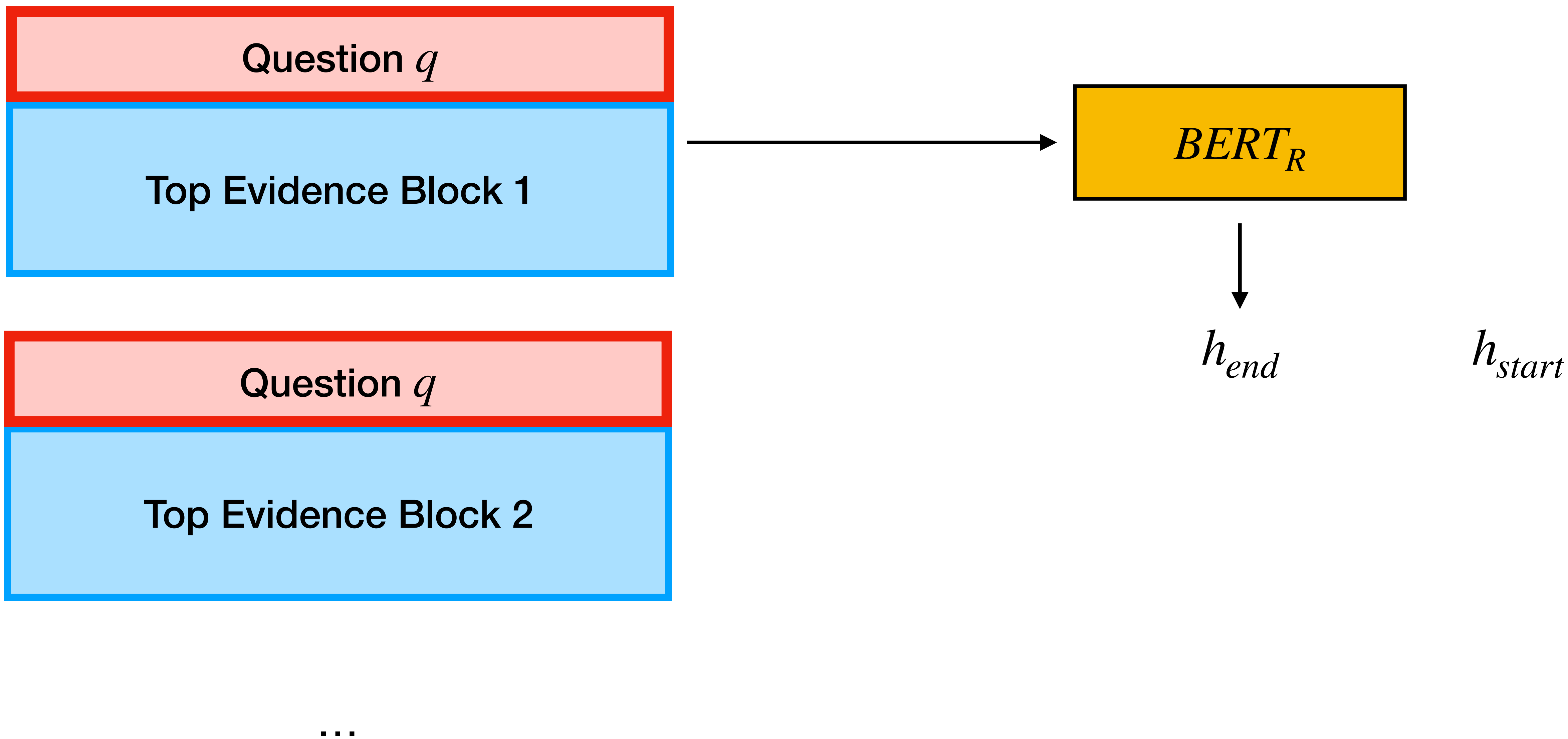


Very similar to the QA model in the BERT paper!

Reader (QA)

2. Pass each (question, block) pair from the retriever through the $BERT_R$ model to find the <u>end</u> token hidden state

