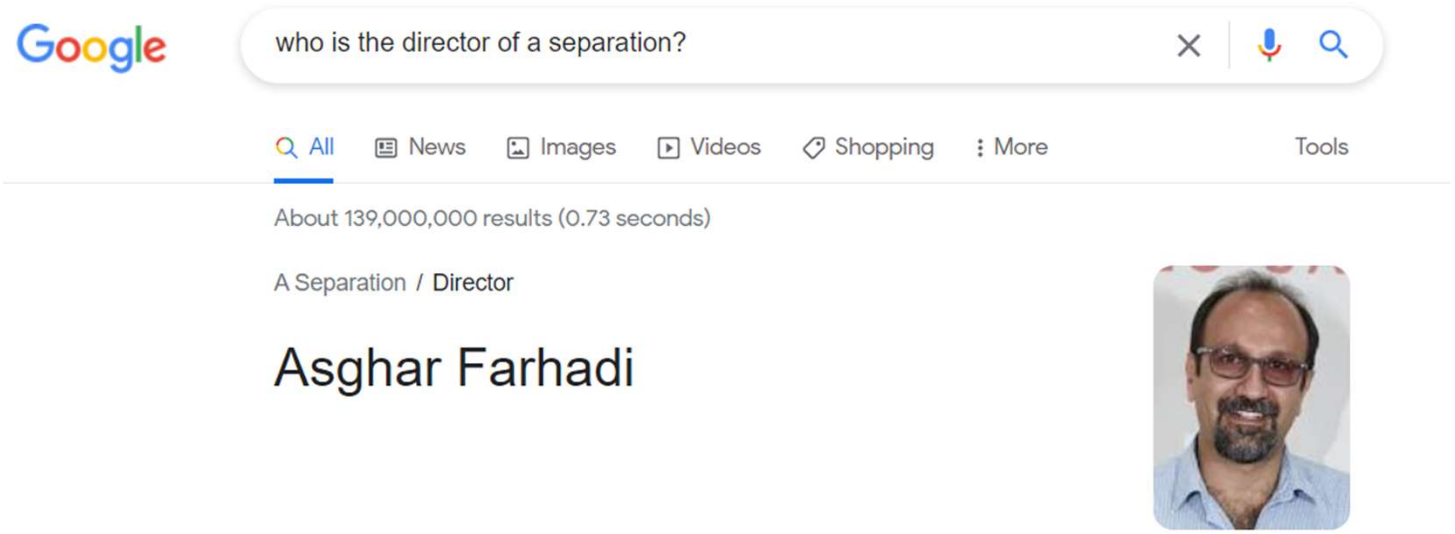
TeIAS | Tehran Institute for Advanced Studies

بنام خداوند جان و خرد

Question Answering

Mohammad Taher Pilehvar



who is the director of a separation?

About 139,000,000 results (0.73 seconds)

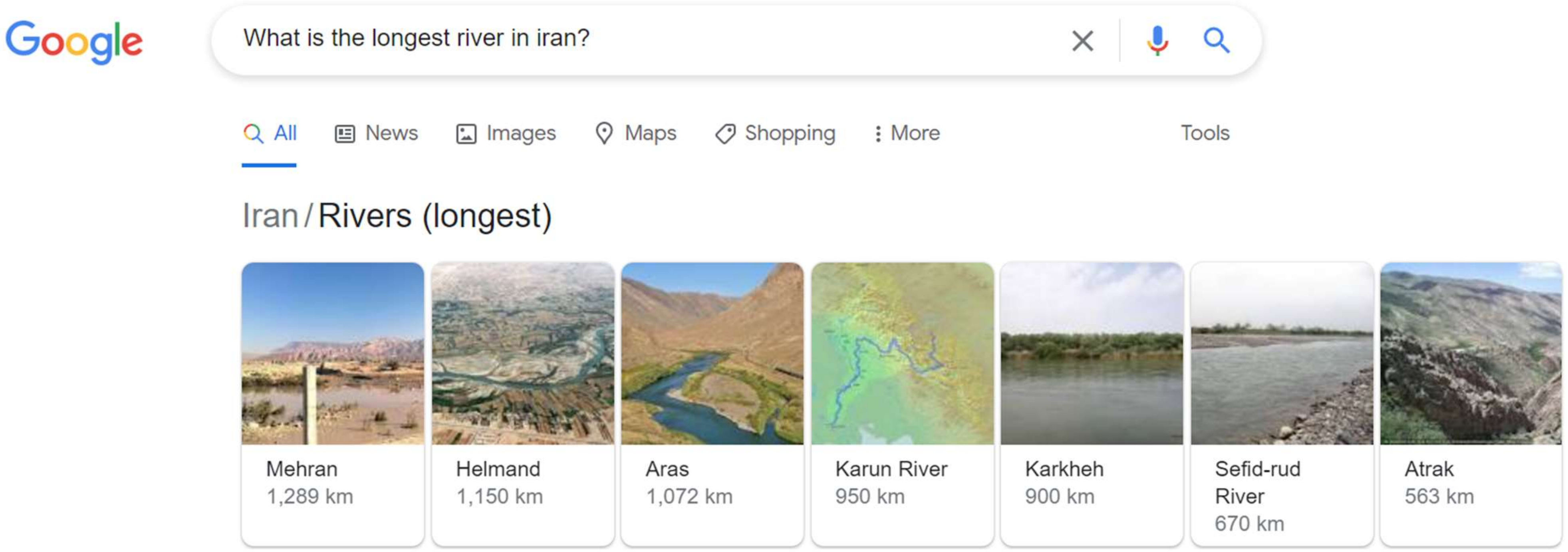
A Separation / Director

Asghar Farhadi

Natural Language Processing 1400
https://teias-courses.github.io/nlp00/

Most materials of these slides are taken from Stanford’s CS224N course

Question Answering

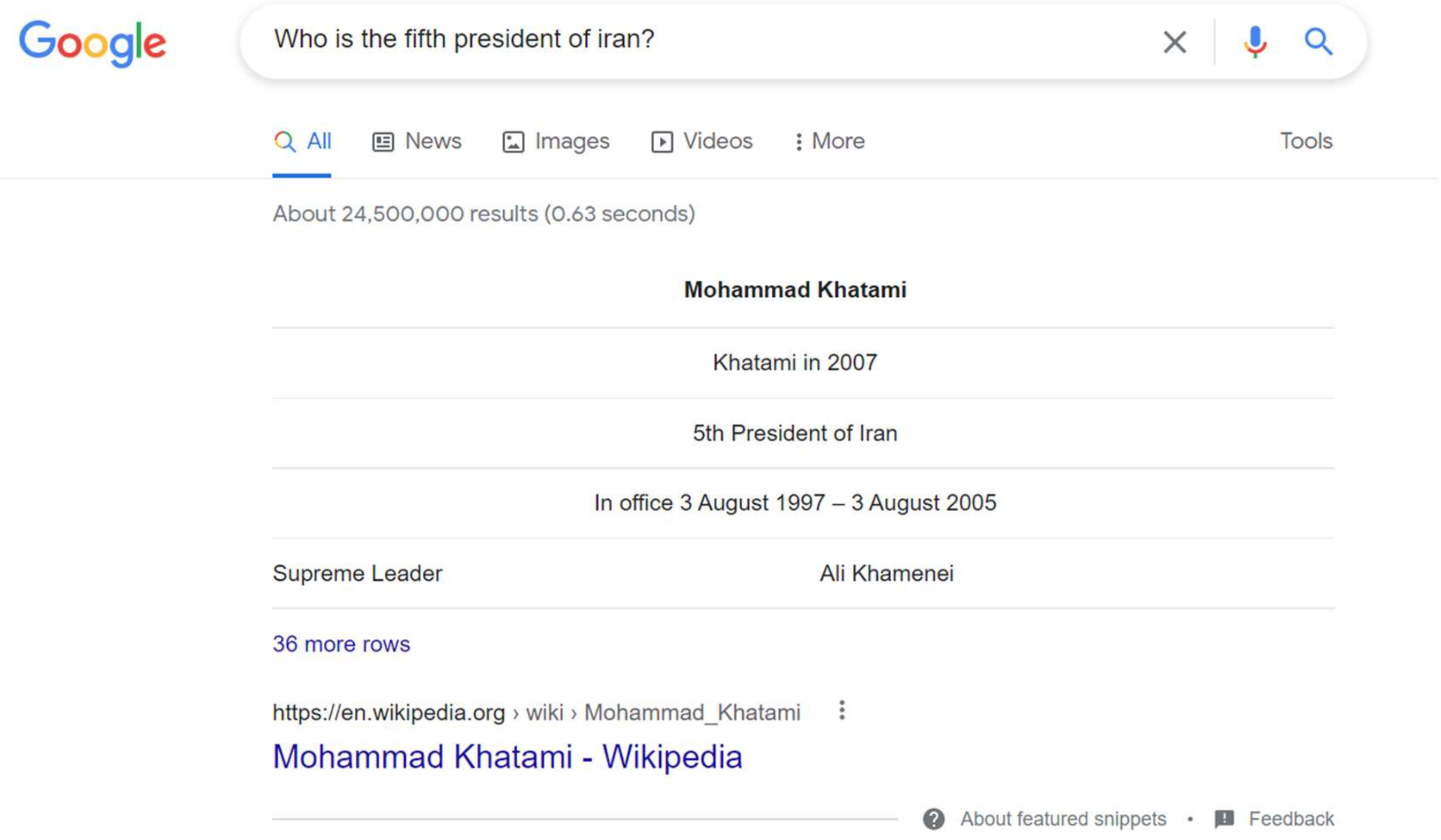


What is the longest river in iran?

Iran / Rivers (longest)

Mehran	1,289 km
Helmand	1,150 km
Aras	1,072 km
Karun River	950 km
Karkheh	900 km
Sefid-rud River	670 km
Atrak	563 km

Question Answering



Who is the fifth president of iran?

About 24,500,000 results (0.63 seconds)

Mohammad Khatami	
Khatami in 2007	
5th President of Iran	
In office 3 August 1997 – 3 August 2005	
Supreme Leader	Ali Khamenei

36 more rows

https://en.wikipedia.org › wiki › Mohammad_Khatami

Mohammad Khatami - Wikipedia

Motivation: Question answering

- With massive collections of full-text documents, i.e., the web, simply returning relevant documents is of limited use
- Rather, we often want answers to our questions
  - Especially on mobile
  - Or using a digital assistant device, like Alexa, Google Assistant, …
- We can factor this into two parts:
  1. Finding documents that (might) contain an answer
     - Which can be handled by traditional information retrieval/web search
  2. Finding an answer in a paragraph or a document
     - This problem is often termed Reading Comprehension
     - It is what we will focus on today

Machine Comprehension (Burges 2013)

A machine comprehends a passage of text if,

for any question regarding that text that can be answered correctly by a majority of native speakers,

that machine can provide a string which those speakers would agree both:

- answers that question,
- and does not contain information irrelevant to that question.”

