SFU NatLangLab

CMPT 713: Natural Language Processing

Question Answering

Spring 2023

2023-03-16

Adapted from slides from Danqi Chen and Karthik Narasimhan
(with some content from slides from Chris Manning)

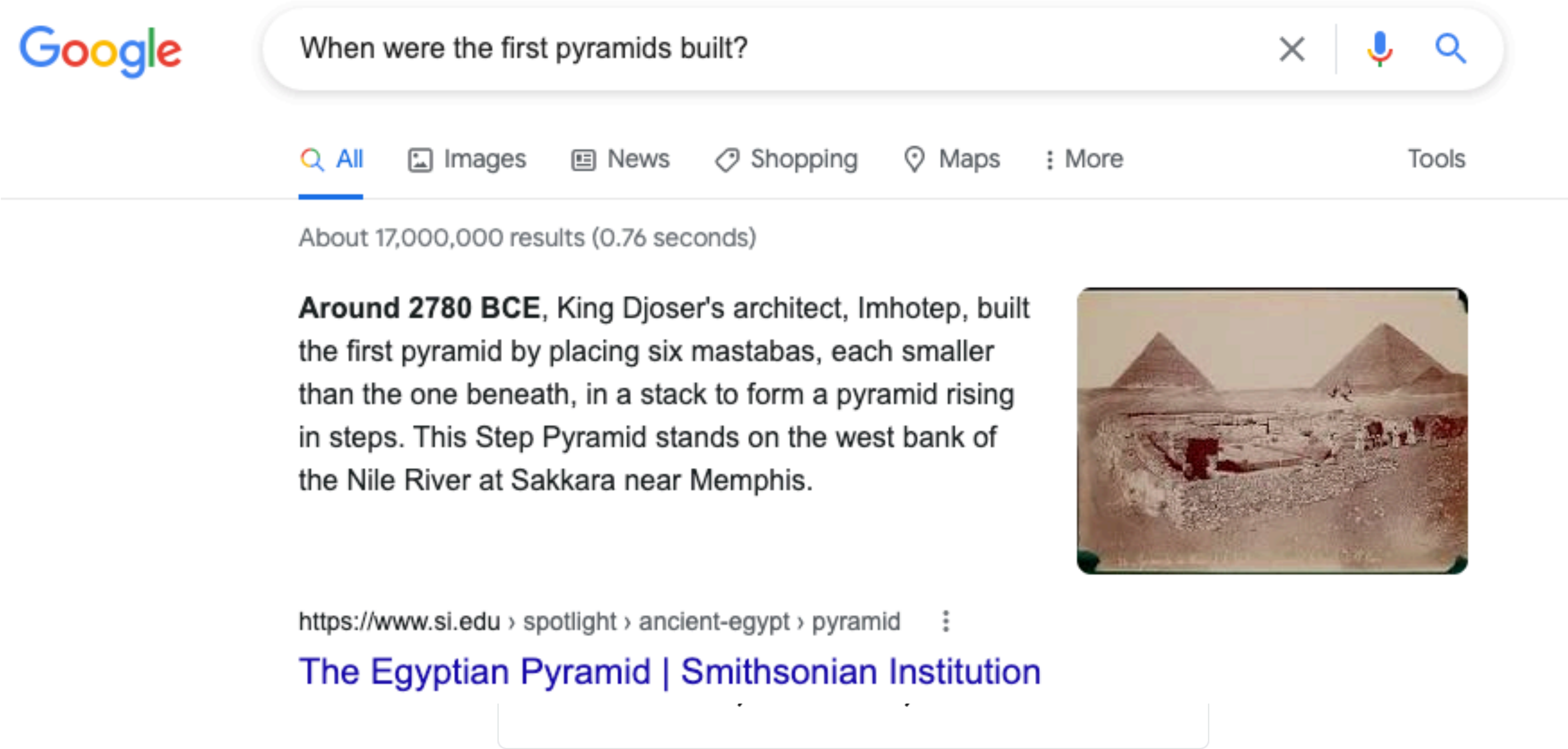
Question Answering

- Goal: build computer systems to answer questions

Question	Answer
When were the first pyramids built?	2630 BC
What's the weather like in Vancouver?	42 F
Where is Einstein's house?	112 Mercer St, Princeton, NJ 08540
Why do we yawn?	When we're bored or tired we don't breathe as deeply as we normally do. This causes a drop in our blood-oxygen levels and yawning helps us counter-balance that.

Question Answering

- You can easily find these answers in google today!



When were the first pyramids built?

About 17,000,000 results (0.76 seconds)

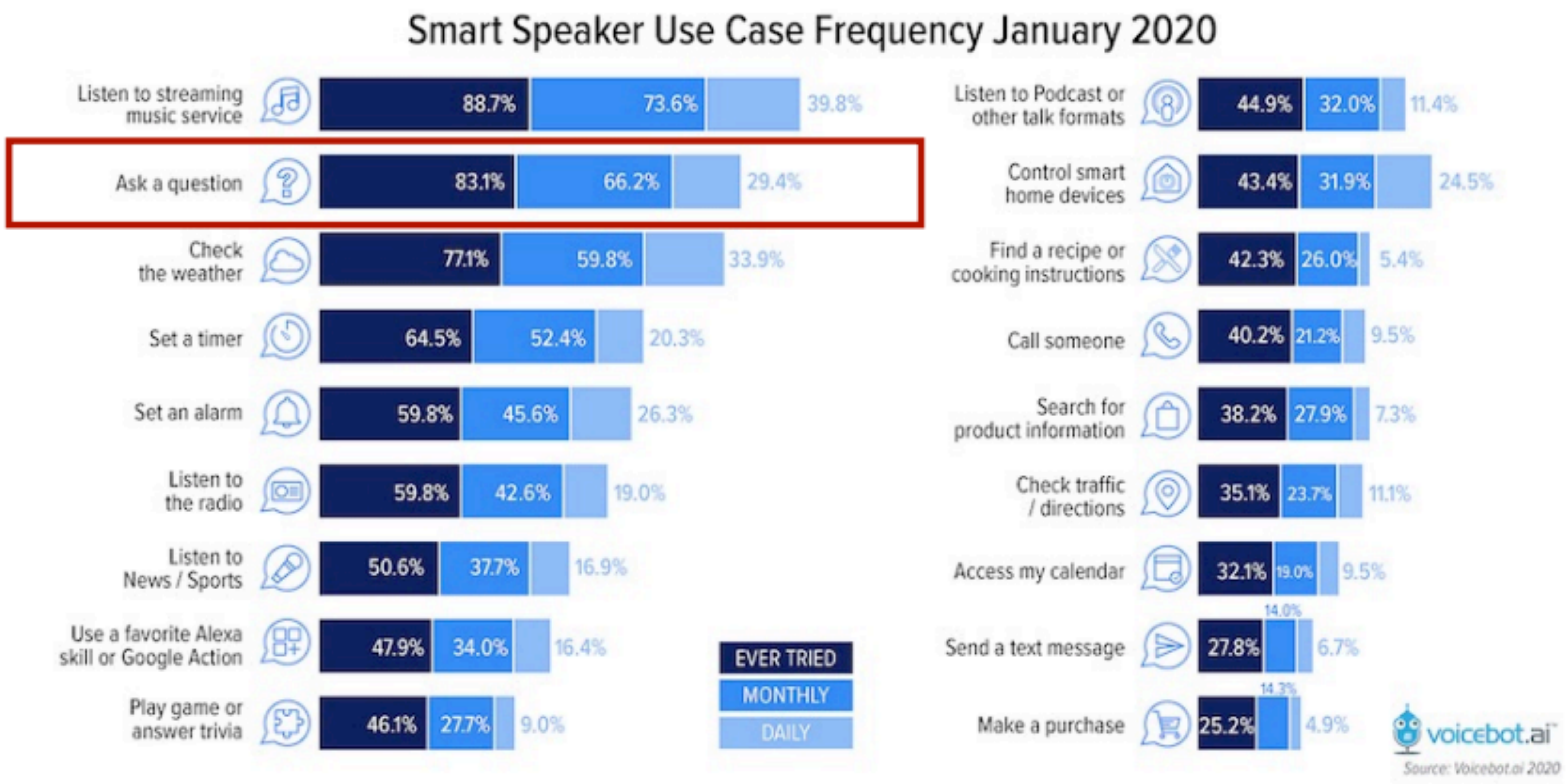
Around 2780 BCE, King Djoser's architect, Imhotep, built the first pyramid by placing six mastabas, each smaller than the one beneath, in a stack to form a pyramid rising in steps. This Step Pyramid stands on the west bank of the Nile River at Sakkara near Memphis.

https://www.si.edu › spotlight › ancient-egypt › pyramid

The Egyptian Pyramid | Smithsonian Institution

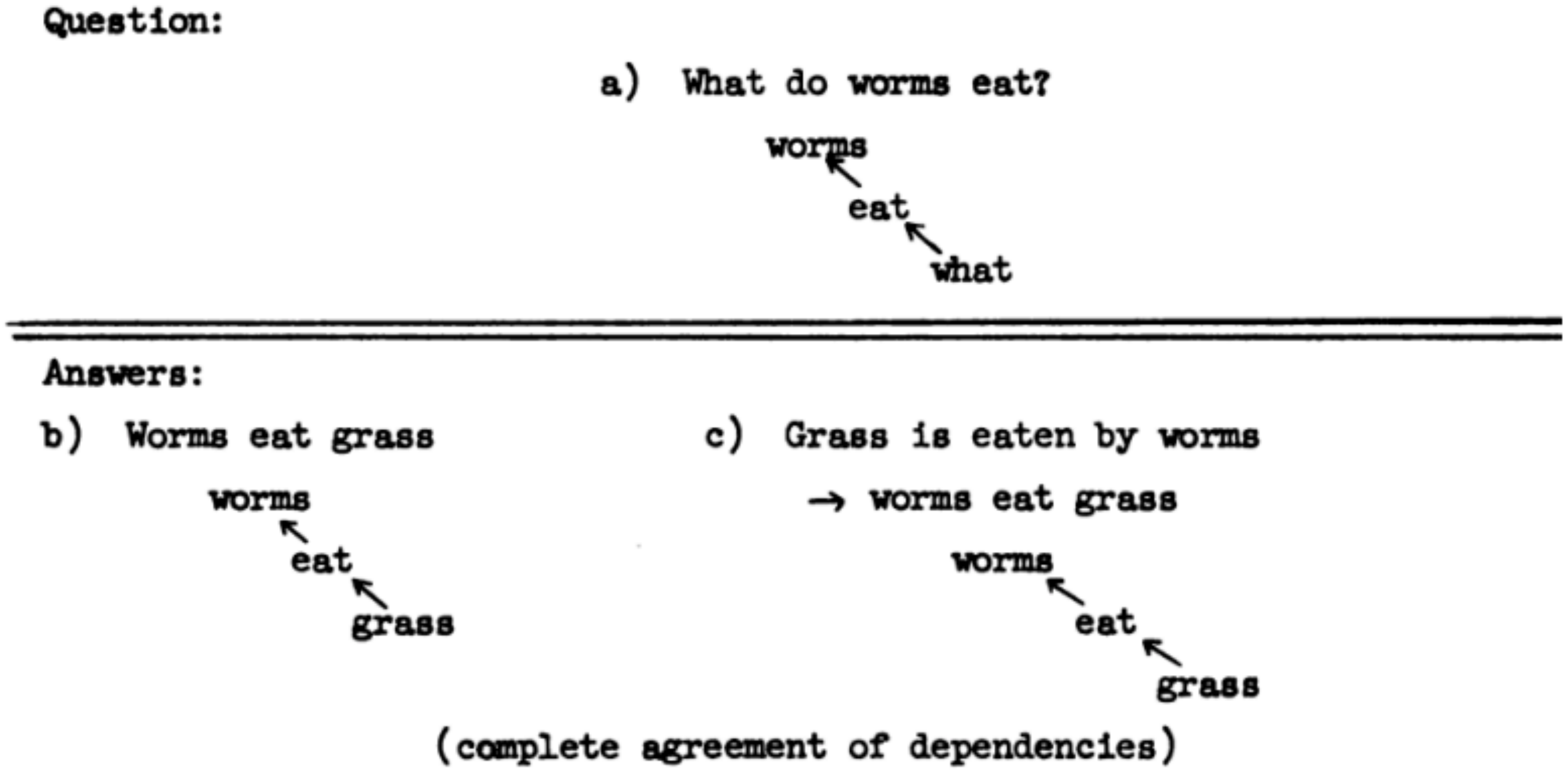
Practical application

- People ask lots of questions to Digital Personal Assistants:



Question answer has a long history

Earliest QA system dated back to the 1960s!



Question:

a) What do worms eat?

worms
↖
eat
↖
what

Answers:

b) Worms eat grass

worms
↖
eat
↖
grass

c) Grass is eaten by worms

→ worms eat grass

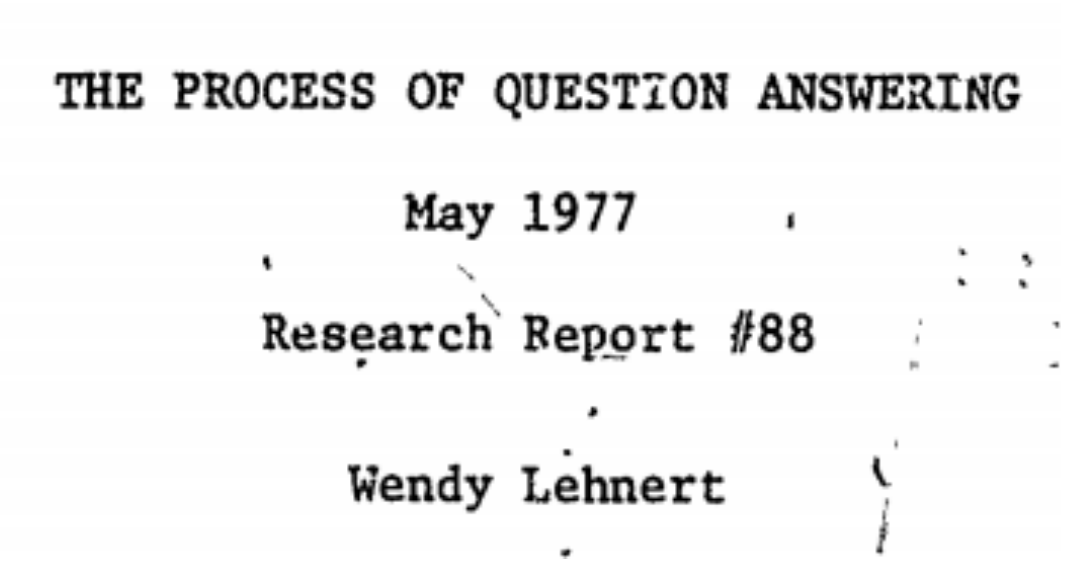
worms
↖
eat
↖
grass

(complete agreement of dependencies)

Indexing and dependency logic for answering english questions
(Simmons et al, 1964)

Why care about question answering?

- Lots of immediate applications: search engines, dialogue systems
- Question answering is an important testbed for evaluating how well compute systems understand human language

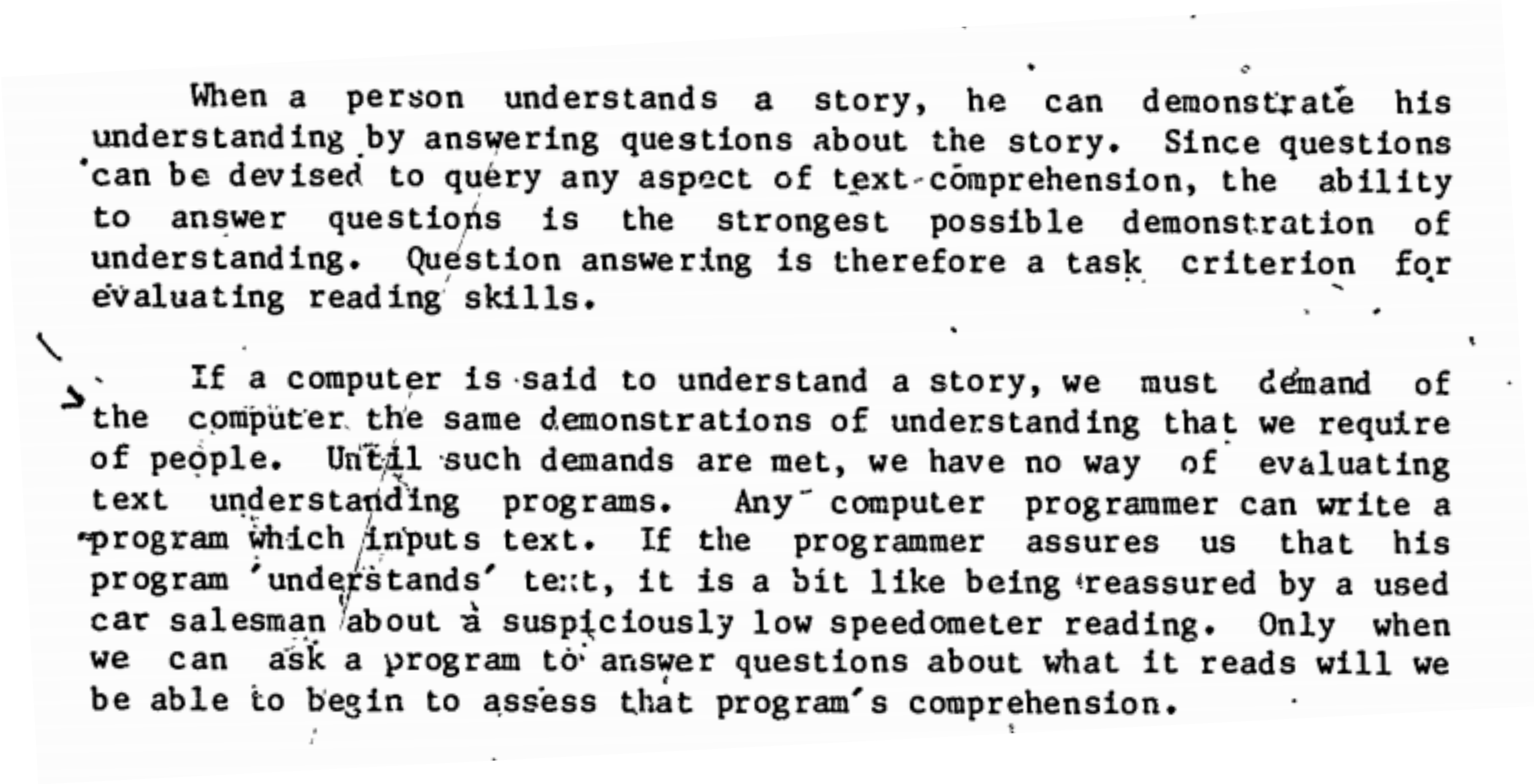


THE PROCESS OF QUESTION ANSWERING

May 1977

Research Report #88

Wendy Lehnert



When a person understands a story, he can demonstrate his understanding by answering questions about the story. Since questions can be devised to query any aspect of text comprehension, the ability to answer questions is the strongest possible demonstration of understanding. Question answering is therefore a task criterion for evaluating reading skills.

If a computer is said to understand a story, we must demand of the computer the same demonstrations of understanding that we require of people. Until such demands are met, we have no way of evaluating text understanding programs. Any computer programmer can write a program which inputs text. If the programmer assures us that his program 'understands' text, it is a bit like being reassured by a used car salesman about a suspiciously low speedometer reading. Only when we can ask a program to answer questions about what it reads will we be able to begin to assess that program's comprehension.

"Since questions can be devised to query **any aspect** of text comprehension, the ability to answer questions is the **strongest possible demonstration of understanding**."

IBM Watson beats Jeopardy Champions

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

IBM Watson beats Jeopardy Champions

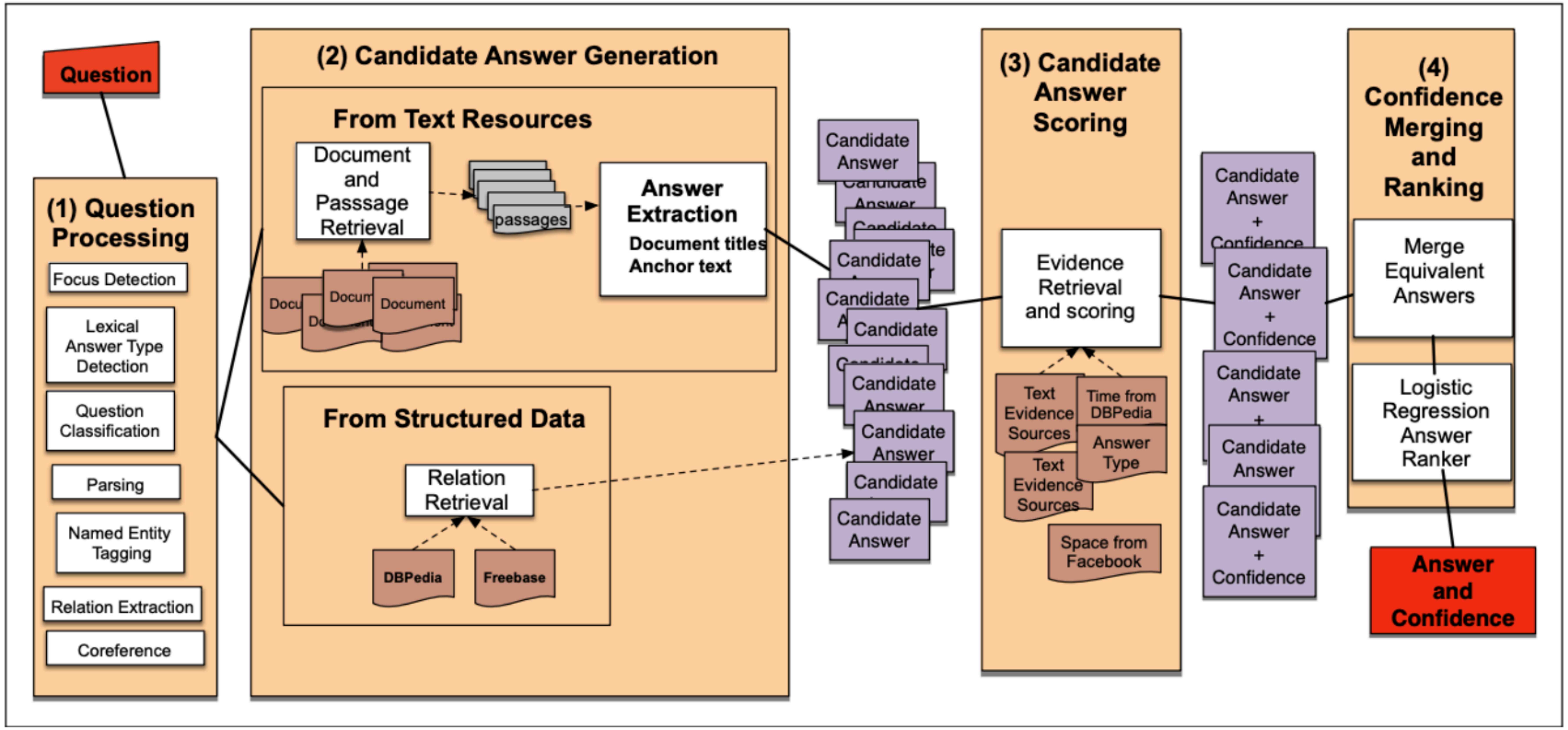


Image credit: J & M, edition 3

(1) Question processing, (2) Candidate answer generation, (3) Candidate answer scoring, and (4) Confidence merging and ranking.

QA Taxonomy

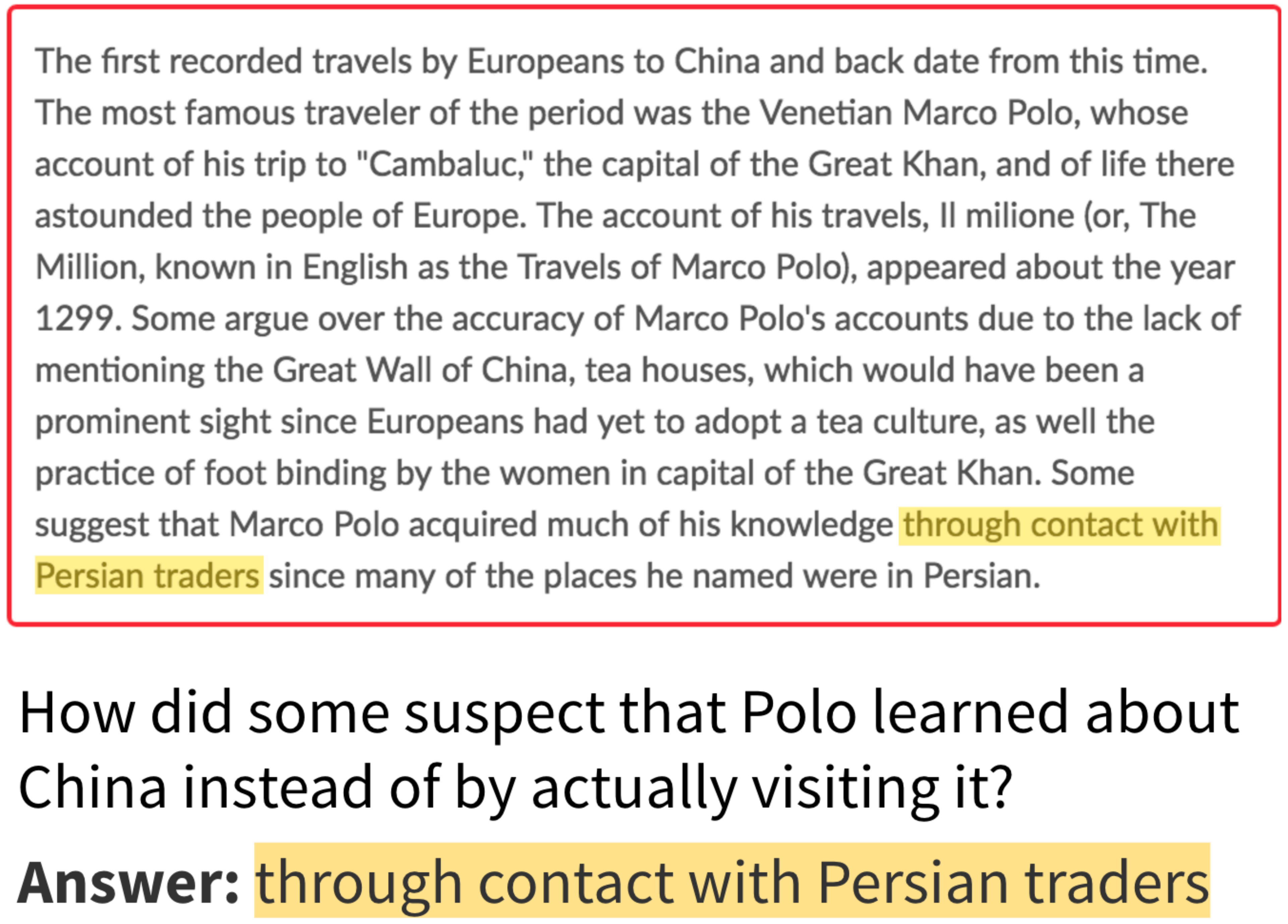
- Context (and available information sources)
  - A passage, a document, a large collection of documents, all web documents
  - Knowledge base
  - Semi-structured tables
  - Images

- Question type
  - Factoid vs non-factoid
  - Open-domain vs closed-domain
  - Simple vs compositional

- Answer type
  - A short span of text
  - A paragraph
  - Yes/No
  - A database entry
  - A list

Textual Question Answering

Also called “Reading Comprehension”



The first recorded travels by Europeans to China and back date from this time. The most famous traveler of the period was the Venetian Marco Polo, whose account of his trip to "Cambaluc," the capital of the Great Khan, and of life there astounded the people of Europe. The account of his travels, Il milione (or, The Million, known in English as the Travels of Marco Polo), appeared about the year 1299. Some argue over the accuracy of Marco Polo's accounts due to the lack of mentioning the Great Wall of China, tea houses, which would have been a prominent sight since Europeans had yet to adopt a tea culture, as well the practice of foot binding by the women in capital of the Great Khan. Some suggest that Marco Polo acquired much of his knowledge through contact with Persian traders since many of the places he named were in Persian.

How did some suspect that Polo learned about China instead of by actually visiting it?