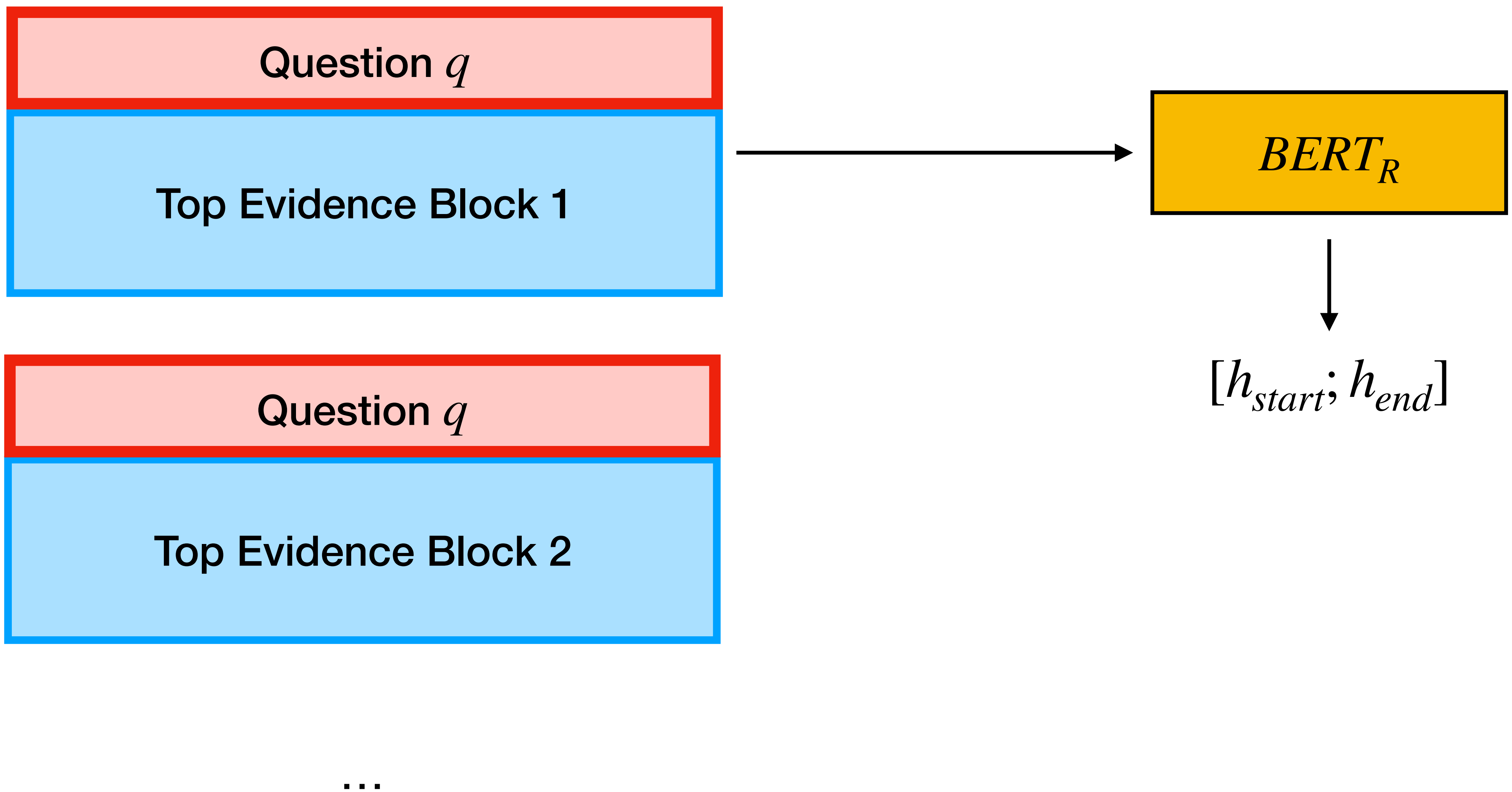
Top-k Evidence Blocks Concat. with Question



Reader (QA)