MCTest dataset

Passage (P) + Question (Q) -> Answer (A)

```
Alyssa got to the beach after a long trip. She's from Charlotte. She
traveled from Atlanta. She's now in Miami. She went to Miami to visit
some friends. But she wanted some time to herself at the beach, so she
went there first. After going swimming and laying out, she went to her
friend Ellen's house. Ellen greeted Alyssa and they both had some
lemonade to drink. Alyssa called her friends Kristin and Rachel to
meet at Ellen's house. The girls traded stories and caught up on their
lives. It was a happy time for everyone. The girls went to a
restaurant for dinner. The restaurant had a special on catfish. Alyssa
enjoyed the restaurant's special. Ellen ordered a salad. Kristin had
soup. Rachel had a steak.
```

```
What city is Alyssa in? { "A": "trip",
                          "B": "Miami",
                          "C": "Atlanta",
                          "D": "beach" }
```

https://huggingface.co/datasets/sagnikrayc/mctest

A brief history of Reading Comprehension

- Much early NLP work attempted reading comprehension
  - Schank, Abelson, Lehnert et al. c. 1977 – “Yale A.I. Project”
- Revived by Lynette Hirschman in 1999:
  - Could NLP systems answer human reading comprehension questions for 3rd to 6th graders? Simple methods attempted.
- Revived again by Chris Burges in 2013 with MCTest
  - Again answering questions over simple story texts
- Floodgates opened in 2015/16 with the production of large datasets which permit supervised neural systems to be built
  - Hermann et al. (NIPS 2015) DeepMind CNN/DM dataset
  - Rajpurkar et al. (EMNLP 2016) SQuAD
  - MS MARCO, TriviaQA, RACE, NewsQA, NarrativeQA, …

A brief history of Open-domain QA

- 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

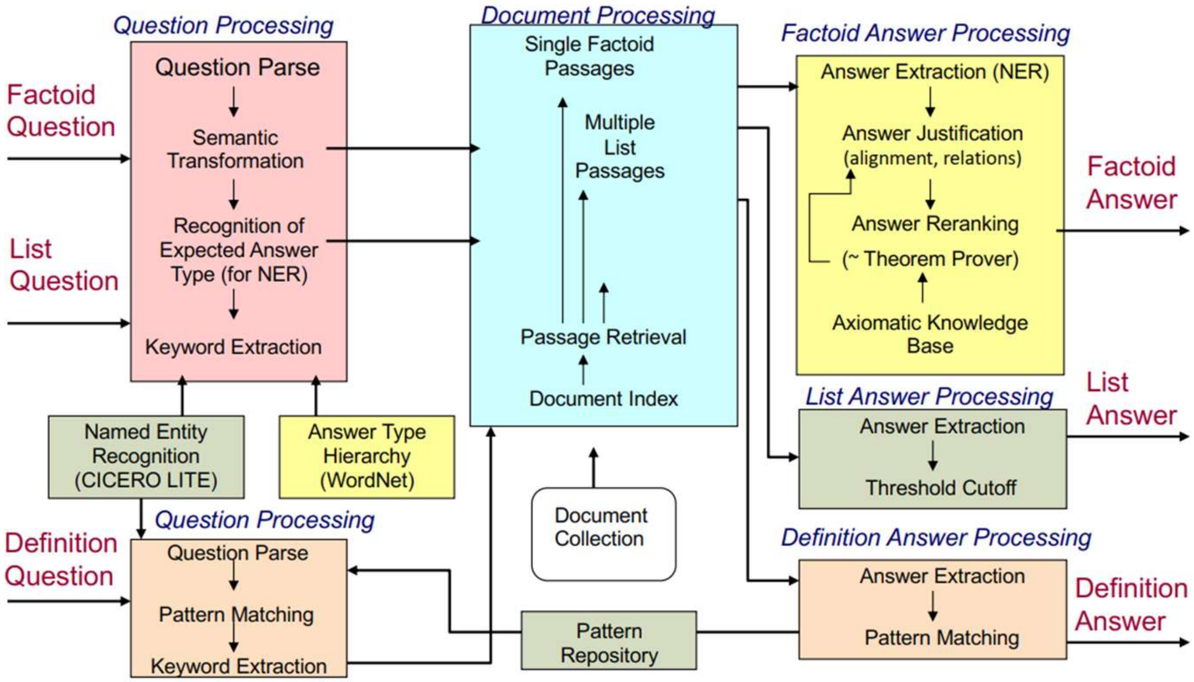
- The NIST TREC QA track begun in 1999 first rigorously investigated answering fact questions over a **large collection of documents**

- 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. 2016) uses IR followed by neural reading comprehension to bring deep learning to Open-domain QA

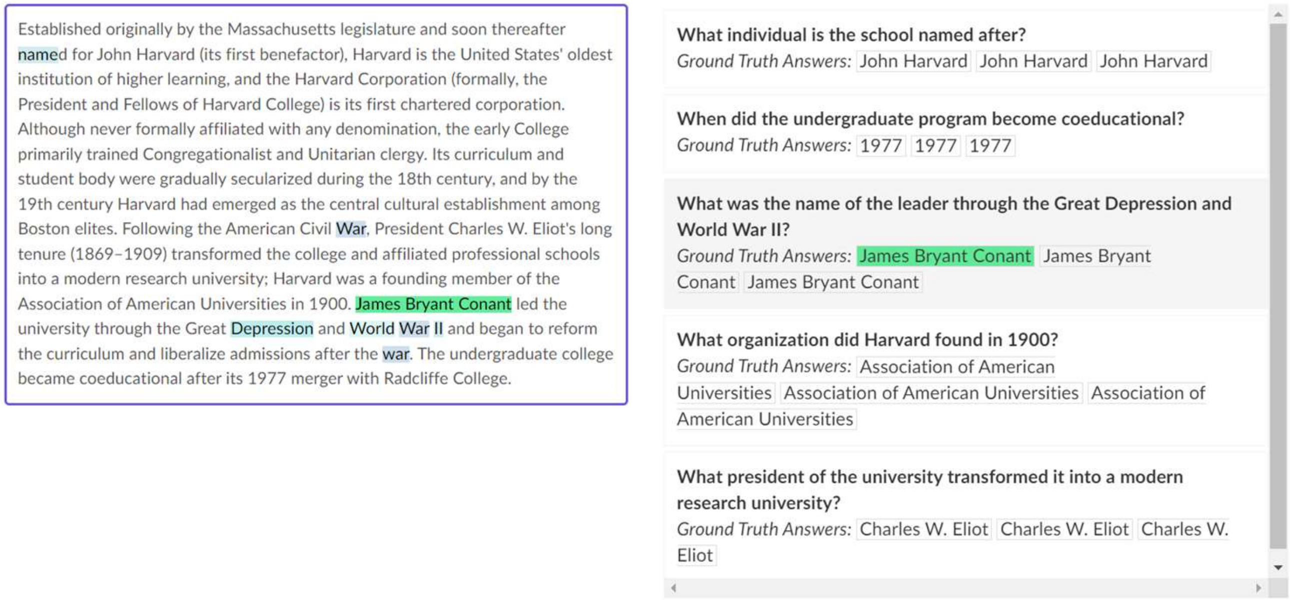
A brief history of Open-domain QA

- Architecture of LCC (Harabagiu/Moldovan) QA system, circa 2003
- Complex systems but they did work fairly well on “factoid” questions



Stanford QA Dataset (SQuAD)

- 100K examples. Extractive question answering (answer span of text)



Established originally by the Massachusetts legislature and soon thereafter named for John Harvard (its first benefactor), Harvard is the United States' oldest institution of higher learning, and the Harvard Corporation (formally, the President and Fellows of Harvard College) is its first chartered corporation. Although never formally affiliated with any denomination, the early College primarily trained Congregationalist and Unitarian clergy. Its curriculum and student body were gradually secularized during the 18th century, and by the 19th century Harvard had emerged as the central cultural establishment among Boston elites. Following the American Civil War, President Charles W. Eliot's long tenure (1869–1909) transformed the college and affiliated professional schools into a modern research university; Harvard was a founding member of the Association of American Universities in 1900. James Bryant Conant led the university through the Great Depression and World War II and began to reform the curriculum and liberalize admissions after the war. The undergraduate college became coeducational after its 1977 merger with Radcliffe College.

What individual is the school named after?
Ground Truth Answers: John Harvard John Harvard John Harvard

When did the undergraduate program become coeducational?
Ground Truth Answers: 1977 1977 1977

What was the name of the leader through the Great Depression and World War II?
Ground Truth Answers: James Bryant Conant James Bryant Conant James Bryant Conant

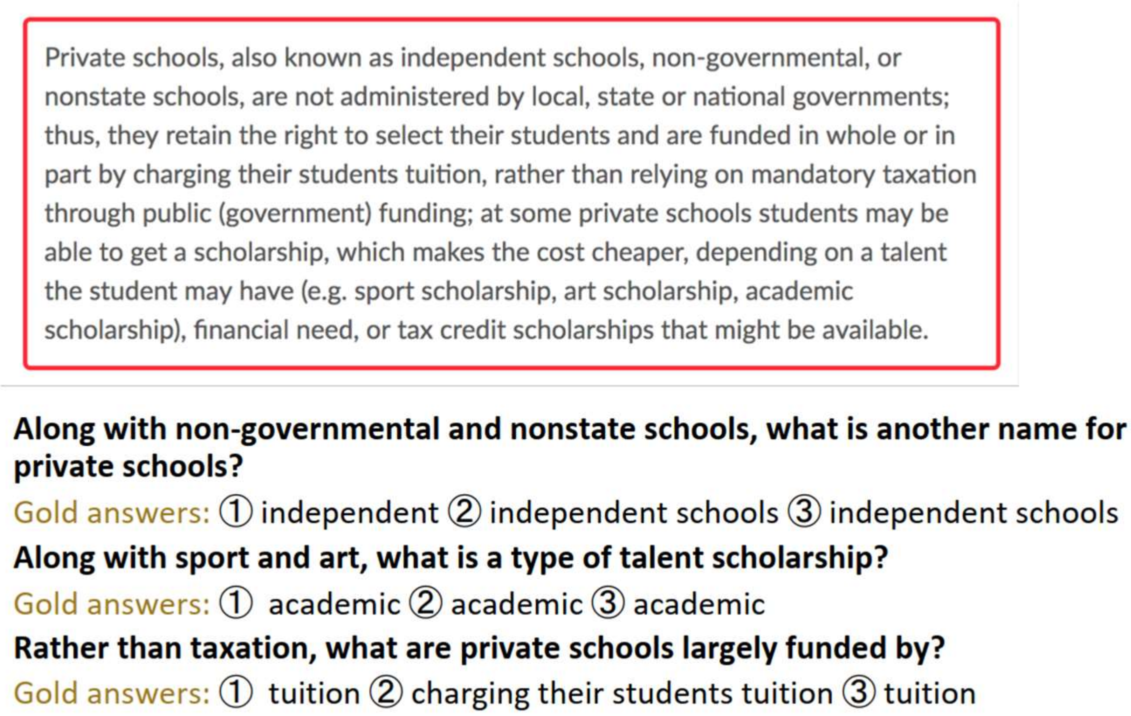
What organization did Harvard found in 1900?
Ground Truth Answers: Association of American Universities Association of American Universities Association of American Universities

What president of the university transformed it into a modern research university?
Ground Truth Answers: Charles W. Eliot Charles W. Eliot Charles W. Eliot

https://rajpurkar.github.io/SQuAD-explorer/

Stanford QA Dataset (SQuAD)

- 100K examples. Extractive question answering (answer span of text)



Private schools, also known as independent schools, non-governmental, or nonstate schools, are not administered by local, state or national governments; thus, they retain the right to select their students and are funded in whole or in part by charging their students tuition, rather than relying on mandatory taxation through public (government) funding; at some private schools students may be able to get a scholarship, which makes the cost cheaper, depending on a talent the student may have (e.g. sport scholarship, art scholarship, academic scholarship), financial need, or tax credit scholarships that might be available.