Answer: through contact with Persian traders

(Rajpurkar et al, 2016): SQuAD: 100,000+ Questions for Machine Comprehension of Text

Textual Question Answering

James the Turtle was always getting in trouble. Sometimes he'd reach into the freezer and empty out all the food. Other times he'd sled on the deck and get a splinter. His aunt Jane tried as hard as she could to keep him out of trouble, but he was sneaky and got into lots of trouble behind her back.

One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.

His aunt was waiting for him in his room. She told James that she loved him, but he would have to start acting like a well-behaved turtle.

After about a month, and after getting into lots of trouble, James finally made up his mind to be a better turtle.

1) What is the name of the trouble making turtle?

A) Fries

B) Pudding

C) James

D) Jane

2) What did James pull off of the shelves in the grocery store?

A) pudding

B) fries

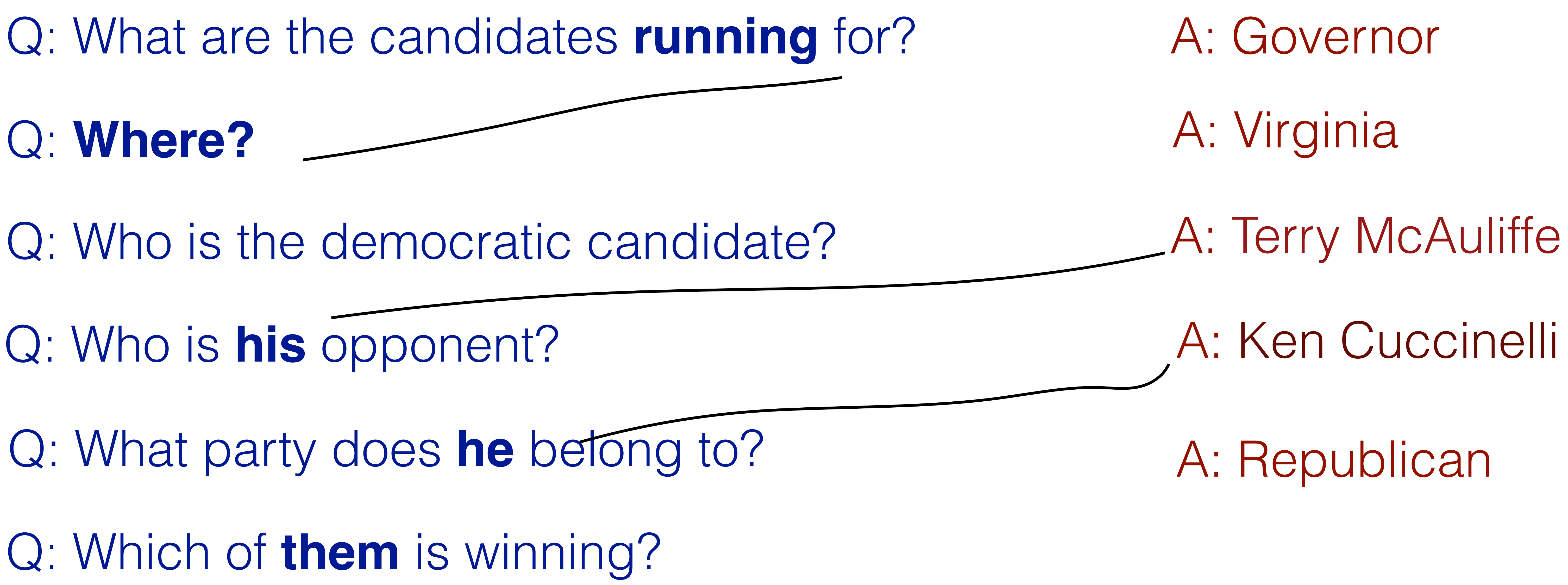
C) food

D) splinters

(Richardson et al, 2013): MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text

Conversational Question Answering

The Virginia governor's race, billed as the marquee battle of an otherwise anticlimactic 2013 election cycle, is shaping up to be a foregone conclusion. Democrat Terry McAuliffe, the longtime political fixer and moneyman, hasn't trailed in a poll since May. Barring a political miracle, Republican Ken Cuccinelli will be delivering a concession speech on Tuesday evening in Richmond. In recent ...



Q: What are the candidates **running** for? A: Governor

Q: **Where?** A: Virginia

Q: Who is the democratic candidate? A: Terry McAuliffe

Q: Who is **his** opponent? A: Ken Cuccinelli

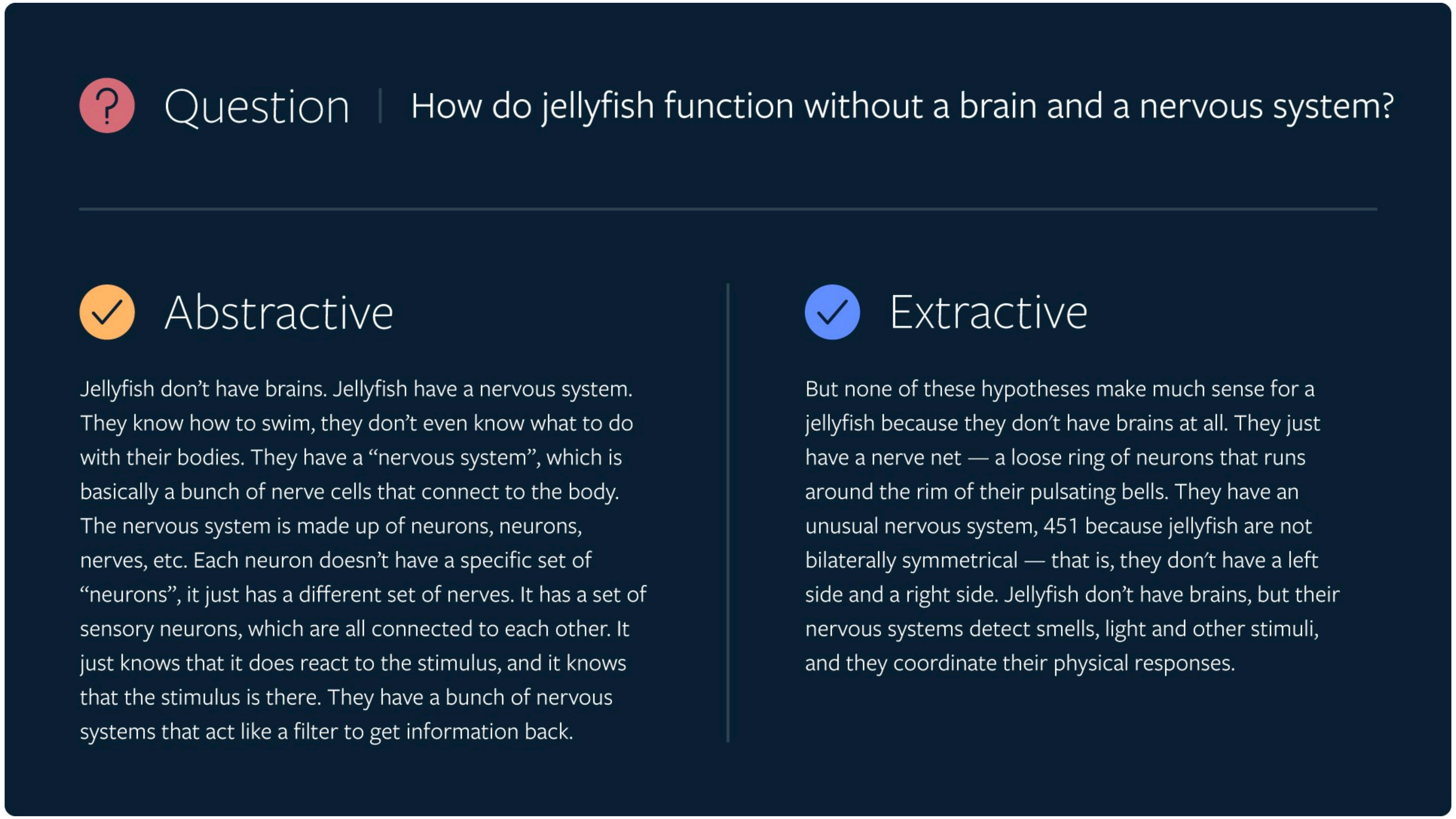
Q: What party does **he** belong to? A: Republican

Q: Which of **them** is winning?

(Reddy & Chen et al, 2019): CoQA: A Conversational Question Answering Challenge

Long-form Question Answering

Abstractive:
Answer made up of novel words and sentences composed through paraphrasing



Question | How do jellyfish function without a brain and a nervous system?

Abstractive

Jellyfish don't have brains. Jellyfish have a nervous system. They know how to swim, they don't even know what to do with their bodies. They have a "nervous system", which is basically a bunch of nerve cells that connect to the body. The nervous system is made up of neurons, neurons, nerves, etc. Each neuron doesn't have a specific set of "neurons", it just has a different set of nerves. It has a set of sensory neurons, which are all connected to each other. It just knows that it does react to the stimulus, and it knows that the stimulus is there. They have a bunch of nervous systems that act like a filter to get information back.

Extractive

But none of these hypotheses make much sense for a jellyfish because they don't have brains at all. They just have a nerve net — a loose ring of neurons that runs around the rim of their pulsating bells. They have an unusual nervous system, 451 because jellyfish are not bilaterally symmetrical — that is, they don't have a left side and a right side. Jellyfish don't have brains, but their nervous systems detect smells, light and other stimuli, and they coordinate their physical responses.

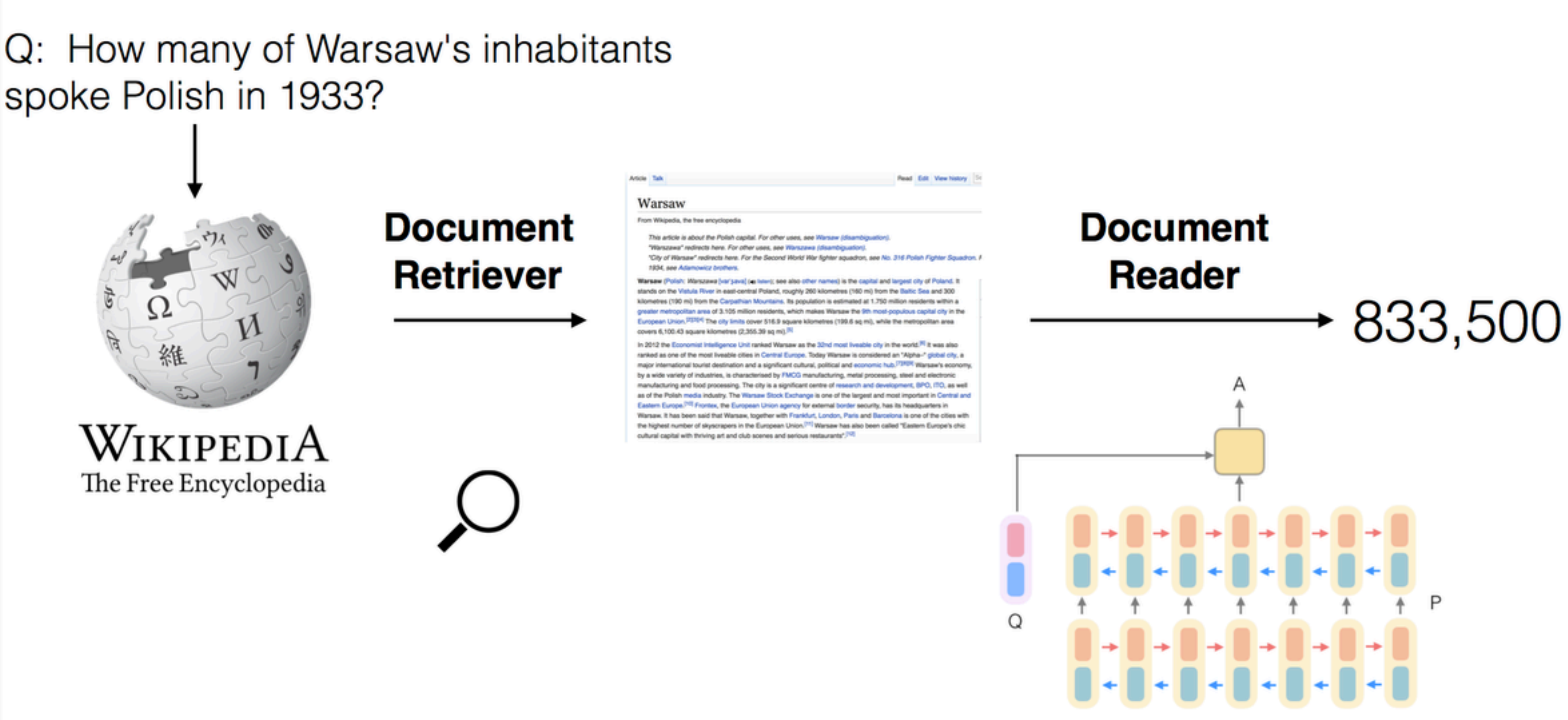
Extractive:
Select excerpts (extracts) and concatenate them to form the answer.

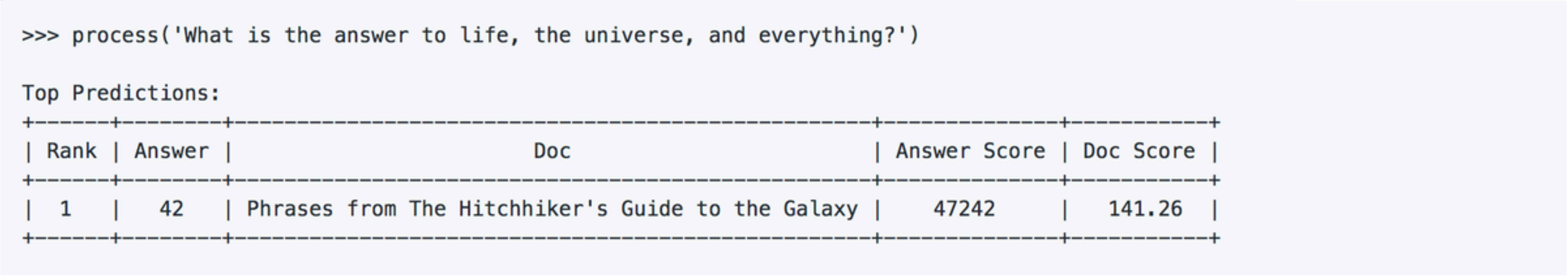
https://ai.facebook.com/blog/longform-qa/
(Fan et al, 2019): ELI5: Long Form Question Answering

Open-domain Question Answering

DrQA

- Factored into two parts:
  - Find documents that might contain an answer (handled with traditional information retrieval)
  - Finding an answer in a paragraph or a document (reading comprehension)





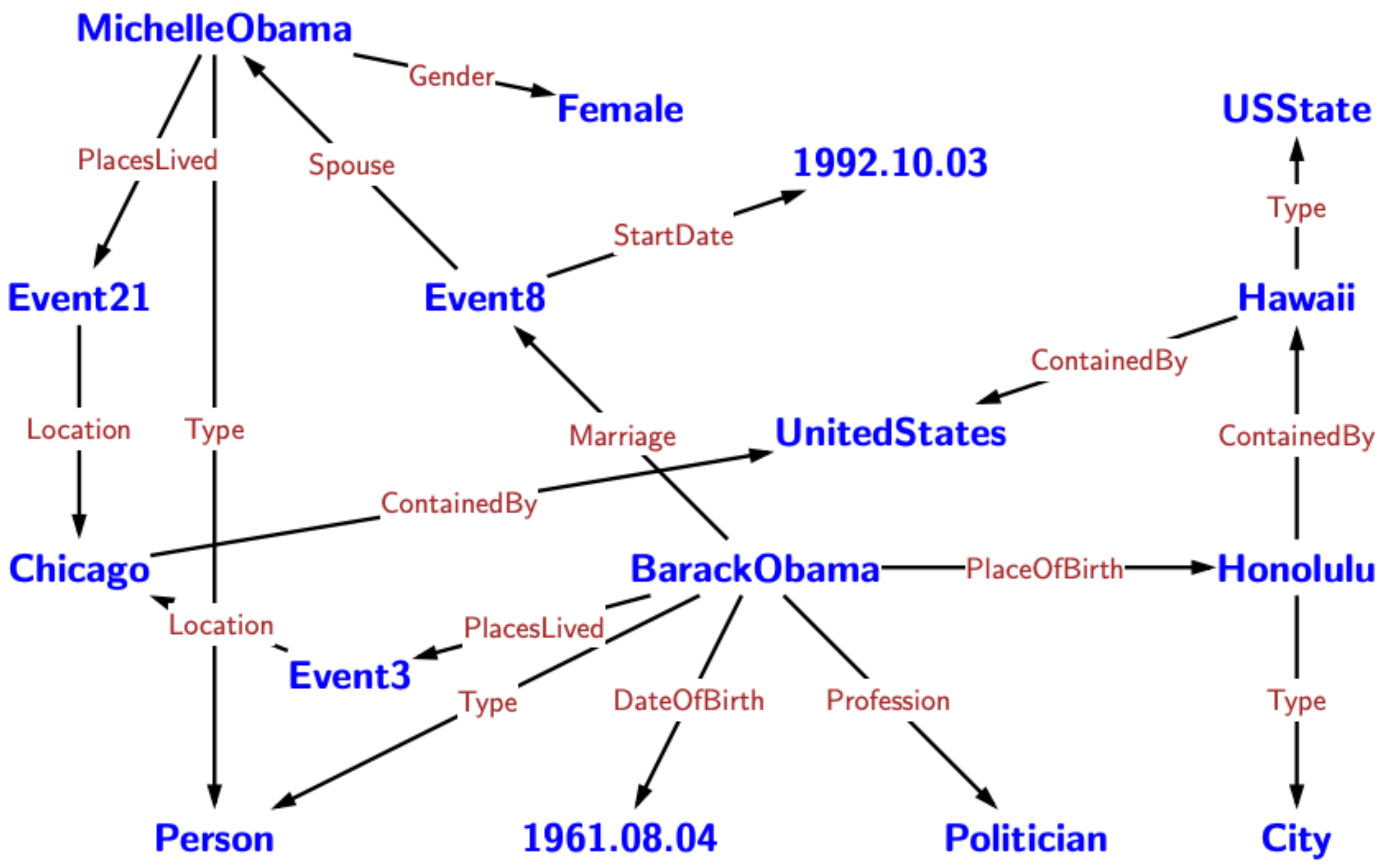
```
>>> process('What is the answer to life, the universe, and everything?')

Top Predictions:
+------+--------+---------------------------------------------+--------------+-----------+
| Rank | Answer |                     Doc                     | Answer Score | Doc Score |
+------+--------+---------------------------------------------+--------------+-----------+
|  1   |   42   | Phrases from The Hitchhiker's Guide to the Galaxy |    47242     |   141.26  |
+------+--------+---------------------------------------------+--------------+-----------+
```

(Chen et al, 2017): Reading Wikipedia to Answer Open-Domain Questions

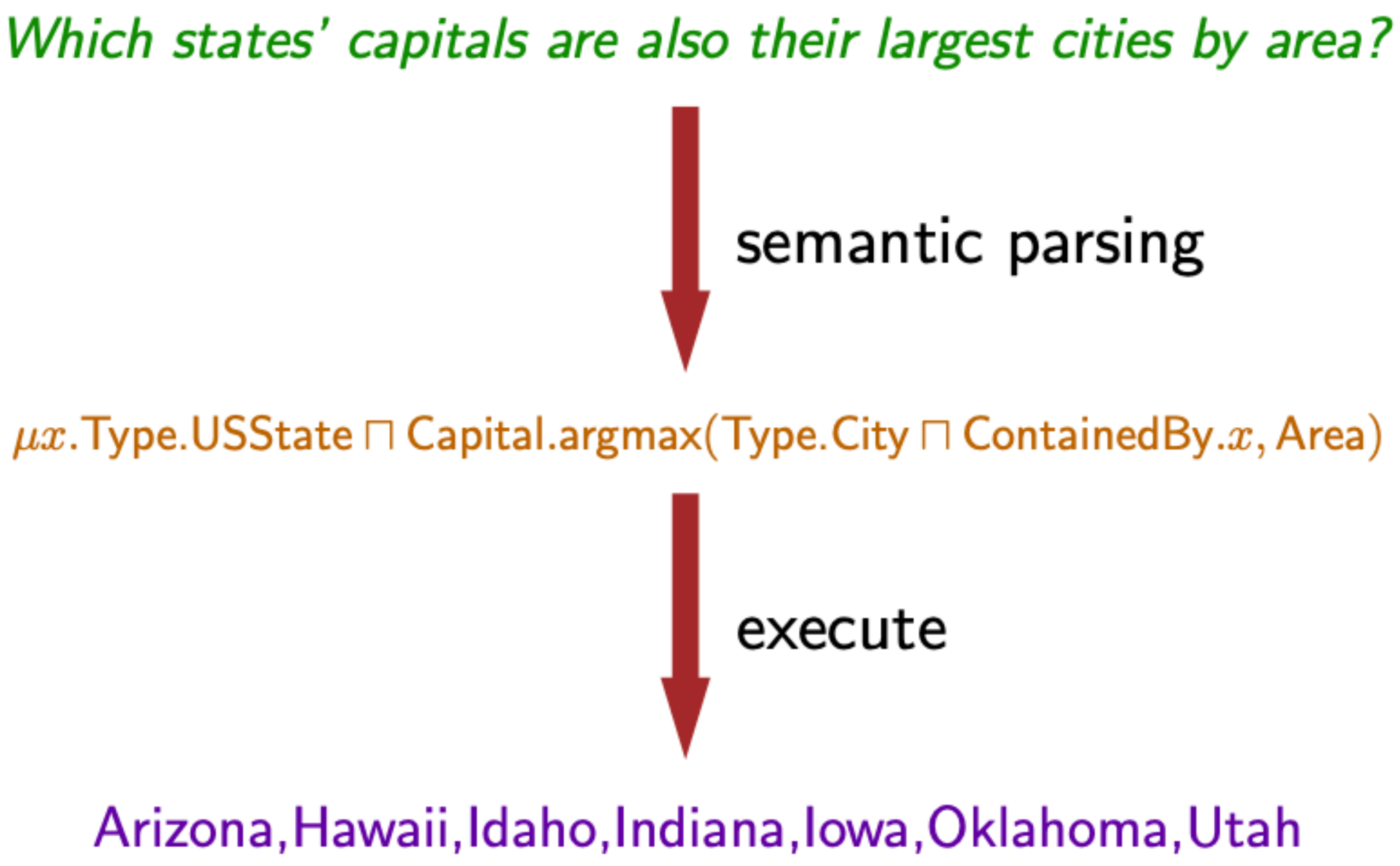
Knowledge Base Question Answering

Freebase™

100M entities (nodes) 1B assertions (edges)



Structured knowledge representation



Which states' capitals are also their largest cities by area?

semantic parsing

$\mu x.\text{Type.USState} \sqcap \text{Capital.argmax}(\text{Type.City} \sqcap \text{ContainedBy}.x, \text{Area})$

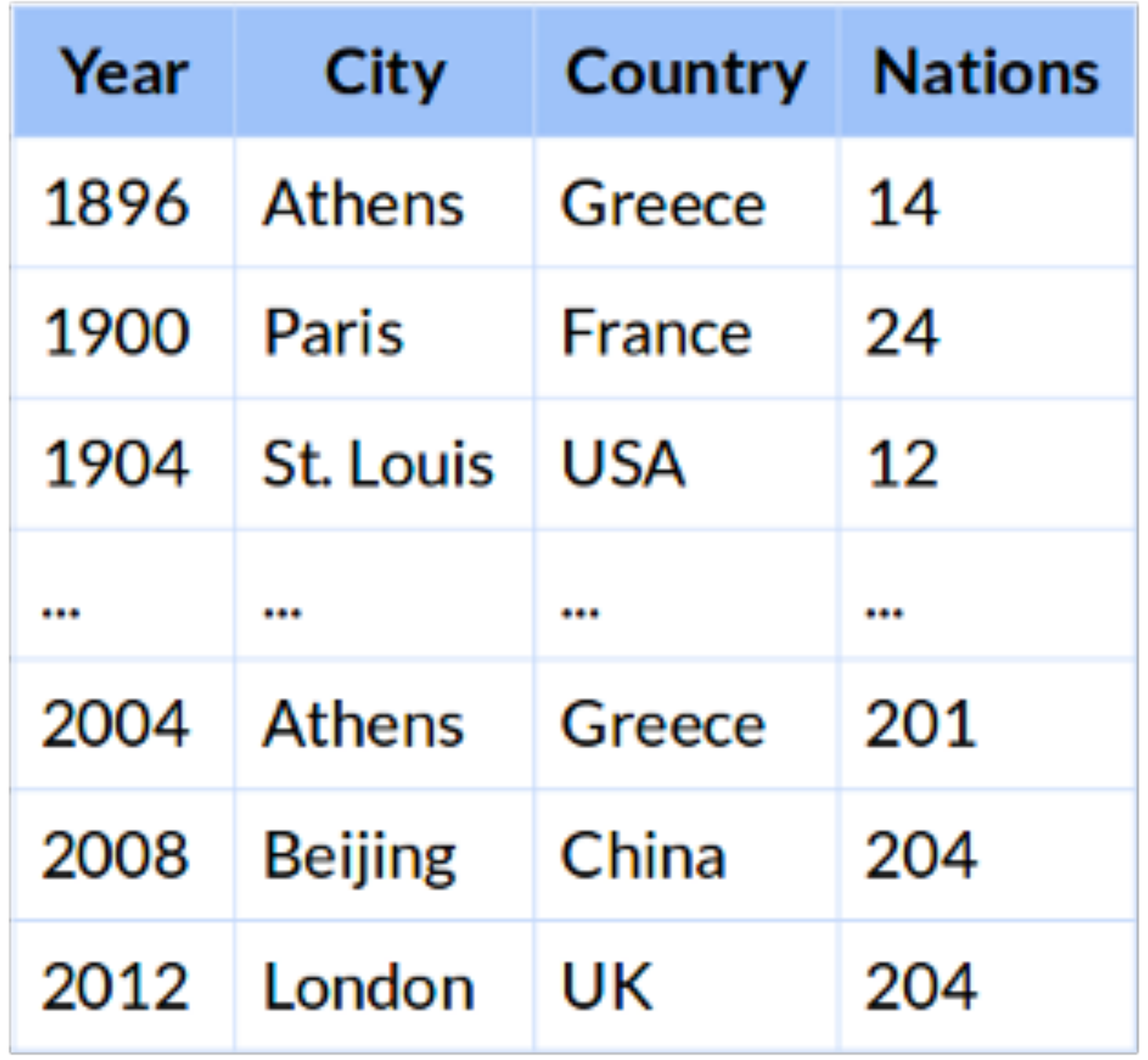
execute

Arizona,Hawaii,Idaho,Indiana,Iowa,Oklahoma,Utah

QA via semantic parsing

(Berant et al, 2013): Semantic Parsing on Freebase from Question-Answer Pairs

Table-based Question Answering



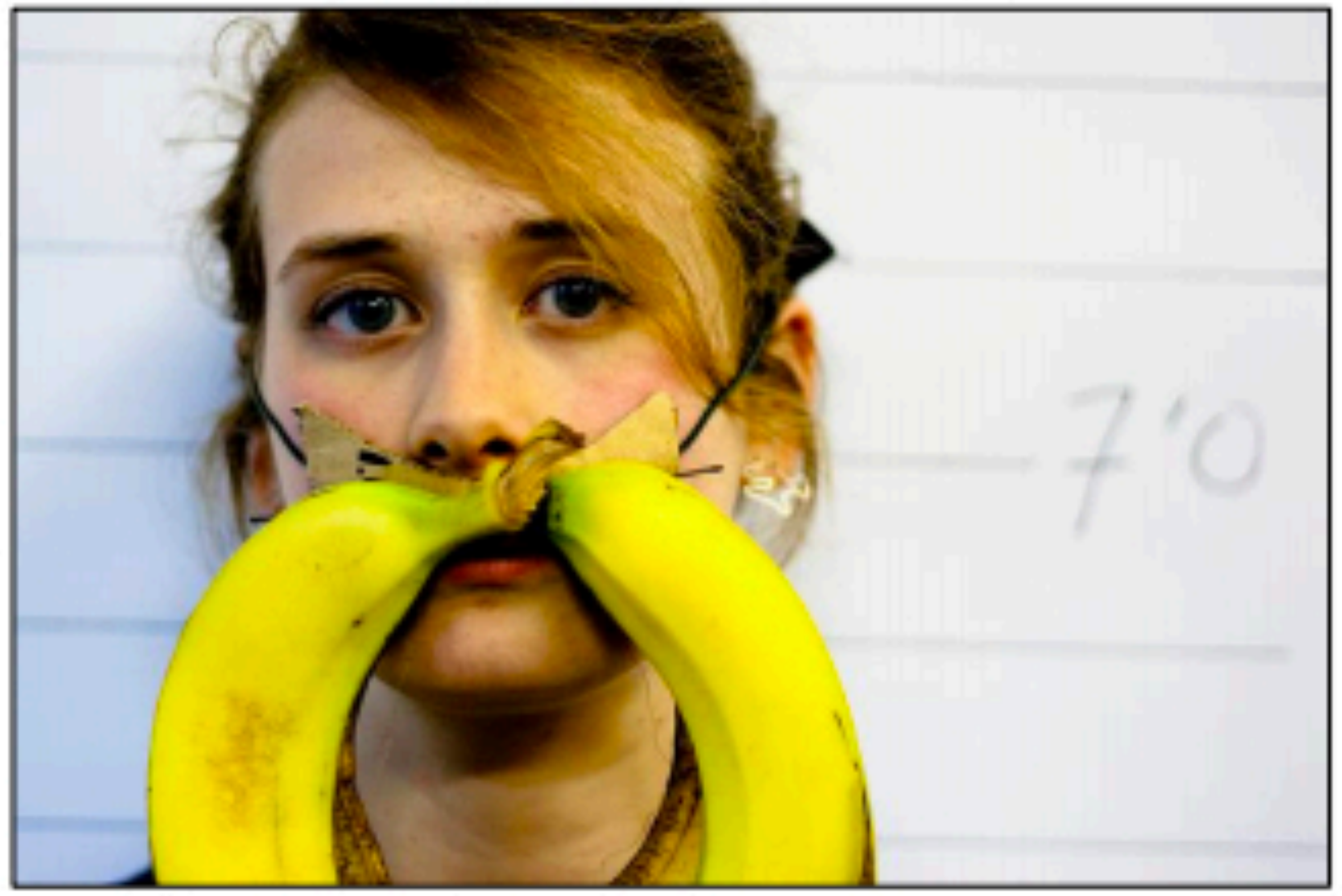
Year	City	Country	Nations
1896	Athens	Greece	14
1900	Paris	France	24
1904	St. Louis	USA	12
...	...	...	...
2004	Athens	Greece	201
2008	Beijing	China	204
2012	London	UK	204

x = Greece held its last Summer Olympics in which year?