3. Concatenate the h_{start} and h_{end} hidden states into a single vector

Top-k Evidence Blocks Concat. with Question

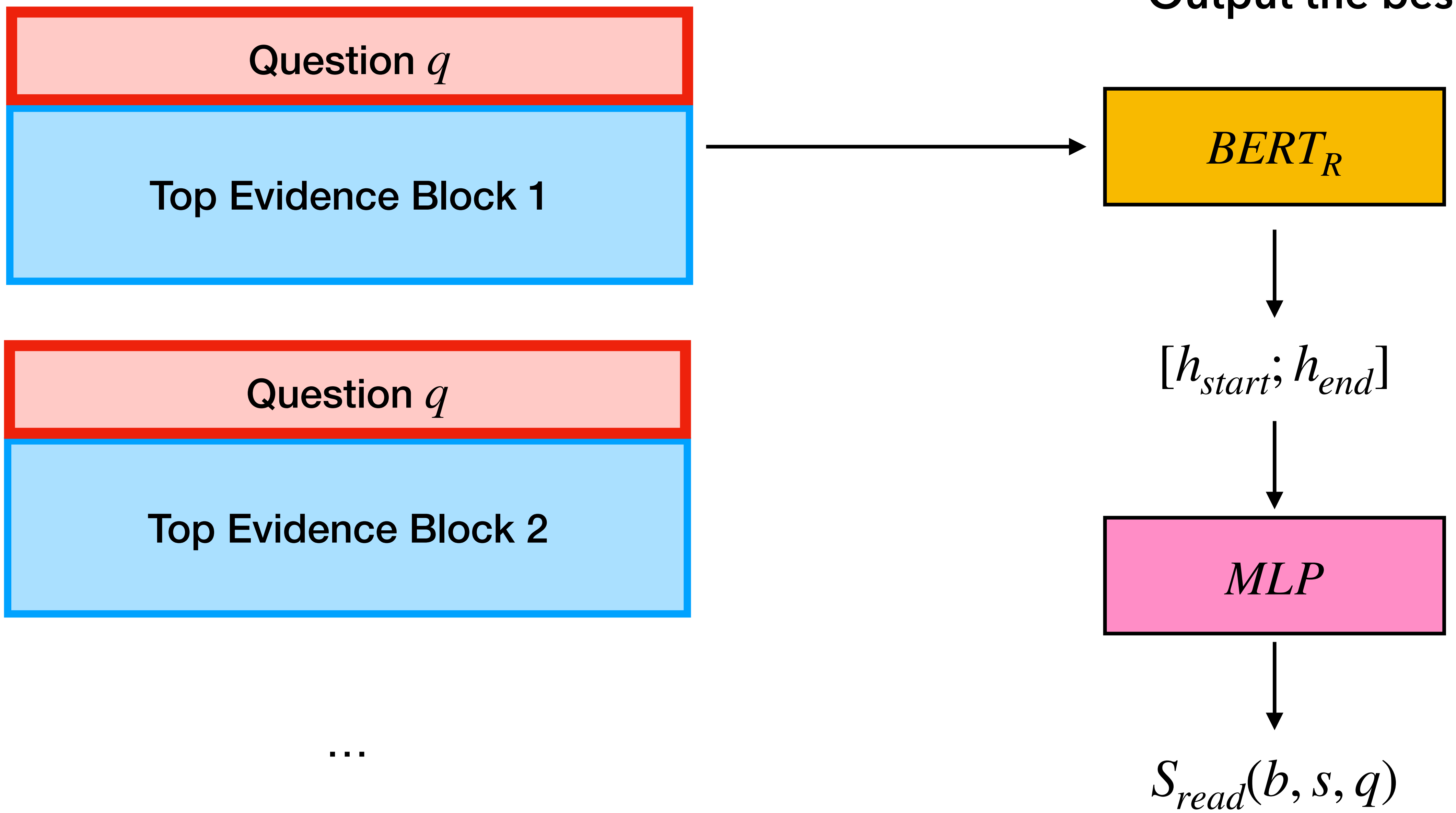


Reader (QA)

Top-k Evidence Blocks
Concat. with Question

4. Calculate the S_{read} score for a given span s in block b for question q.

Output the best scoring span



Inference & Learning Challenges

1.) The search space is **huge**! There's around 13 *million* evidence blocks to choose from.
(English Wikipedia ~*5 million* articles)

2.) How to pick relevant evidence blocks is *latent* (not explicitly learned) since we're learning on the right answer.

Example	Supportive Evidence	Spurious Ambiguity
Q: Who is credited with developing the XY coordinate plane? **A**: René Descartes	...invention of Cartesian coordinates by **René Descartes** revolutionized...	...**René Descartes** was born in La Haye en Touraine, France...
Q: How many districts are in the state of Alabama? **A**: seven	...Alabama is currently divided into **seven** congressional districts, each represented by ...	...Alabama is one of **seven** states that levy a tax on food at the same rate as other goods...

"Spuriously ambiguous derivations"

Inference & Learning Challenges

1.) The search space is **huge**! There's around 13 *million* evidence blocks to choose from.
(English Wikipedia ~*5 million* articles)

2.) How to pick relevant evidence blocks is *latent* (not explicitly learned) since we're learning on the right answer.

Solved using Inverse Cloze Task (ICT) Pre-training!

Example	Supportive Evidence	Spurious Ambiguity
Q: Who is credited with developing the XY coordinate plane? **A**: René Descartes	...invention of Cartesian coordinates by **René Descartes** revolutionized...	...**René Descartes** was born in La Haye en Touraine, France...
Q: How many districts are in the state of Alabama? **A**: seven	...Alabama is currently divided into **seven** congressional districts, each represented by ...	...Alabama is one of **seven** states that levy a tax on food at the same rate as other goods...

"Spuriously ambiguous derivations"

Cloze Task

Predict the masked-out *sentence* based on its *context*

Given

[CLS]
…Zebras have four gaits: walk, trot, canter, and gallop.

When chased, a zebra will zig-zag from side to side…
[SEP]

Choices

[CLS] They are generally slower than horses, but their great stamina helps them outrun predators [SEP]

[CLS] Gagarin was further selected for an elite training group known as the Sochi Six [SEP]

…

Inverse Cloze Task

Predict the masked-out *context* based on its *sentence*

Given

[CLS]

____

They are generally slower than horses, but their great stamina helps them outrun predators.

____

[SEP]

Choices

[CLS] …Zebras have four gaits: walk, trot, canter and gallop. When chased, a zebra will zig-zag from size to side… [SEP]

[CLS]…Gagarin was further selected for an elite training group known as the Sochi Six… [SEP]

…

Question: "pseudo-query"

Answer: "psuedo evidence text"

How often do we mask the pseudo-query?