Along with non-governmental and nonstate schools, what is another name for private schools?

Gold answers: ① independent ② independent schools ③ independent schools

Along with sport and art, what is a type of talent scholarship?

Gold answers: ① academic ② academic ③ academic

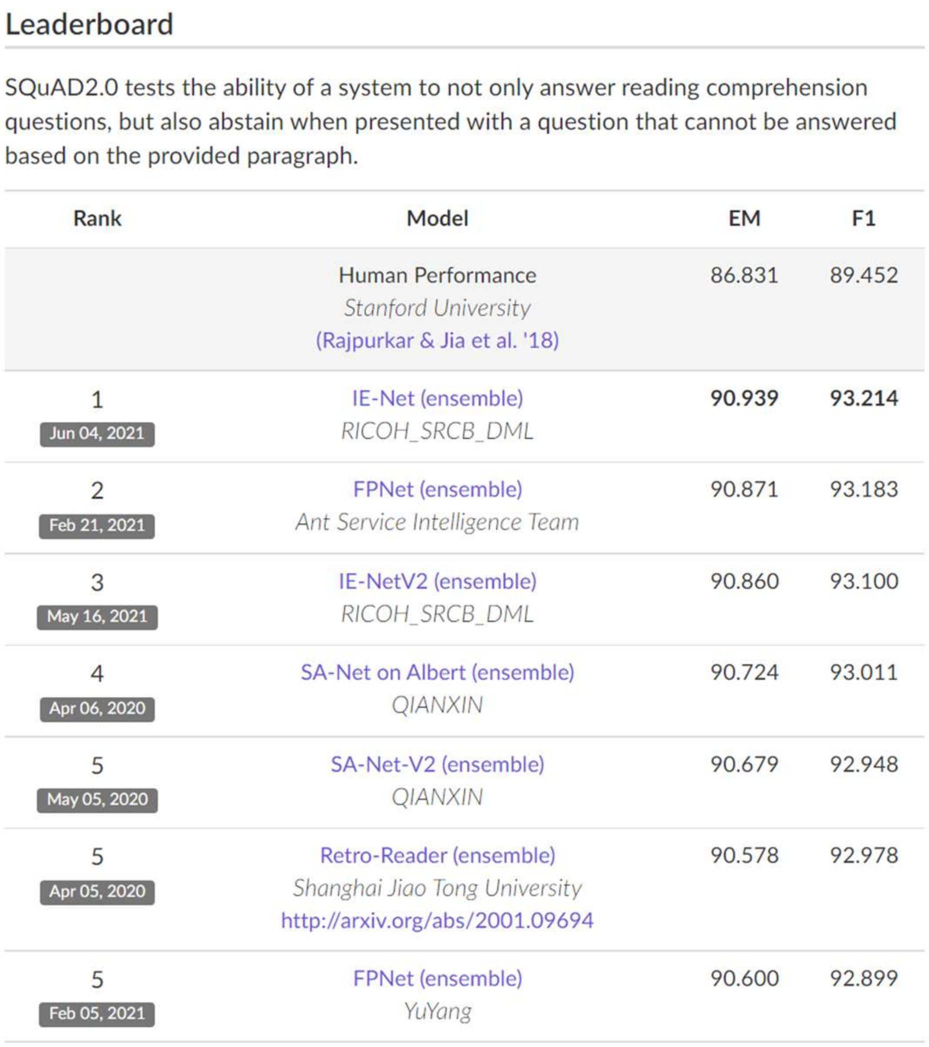
Rather than taxation, what are private schools largely funded by?

Gold answers: ① tuition ② charging their students tuition ③ tuition

SQuAD evaluation

- Authors collected 3 gold answers
- Systems are scored on two metrics:
  - Exact match: 1/0 accuracy on whether you match one of the 3 answers
  - F1: Take system and each gold answer as bag of words, evaluate Precision = $\frac{TP}{TP+FP}$, Recall = $\frac{TP}{TP+FN}$, harmonic mean F1 = $\frac{2PR}{P+R}$
    Score is (macro-)average of per-question F1 scores
- F1 measure is seen as more reliable and taken as primary
  - It's less based on choosing exactly the same span that humans chose, which is susceptible to various effects, including line breaks
- Both metrics ignore punctuation and articles (**a**, **an**, **the** only)

SQuAD



Leaderboard

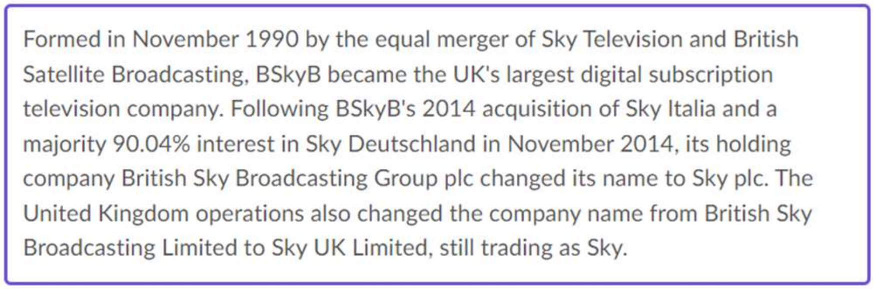
SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jun 04, 2021	IE-Net (ensemble) *RICOH_SRCB_DML*	**90.939**	**93.214**
2 Feb 21, 2021	FPNet (ensemble) *Ant Service Intelligence Team*	90.871	93.183
3 May 16, 2021	IE-NetV2 (ensemble) *RICOH_SRCB_DML*	90.860	93.100
4 Apr 06, 2020	SA-Net on Albert (ensemble) *QIANXIN*	90.724	93.011
5 May 05, 2020	SA-Net-V2 (ensemble) *QIANXIN*	90.679	92.948
5 Apr 05, 2020	Retro-Reader (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694	90.578	92.978
5 Feb 05, 2021	FPNet (ensemble) *YuYang*	90.600	92.899

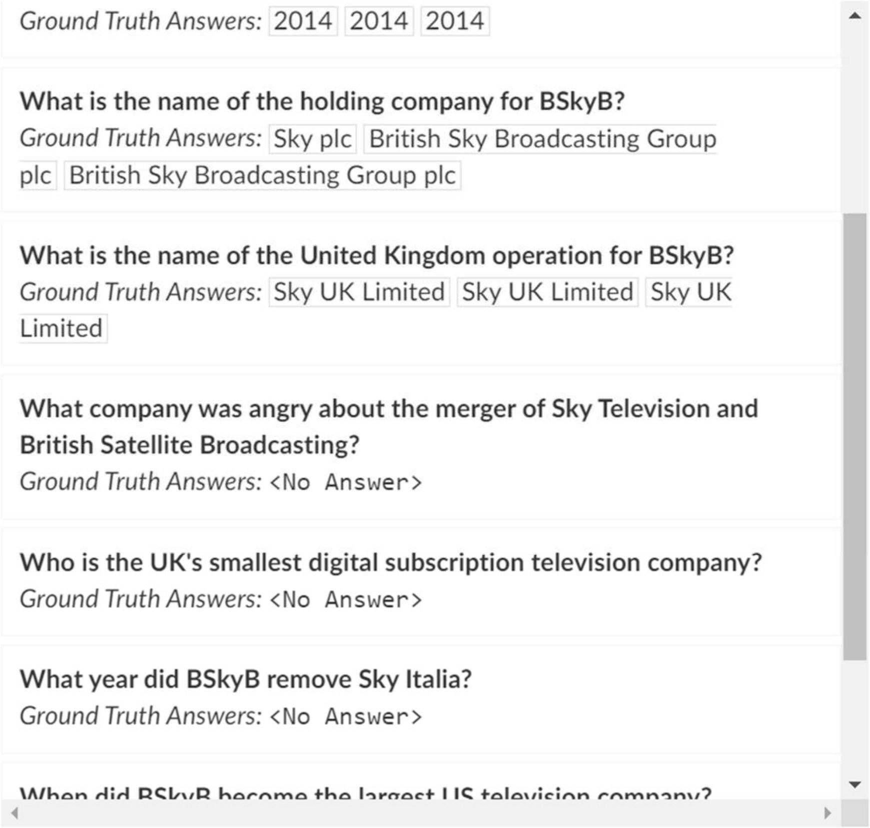
SQuAD limitations

- A defect of SQuAD 1.0 is that all questions have an answer in the paragraph
- Systems (implicitly) rank candidates and choose the best one
- You don’t have to judge whether a span answers the question
- In SQuAD 2.0, 1/3 of the training questions have no answer, and about 1/2 of the dev/test questions have no answer
  - For NoAnswer examples, NoAnswer receives a score of 1, and any other response gets 0, for both exact match and F1
- Simplest system approach to SQuAD 2.0:
  - Have a threshold score for whether a span answers a question
- Or you could have a second component that confirms answering
  - Like Natural Language Inference (NLI) or “Answer validation”

SQuAD 2.0



Formed in November 1990 by the equal merger of Sky Television and British Satellite Broadcasting, BSkyB became the UK's largest digital subscription television company. Following BSkyB's 2014 acquisition of Sky Italia and a majority 90.04% interest in Sky Deutschland in November 2014, its holding company British Sky Broadcasting Group plc changed its name to Sky plc. The United Kingdom operations also changed the company name from British Sky Broadcasting Limited to Sky UK Limited, still trading as Sky.



Ground Truth Answers: 2014 2014 2014

What is the name of the holding company for BSkyB?
Ground Truth Answers: Sky plc British Sky Broadcasting Group plc British Sky Broadcasting Group plc

What is the name of the United Kingdom operation for BSkyB?
Ground Truth Answers: Sky UK Limited Sky UK Limited Sky UK Limited

What company was angry about the merger of Sky Television and British Satellite Broadcasting?
Ground Truth Answers: `<No Answer>`

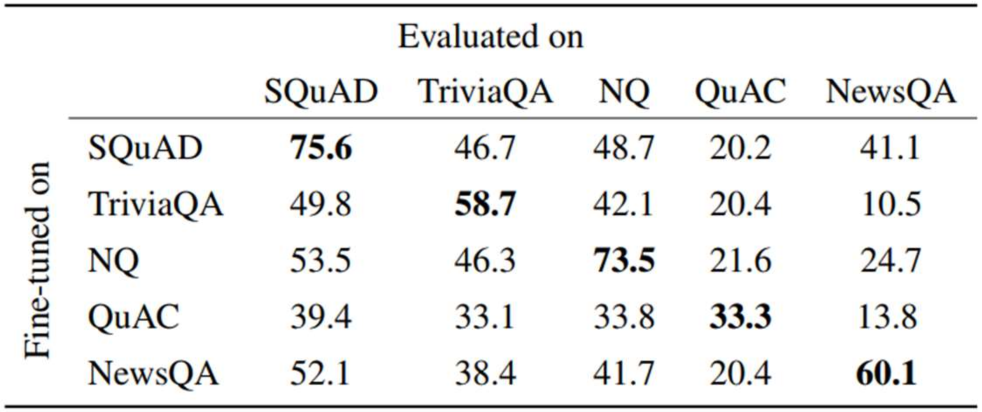
Who is the UK's smallest digital subscription television company?
Ground Truth Answers: `<No Answer>`

What year did BSkyB remove Sky Italia?
Ground Truth Answers: `<No Answer>`

When did BSkyB become the largest US television company?