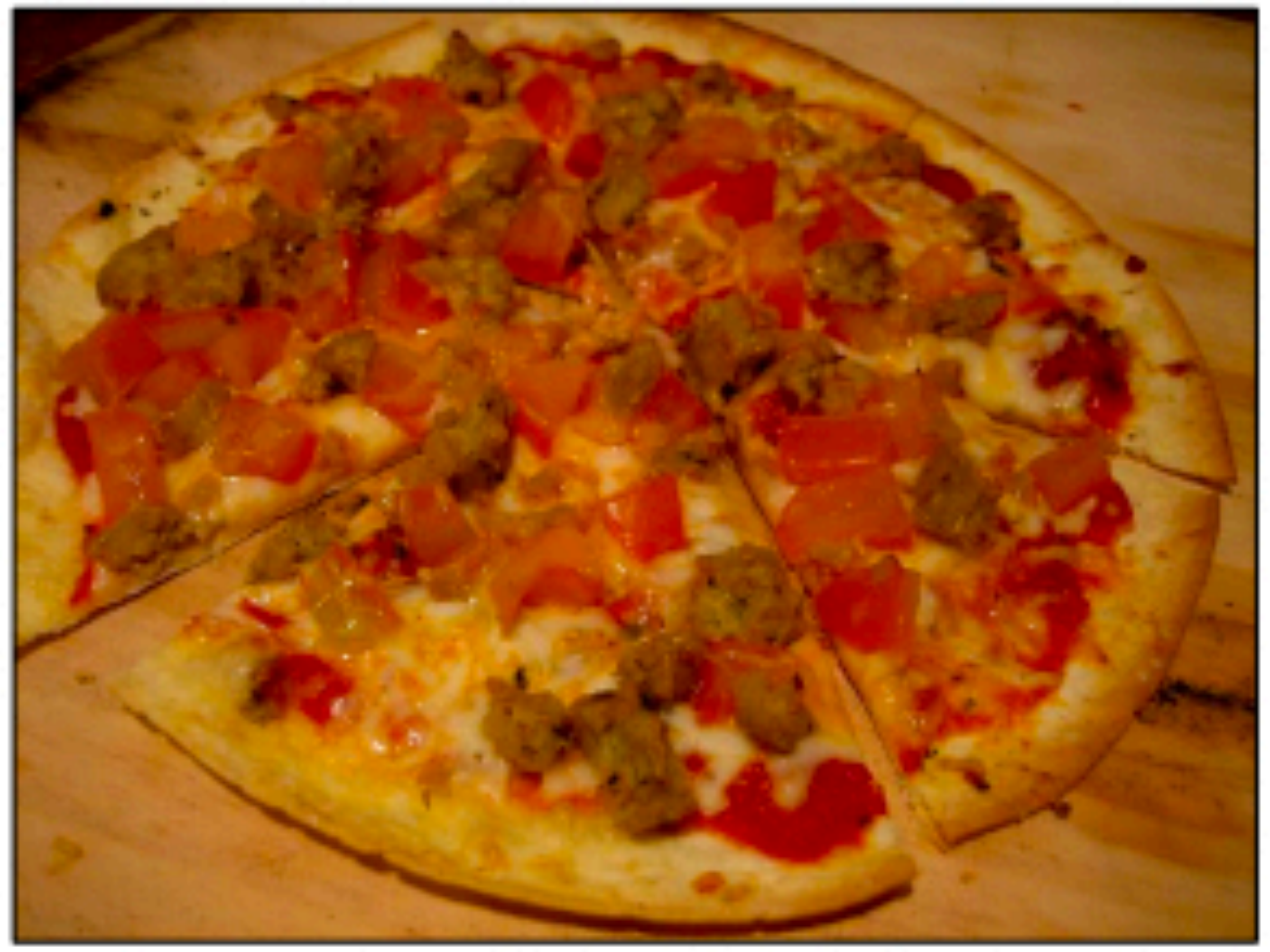
y = 2004

(Pasupat and Liang, 2015): Compositional Semantic Parsing on Semi-Structured Tables.

Visual Question Answering



What color are her eyes?
What is the mustache made of?



How many slices of pizza are there?
Is this a vegetarian pizza?

(Antol et al, 2015): Visual Question Answering

Reading Comprehension

Why do we care about this problem?

- Useful for many practical applications
- Reading comprehension is an important testbed for evaluating how well computer systems understand human language
  - Wendy Lehnert 1977: “Since questions can be devised to query any aspect of text comprehension,the ability to answer questions is the strongest possible demonstration of understanding.”
- Many other NLP tasks can be reduced to a reading comprehension problem:

Information extraction

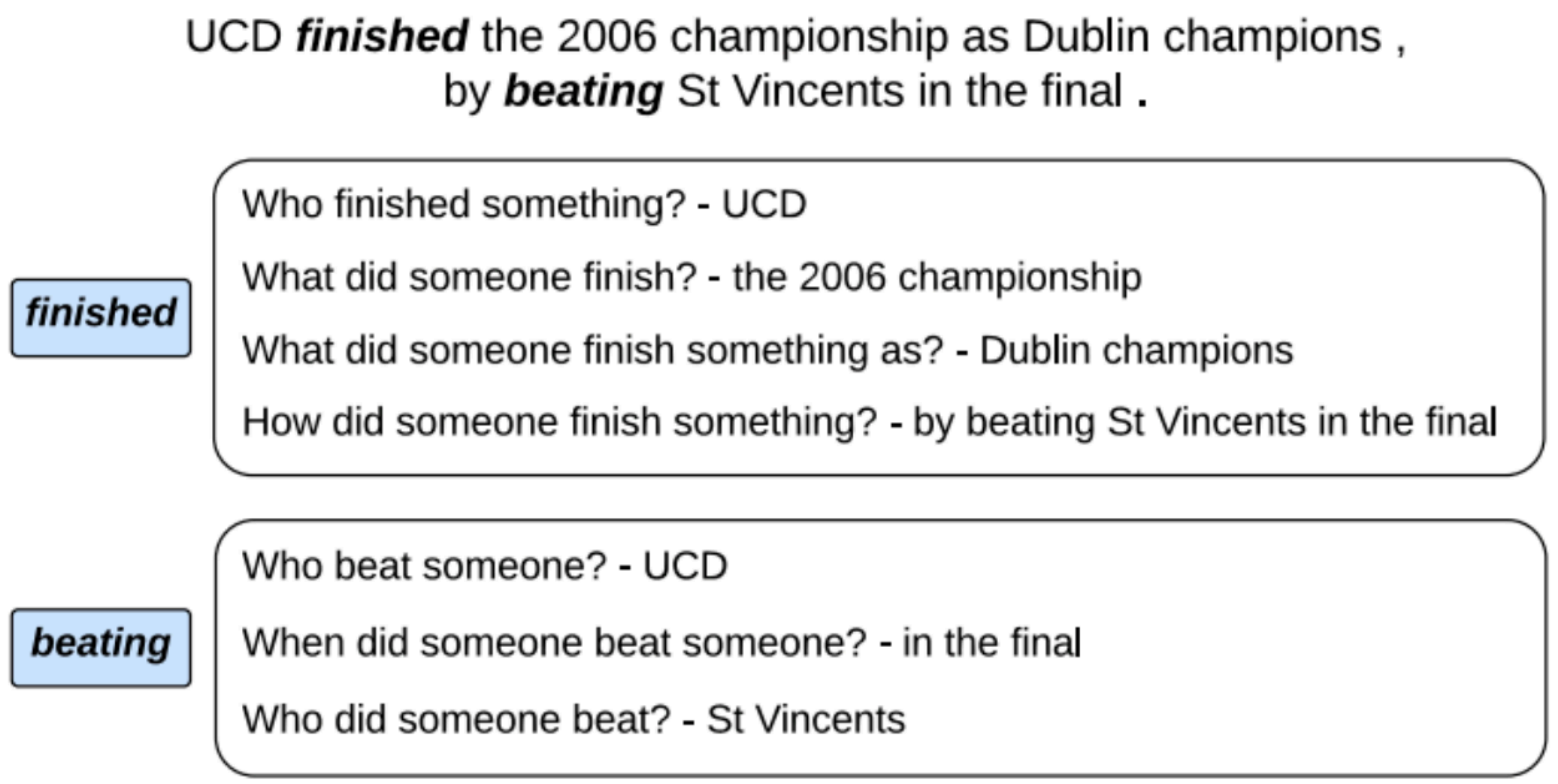
(Barack Obama, educated_at, ?)

Question: Where did Barack Obama graduate from?

Passage: Obama was born in Honolulu, Hawaii. After graduating from Columbia University in 1983, he worked as a community organizer in Chicago.

(Levy et al. 2017)

Semantic role labeling



UCD ***finished*** the 2006 championship as Dublin champions , by ***beating*** St Vincents in the final .

finished

Who finished something? - UCD

What did someone finish? - the 2006 championship

What did someone finish something as? - Dublin champions

How did someone finish something? - by beating St Vincents in the final

beating

Who beat someone? - UCD

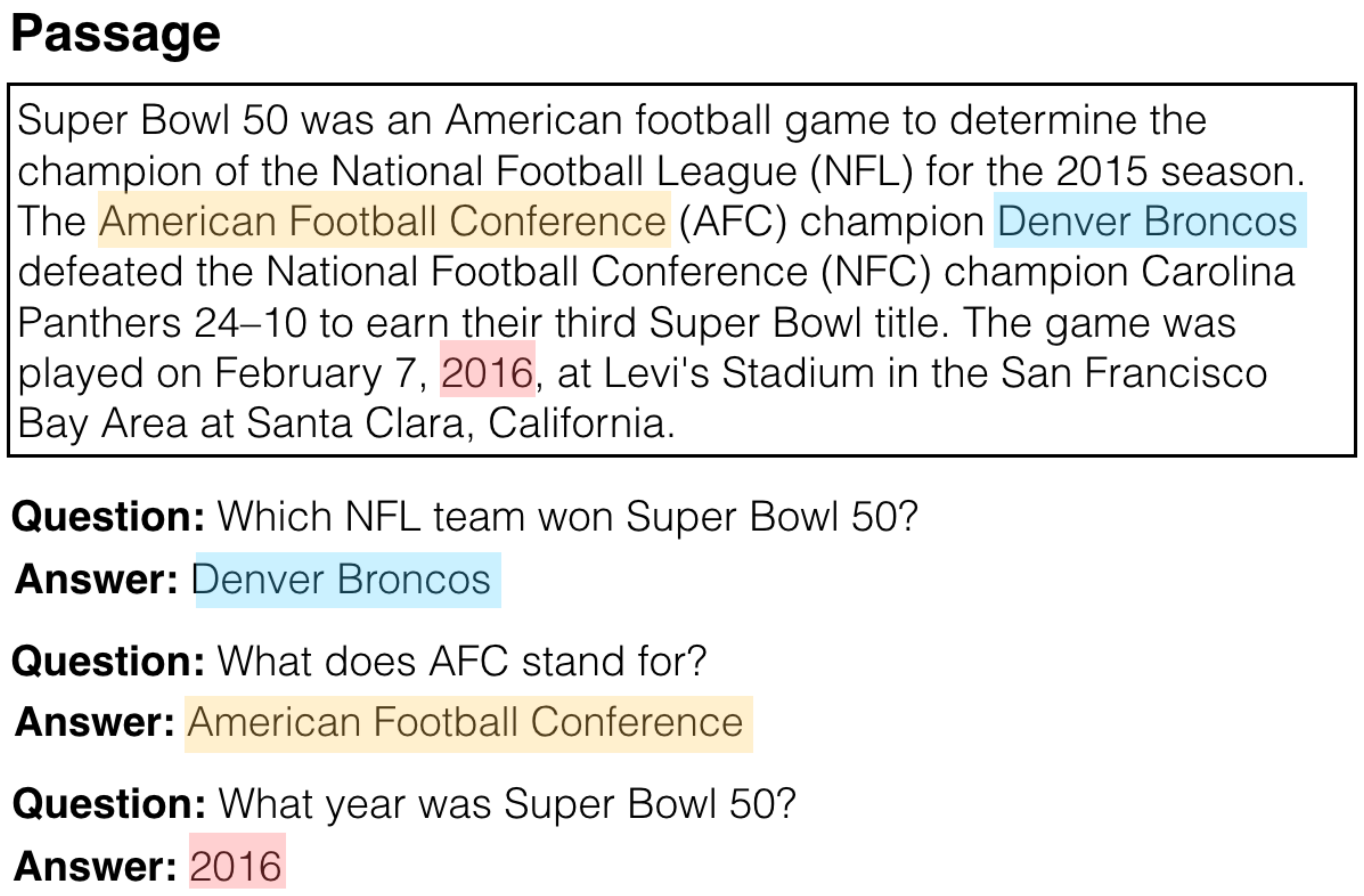
When did someone beat someone? - in the final

Who did someone beat? - St Vincents

(He et al. 2015)

Slide credit: John Hewitt

Stanford Question Answering Dataset (SQuAD)



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

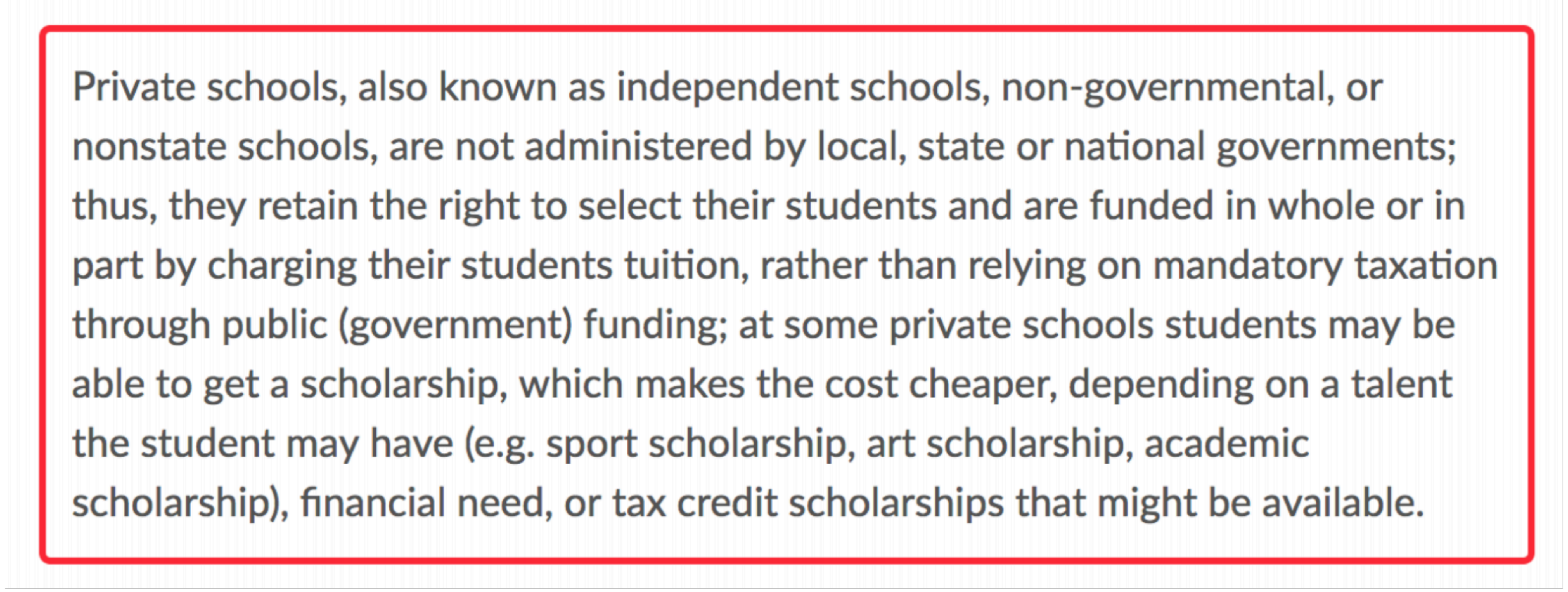
Question: What year was Super Bowl 50?
Answer: 2016

SQuAD 2.0:
Have classifier/threshold to decide whether to take the most likely prediction as answer

- (passage, question, answer) triples
- Passage is from Wikipedia (~100-500 words), question is crowd-sourced
- Answer must be a span of text in the passage (aka. "extractive question answering")
- SQuAD 1.1: 100k answerable questions, SQuAD 2.0: another 50k unanswerable questions

https://stanford-qa.com
(Rajpurkar et al, 2016): SQuAD: 100,000+ Questions for Machine Comprehension of Text

Stanford Question Answering Dataset (SQuAD)



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

3 gold answers are collected for each question

Slide credit: Chris Manning

Stanford Question Answering Dataset (SQuAD)

SQuAD 1.1 evaluation:

- Two metrics: exact match (EM) and F1
  - Exact match: 1/0 accuracy on whether you match one of the three answers
  - F1: take each gold answer and system output as bag of words, compute precision, recall and harmonic mean. Take the max of the three scores.
- Final exact match and F1 are average of instance exact and F1 scores
- Estimated human performance: EM = 82.3, F1 = 91.2

Example

Q: What did Tesla do in December 1878?

A: {left Graz, left Graz, left Graz and severed all relations with his family}

Prediction: {left Graz and served}

Exact match: max{0, 0, 0} = 0

F1: max{0.67, 0.67, 0.61} = 0.67

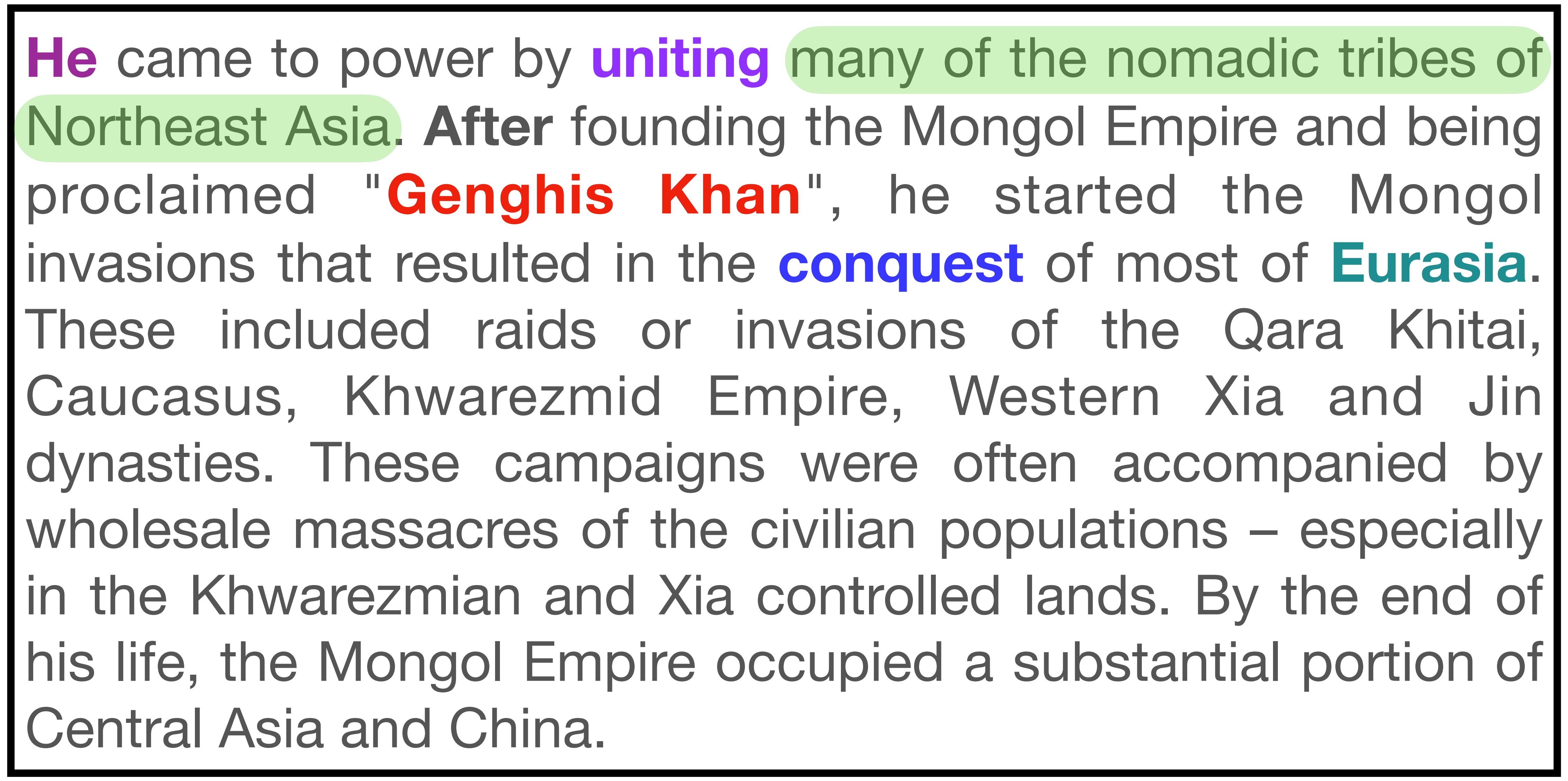
(Rajpurkar et al, 2016): SQuAD: 100,000+ Questions for Machine Comprehension of Text

Other datasets

- TriviaQA: Questions and answers by trivia enthusiasts. Independently collected web paragraphs that contain the answer and seem to discuss question, but no human verification that paragraph supports answer to question

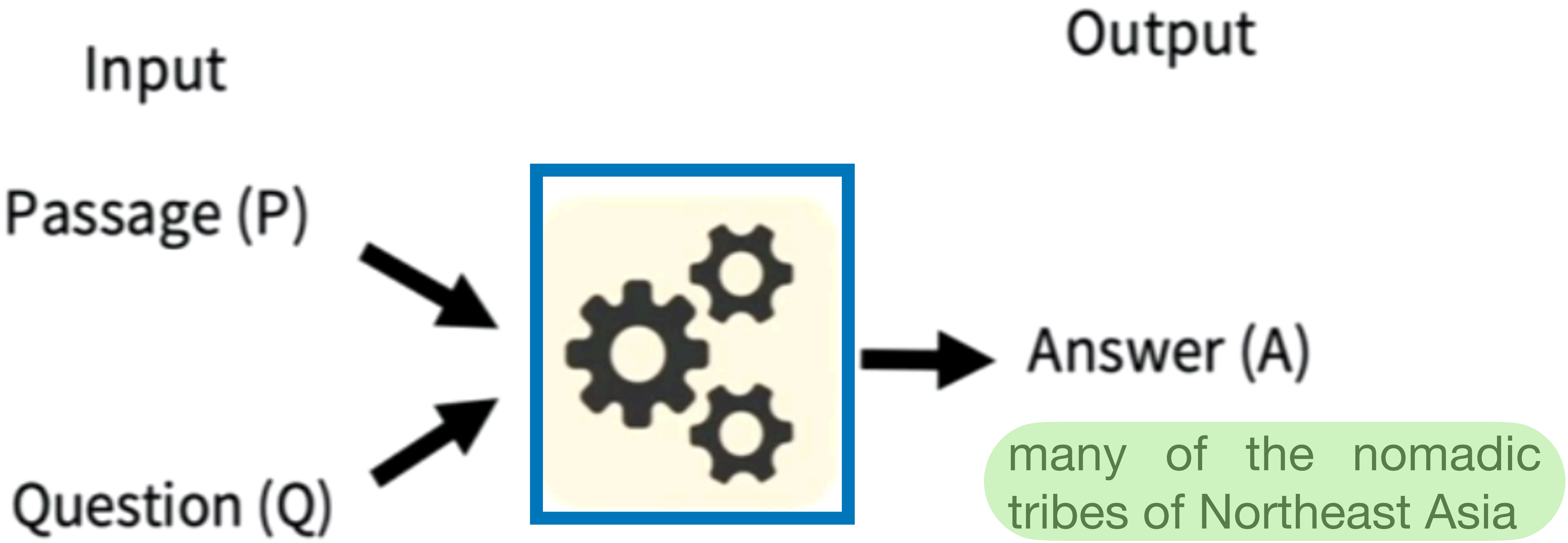
- Natural Questions: Question drawn from frequently asked Google search questions. Answers from Wikipedia paragraphs. Answer can be substring, yes, no, or NOT_PRESENT. Verified by human annotation.
- HotpotQA. Constructed questions to be answered from the whole of Wikipedia which involve getting information from two pages to answer a multistep query:
  - Q: Which novel by the author of "Armada" will be adapted as a feature film by Steven Spielberg?
  - A: Ready Player One

Models for Reading Comprehension



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.