What if we masked the pseudo-query 100% of the time?
- Trouble learning basic **word overlap** between evidence and query

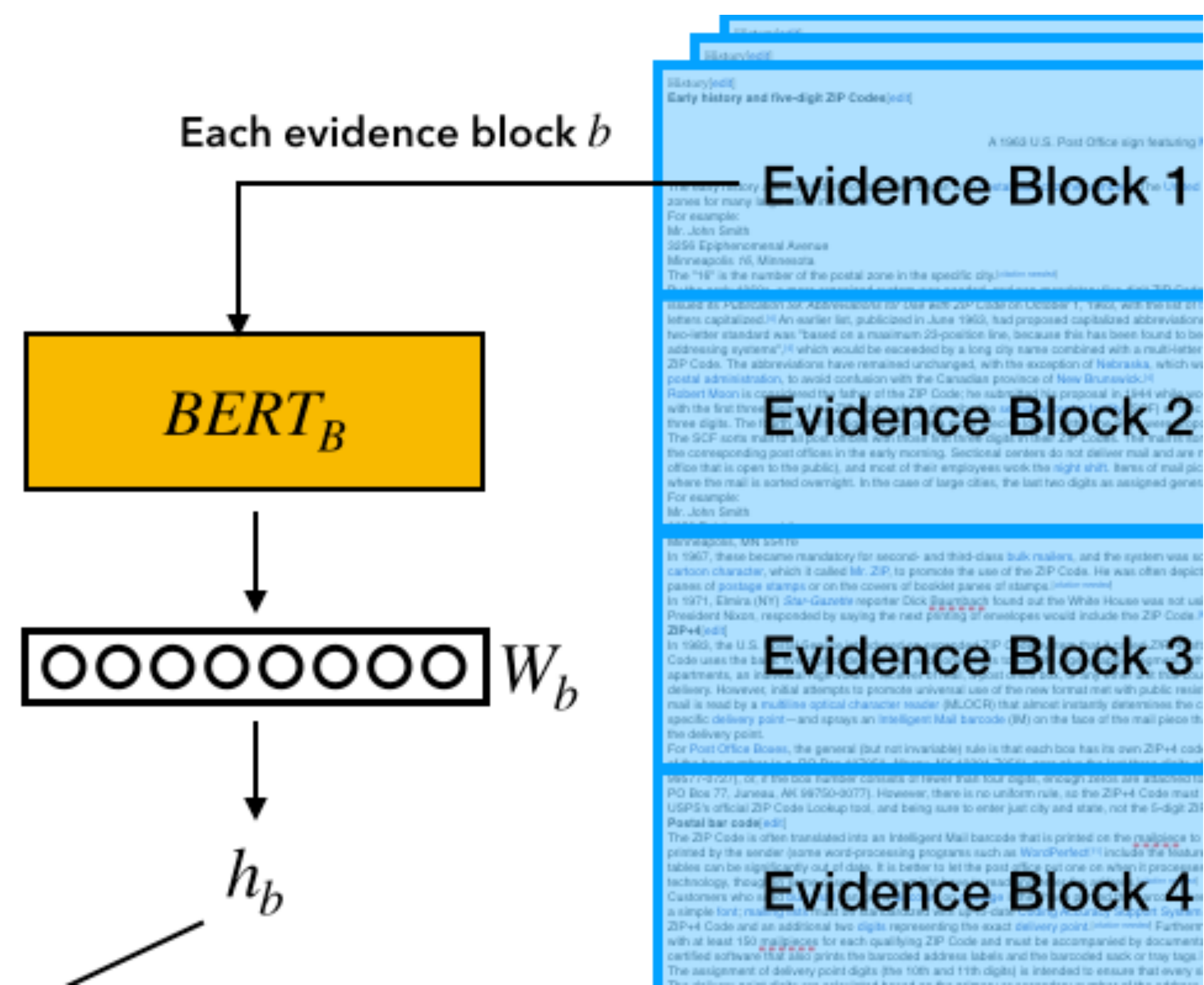
What if we masked the pseudo-query 0% of the time?
- Task becomes **trivial** and doesn't learn much! Find the evidence with the query in it

In response, ORQA removes the pseudo-question sentence *90%* of the time.
- 90% of the time, focus on abstract representations
- 10% of the time, focus on word matching

"They are generally slower than horses, but their great stamina helps them outrun predators."