QA: Lack of cross-domain robustness

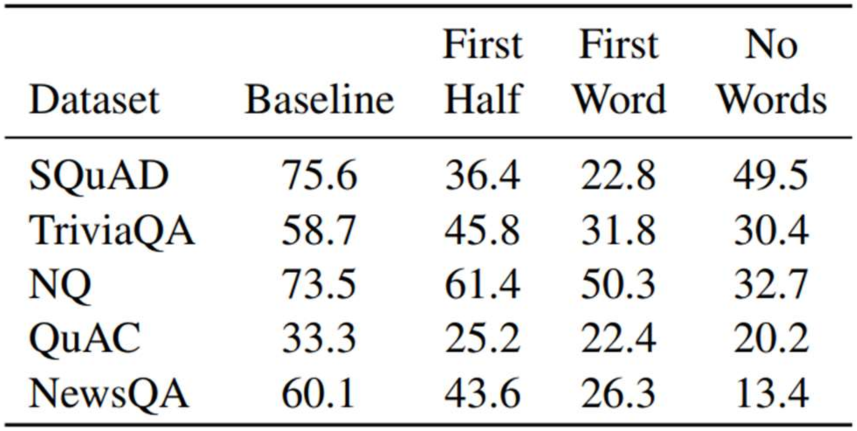
- Sen and Saffari (2020): What do Models Learn from Question Answering Datasets?



		Evaluated on				
		SQuAD	TriviaQA	NQ	QuAC	NewsQA
Fine-tuned on	SQuAD	**75.6**	46.7	48.7	20.2	41.1
	TriviaQA	49.8	**58.7**	42.1	20.4	10.5
	NQ	53.5	46.3	**73.5**	21.6	24.7
	QuAC	39.4	33.1	33.8	**33.3**	13.8
	NewsQA	52.1	38.4	41.7	20.4	**60.1**

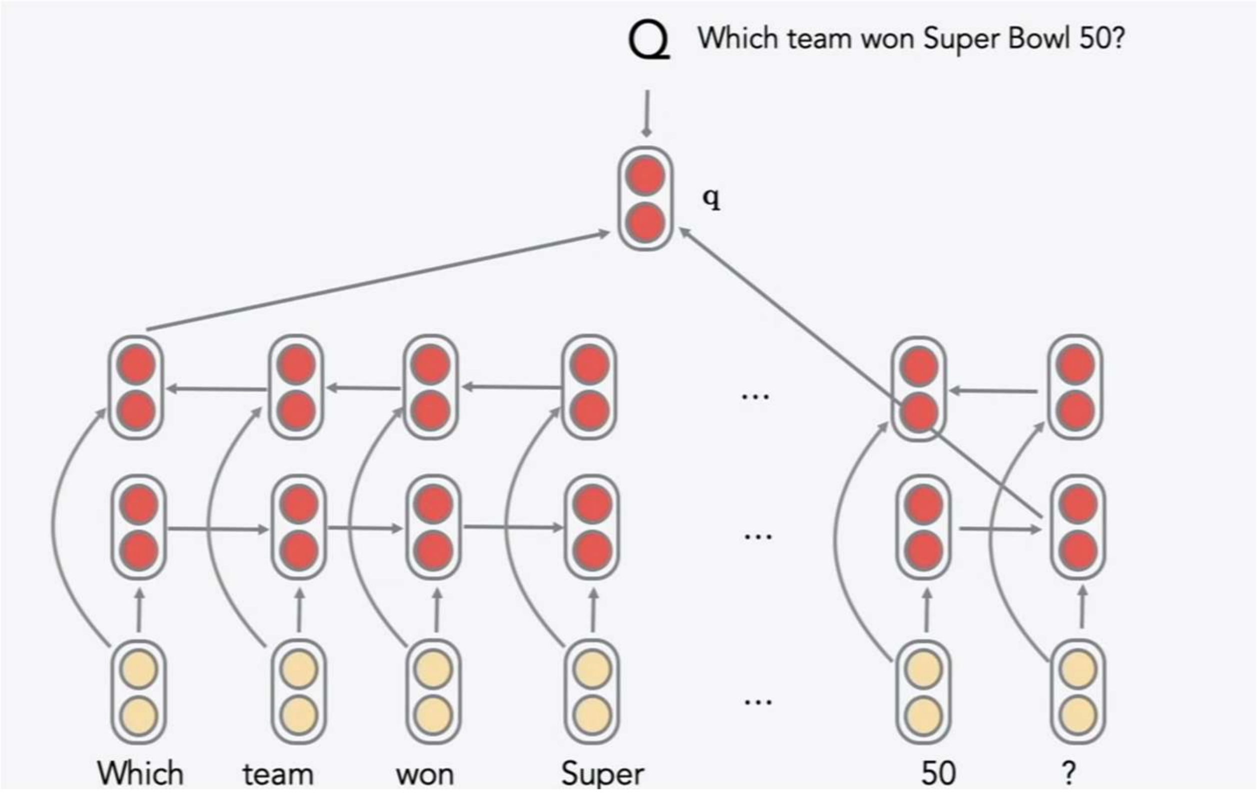
QA: Annotation bias (superficial cues)

- Sen and Saffari (2020): What do Models Learn from Question Answering Datasets?

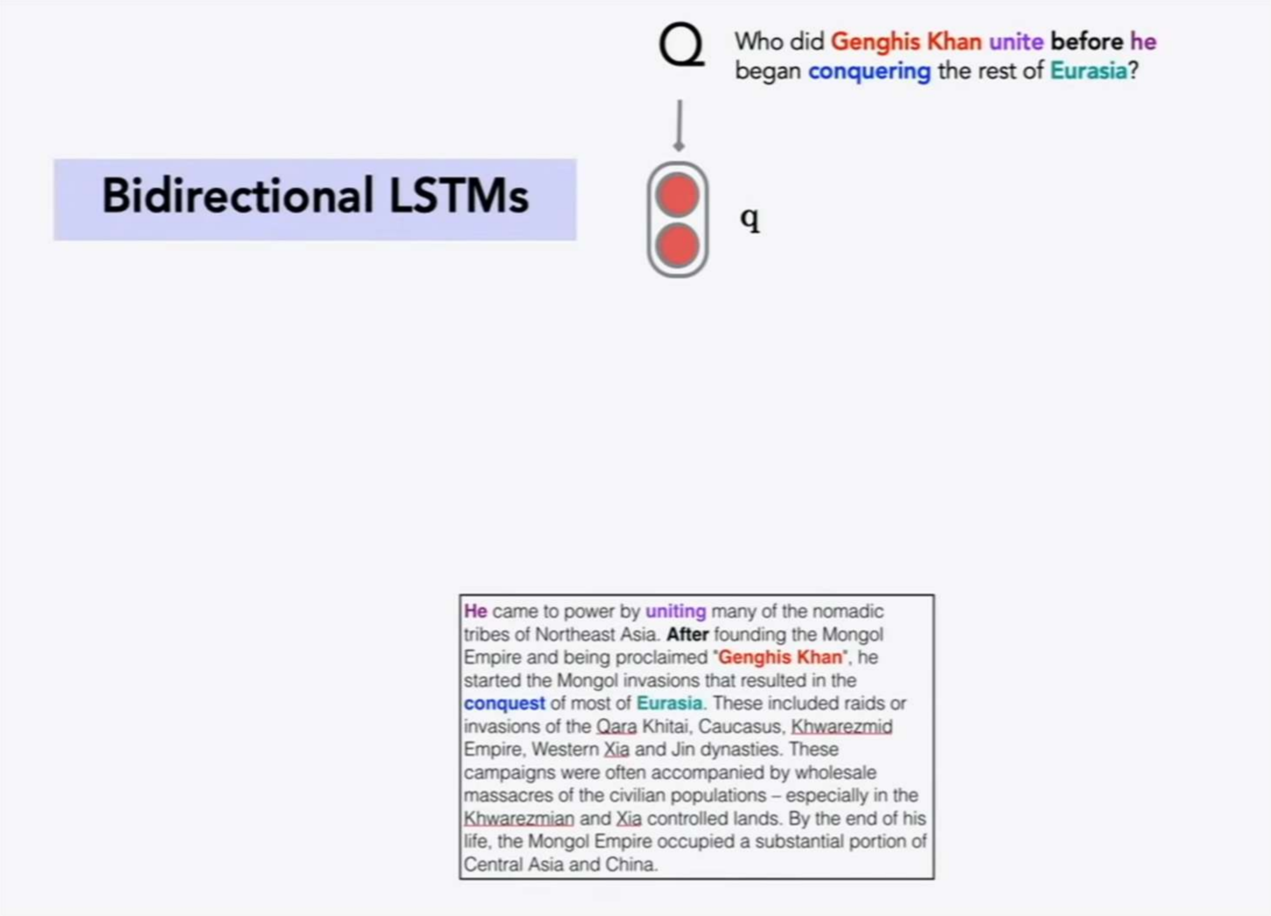


Dataset	Baseline	First Half	First Word	No Words
SQuAD	75.6	36.4	22.8	49.5
TriviaQA	58.7	45.8	31.8	30.4
NQ	73.5	61.4	50.3	32.7
QuAC	33.3	25.2	22.4	20.2
NewsQA	60.1	43.6	26.3	13.4