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



Feature-based models (2016)

- Generate a list of candidate answers $\{a_1, a_2, \ldots, a_M\}$
  - Considered only the constituents in parse trees

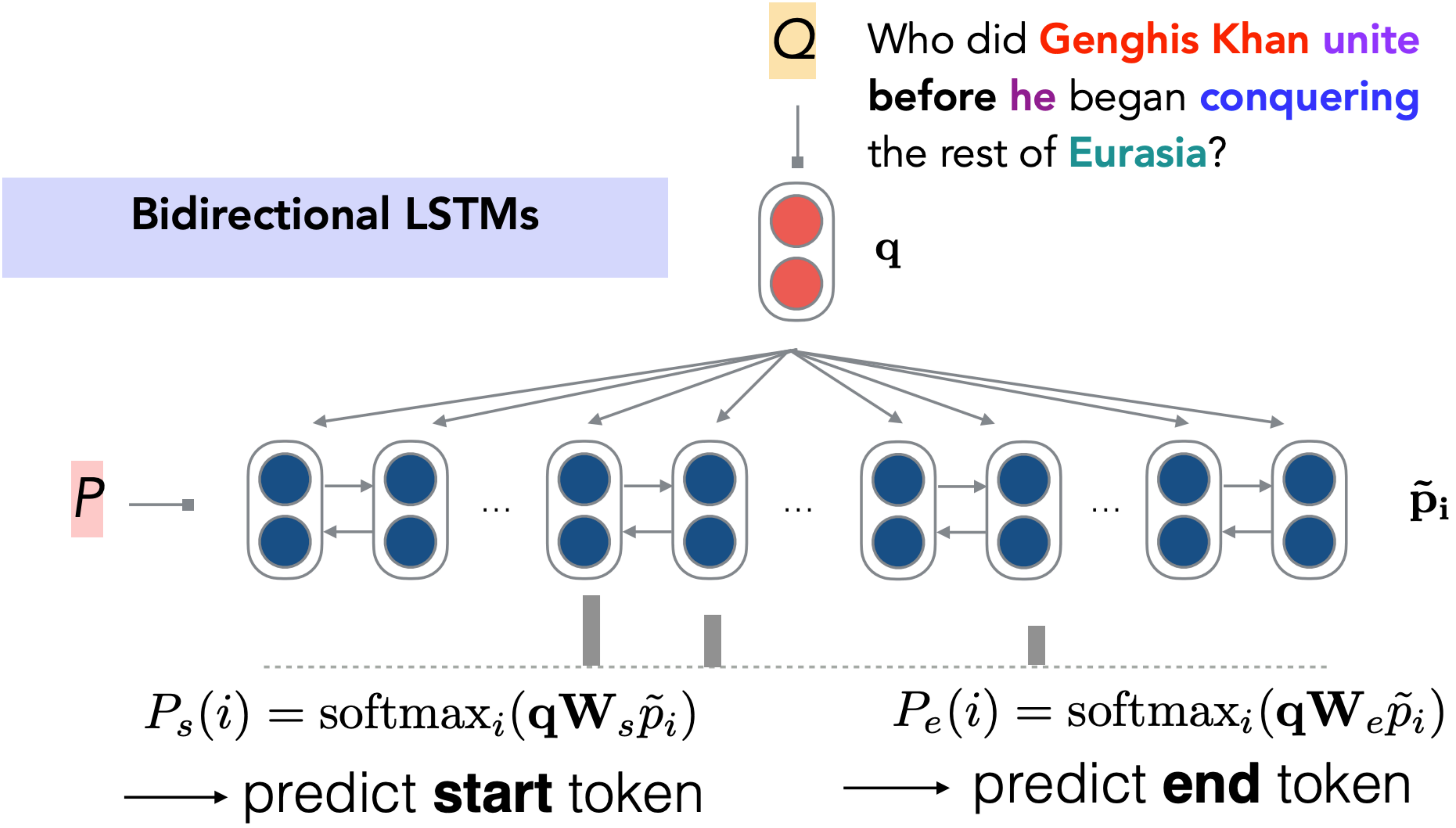
- Define a feature vector $\phi(p, q, a_i) \in \mathbb{R}^d$:
  - Word/bigram frequencies
  - Parse tree matches
  - Dependency labels, length, part-of-speech tags

- Apply a (multi-class) logistic regression model

(Rajpurkar et al, 2016): SQuAD: 100,000+ Questions for Machine Comprehension of Text

Neural models for reading comprehension (after 2016)

- LSTM-based models with attention (2016-2018)



Q
Who did Genghis Khan unite before he began conquering the rest of Eurasia?
Bidirectional LSTMs
q
P
...
$\tilde{\mathbf{p}}_i$

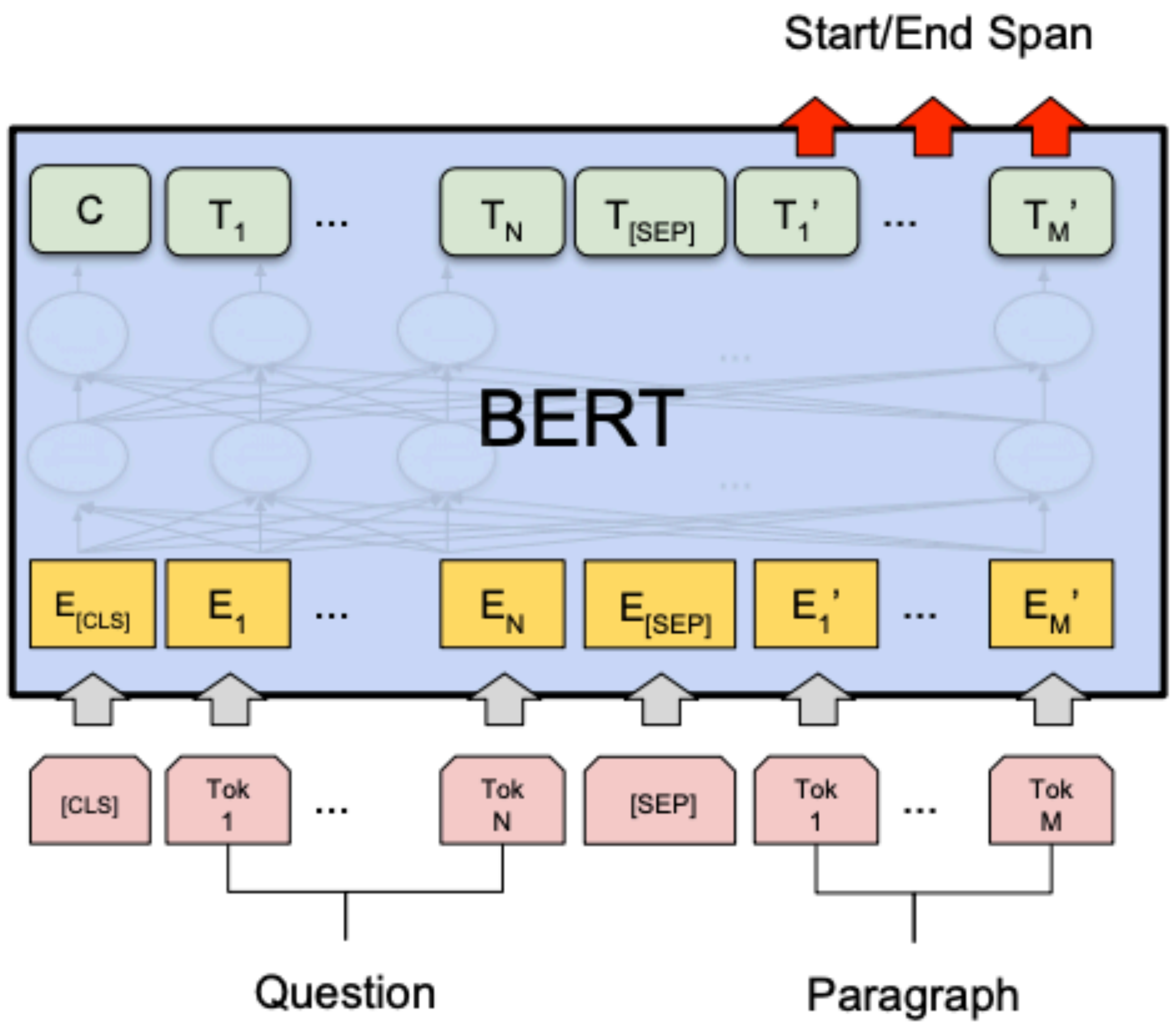
$P_s(i) = \text{softmax}_i(\mathbf{q}\mathbf{W}_s\tilde{p}_i)$ → predict **start** token

$P_e(i) = \text{softmax}_i(\mathbf{q}\mathbf{W}_e\tilde{p}_i)$ → predict **end** token

Chen et al, 2016

Attentive Reader (Hermann et al., 2015), Stanford Attentive Reader (Chen et al., 2016), Match-LSTM (Wang et al., 2017), BiDAF (Seo et al., 2017), Dynamic coattention network (Xiong et al., 2017), DrQA (Chen et al., 2017), R-Net (Wang et al., 2017), ReasoNet (Shen et al., 2017)...

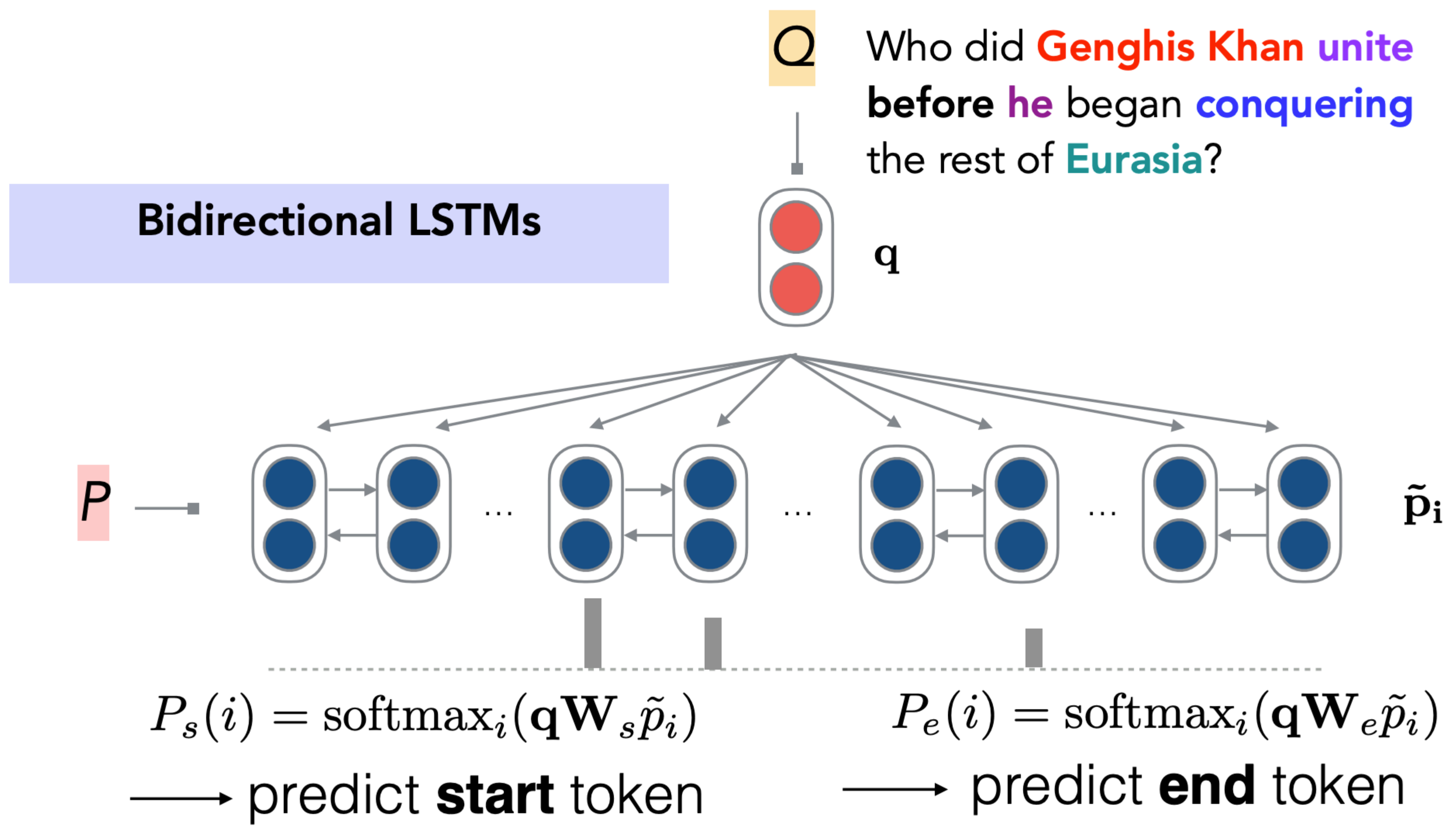
- Fine-tuning BERT-like models for reading comprehension (2019+)



Devlin et al, 2018

Stanford Attentive Reader (Chen, Bolten, and Manning, 2016)

- Simple model with good performance
- Encode the question and passage word embeddings and BiLSTM encoders
- Use attention to predict start and end span



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

Bidirectional LSTMs

$\mathbf{q}$

P

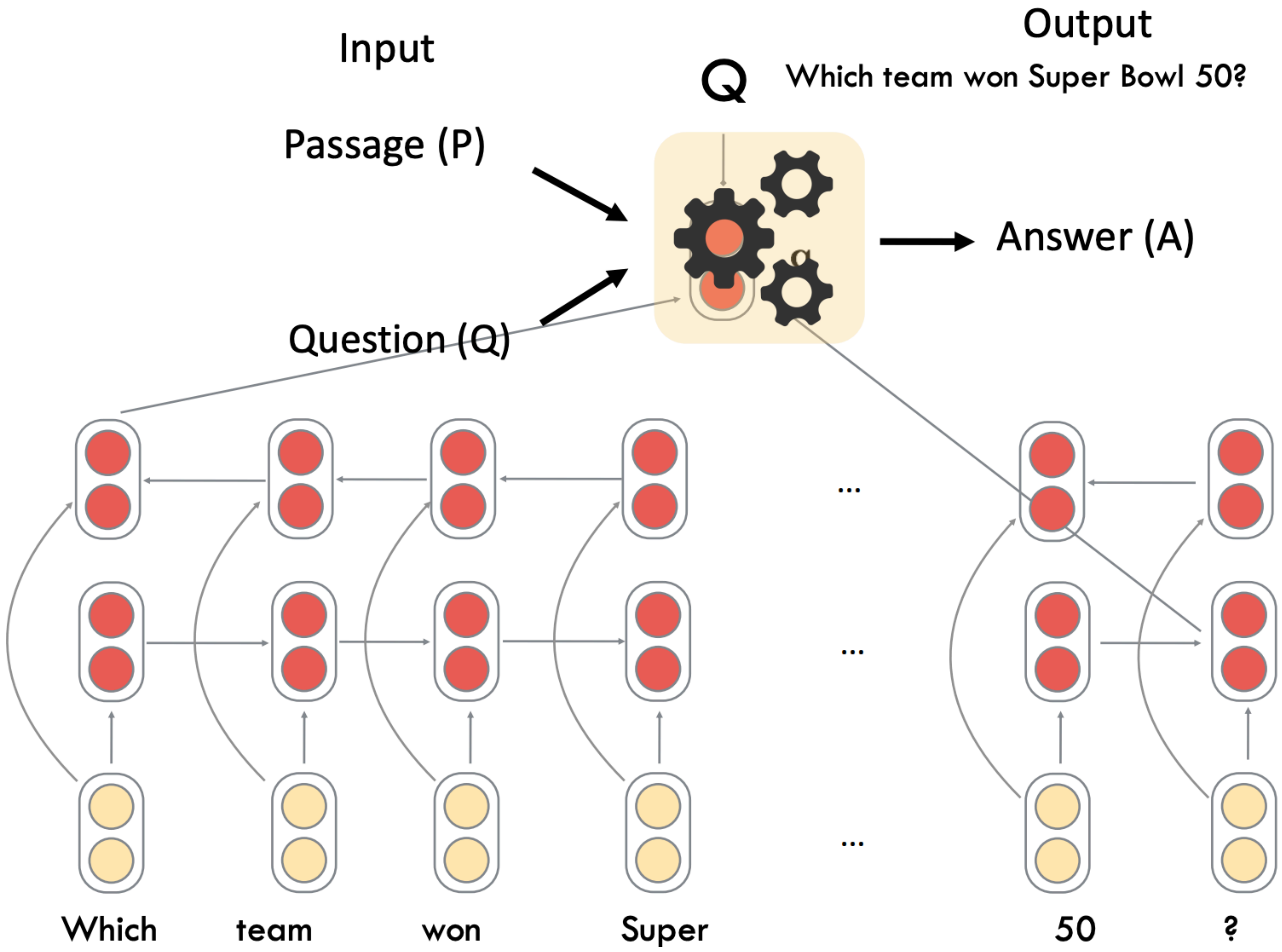
$\tilde{\mathbf{p}}_\mathbf{i}$

$P_s(i) = \text{softmax}_i(\mathbf{q}\mathbf{W}_s\tilde{p}_i)$ ⟶ predict **start** token

$P_e(i) = \text{softmax}_i(\mathbf{q}\mathbf{W}_e\tilde{p}_i)$ ⟶ predict **end** token

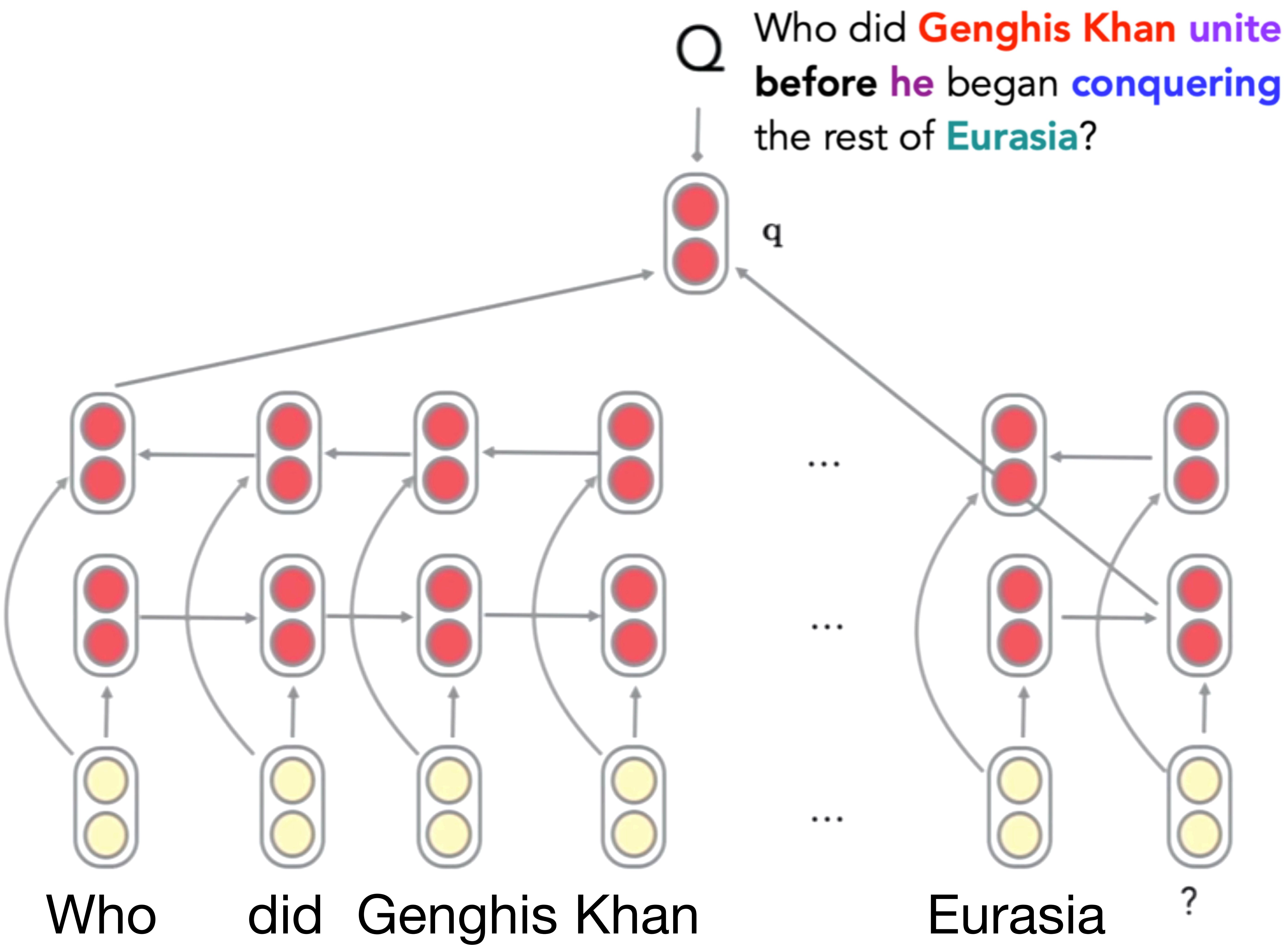
Also used in DrQA (Chen et al, 2017)

Stanford Attentive Reader



Slide credit: Chris Manning

Stanford Attentive Reader
Question Encoder



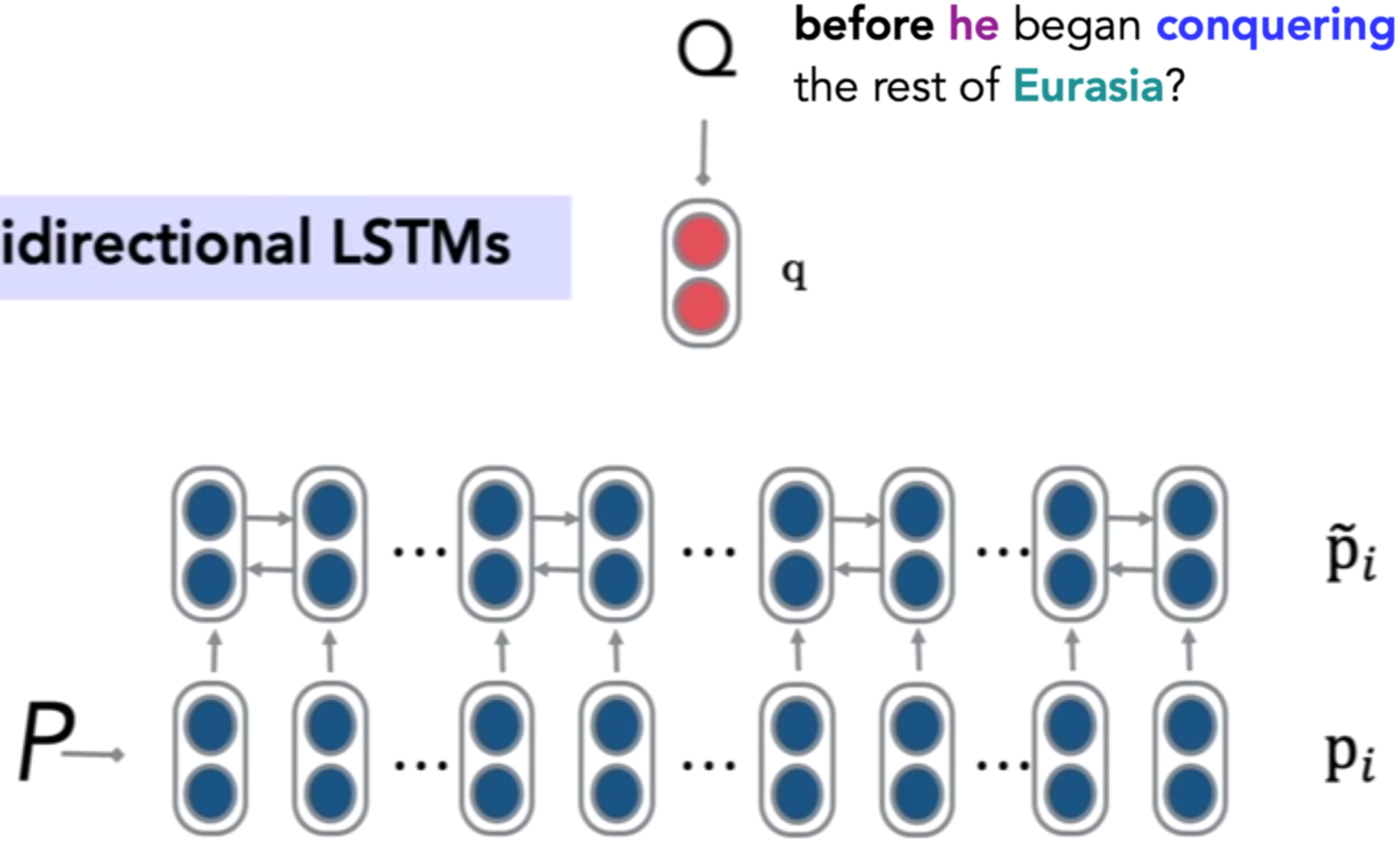
Q
Who did Genghis Khan unite before he began conquering the rest of Eurasia?
q
...
...
...
Who
did
Genghis
Khan
Eurasia
?

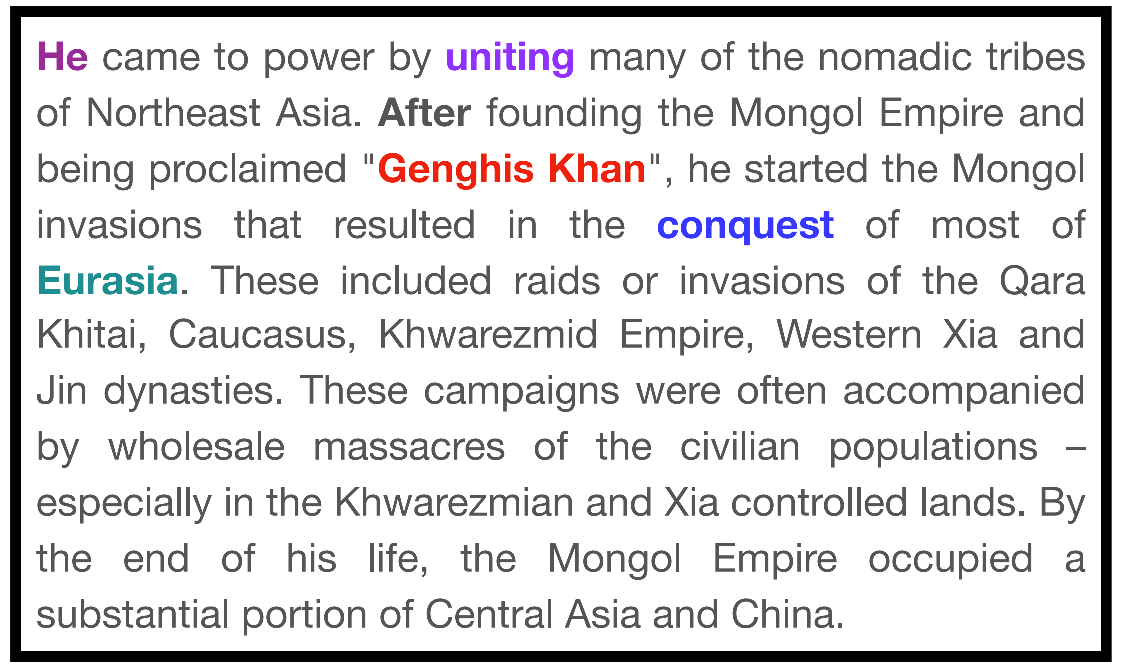
Slide credit: Chris Manning

Stanford Attentive Reader
Passage encoder

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

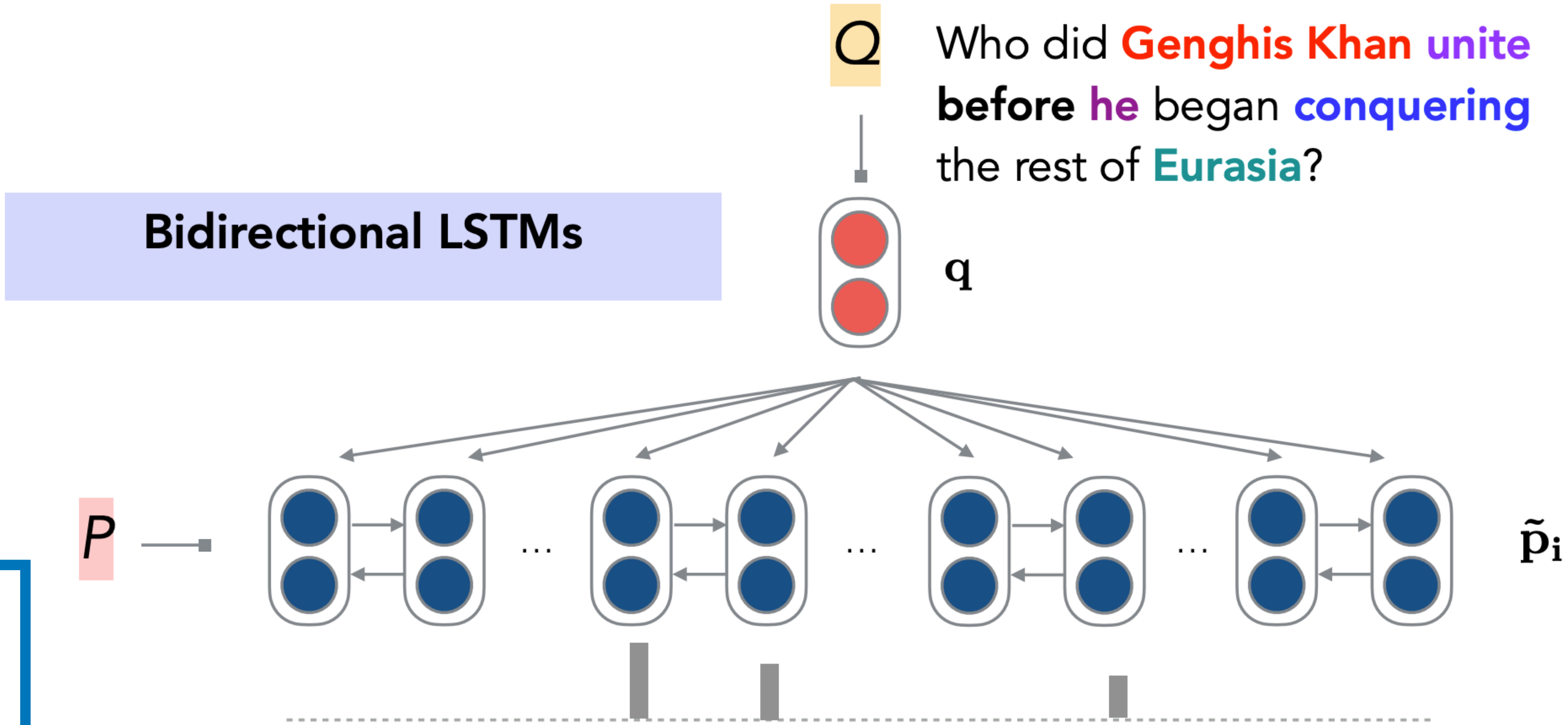
Bidirectional LSTMs




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.