Retrieval Pre-Training & Inference

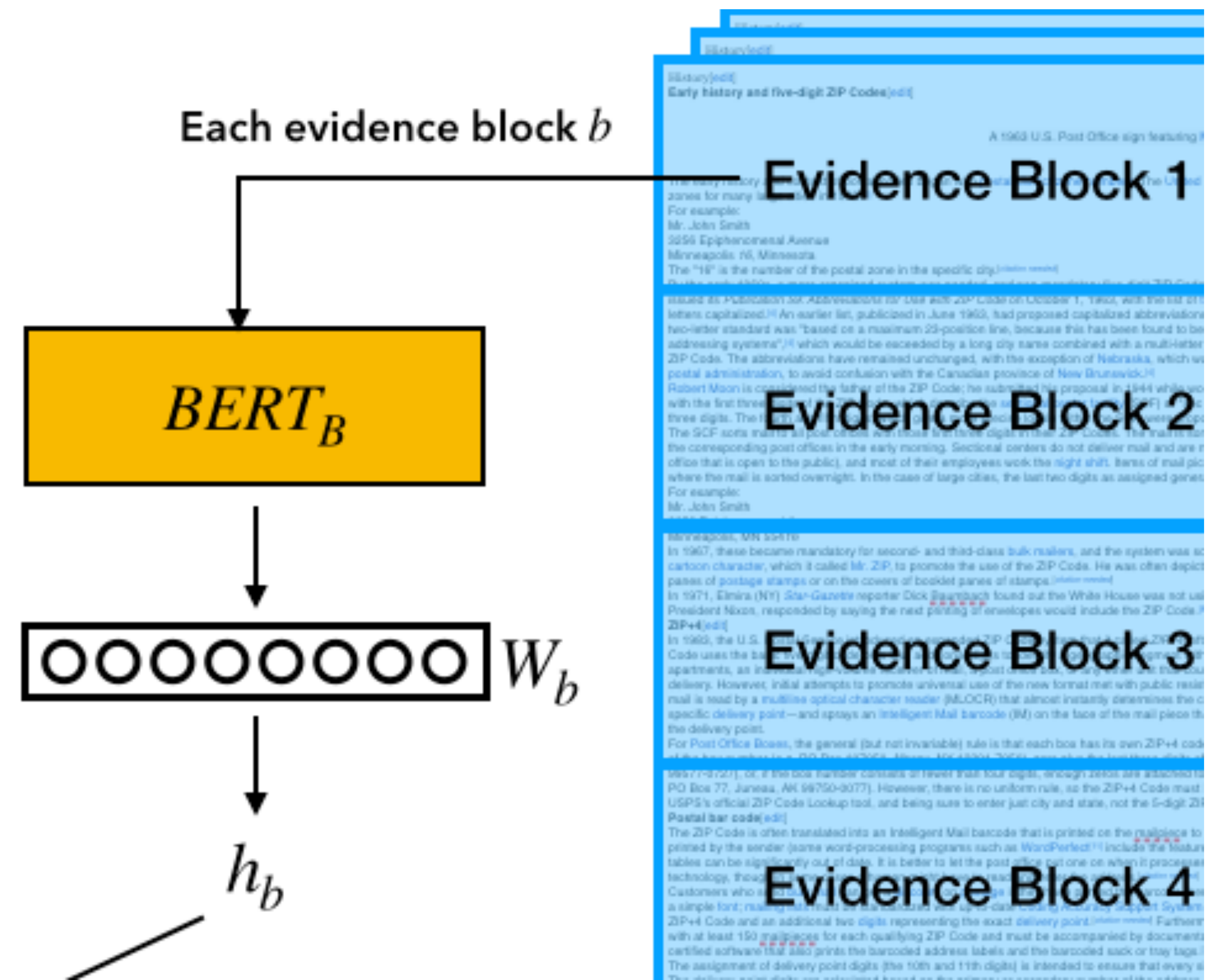
- Pre-trains the IR sub-model on the Inverse Cloze Task (ICT) with sentences
- Masks the sentences 90% of the time
- Freeze the $BERT_B(b)$ model afterwards
- Pre-compute the hidden representations for all of the evidence blocks (h_b) into a giant index
- Beam-search on k top blocks
- **Solves large search space problem!**



Retrieval Pre-Training & Inference

- Pre-trains the IR sub-model on the Inverse Cloze Task (ICT) with sentences
- Masks the sentences 90% of the time
- Freeze the $BERT_B(b)$ model afterwards
- **Pre-compute the hidden representations for all of the evidence blocks (h_b) into a giant index**
- Beam-search on k top blocks

This is really difficult too!
- Uses Locality Sensitive Hashing (LSH) to quickly find maximum inner products!
- Really important to make finding the best evidence blocks efficient!



Fine-Tuning & Learning

Probability distribution of any span s in any top-k block b given question q:

$$S(b, s, q) = S_{retr}(b, q) + S_{read}(b, s, q)$$

$$P(b, s|q) = \frac{\exp(S(b, s, q))}{\sum_{b' \in \text{TOP}(k)} \sum_{s' \in b'} \exp(S(b', s', q))}$$

Softmax over every span in the top-k blocks

Given a gold answer a, we find all spans that *exactly match* a and optimize their marginal log-likelihood:

$$L_{\text{full}}(q, a) = -\log \sum_{b \in \text{TOP}(k)} \sum_{s \in b,\ a = \text{TEXT}(s)} P'(b, s|q)$$

k ~ 5

Exact match!