Stanford Attentive Reader



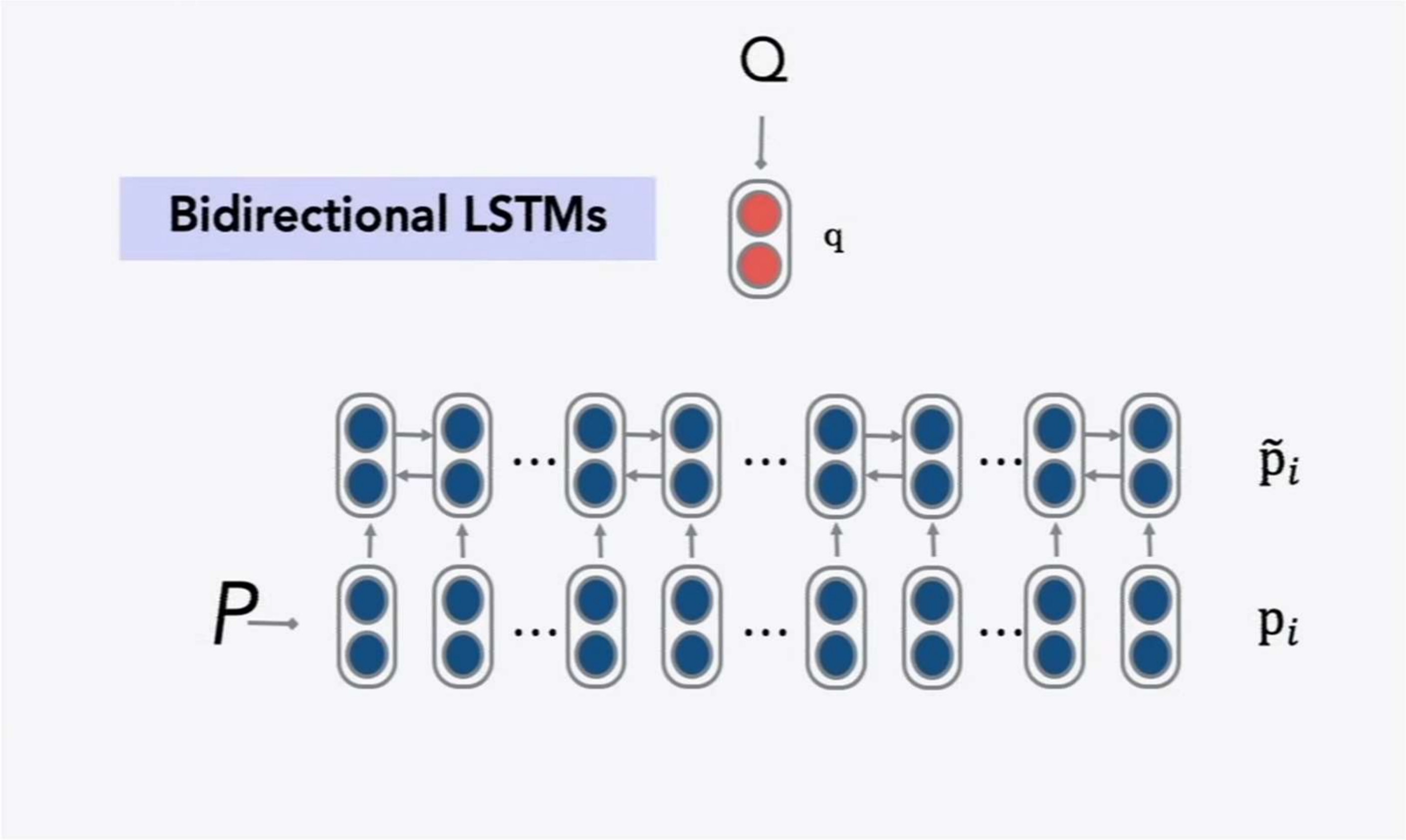
Stanford Attentive Reader



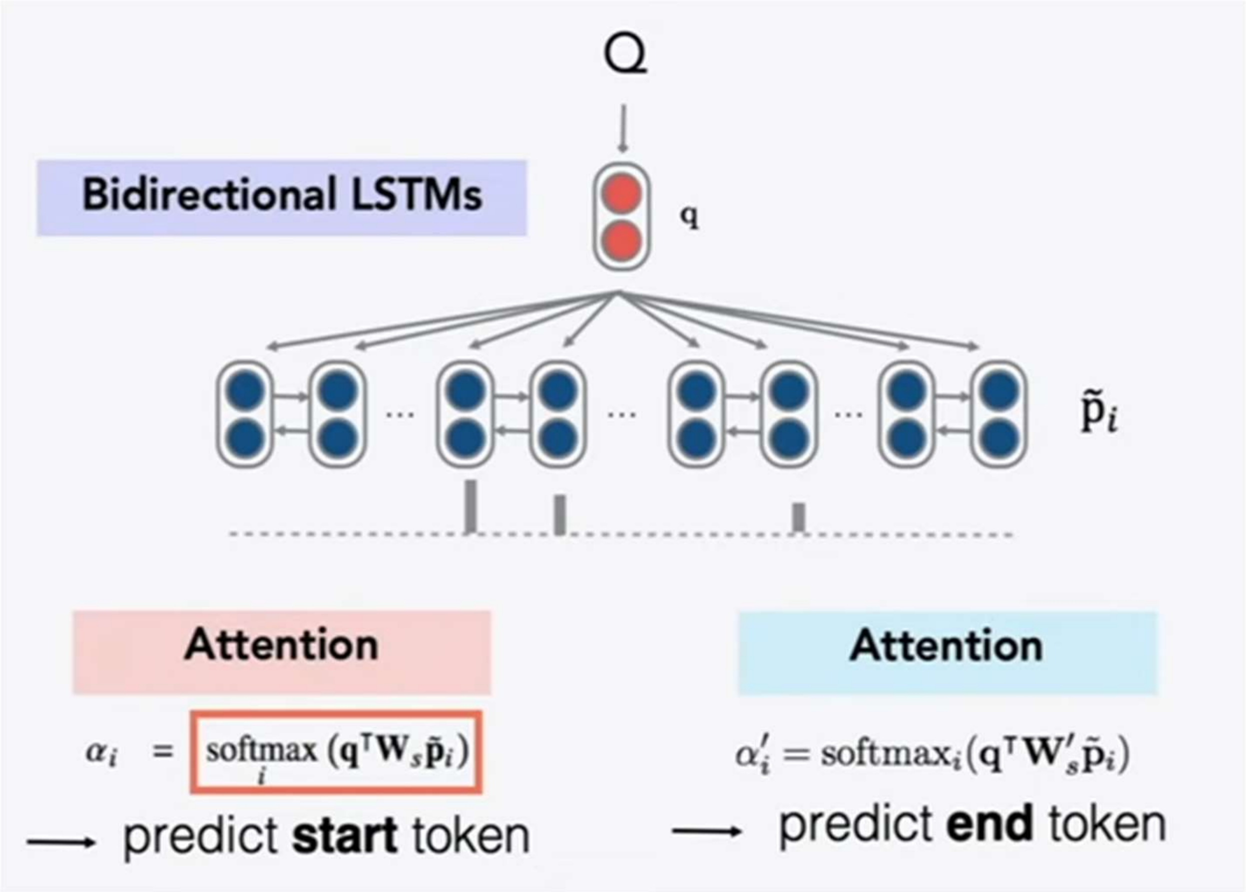
Who did **Genghis Khan unite before he** began **conquering** the rest of **Eurasia**?

He came to power by **uniting** many of the nomadic tribes of Northeast Asia. **After** founding the Mongol Empire and being proclaimed "**Genghis Khan**", he started the Mongol invasions that resulted in the **conquest** of most of **Eurasia**. These included raids or invasions of the Qara Khitai, Caucasus, Khwarezmid Empire, Western Xia and Jin dynasties. These campaigns were often accompanied by wholesale massacres of the civilian populations – especially in the Khwarezmian and Xia controlled lands. By the end of his life, the Mongol Empire occupied a substantial portion of Central Asia and China.

Stanford Attentive Reader



Stanford Attentive Reader



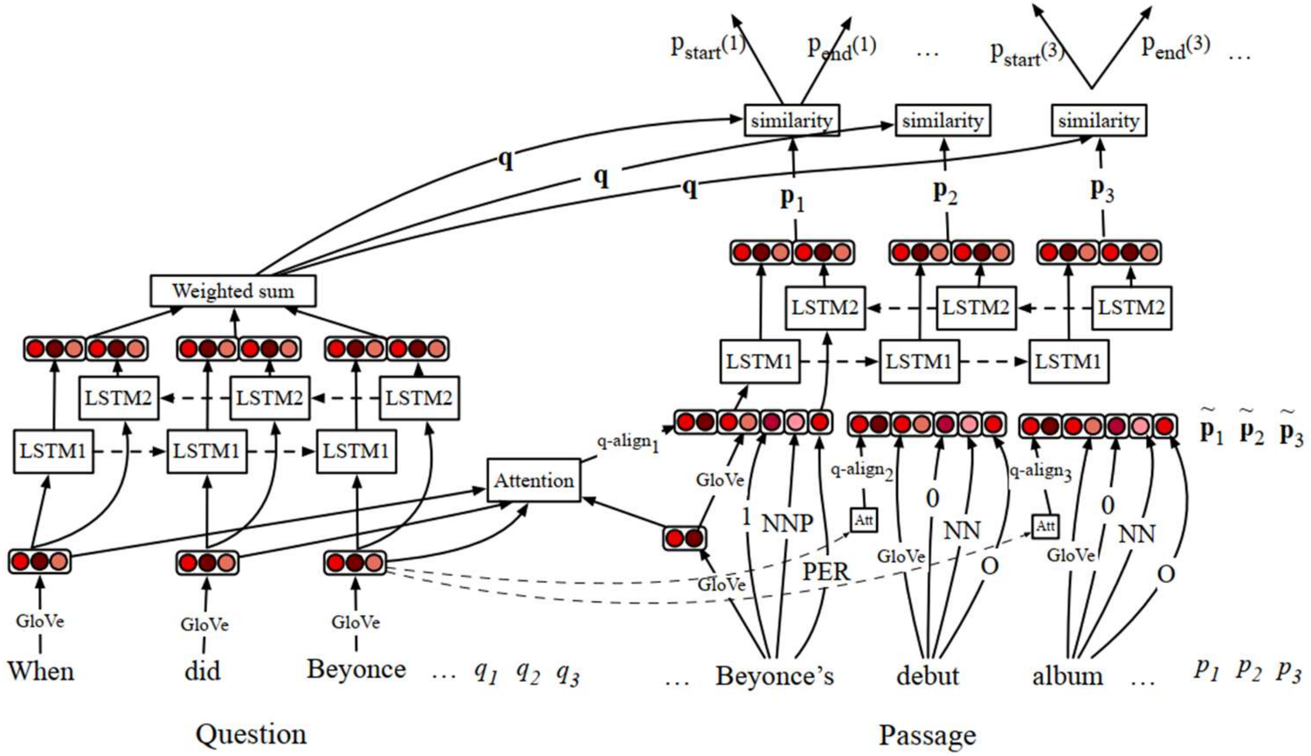
$$\alpha_i = \boxed{\underset{i}{\text{softmax}}\,(\mathbf{q}^\top \mathbf{W}_s \tilde{\mathbf{p}}_i)}$$

→ predict **start** token

$$\alpha'_i = \text{softmax}_i(\mathbf{q}^\top \mathbf{W}'_s \tilde{\mathbf{p}}_i)$$

→ predict **end** token

Stanford Attentive Reader++

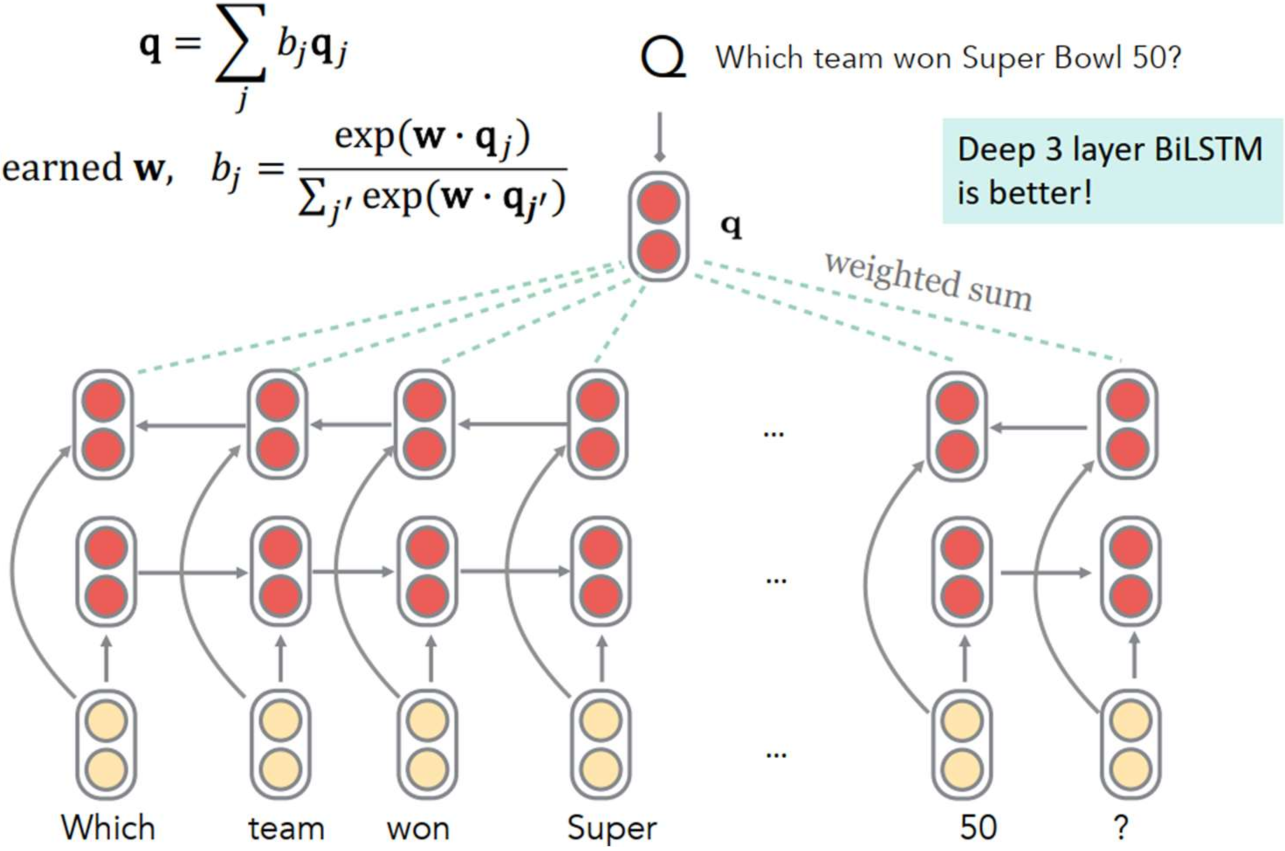


Training objective: $\mathcal{L} = -\sum \log P^{(\text{start})}(a_{\text{start}}) - \sum \log P^{(\text{end})}(a_{\text{end}})$