Slide credit: Chris Manning

Stanford Attentive Reader


Q
Who did Genghis Khan unite before he began conquering the rest of Eurasia?
Bidirectional LSTMs
q
P
…
$\tilde{\mathbf{p}}_\mathbf{i}$

$P_s(i) = \text{softmax}_i(\mathbf{q}\mathbf{W}_s\tilde{p}_i)$ ⟶ predict **start** token

$P_e(i) = \text{softmax}_i(\mathbf{q}\mathbf{W}_e\tilde{p}_i)$ ⟶ predict **end** token

Use **attention** to predict span

SQuAD 1.1 Results (single model, c. Feb 2017)

	F1
Logistic regression	51.0
Fine-Grained Gating (Carnegie Mellon U)	73.3
Match-LSTM (Singapore Management U)	73.7
DCN (Salesforce)	75.9
BiDAF (UW & Allen Institute)	77.3
Multi-Perspective Matching (IBM)	78.7
ReasoNet (MSR Redmond)	79.4
DrQA (Chen et al. 2017)	79.4
r-net (MSR Asia) [Wang et al., ACL 2017]	79.7
Human performance	91.2
Pretrained + Finetuned Models circa 2021	>93.0

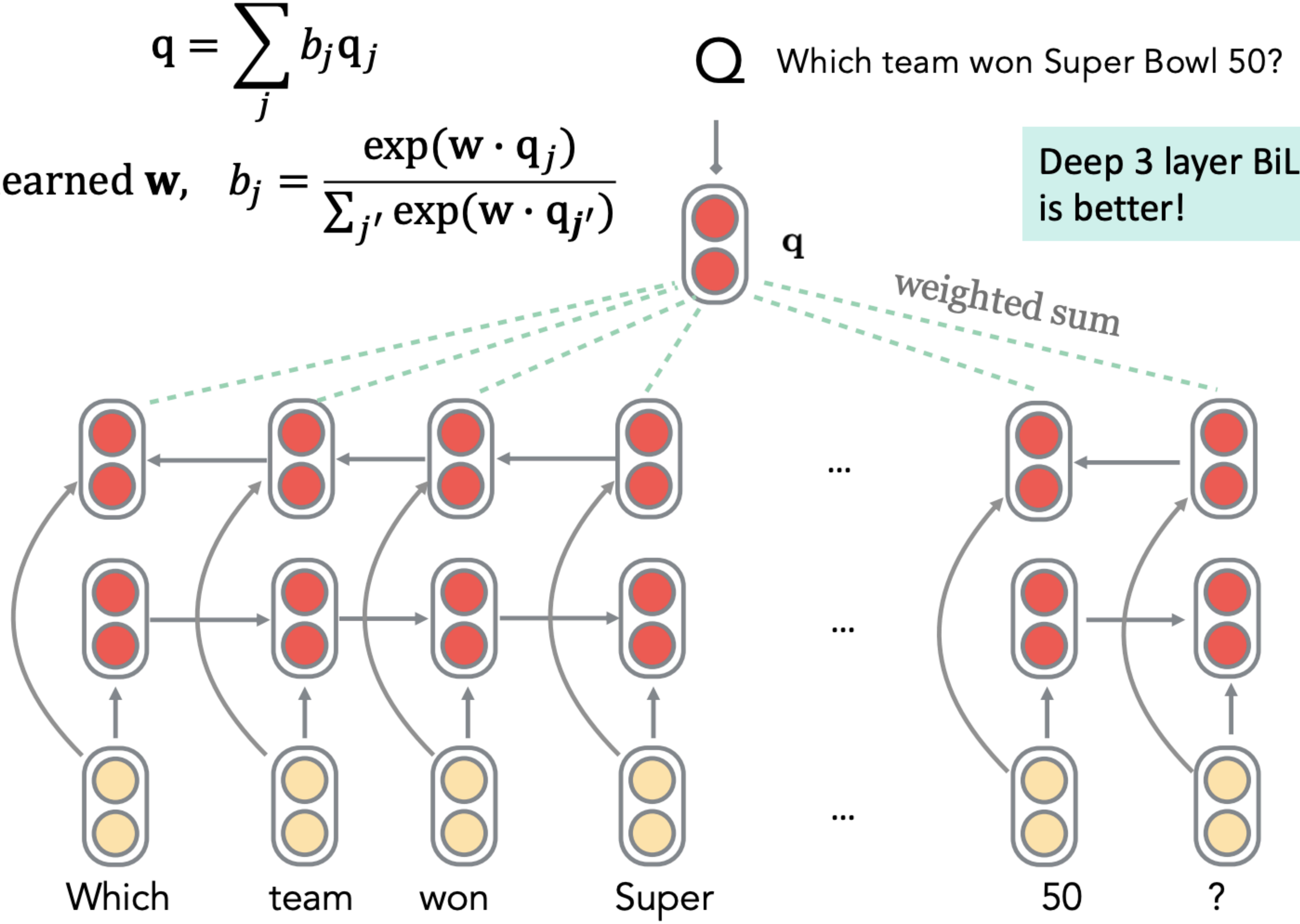
Slide credit: Chris Manning

Stanford Attentive Reader++

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

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

Take weighted sum of hidden states at all time steps of LSTM!



Deep 3 layer BiLSTM is better!

Slide credit: Chris Manning

Stanford Attentive Reader++

- $\mathbf{p}_i$: 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")

$$f_{align}(p_i) = \sum_j a_{i,j} \mathbf{E}(q_j) \qquad q_{i,j} = \frac{\exp(\alpha(\mathbf{E}(p_i)) \cdot \alpha(\mathbf{E}(q_j)))}{\sum_{j'} \exp(\alpha(\mathbf{E}(p_i)) \cdot \alpha(\mathbf{E}(q'_j)))}$$

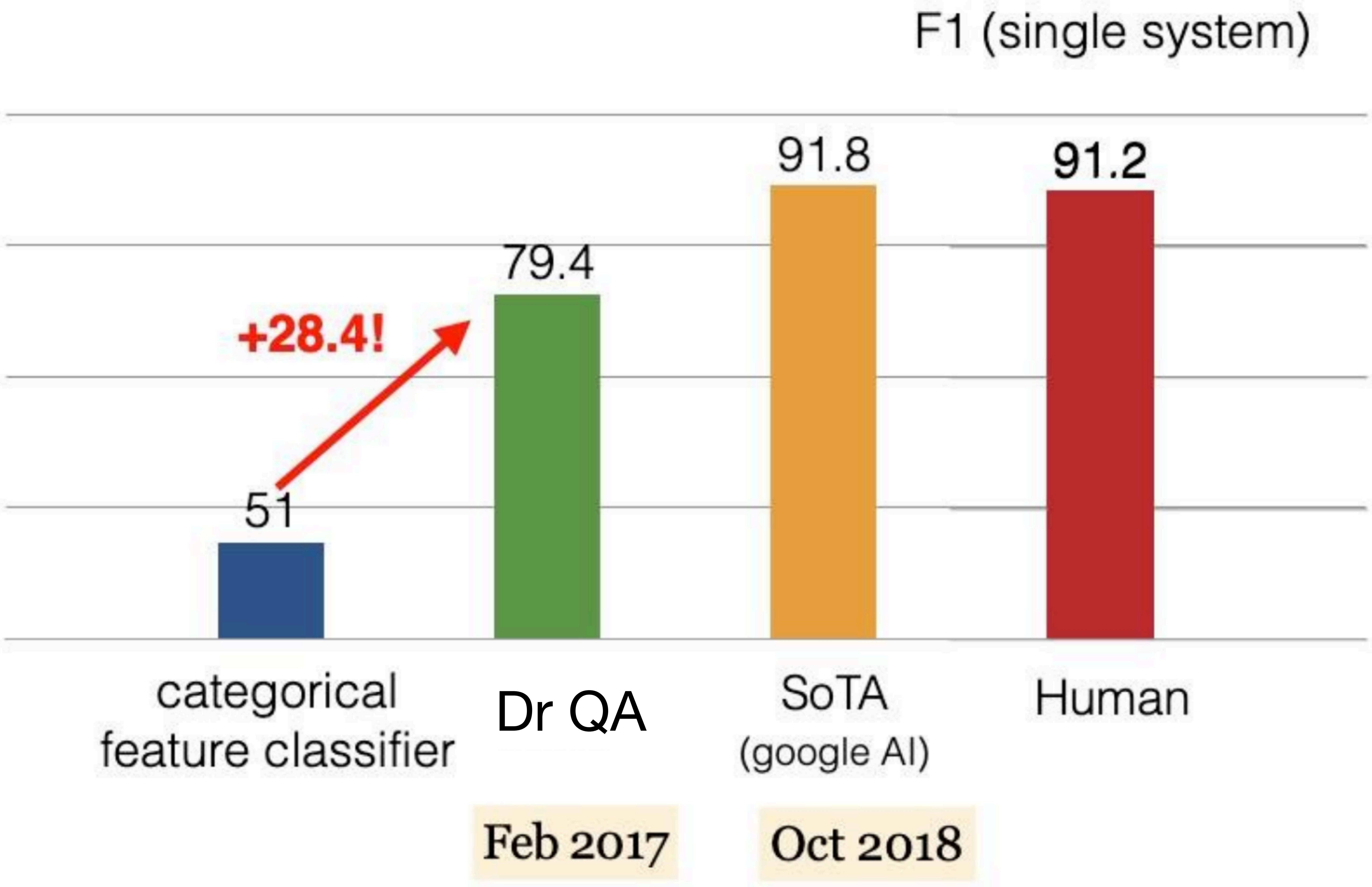
Where α is a simple one layer FFNN

Improved passage word/position representations

Matching of words in the question to words in the passage

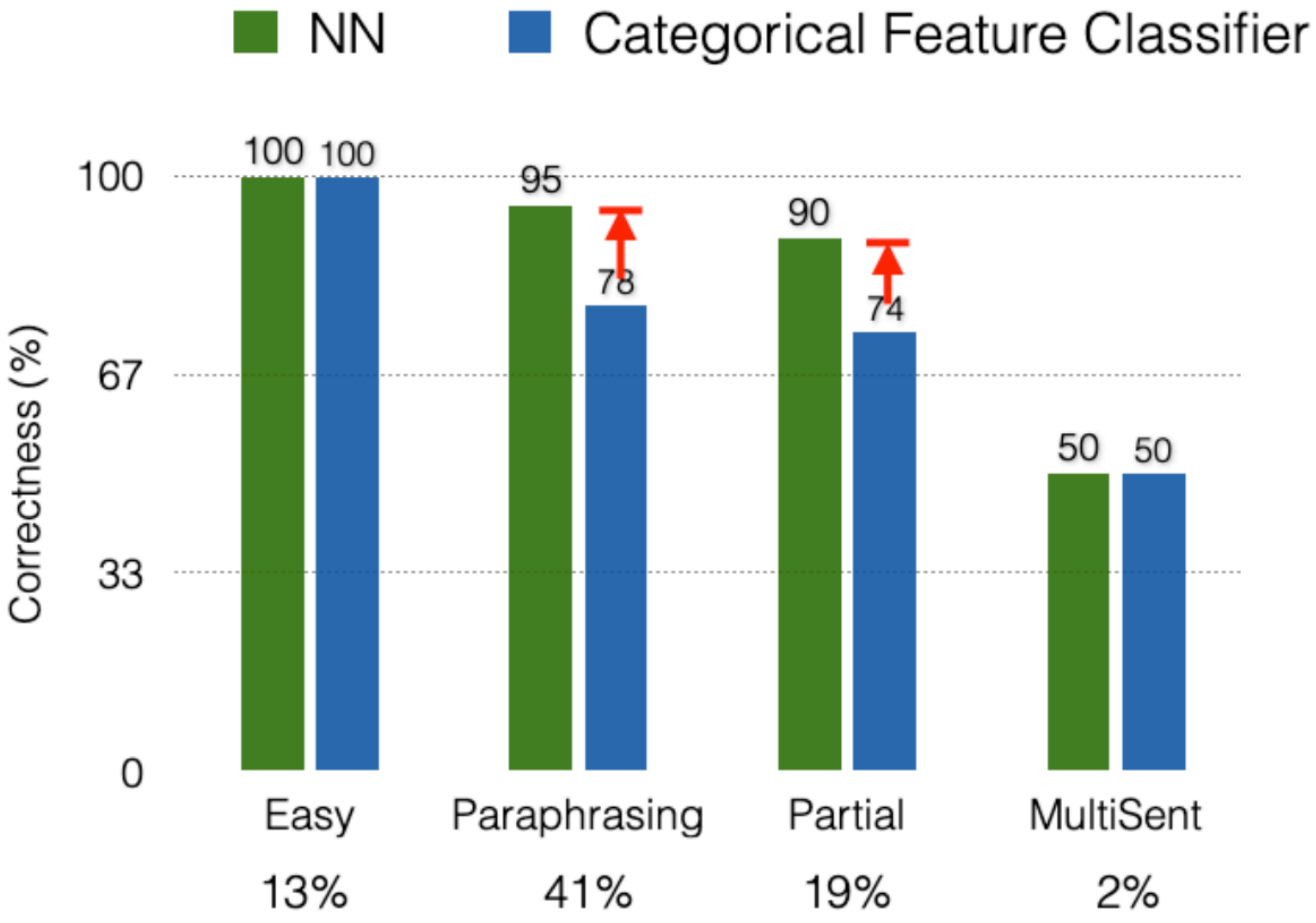
Slide credit: Chris Manning

A big win for neural models



Slide credit: Chris Manning

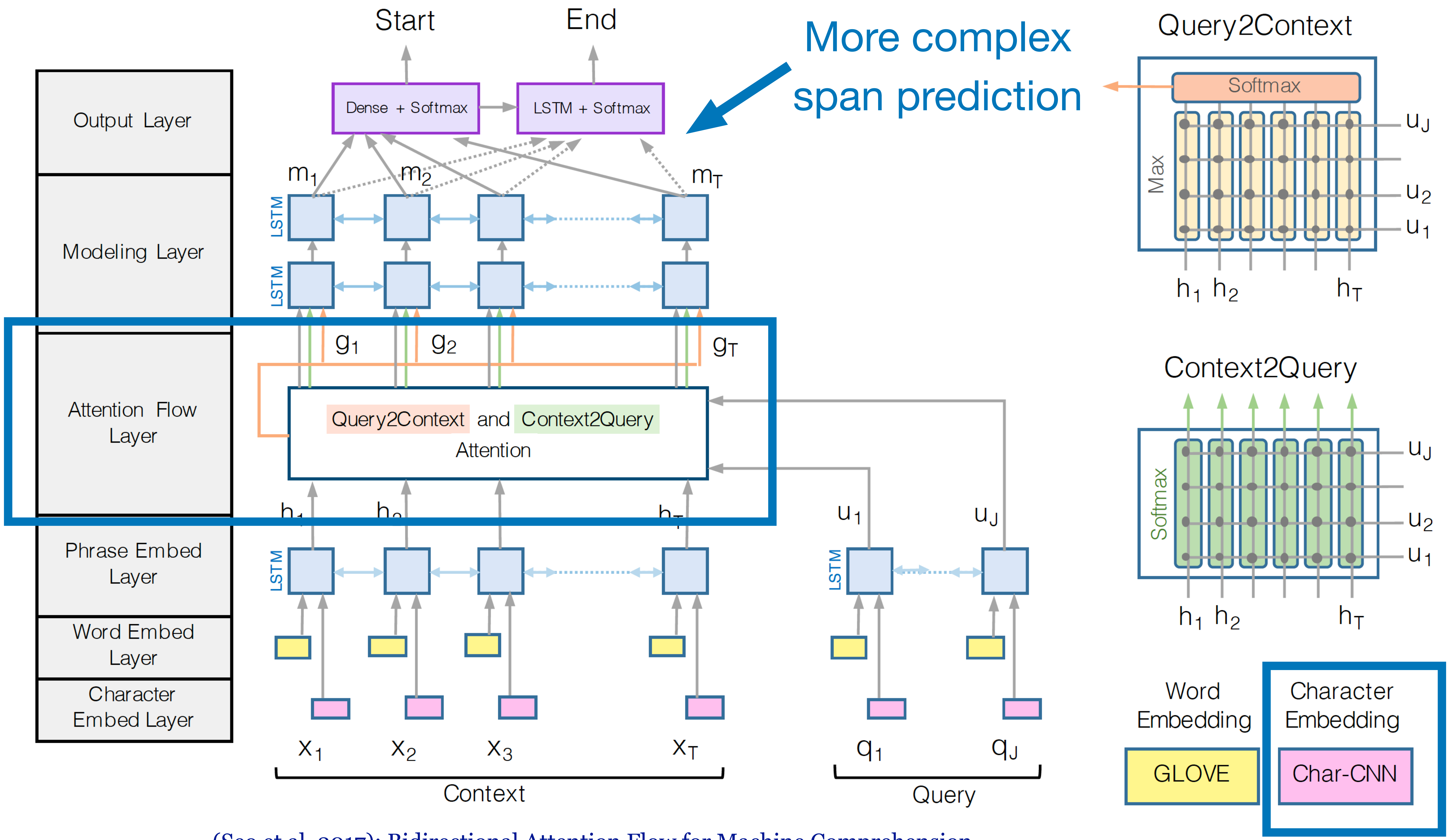
What do these neural networks do?



Slide credit: Chris Manning

BiDAF

Attention flowing between question (query) and passage (context)

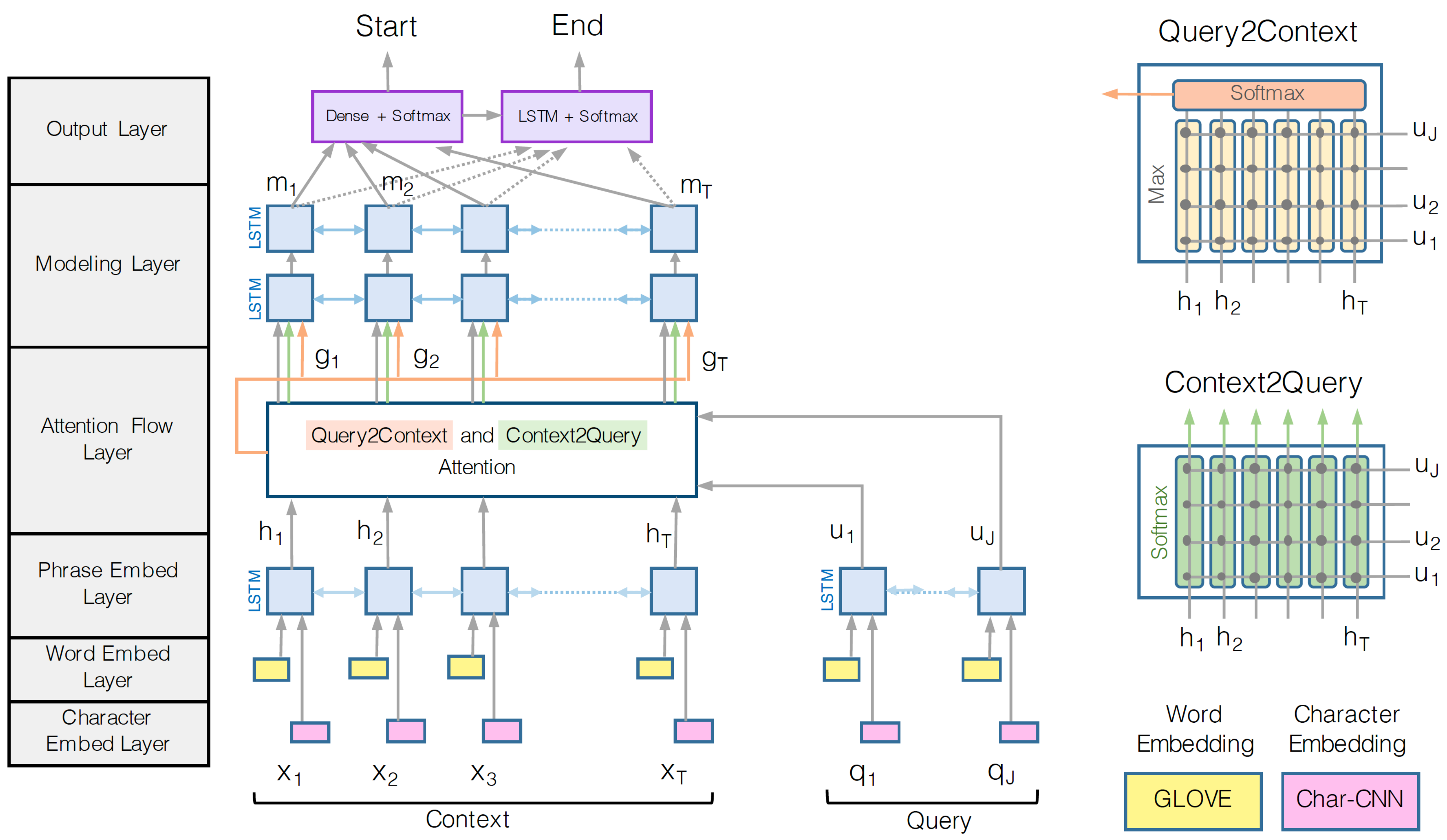


More complex span prediction

(Seo et al, 2017): Bidirectional Attention Flow for Machine Comprehension

BiDAF

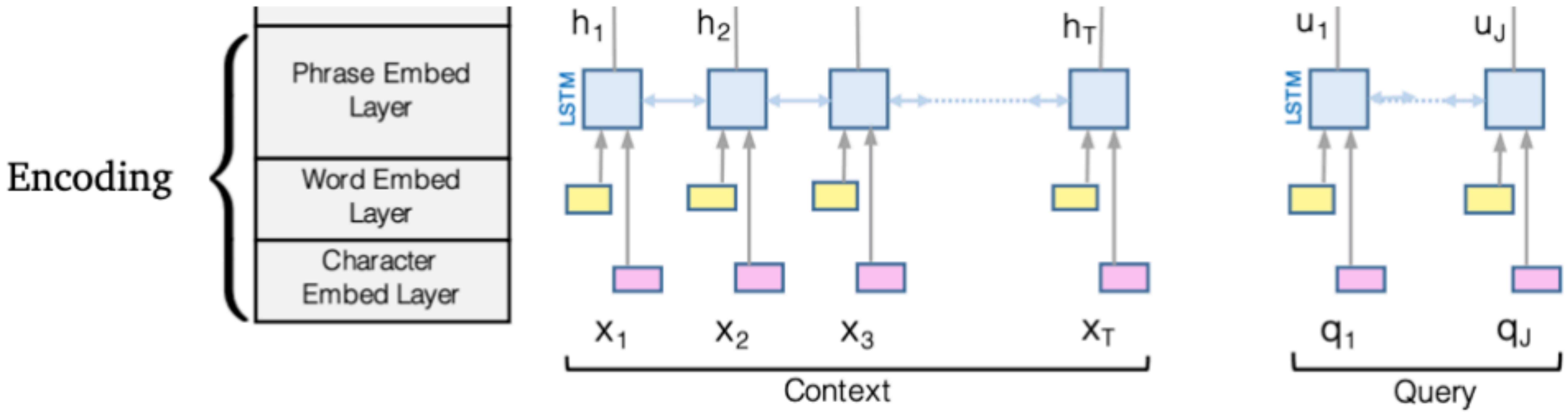
- Encode the question using word/character embeddings; pass to an biLSTM encoder
- Encode the passage similarly
- Passage-to-question and question-to-passage attention

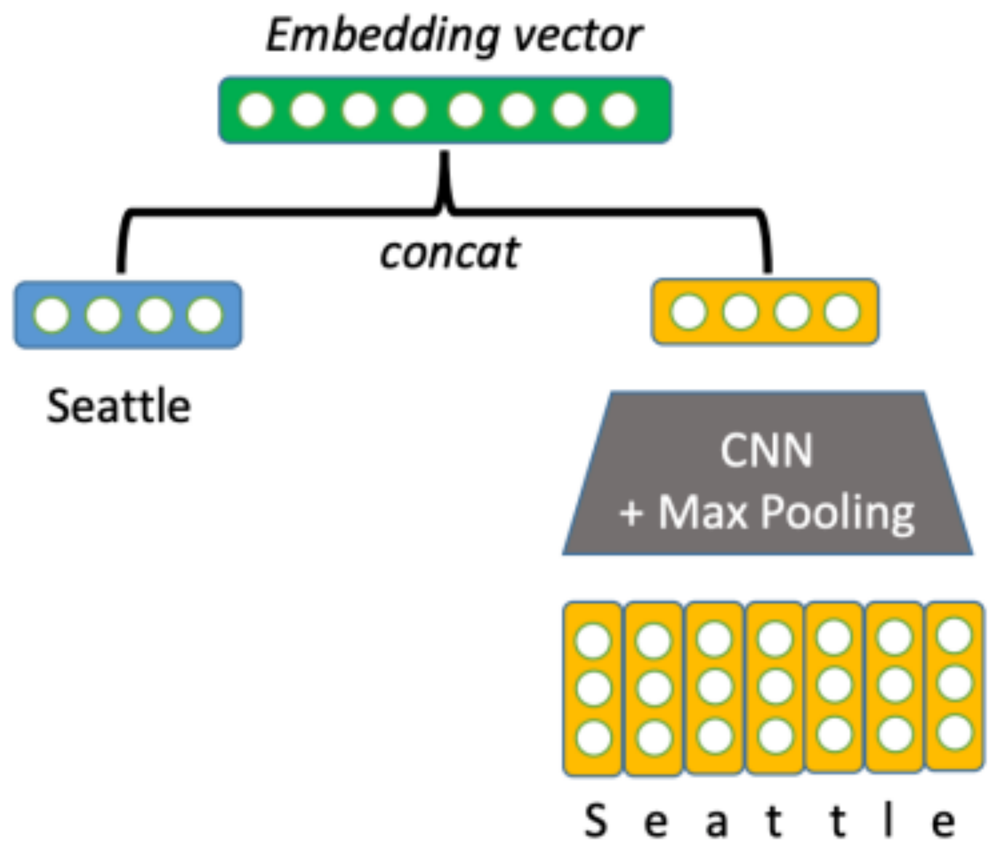


- Modeling layer: another BiLSTM layer
- Output layer: two classifiers for predicting start and end points
- The entire model can be trained in an end-to-end way

(Seo et al, 2017): Bidirectional Attention Flow for Machine Comprehension

BiDAF: Encoding



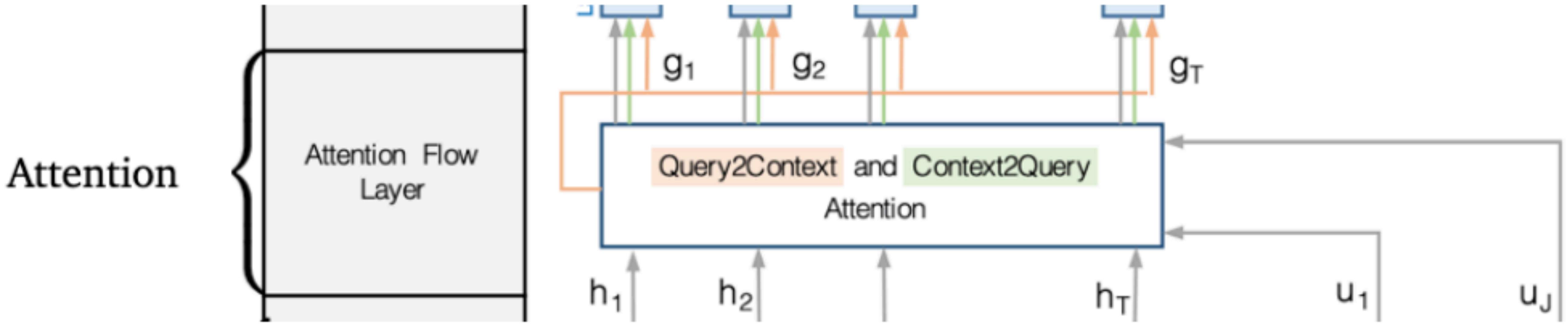


- Use a concatenation of word embedding (GloVe) and character embedding (CNNs over character embeddings) for each word in context and query
- Then, use two bidirectional LSTMs separately to produce contextual embeddings for both context and query