Fine-Tuning & Learning

Early Learning consider a larger set of c evidence blocks and update the retrieval score:

Softmax over the top-c blocks

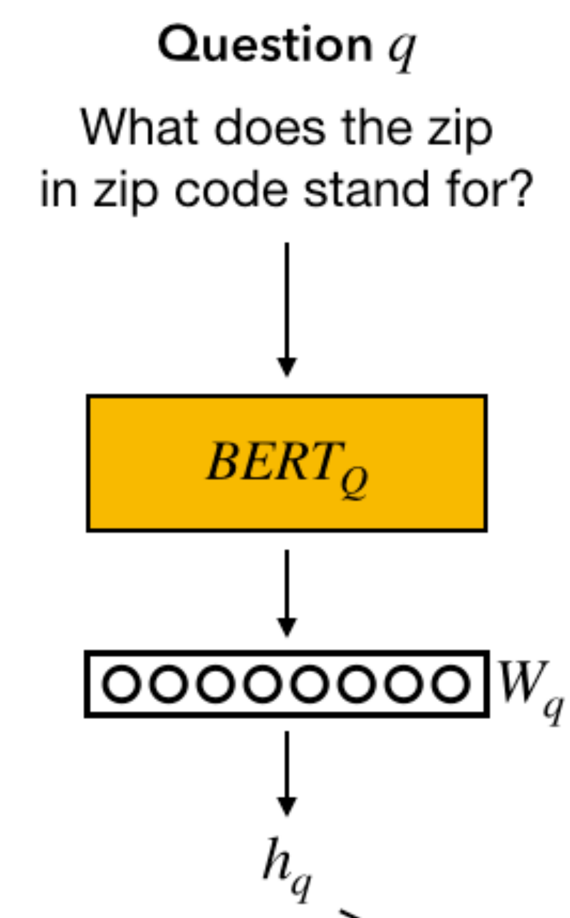
$$P_{\text{early}}(b|q) = \frac{\exp(S_{retr}(b,q))}{\sum_{b' \in \text{TOP}(c)} \exp(S_{retr}(b',q))}$$

$$L_{\text{early}}(q,a) = -\log \sum_{b \in \text{TOP}(c),\ a \in \text{TEXT}(b)} P_{\text{early}}(b|q)$$

Contains!

Provides additional training to the $BERT_Q(q)$ encoder!

$c \sim 5{,}000$



Fine-Tuning & Learning

Final Loss is the combination of both L_{full} and L_{early}

$$L(q, a) = L_{\text{early}}(q, a) + L_{\text{full}}(q, a)$$

If the answer isn't found in the top-k blocks, then discard the example.

Because ICT pre-training is such an effective strategy, only < 10% of examples are discarded!

Model Overview

Trained Information Retrieval picks the top-k scoring evidence blocks from all Wikipedia documents by taking the inner product between the question and block encodings

Trained Reader uses beam-search to find the answer *span* within the top-k evidence blocks

Inverse Cloze Task Pre-training initializes the block encoder weights to a sufficient starting point

Pre-computed Block Encoding Index computes all the encodings for each evidence block after ICT pre-training

Fine-Tuning on each task helps train the reader and the question encoder

Early Updates help train the question encoder by calculating loss on the top-c evidence blocks

ORQA

1. Introduction & Motivation
2. Model
3. **Evaluation**
4. Analysis

Datasets

Evaluated on DrQA

Natural Questions (Kwiatkowski et al., 2019) from Google Search API, discard evidence and long answers

WebQuestions (Berant et al., 2013) from Google Suggest, only keep string representations

CuratedTrec (Baudis and Sedivy, 2015) QA data from TREC

TriviaQA (Joshi et al., 2017) is a trivia QA collection from the web, discard evidence

SQuAD (Rajpurkar et al., 2016) is a well-known QA dataset, discard given evidence

Dataset	Train	Dev	Test	Example Question	Example Answer
Natural Questions	79168	8757	3610	What does the zip in zip code stand for?	Zone Improvement Plan
WebQuestions	3417	361	2032	What airport is closer to downtown Houston?	William P. Hobby Airport
CuratedTrec	1353	133	694	What metal has the highest melting point?	Tungsten
TriviaQA	78785	8837	11313	What did L. Fran Baum, author of The Wonderful Wizard of Oz, call his home in Hollywood?	Ozcot
SQuAD	78713	8886	10570	Other than the Automobile Club of Southern California, what other AAA Auto Club chose to simplify the divide?	California State Automobile Association

Datasets Biases

Natural Questions, **WebQuestions**, and **CuratedTrec** all have tool-assisted answers - bias towards the tool

TriviaQA and **SQuAD** question writers are aware of the answers

Dataset	Question writer knows answer	Question writer knows evidence	Tool-assisted answer
Natural Questions			✓
WebQuestions			✓
CuratedTrec			✓
TriviaQA	✓		
SQuAD	✓	✓	

Datasets Biases

SQuAD Paragraph

The largest living species is the emperor penguin (*Aptenodytes forsteri*): on average, adults are about 1.1 m (3 ft 7 in) tall and weigh 35 kg (77 lb). The smallest penguin species is the little blue penguin (*Eudyptula minor*), also known as the fairy penguin, which stands around 40 cm (16 in) tall and weighs 1 kg (2.2 lb). Among extant penguins, larger penguins inhabit colder regions, while smaller penguins are generally found in temperate or even tropical climates.

Question + Answer?

Datasets Biases

SQuAD Paragraph

The largest living species is the emperor penguin (*Aptenodytes forsteri*): on average, adults are about 1.1 m (3 ft 7 in) tall and weigh 35 kg (77 lb). The smallest penguin species is the little blue penguin (*Eudyptula minor*), also known as the fairy penguin, which stands around 40 cm (16 in) tall and weighs 1 kg (2.2 lb). Among extant penguins, larger penguins inhabit colder regions, while smaller penguins are generally found in temperate or even tropical climates.