Stanford Attentive Reader++

$$\mathbf{q} = \sum_j b_j \mathbf{q}_j$$

For learned $\mathbf{w}$, $\quad b_j = \dfrac{\exp(\mathbf{w} \cdot \mathbf{q}_j)}{\sum_{j'} \exp(\mathbf{w} \cdot \mathbf{q}_{j'})}$



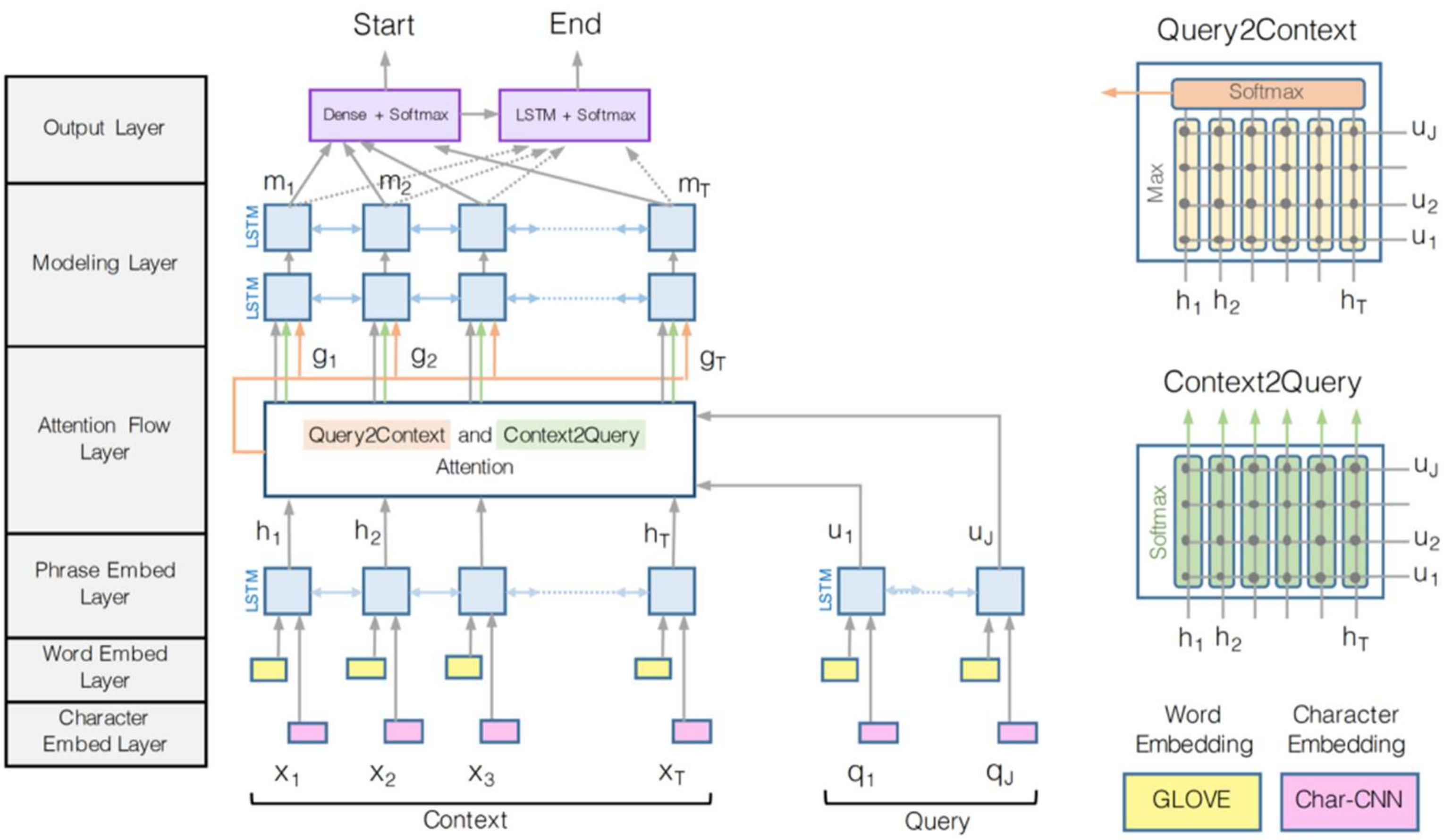
Deep 3 layer BiLSTM is better!

Stanford Attentive Reader++

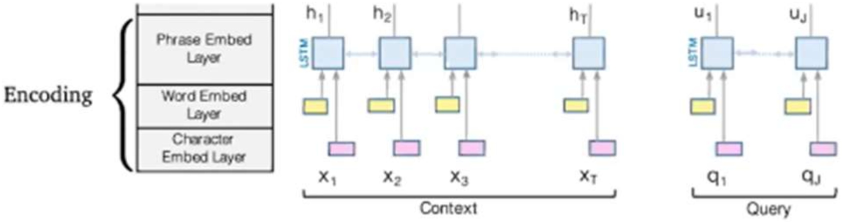
- Vector representation of each token in passage
- Made from concatenation of
  - Word embedding (GloVe 300d)
  - Linguistic features: POS & NER tags, one-hot encoded
  - Term frequency (unigram probability)
  - Exact match: whether the word appears in the question
  - 3 binary features: exact, uncased, lemma

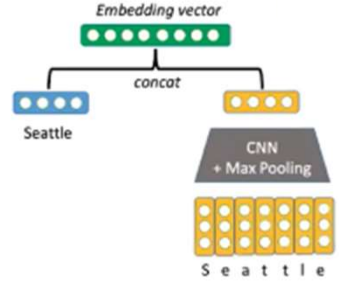
- Aligned question embedding (“car” vs “vehicle”)

BiDAF: Bi-Directional Attention Flow for Machine Comprehension



BiDAF: Bi-Directional Attention Flow for Machine Comprehension





- Use a concatenation of word embedding (GloVe) and character embedding (CNNs over character embeddings) for each word in context and query.

$$e(c_i) = f([\text{GloVe}(c_i); \text{charEmb}(c_i)]) \qquad e(q_i) = f([\text{GloVe}(q_i); \text{charEmb}(q_i)])$$

f: high-way networks omitted here

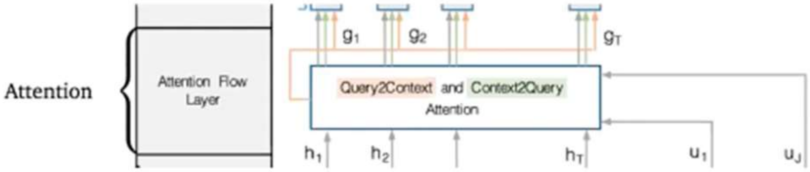
- Then, use two **bidirectional** LSTMs separately to produce contextual embeddings for both context and query.

$$\overrightarrow{\mathbf{c}}_i = \text{LSTM}(\overrightarrow{\mathbf{c}}_{i-1}, e(c_i)) \in \mathbb{R}^H \qquad \overrightarrow{\mathbf{q}}_i = \text{LSTM}(\overrightarrow{\mathbf{q}}_{i-1}, e(q_i)) \in \mathbb{R}^H$$

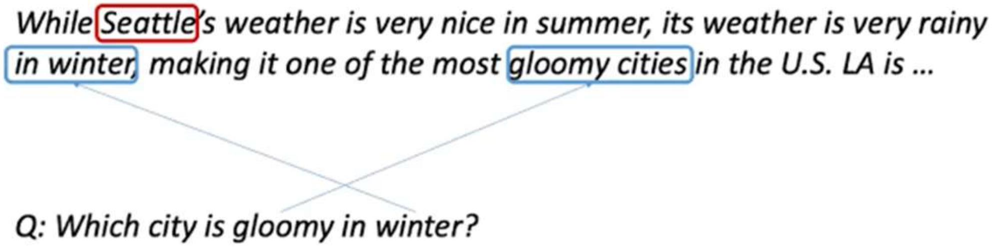
$$\overleftarrow{\mathbf{c}}_i = \text{LSTM}(\overleftarrow{\mathbf{c}}_{i+1}, e(c_i)) \in \mathbb{R}^H \qquad \overleftarrow{\mathbf{q}}_i = \text{LSTM}(\overleftarrow{\mathbf{q}}_{i+1}, e(q_i)) \in \mathbb{R}^H$$

$$\mathbf{c}_i = [\overrightarrow{\mathbf{c}}_i; \overleftarrow{\mathbf{c}}_i] \in \mathbb{R}^{2H} \qquad \mathbf{q}_i = [\overrightarrow{\mathbf{q}}_i; \overleftarrow{\mathbf{q}}_i] \in \mathbb{R}^{2H}$$

BiDAF: Bi-Directional Attention Flow for Machine Comprehension



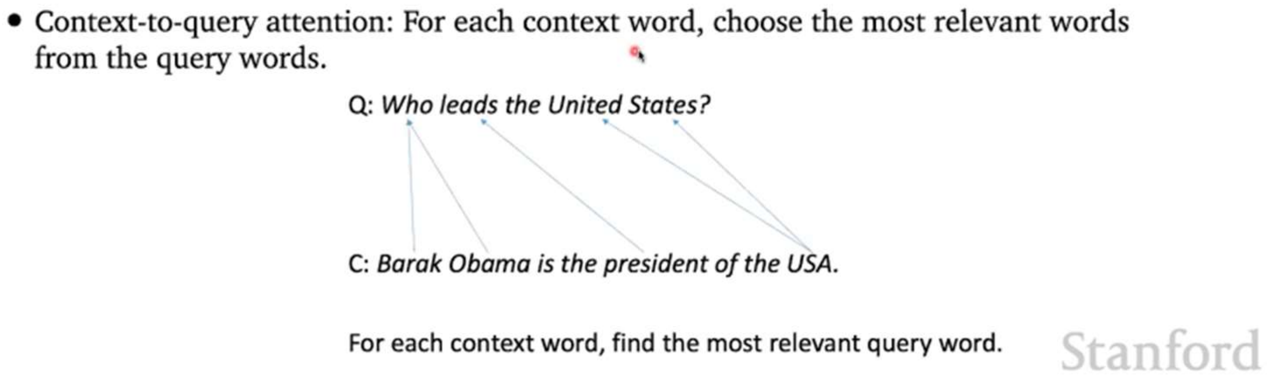
- Query-to-context attention: choose the context words that are most relevant to one of query words.