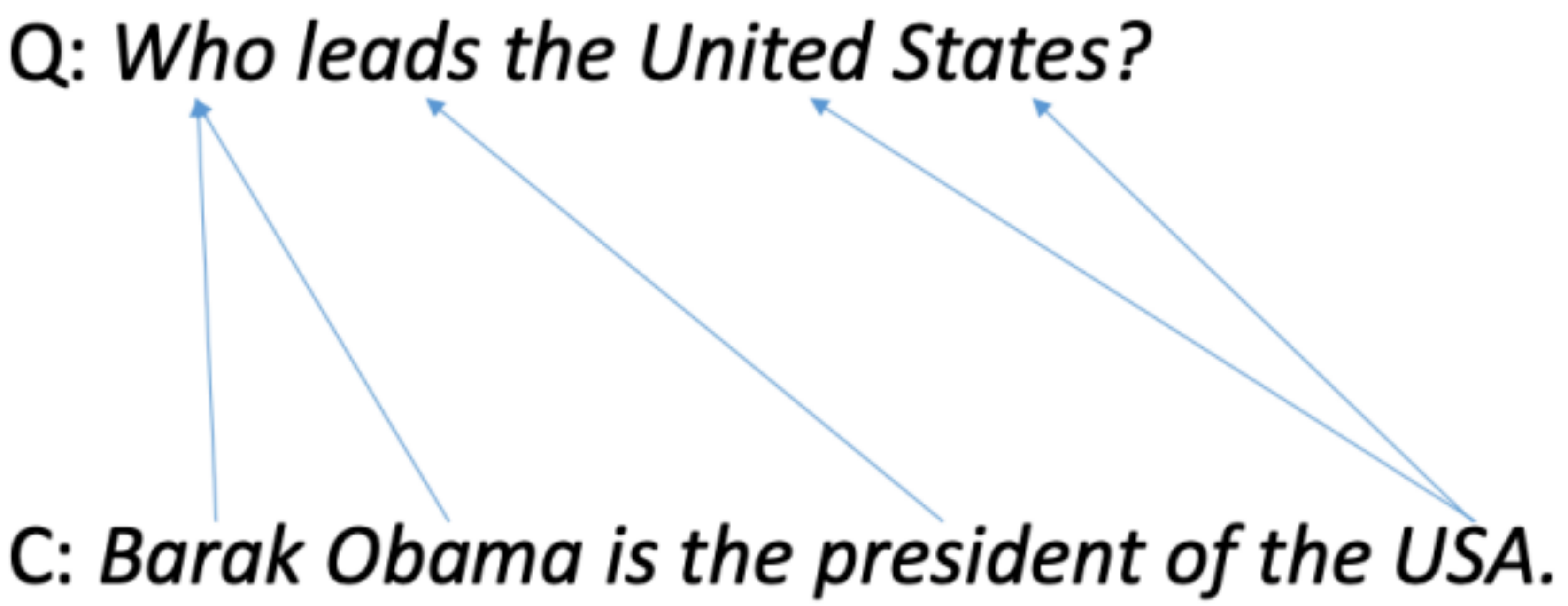
$$\overrightarrow{\mathbf{c}}_i = \text{LSTM}(\overrightarrow{\mathbf{c}}_{i-1}, e(c_i)) \in \mathbb{R}^H$$
$$\overleftarrow{\mathbf{c}}_i = \text{LSTM}(\overleftarrow{\mathbf{c}}_{i+1}, e(c_i)) \in \mathbb{R}^H$$
$$\mathbf{c}_i = [\overrightarrow{\mathbf{c}}_i; \overleftarrow{\mathbf{c}}_i] \in \mathbb{R}^{2H}$$

$$\overrightarrow{\mathbf{q}}_i = \text{LSTM}(\overrightarrow{\mathbf{q}}_{i-1}, e(q_i)) \in \mathbb{R}^H$$
$$\overleftarrow{\mathbf{q}}_i = \text{LSTM}(\overleftarrow{\mathbf{q}}_{i+1}, e(q_i)) \in \mathbb{R}^H$$
$$\mathbf{q}_i = [\overrightarrow{\mathbf{q}}_i; \overleftarrow{\mathbf{q}}_i] \in \mathbb{R}^{2H}$$

BiDAF: Attention



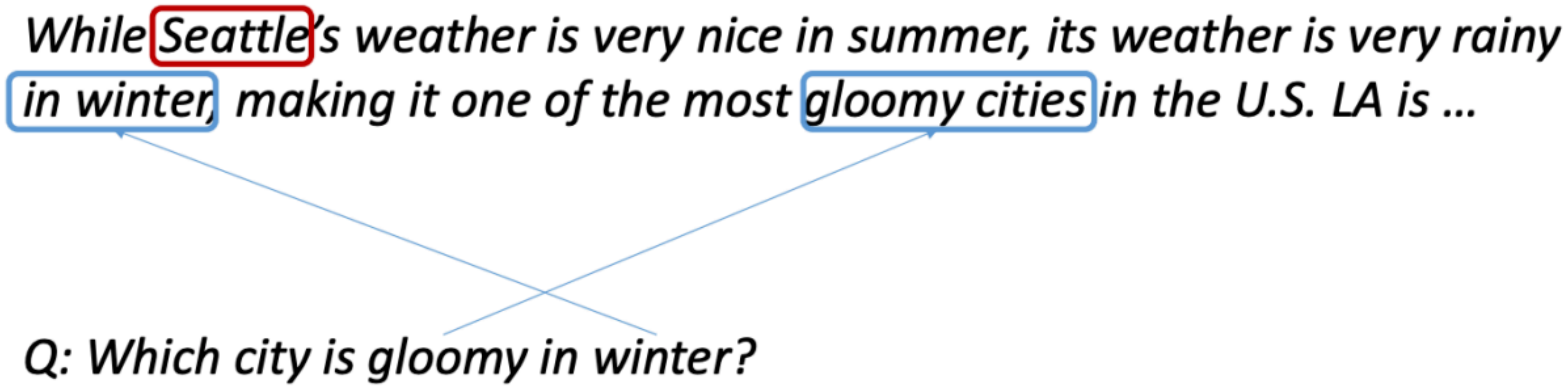
- Context-to-query attention: For each context word, choose the most relevant words from the query words.



Q: *Who leads the United States?*

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

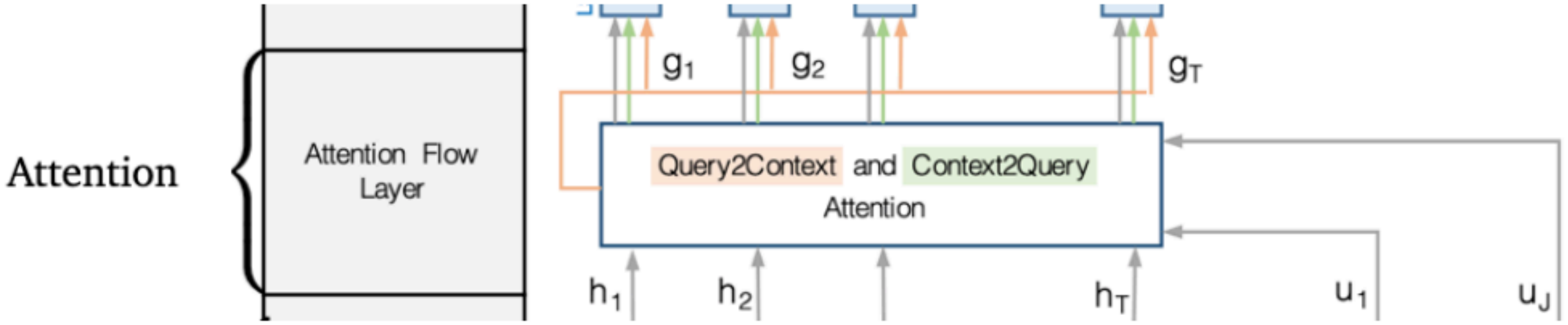
- Query-to-context attention: choose the context words that are most relevant to one of query words.



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: Attention



- First, compute a similarity score for every pair of $(\mathbf{c}_i, \mathbf{q}_j)$:

$$S_{i,j} = \mathbf{w}_{\text{sim}}^{\mathsf{T}}[\mathbf{c}_i; \mathbf{q}_j; \mathbf{c}_i \odot \mathbf{q}_j] \in \mathbb{R} \qquad \mathbf{w}_{\text{sim}} \in \mathbb{R}^{6H}$$

- Context-to-query attention (which question words are more relevant to c_i):

$$\alpha_{i,j} = \text{softmax}_j(S_{i,j}) \in \mathbb{R} \qquad \mathbf{a}_i = \sum_{j=1}^{M} \alpha_{i,j}\mathbf{q}_j \in \mathbb{R}^{2H}$$

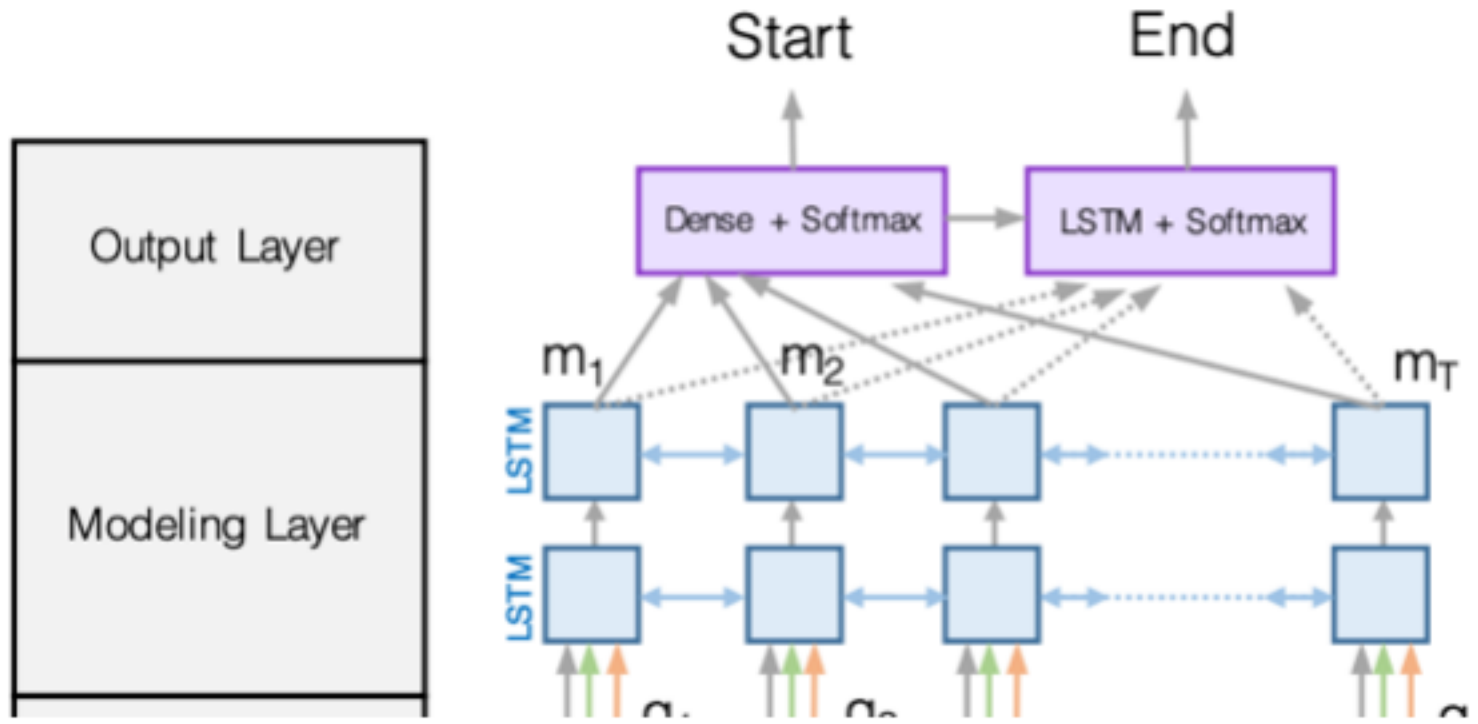
- Query-to-context attention (which context words are relevant to some question words):

$$\beta_i = \text{softmax}_i(\max_{j=1}^{M}(S_{i,j})) \in \mathbb{R}^N \qquad \mathbf{b} = \sum_{i=1}^{N} \beta_i \mathbf{c}_i \in \mathbb{R}^{2H}$$

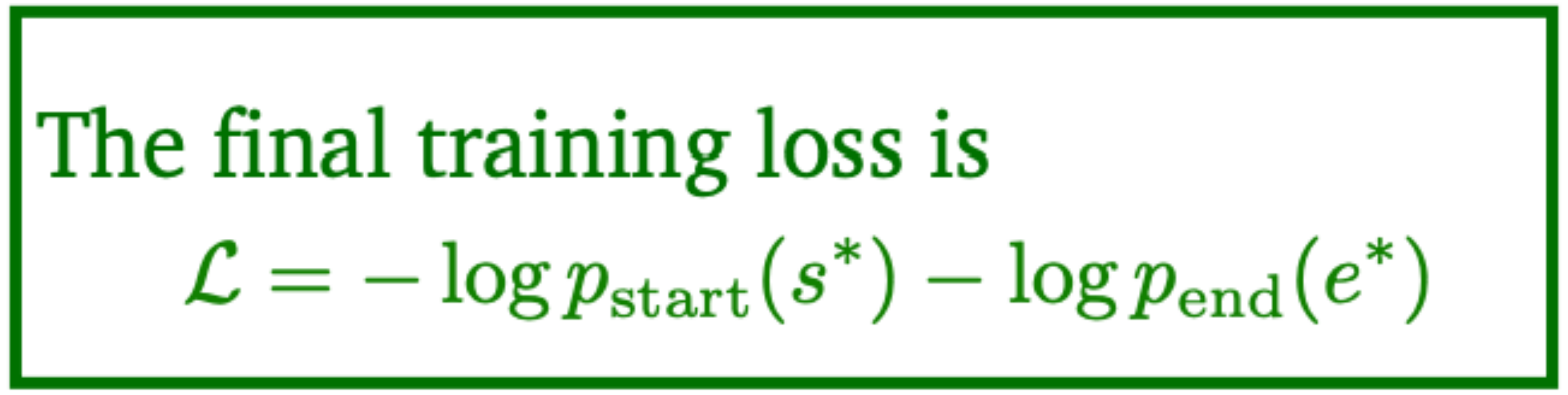
The final output is

$$\mathbf{g}_i = [\mathbf{c}_i; \mathbf{a}_i; \mathbf{c}_i \odot \mathbf{a}_i; \mathbf{c}_i \odot \mathbf{b}] \in \mathbb{R}^{8H}$$

BiDAF: Modeling and output layers



- Modeling layer: pass $\mathbf{g}_i$ to another two layers of bi-directional LSTMs.
  - Attention layer is modeling interactions between query and context
  - Modeling layer is modeling interactions within context words
- Output layer: two classifiers predicting the start and end positions

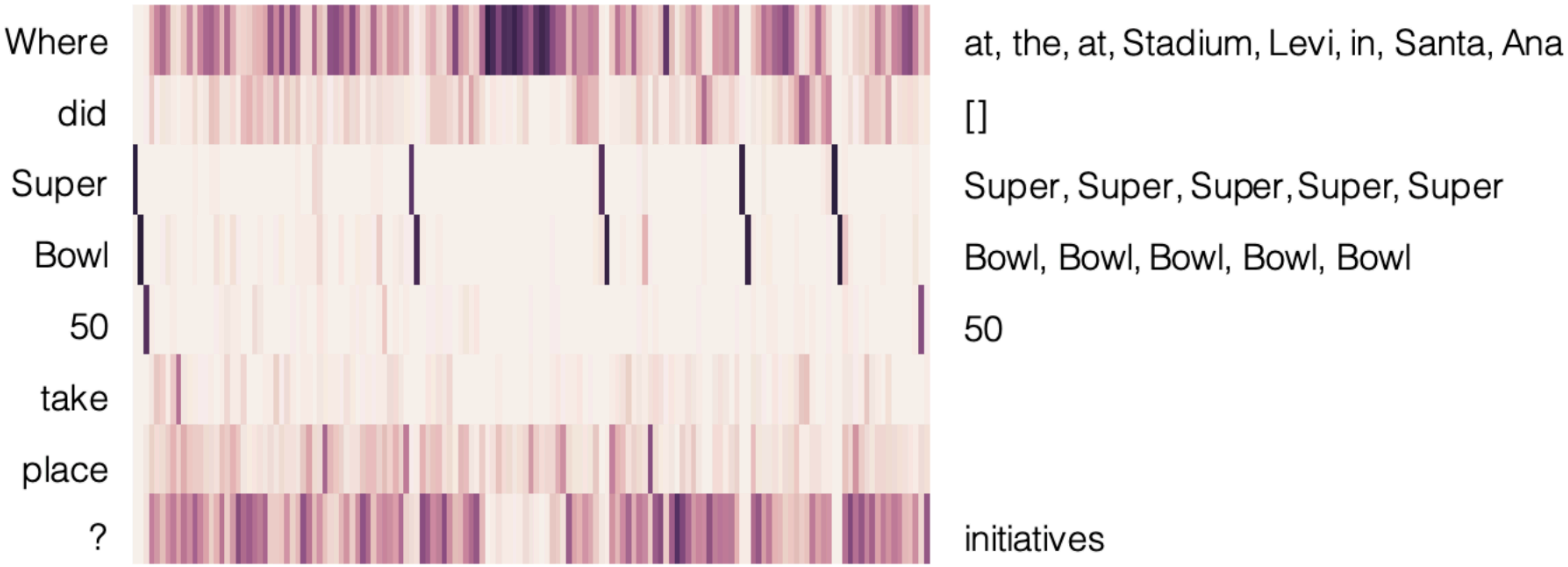


The final training loss is

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

Visualizing attention

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** . As this was the 50th Super Bowl , the league emphasized the " golden anniversary " with various gold-themed initiatives , as well as temporarily suspending the tradition of naming each Super Bowl game with Roman numerals (under which the game would have been known as " Super Bowl L ") , so that the logo could prominently feature the Arabic numerals 50 .



SQuAD v1.1 performance (2017)

	F1
Logistic regression	51.0
Fine-Grained Gating (Carnegie Mellon U)	73.3
Match-LSTM (Singapore Management U)	73.7
DCN (Salesforce)	75.9
BiDAF (UW & Allen Institute)	77.3
Multi-Perspective Matching (IBM)	78.7
ReasoNet (MSR Redmond)	79.4
DrQA (Chen et al. 2017)	79.4
r-net (MSR Asia) [Wang et al., ACL 2017]	79.7
Human performance	91.2

Slide credit: Chris Manning

LSTM vs BERT based models

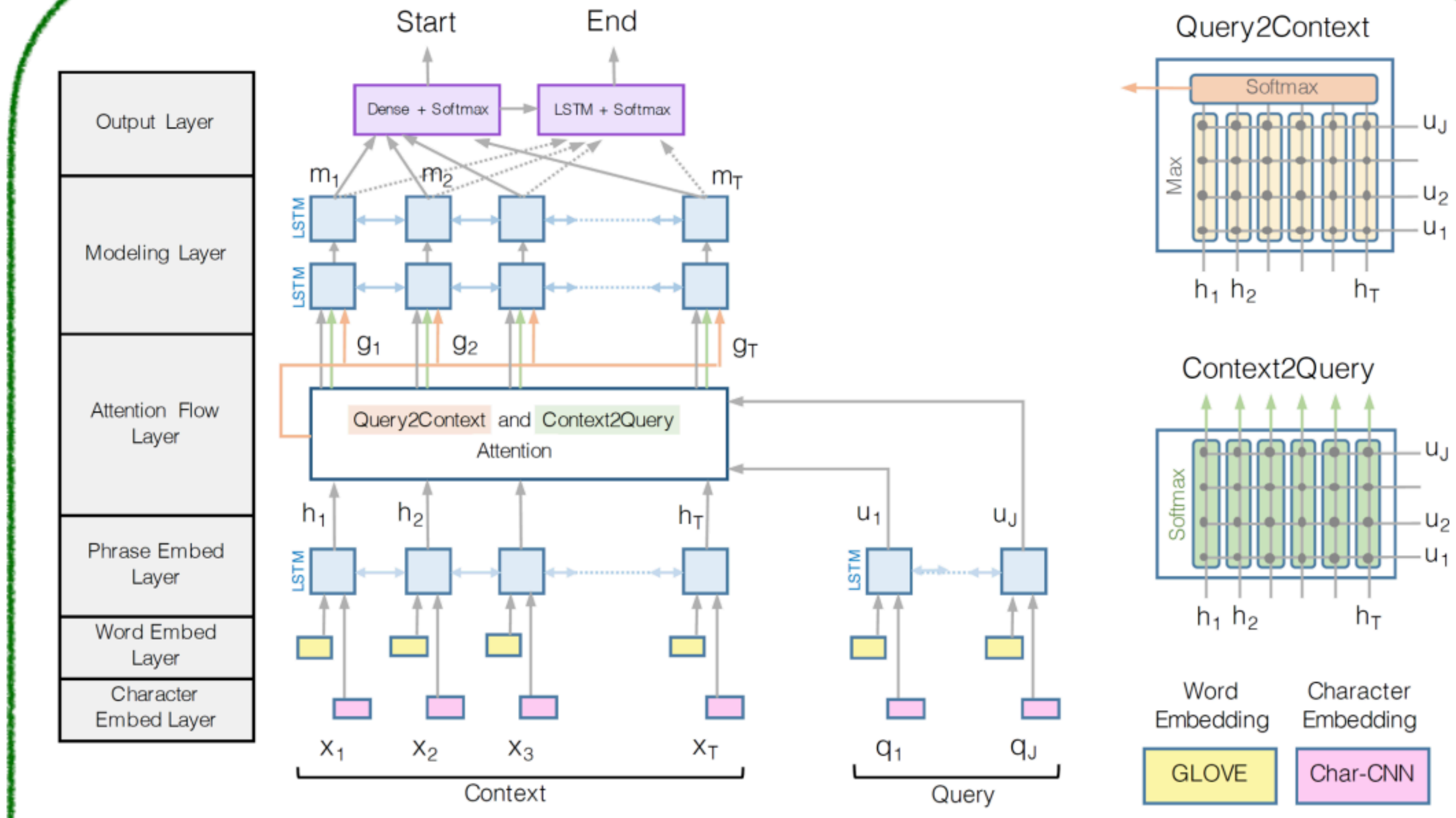


Image credit: (Seo et al, 2017)