Question + Answer

Q: What is the smallest penguin species?

A: Highlighted

Very high word overlap between the question and the paragraph!

Baseline Models

BM25+BERT is like the 2019 version of DrQA
- BM25 is an updated version of TF-IDF
- BERT is an updated version of DocumentReader

NNLM is a context-*independent* embedding from feed-forward language models

ELMo is a context-*dependent* bidirectional LSTM language model

All models use the **same BERT-based reader as ORQA**

NNLM and ELMo both use the same scoring heuristic as ORQA for retrieval

Results

Performed well on datasets with low question-evidence overlap!

	Model	BM25 +BERT	NNLM +BERT	ELMo +BERT	ORQA
Dev	Natural Questions	24.8	3.2	3.6	**31.3**
	WebQuestions	20.8	9.1	17.7	**38.5**
	CuratedTrec	27.1	6.0	8.3	**36.8**
	TriviaQA	**47.2**	7.3	6.0	45.1
	SQuAD	**28.1**	2.8	1.9	26.5
Test	Natural Questions	26.5	4.0	4.7	**33.3**
	WebQuestions	17.7	7.3	15.6	**36.4**
	CuratedTrec	21.3	4.5	6.8	**30.1**
	TriviaQA	**47.1**	7.1	5.7	45.0
	SQuAD	**33.2**	3.2	2.3	20.2

Performed at-par with BM25 on SQuAD and TriviaQA

Generally poor baselines

Results Takeaways

- **Successful End-to-End Training**… when there isn't "bias" in the dataset
- Previous neural retrieval methods (NNLM, ELMo-based) were very bad, but ORQA does a lot better
- ICT pre-trained retriever outperforms BM25 by 6 - 19 points depending on the dataset
- 128-dimensional vector may be too small to represent every word in the evidence
- SQuAD's 100k questions are derived from only 536 documents! Good retrievals are highly correlated between examples

ORQA

1. Introduction & Motivation

2. Model

3. Evaluation

4. **Analysis**

Strongly Supervised Model Comparison

DrQA is the state of the art unsupervised open-domain question-answering method

BERT_Serini is another BERT-based model using BM25 that splits on paragraphs instead of blocks (i.e. more evidence blocks)

BM25+BERT is the best performing model from the results

Evaluate on just the SQuAD testing dataset

Results

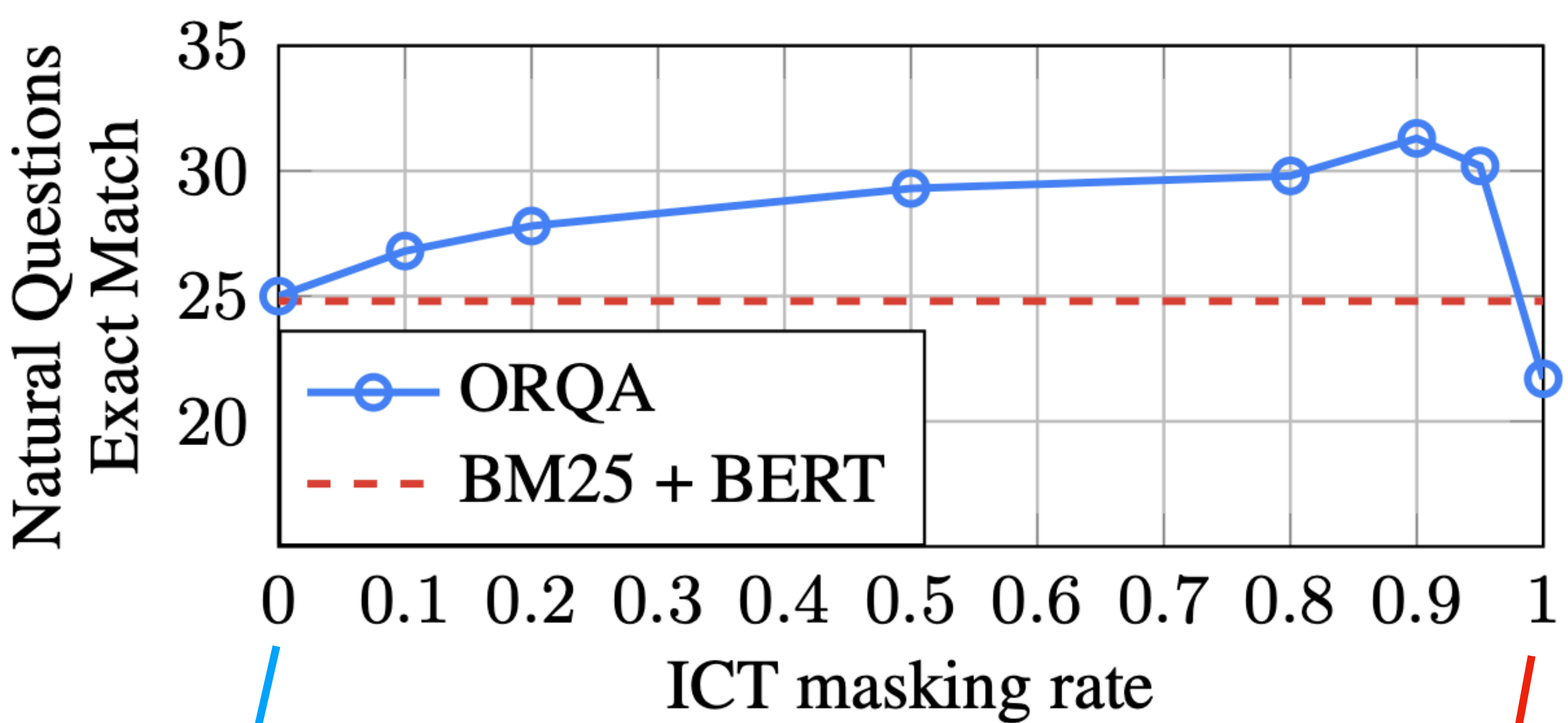
Excellent on an evidence retrieved efficiency level!