While Seattle's weather is very nice in summer, its weather is very rainy in winter, making it one of the most gloomy cities in the U.S. LA is ...

Q: Which city is gloomy in winter?

BiDAF: Bi-Directional Attention Flow for Machine Comprehension



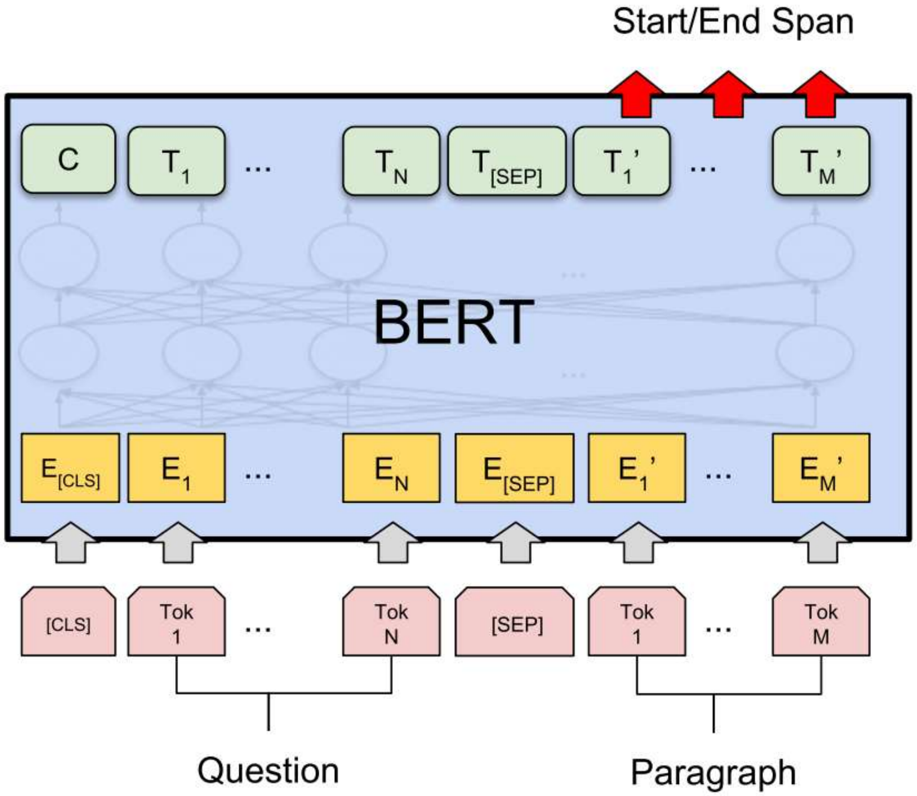
- Context-to-query attention: For each context word, choose the most relevant words from the query words.

Q: *Who leads the United States?*

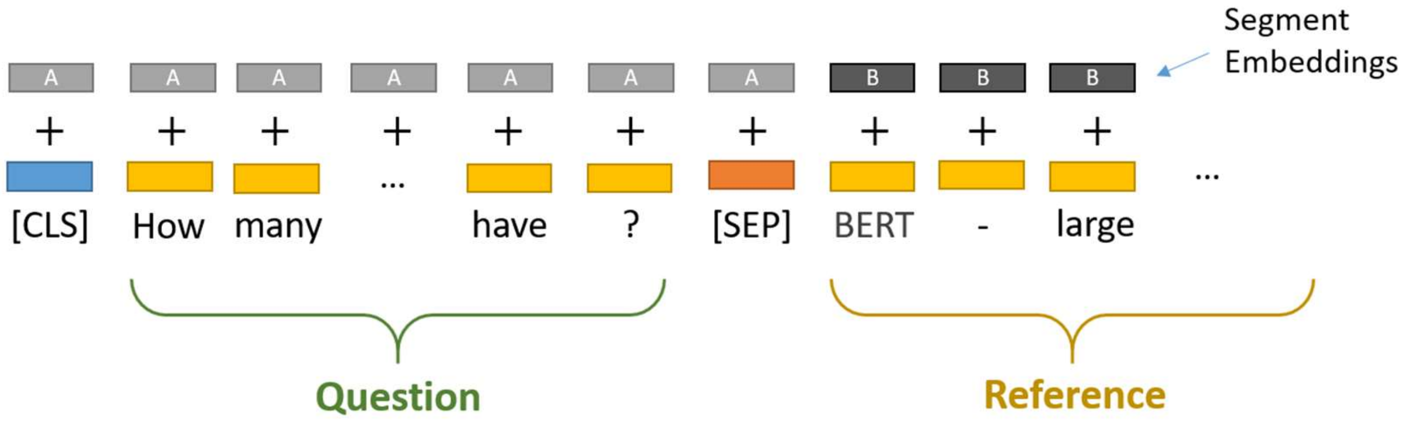
C: *Barak Obama is the president of the USA.*

For each context word, find the most relevant query word.

BERT for Question Answering



BERT for Question Answering



Question: `How many parameters does BERT-large have?`

Reference Text:

```
BERT-large is really big... it has 24 layers
and an embedding size of 1,024, for a total
of 340M parameters! Altogether it is 1.34GB,
so expect it to take a couple minutes to
download to your Colab instance.
```

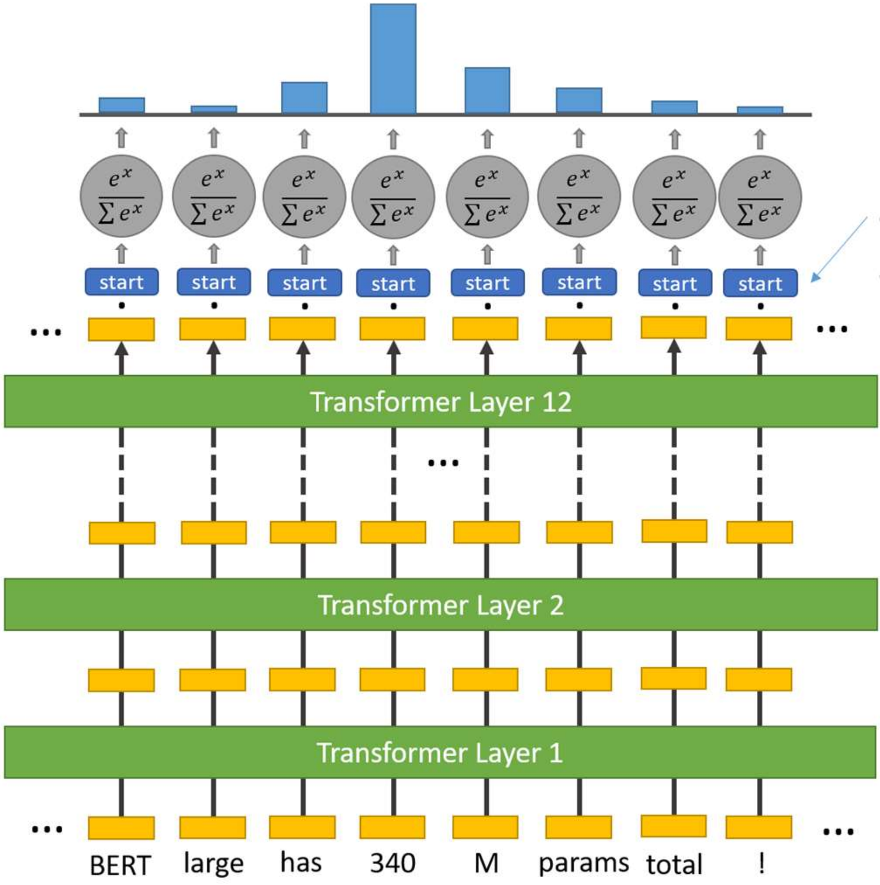
BERT for Question Answering

$$\mathcal{L} = -\log p_{\text{start}}(s^*) - \log p_{\text{end}}(e^*)$$

$$p_{\text{start}}(i) = \text{softmax}_i(\mathbf{w}_{\text{start}}^{\mathsf{T}} \mathbf{h}_i)$$

$$p_{\text{end}}(i) = \text{softmax}_i(\mathbf{w}_{\text{end}}^{\mathsf{T}} \mathbf{h}_i)$$

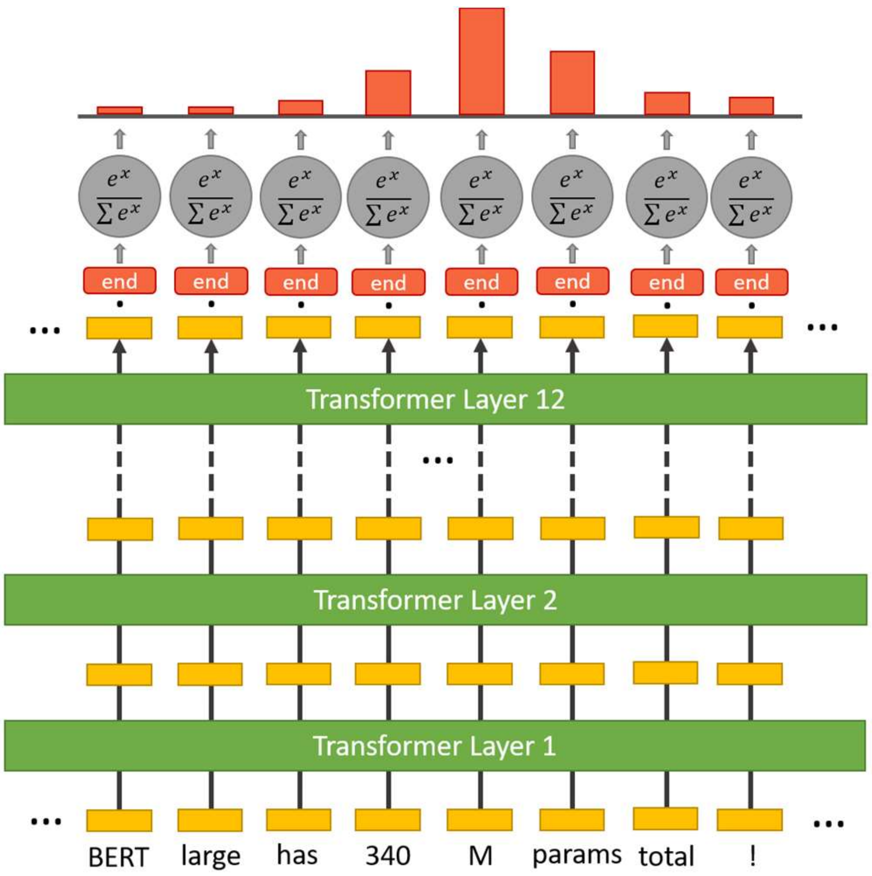
where $\mathbf{h}_i$ is the hidden vector of c_i, returned by BERT



This length 768 vector is the ***weights*** *for the start token classifier.*
The ***same weights*** *are applied to* ***every position****.*