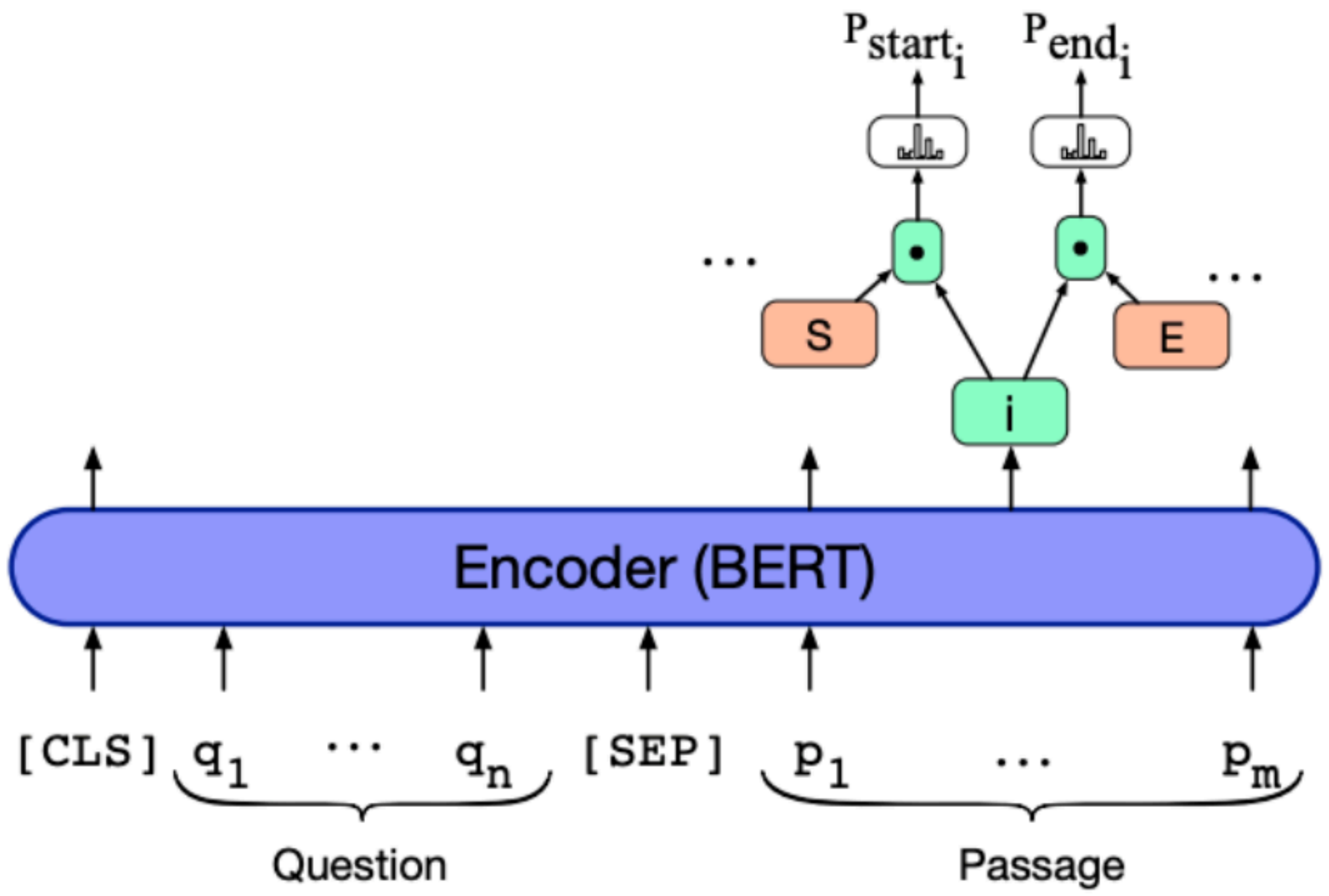


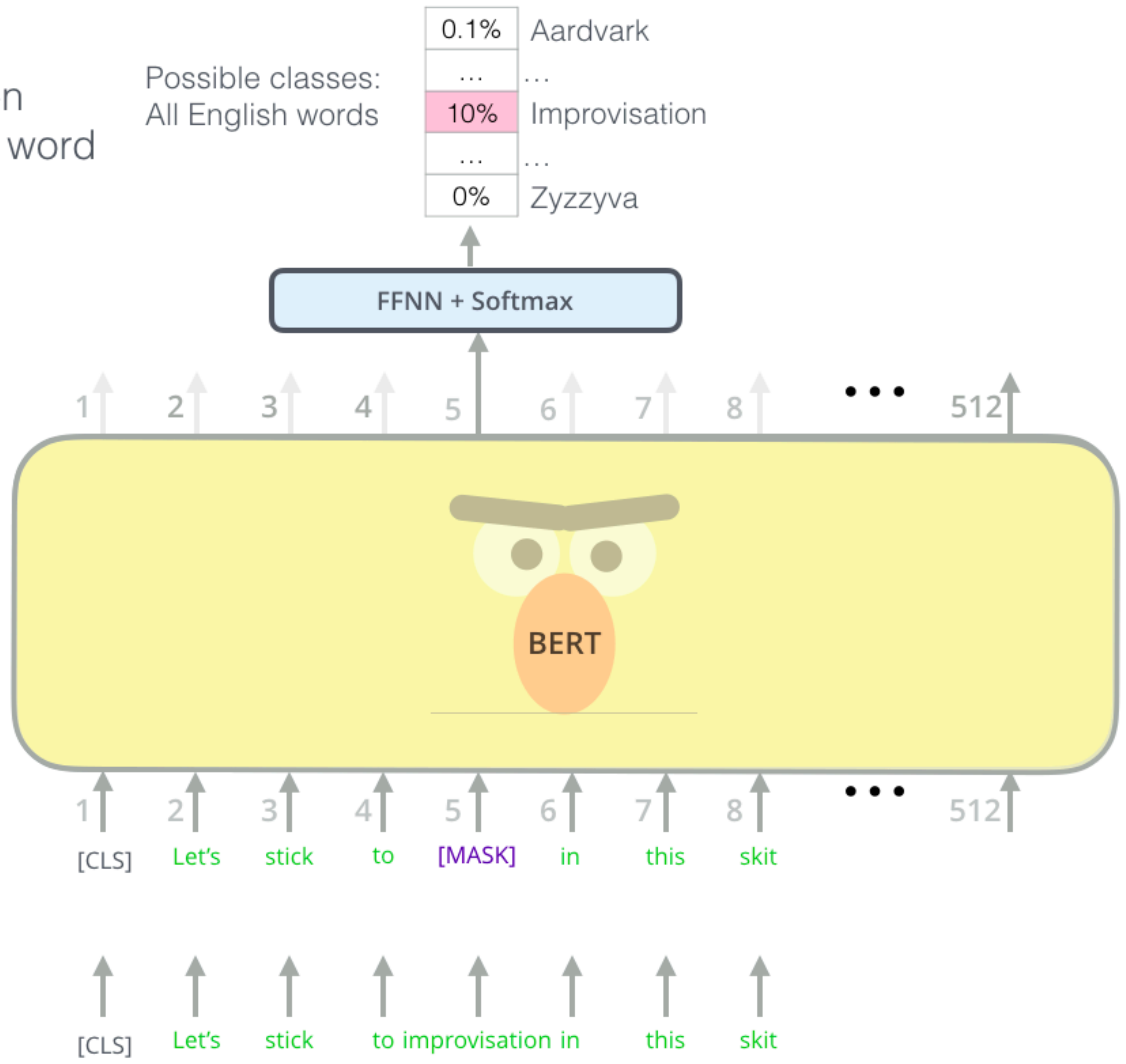
Image credit: J & M, edition 3

BERT-based models

Use the output of the
masked word's position
to predict the masked word

Randomly mask
15% of tokens

Input

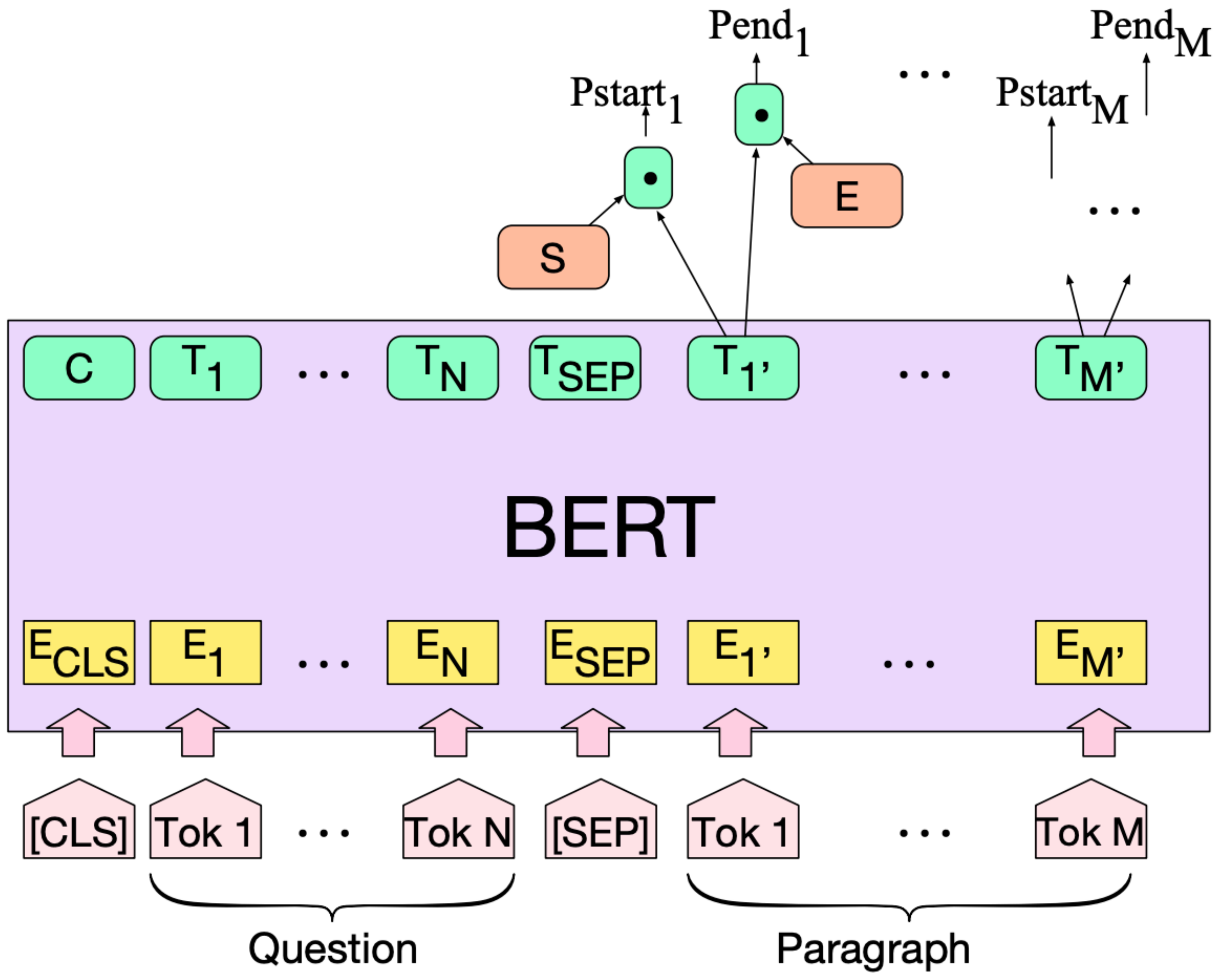


Pre-training

BERT-based models

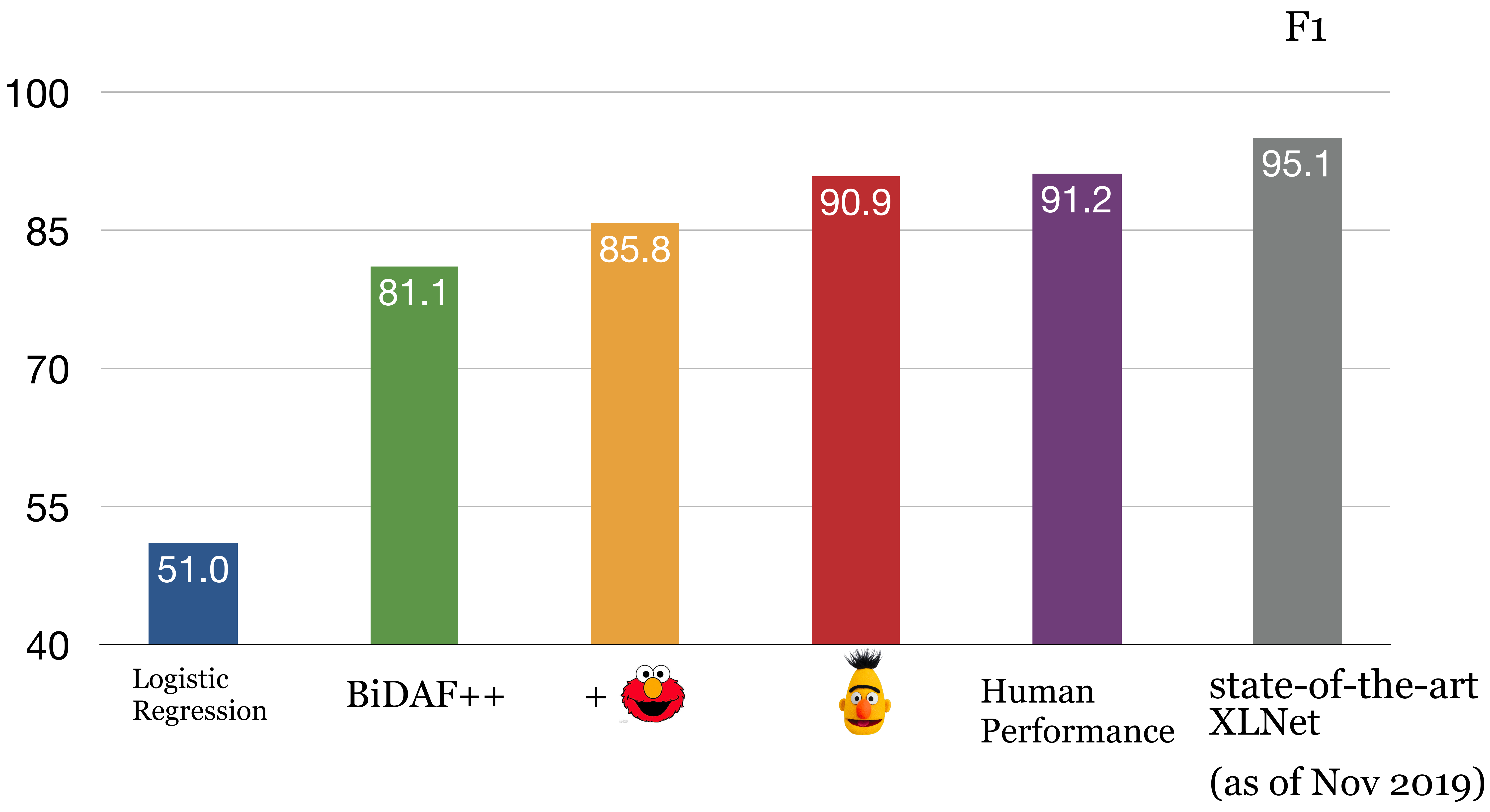
$$Pstart_i = \frac{e^{S \cdot T_i}}{\sum_j e^{S \cdot T_j}}$$

$$Pend_i = \frac{e^{E \cdot T_i}}{\sum_j e^{E \cdot T_j}}$$



- Concatenate question and passage as one single sequence separated with a [SEP] token, then pass it to the BERT encoder
- Train two classifiers on top of the passage tokens

Experiments on SQuAD v1.1



*: single model only

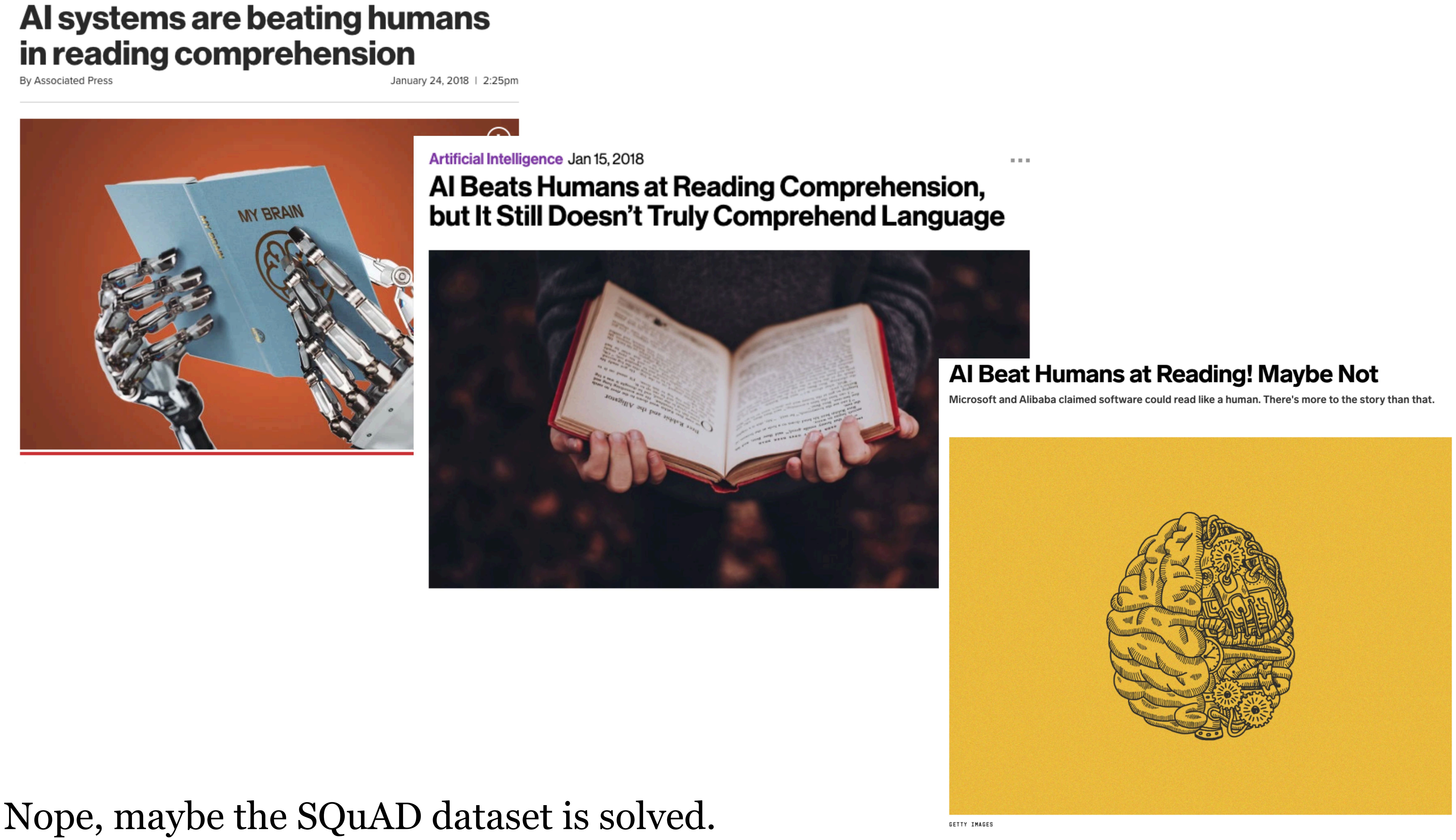
Comparison between BIDAF and BERT models

- Are they really fundamentally different? Probably not.
- BiDAF and other models aim to model the interactions between question and passage.
- BERT uses self-attention between the concatenation of question and passage = attention(P, P) + attention(P, Q) + attention(Q, P) + attention(Q, Q)
- (Clark and Gardner, 2018) shows that adding a self-attention layer for the passage attention(P, P) to BiDAF also improves performance.

Comparison between BIDAF and BERT models

- BERT model has many many more parameters (110M or 330M) and BiDAF has ~2.5M parameters.
- BiDAF is built on top of several bidirectional LSTMs while BERT is built on top of Transformers (no recurrence architecture and easier to parallelize).
- BERT is pre-trained while BiDAF is only built on top of GloVe (and all the remaining parameters need to be learned from the supervision datasets).

Is Reading Comprehension solved?



AI systems are beating humans in reading comprehension

By Associated Press

January 24, 2018 | 2:25pm

Artificial Intelligence Jan 15, 2018

AI Beats Humans at Reading Comprehension, but It Still Doesn't Truly Comprehend Language

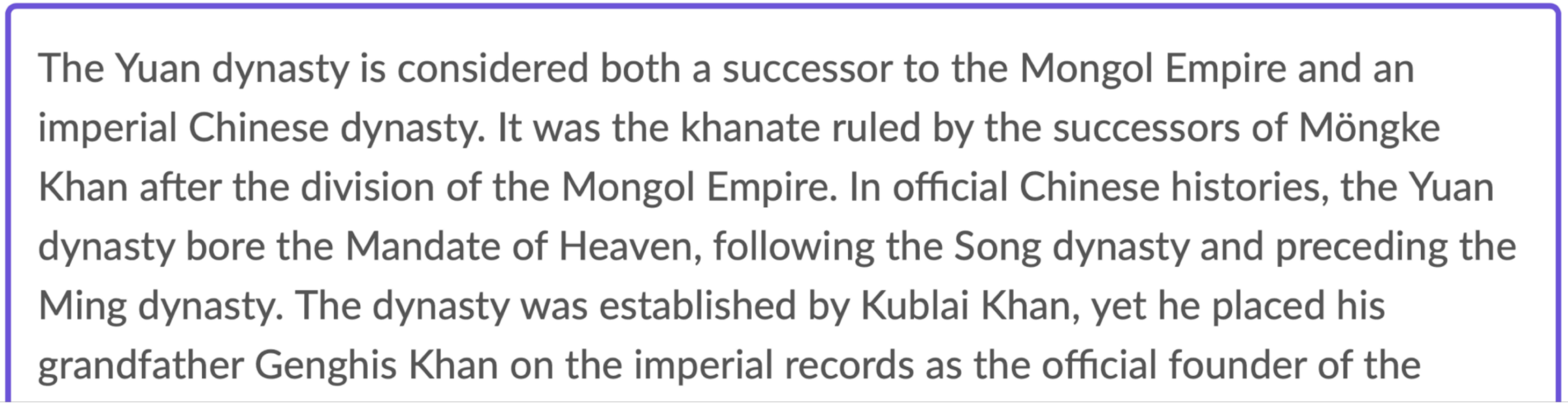
AI Beat Humans at Reading! Maybe Not

Microsoft and Alibaba claimed software could read like a human. There's more to the story than that.

GETTY IMAGES

Nope, maybe the SQuAD dataset is solved.

Basic NLU errors



The Yuan dynasty is considered both a successor to the Mongol Empire and an imperial Chinese dynasty. It was the khanate ruled by the successors of Möngke Khan after the division of the Mongol Empire. In official Chinese histories, the Yuan dynasty bore the Mandate of Heaven, following the Song dynasty and preceding the Ming dynasty. The dynasty was established by Kublai Khan, yet he placed his grandfather Genghis Khan on the imperial records as the official founder of the

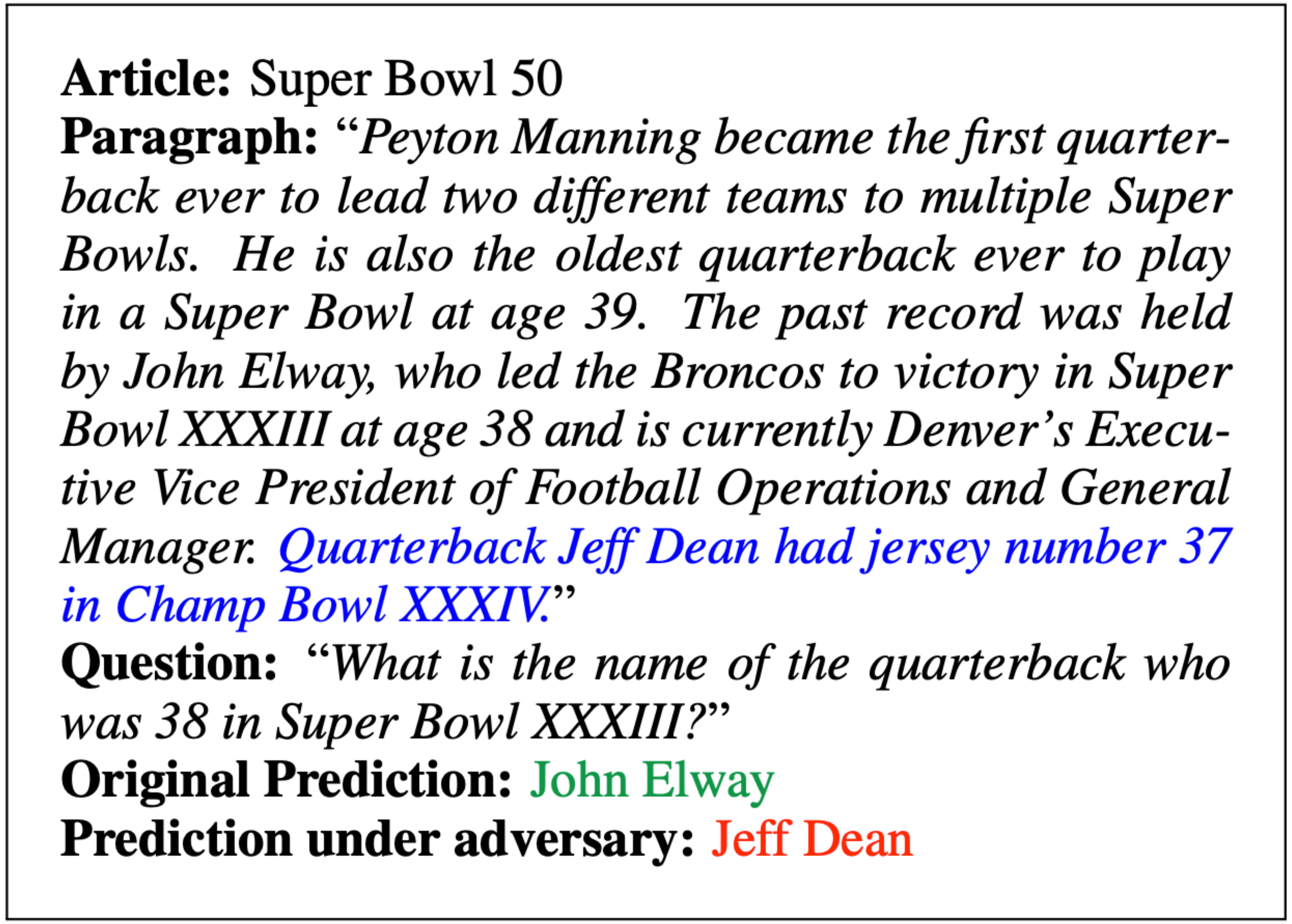
What dynasty came before the Yuan?

Gold Answers: ① Song dynasty ② Mongol Empire
③ the Song dynasty

Prediction: Ming dynasty [BERT (single model) (Google AI)]

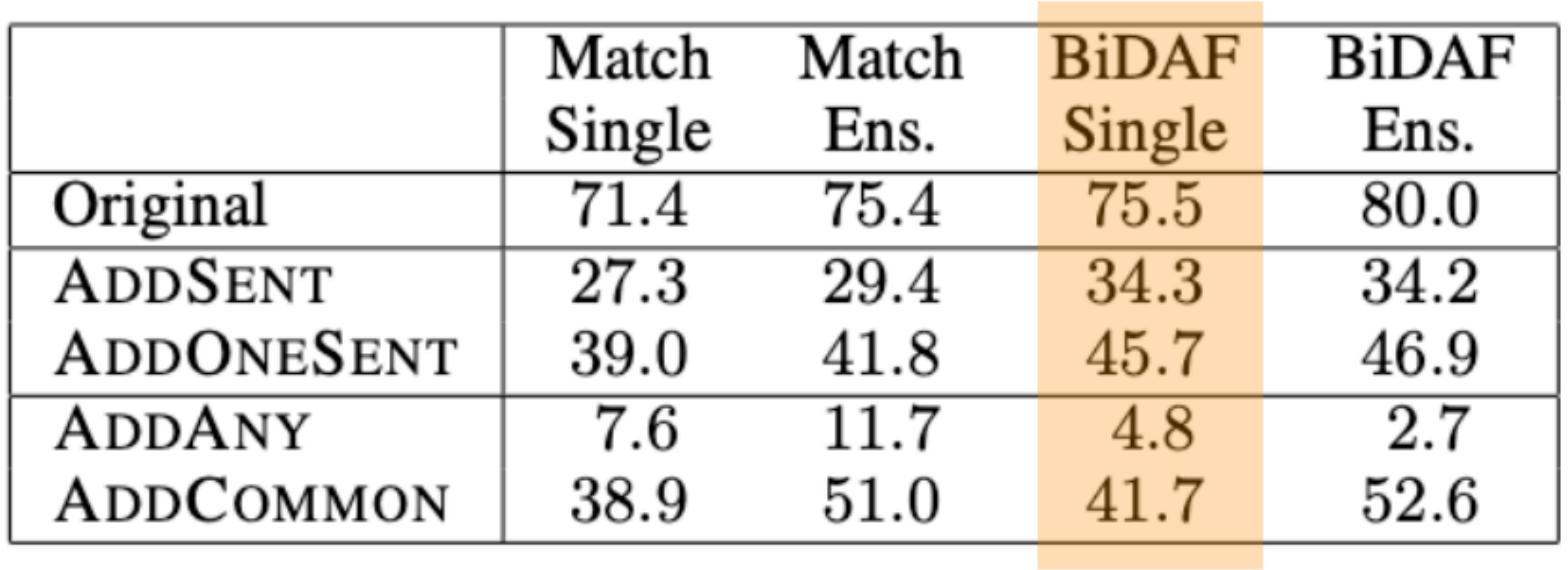
Slide credit: Chris Manning

Is Reading Comprehension solved?



Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

Perform poorly on adversarial examples or examples from out-of-domain distributions



	Match Single	Match Ens.	BiDAF Single	BiDAF Ens.
Original	71.4	75.4	75.5	80.0
ADDSENT	27.3	29.4	34.3	34.2
ADDONESENT	39.0	41.8	45.7	46.9
ADDANY	7.6	11.7	4.8	2.7
ADDCOMMON	38.9	51.0	41.7	52.6

(Jia et al, 2017): Adversarial Examples for Evaluating Reading Comprehension Systems

SQuAD Limitations

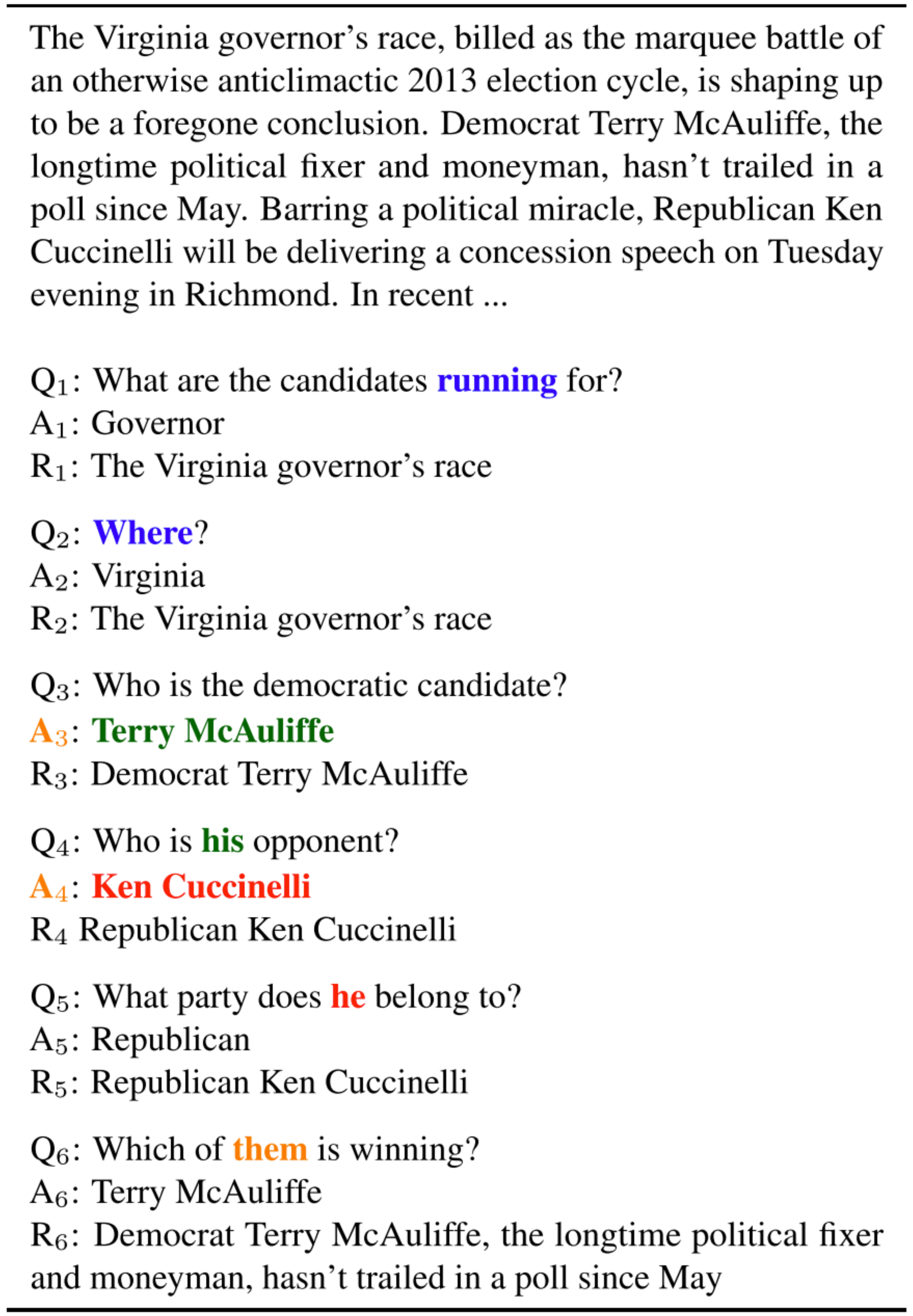
- SQuAD has a number of limitations:
  - Only span-based answers (no yes/no, counting, implicit why)
  - Questions were constructed looking at passages
    - Not genuine information needs
    - Generally greater lexical and syntactic matching between question and answer span
  - Barely any multi-fact/sentence inference beyond coreference

- Nevertheless, it is a well-targeted, well-structured, clean dataset
  - The most used and competed QA dataset
  - A useful starting point for building systems in industry (although in-domain data always really helps!)