Model	Evidence Retrieved	SQuAD
DrQA	5 documents	27.1
DrQA (DS)	5 documents	28.4
DrQA (DS + MTL)	5 documents	29.8
BERTserini	5 documents	19.1
BERTserini	29 paragraphs	36.6
BERTserini	100 paragraphs	38.6
BM25 + BERT (gold deriv.)	5 blocks	34.7

BERT + BM25 + paragraphs instead of blocks

Best performing combo is comparable to SOTA

What if we vary the ICT masking rate?



0% masking rate means only looking for word similarity!

100% masking rate means not looking for word similarity and only using semantic meaning!

Error Comparisons

ORQA > BM25+BERT for separating semantically distinct text with high lexical overlap

Example	ORQA	BM25 + BERT
Q: what is the new orleans saints symbol called **A**: fleur-de-lis	...The team's primary colors are old gold and black; their logo is a simplified **fleur-de-lis**. They played their home games in Tulane Stadium through the 1974 NFL season....	...the **SkyDome** was owned by Sportsco at the time... the sale of the New Orleans Saints with team owner Tom Benson... the Saints became a symbol for that community...
Q: how many senators per state in the us **A**: two	...powers of the Senate are established in Article One of the U.S. Constitution. Each U.S. state is represented by **two** senators...	...The Georgia Constitution mandates a maximum of **56** senators, elected from single-member districts...
Q: when was germany given a permanent seat on the council of the league of nations **A**: 1926	...Under the Weimar Republic, Germany (in fact the "Deutsches Reich" or German Empire) was admitted to the League of Nations through a resolution passed on September 8 **1926**. An additional 15 countries joined later...	...the accession of the German Democratic Republic to the Federal Republic of Germany, it was effective on **3 October 1990**...Germany has been elected as a non-permanent member of the United Nations Security Council...
Q: when was diary of a wimpy kid double down published **A**: November 1, 2016	..."Diary of a Wimpy Kid" first appeared on FunBrain in 2004, where it was read 20 million times. The abridged hardcover adaptation was released on **April 1, 2007**...	Diary of a Wimpy Kid: Double Down is the eleventh book in the "Diary of a Wimpy Kid" series by Jeff Kinney... The book was published on **November 1, 2016**...

BM25+BERT > ORQA for very specific representations better represented by sparse vectors

Conclusion

Significant Contributions

1.) First retriever-reader trained jointly end-to-end using only question-answer pairs

2.) Made possible because of the novel pre-training task: Inverse Cloze Task

3.) Learned retrieval proved to be successful when the question writers don't know the answer ("true" information seeking)

Potential Additions

1.) Only uses 128 dimension vectors, what happens when we increase this?

2.) Can we quantify the bias in TriviaQA and SQuAD?

Discussion Question

(Lee et al, 2019) made a distinction between different types of QA datasets and demonstrated that in some cases, a traditional unsupervised retrieval method (e.g., BM25, TF-IDF) works better while in some other cases, it is more effective to "learn" a retriever.

Can you state the argument and do you agree with it?

Discussion Question

(Lee et al, 2019) made a distinction between different types of QA datasets and demonstrated that in some cases, a traditional unsupervised retrieval method (e.g., BM25, TF-IDF) works better while in some other cases, it is more effective to "learn" a retriever.

Can you state the argument and do you agree with it?

Learned retrievers are better for “true” information-seeking tasks and succeed when question writers don’t know the answer ahead of time.

Do you agree?

Appendix

Hyperparameters and Specifics

What if we vary the ICT masking rate?

In all uses of BERT, they used an uncased base model
- 12 transformer layers
- 768 hidden size
- Default optimizer

h_q and h_b have 128 dimensions - small because they wanted it to run on a single machine

ICT Pre-training
- Learning rate of 10^{-4}
- Batch size 4096

Fine-tuning
- Learning rate of 10^{-5}
- Batch size of 1 on a single machine with 12GB GPU

Answer spans limited to 10 tokens

2 epochs of fine-tuning on large datasets with 20 for smaller ones