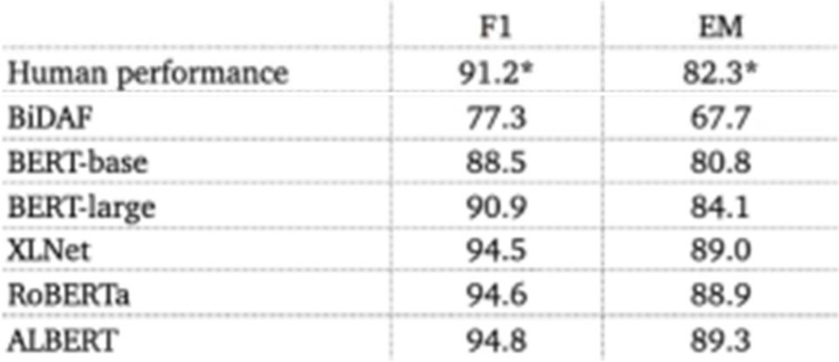
by Chris McCormick

BERT for Question Answering



by Chris McCormick

BERT vs. older models



	F1	EM
Human performance	91.2*	82.3*
BiDAF	77.3	67.7
BERT-base	88.5	80.8
BERT-large	90.9	84.1
XLNet	94.5	89.0
RoBERTa	94.6	88.9
ALBERT	94.8	89.3

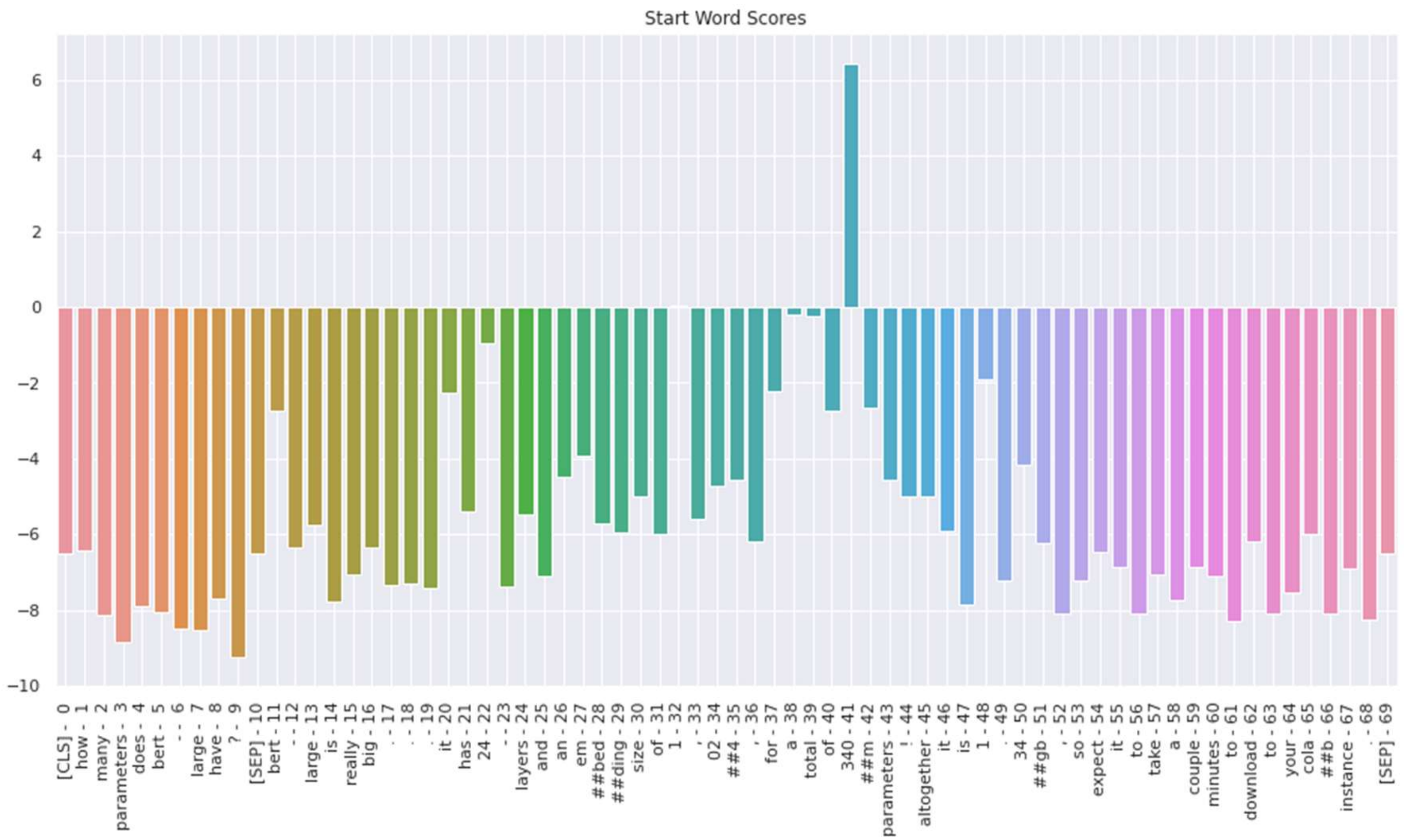
BERT for Question Answering

Fine-tuning BERT for QA

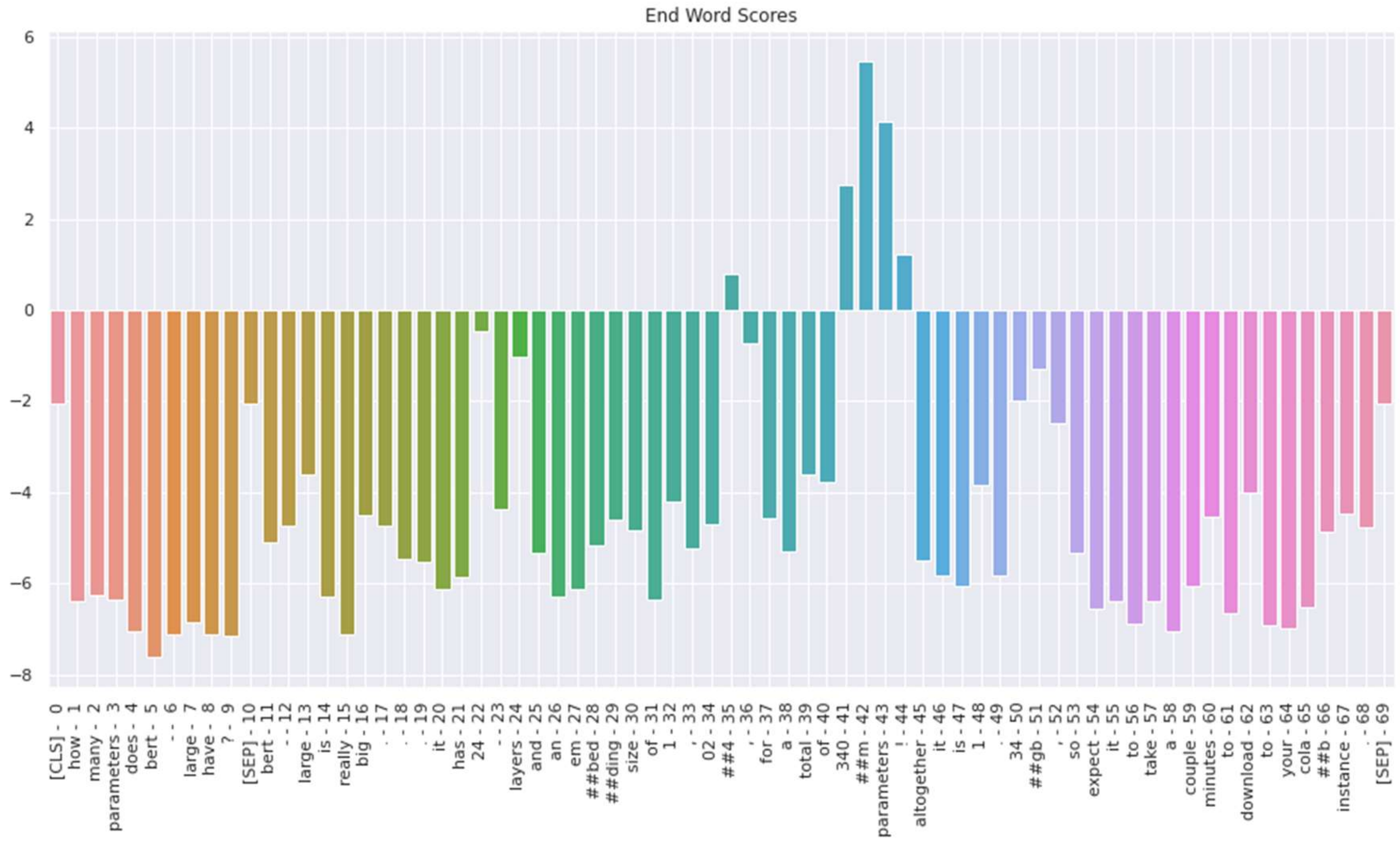
Colab notebook by Chris McCormick

https://colab.research.google.com/drive/1uSlWtJdZmLrI3FCNIlUHFxwAJiSu2J0-#scrollTo=bT5ESKDxfnLf

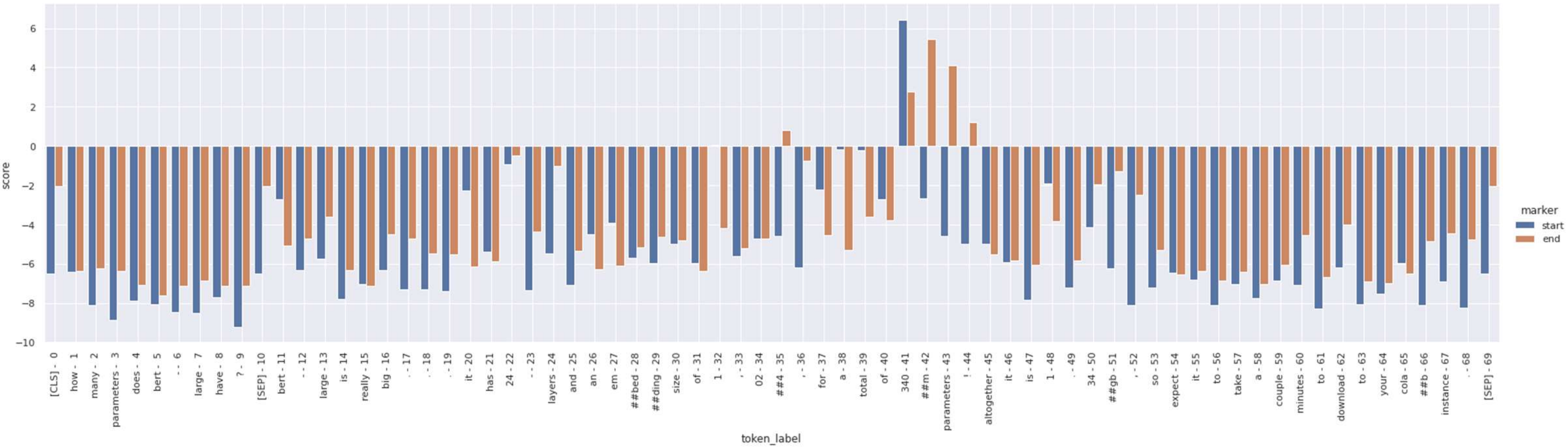
BERT for Question Answering



BERT for Question Answering



BERT for Question Answering



Questions