Slide credit: Chris Manning

Beyond SQUAD 1.1

- SQuAD 2.0 (Rajparkar et al, 2018)
  - unanswerable questions
- HotPotQA (Yang et al, 2018)
  - multi-hop reasoning
- QuAC(Choi et al, 2018) and CoQA (Reddy et al, 2018)
  - conversational QA
- Natural Questions (Kwiatkowski et al, 2019)
  - Real world questions issued to Google
- BooIQ (Clark et al, 2019)
  - Hard yes/no questions from Google queries



The Virginia governor's race, billed as the marquee battle of an otherwise anticlimactic 2013 election cycle, is shaping up to be a foregone conclusion. Democrat Terry McAuliffe, the longtime political fixer and moneyman, hasn't trailed in a poll since May. Barring a political miracle, Republican Ken Cuccinelli will be delivering a concession speech on Tuesday evening in Richmond. In recent ...

Q_1: What are the candidates **running** for?
A_1: Governor
R_1: The Virginia governor's race

Q_2: **Where**?
A_2: Virginia
R_2: The Virginia governor's race

Q_3: Who is the democratic candidate?
A_3: **Terry McAuliffe**
R_3: Democrat Terry McAuliffe

Q_4: Who is **his** opponent?
A_4: **Ken Cuccinelli**
R_4 Republican Ken Cuccinelli

Q_5: What party does **he** belong to?
A_5: Republican
R_5: Republican Ken Cuccinelli

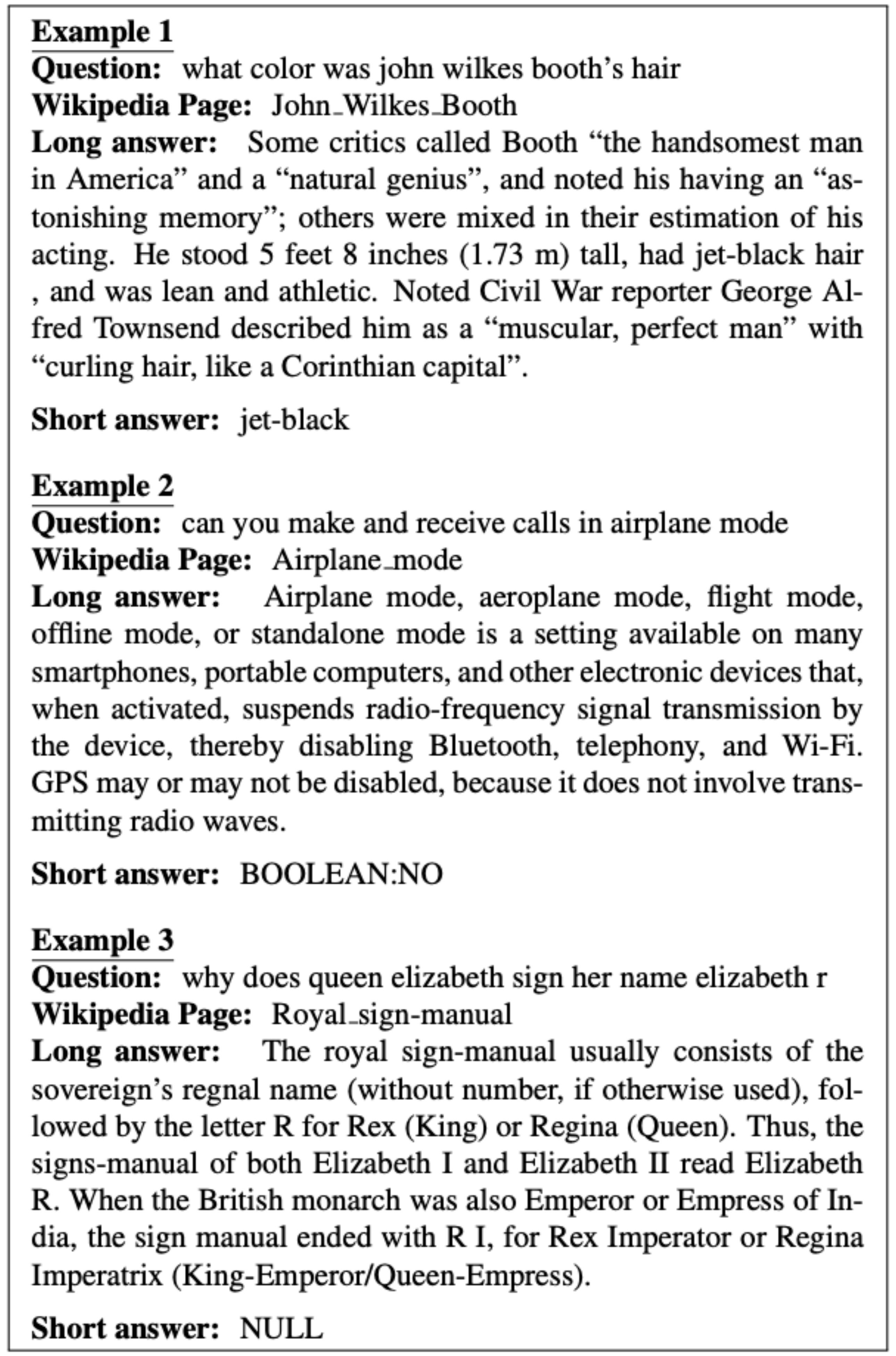
Q_6: Which of **them** is winning?
A_6: Terry McAuliffe
R_6: Democrat Terry McAuliffe, the longtime political fixer and moneyman, hasn't trailed in a poll since May

CoQA (Reddy et al, 2018)

Beyond SQUAD 1.1

Natural Questions

Real world queries to Google



Example 1
Question: what color was john wilkes booth's hair
Wikipedia Page: John_Wilkes_Booth
Long answer: Some critics called Booth "the handsomest man in America" and a "natural genius", and noted his having an "astonishing memory"; others were mixed in their estimation of his acting. He stood 5 feet 8 inches (1.73 m) tall, had jet-black hair , and was lean and athletic. Noted Civil War reporter George Alfred Townsend described him as a "muscular, perfect man" with "curling hair, like a Corinthian capital".

Short answer: jet-black

Example 2
Question: can you make and receive calls in airplane mode
Wikipedia Page: Airplane_mode
Long answer: Airplane mode, aeroplane mode, flight mode, offline mode, or standalone mode is a setting available on many smartphones, portable computers, and other electronic devices that, when activated, suspends radio-frequency signal transmission by the device, thereby disabling Bluetooth, telephony, and Wi-Fi. GPS may or may not be disabled, because it does not involve transmitting radio waves.

Short answer: BOOLEAN:NO

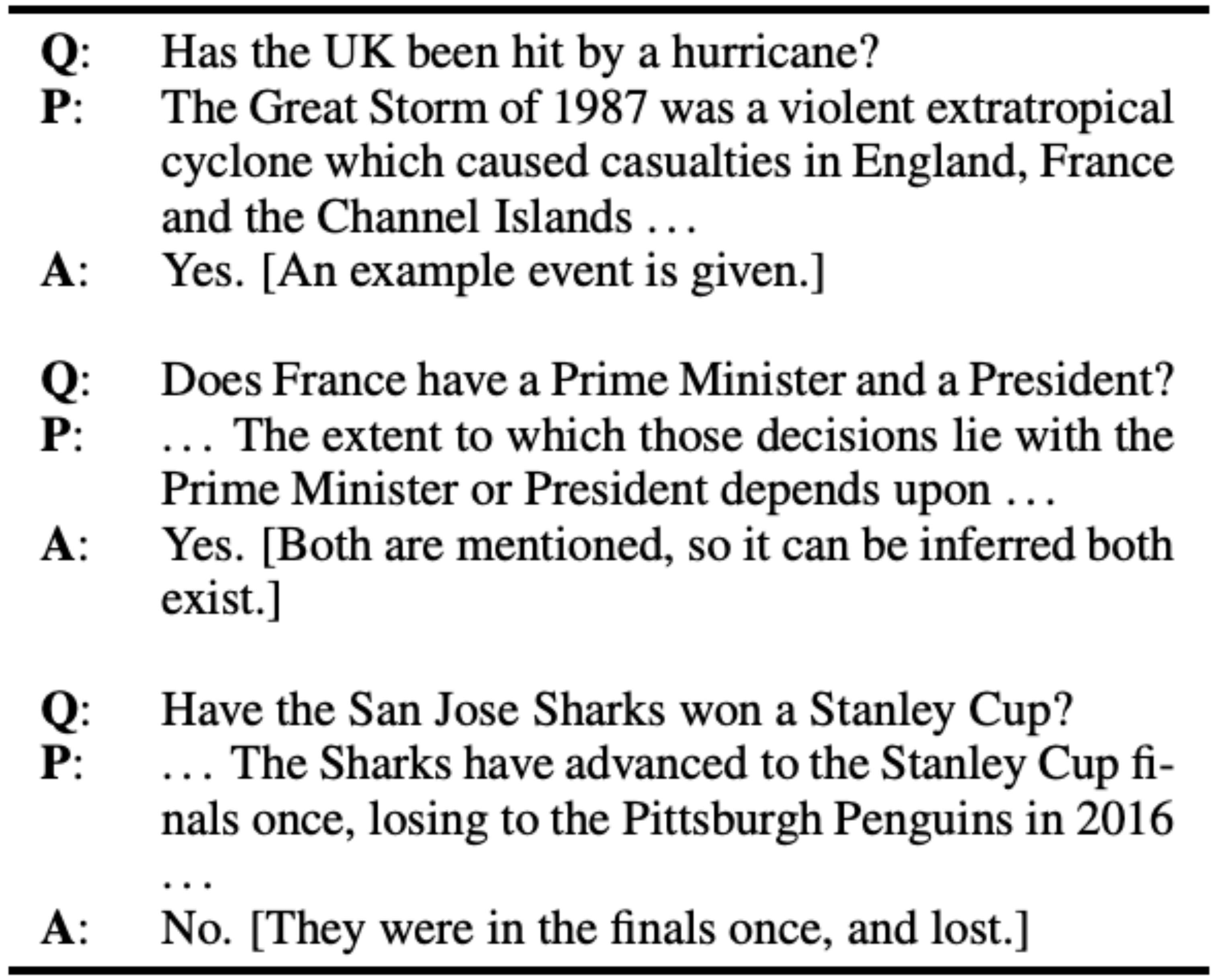
Example 3
Question: why does queen elizabeth sign her name elizabeth r
Wikipedia Page: Royal_sign-manual
Long answer: The royal sign-manual usually consists of the sovereign's regnal name (without number, if otherwise used), followed by the letter R for Rex (King) or Regina (Queen). Thus, the signs-manual of both Elizabeth I and Elizabeth II read Elizabeth R. When the British monarch was also Emperor or Empress of India, the sign manual ended with R I, for Rex Imperator or Regina Imperatrix (King-Emperor/Queen-Empress).

Short answer: NULL

(Kwiatkowski et al, 2019)

BoolQ

Hard yes/no questions from Google queries



Q: Has the UK been hit by a hurricane?
P: The Great Storm of 1987 was a violent extratropical cyclone which caused casualties in England, France and the Channel Islands . . .
A: Yes. [An example event is given.]

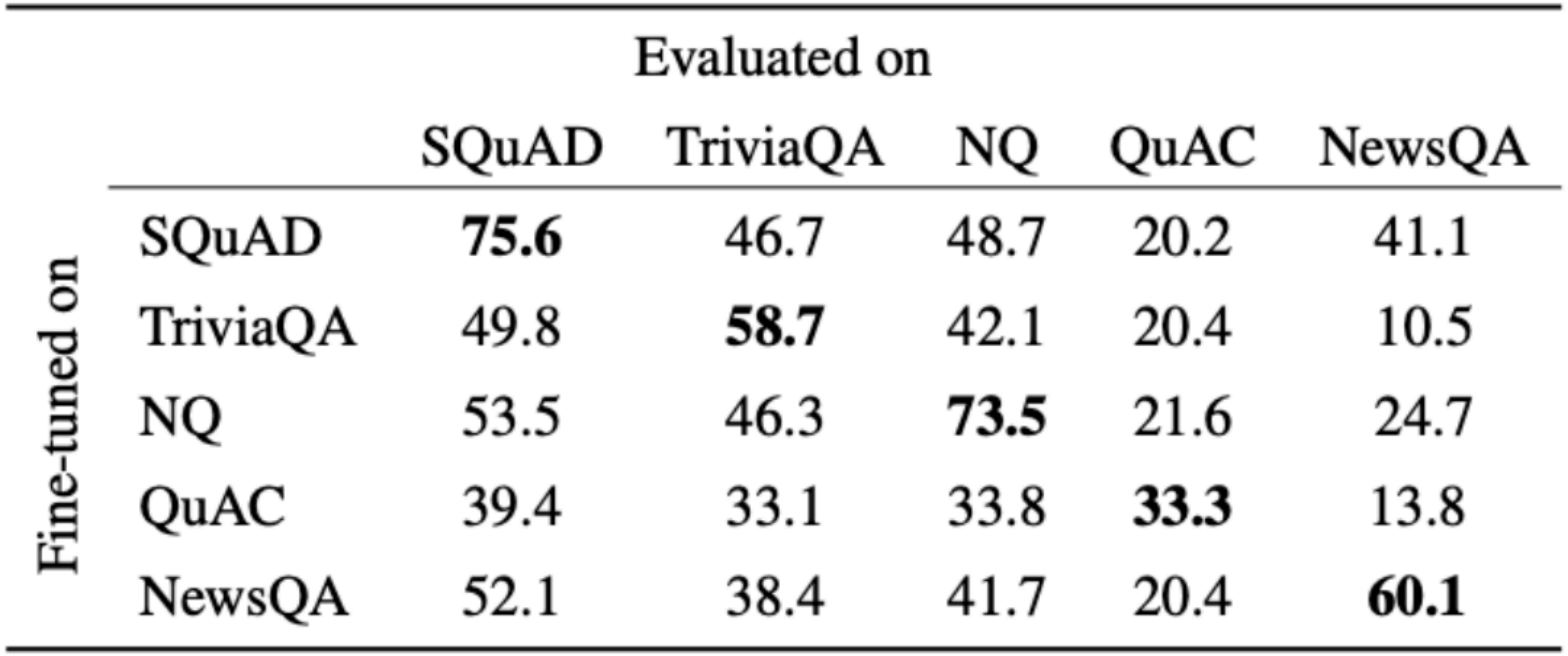
Q: Does France have a Prime Minister and a President?
P: . . . The extent to which those decisions lie with the Prime Minister or President depends upon . . .
A: Yes. [Both are mentioned, so it can be inferred both exist.]

Q: Have the San Jose Sharks won a Stanley Cup?
P: . . . The Sharks have advanced to the Stanley Cup finals once, losing to the Pittsburgh Penguins in 2016 . . .
A: No. [They were in the finals once, and lost.]

(Clark et al, 2019)

Is reading comprehension solved?

- System trained on one dataset can't generalize to other datasets

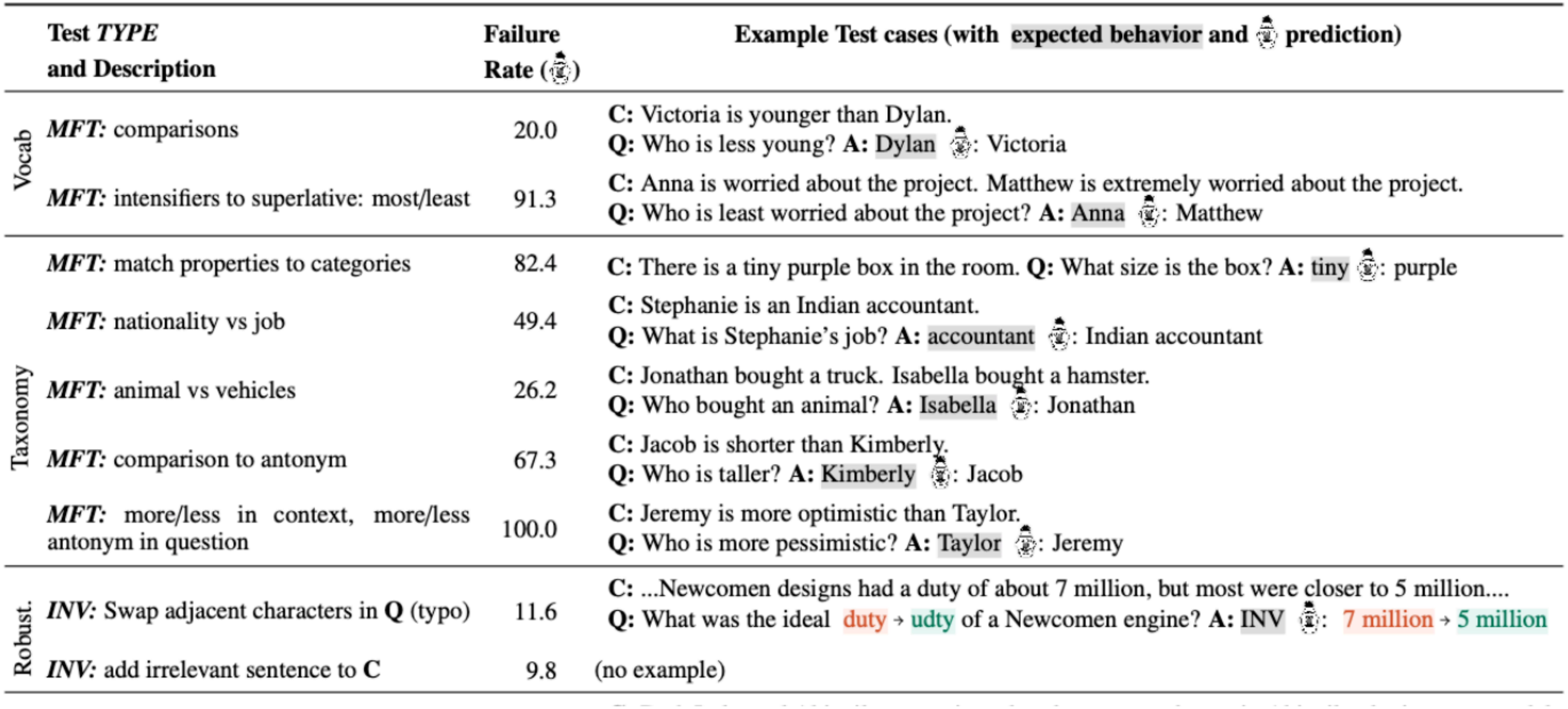


		Evaluated on				
		SQuAD	TriviaQA	NQ	QuAC	NewsQA
Fine-tuned on	SQuAD	**75.6**	46.7	48.7	20.2	41.1
	TriviaQA	49.8	**58.7**	42.1	20.4	10.5
	NQ	53.5	46.3	**73.5**	21.6	24.7
	QuAC	39.4	33.1	33.8	**33.3**	13.8
	NewsQA	52.1	38.4	41.7	20.4	**60.1**

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

Is reading comprehension solved?

BERT-large model trained on SQuAD

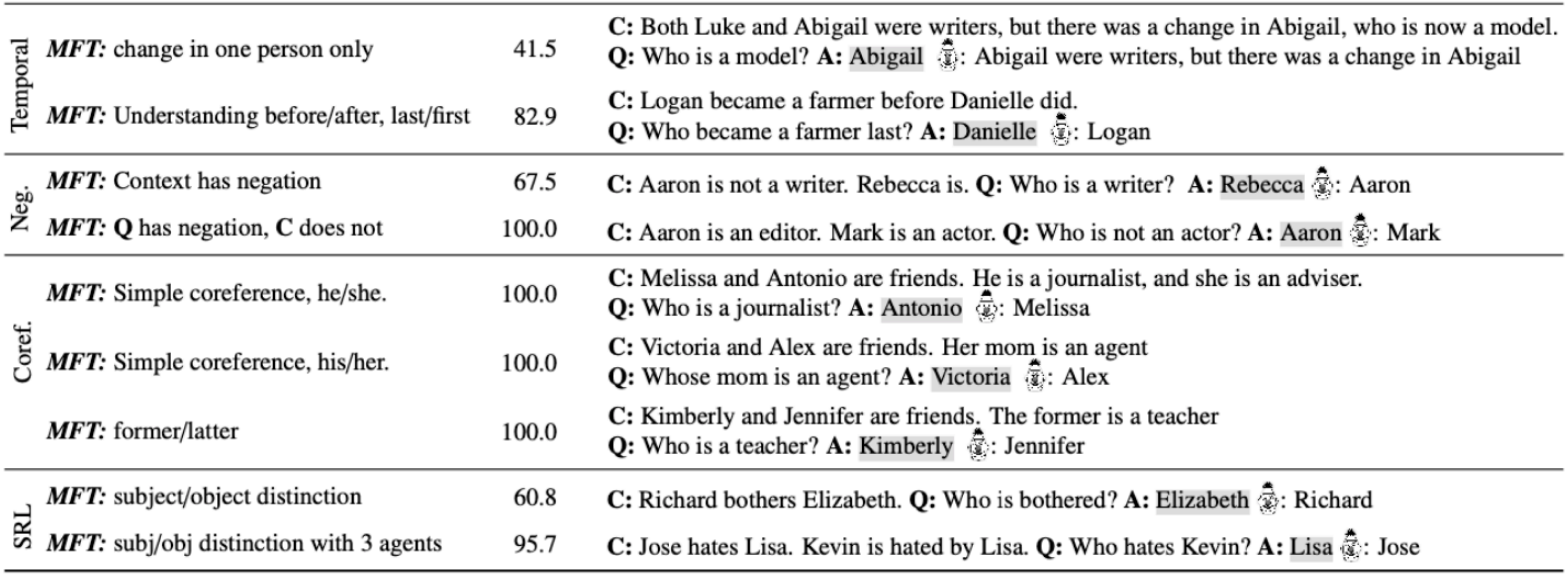


	Test *TYPE* and Description	Failure Rate (🤖)	Example Test cases (with expected behavior and 🤖 prediction)
Vocab	***MFT:*** comparisons	20.0	**C:** Victoria is younger than Dylan. **Q:** Who is less young? **A:** Dylan 🤖: Victoria
	MFT: intensifiers to superlative: most/least	91.3	**C:** Anna is worried about the project. Matthew is extremely worried about the project. **Q:** Who is least worried about the project? **A:** Anna 🤖: Matthew
Taxonomy	***MFT:*** match properties to categories	82.4	**C:** There is a tiny purple box in the room. **Q:** What size is the box? **A:** tiny 🤖: purple
	MFT: nationality vs job	49.4	**C:** Stephanie is an Indian accountant. **Q:** What is Stephanie's job? **A:** accountant 🤖: Indian accountant
	MFT: animal vs vehicles	26.2	**C:** Jonathan bought a truck. Isabella bought a hamster. **Q:** Who bought an animal? **A:** Isabella 🤖: Jonathan
	MFT: comparison to antonym	67.3	**C:** Jacob is shorter than Kimberly. **Q:** Who is taller? **A:** Kimberly 🤖: Jacob
	MFT: more/less in context, more/less antonym in question	100.0	**C:** Jeremy is more optimistic than Taylor. **Q:** Who is more pessimistic? **A:** Taylor 🤖: Jeremy
Robust.	***INV:*** Swap adjacent characters in **Q** (typo)	11.6	**C:** ...Newcomen designs had a duty of about 7 million, but most were closer to 5 million.... **Q:** What was the ideal duty → udty of a Newcomen engine? **A:** INV 🤖: 7 million → 5 million
	INV: add irrelevant sentence to **C**	9.8	(no example)

(Ribeiro et al., 2020): Beyond Accuracy: Behavioral Testing of NLP Models with CheckList

Is reading comprehension solved?

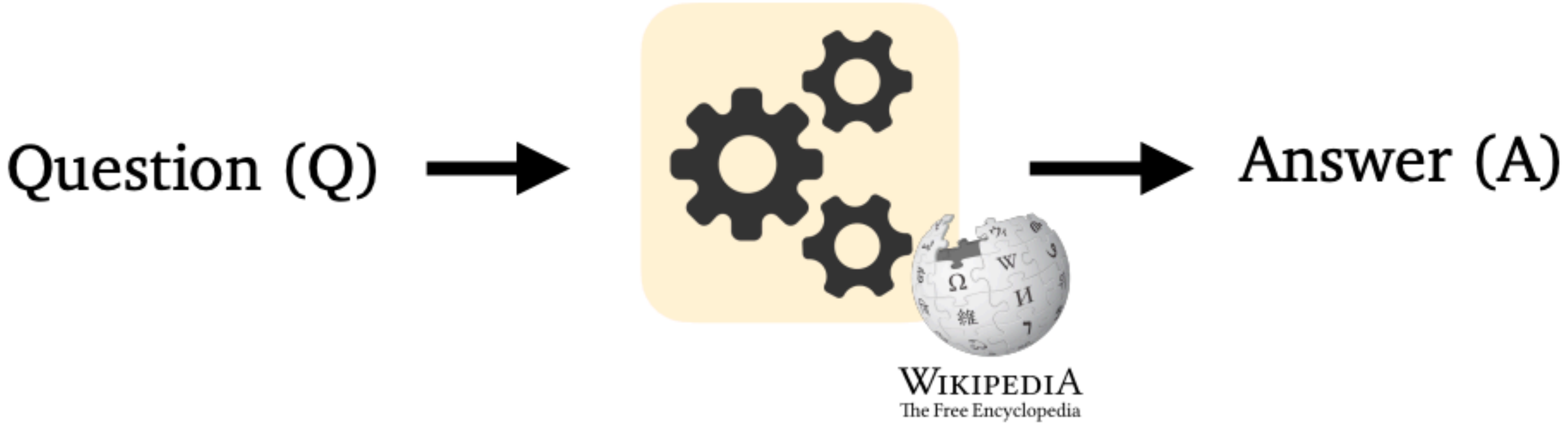
BERT-large model trained on SQuAD



Temporal	***MFT:*** change in one person only	41.5	**C:** Both Luke and Abigail were writers, but there was a change in Abigail, who is now a model. **Q:** Who is a model? **A:** Abigail 🤖: Abigail were writers, but there was a change in Abigail
	MFT: Understanding before/after, last/first	82.9	**C:** Logan became a farmer before Danielle did. **Q:** Who became a farmer last? **A:** Danielle 🤖: Logan
Neg.	***MFT:*** Context has negation	67.5	**C:** Aaron is not a writer. Rebecca is. **Q:** Who is a writer? **A:** Rebecca 🤖: Aaron
	MFT: **Q** has negation, **C** does not	100.0	**C:** Aaron is an editor. Mark is an actor. **Q:** Who is not an actor? **A:** Aaron 🤖: Mark
Coref.	***MFT:*** Simple coreference, he/she.	100.0	**C:** Melissa and Antonio are friends. He is a journalist, and she is an adviser. **Q:** Who is a journalist? **A:** Antonio 🤖: Melissa
	MFT: Simple coreference, his/her.	100.0	**C:** Victoria and Alex are friends. Her mom is an agent **Q:** Whose mom is an agent? **A:** Victoria 🤖: Alex
	MFT: former/latter	100.0	**C:** Kimberly and Jennifer are friends. The former is a teacher **Q:** Who is a teacher? **A:** Kimberly 🤖: Jennifer
SRL	***MFT:*** subject/object distinction	60.8	**C:** Richard bothers Elizabeth. **Q:** Who is bothered? **A:** Elizabeth 🤖: Richard
	MFT: subj/obj distinction with 3 agents	95.7	**C:** Jose hates Lisa. Kevin is hated by Lisa. **Q:** Who hates Kevin? **A:** Lisa 🤖: Jose

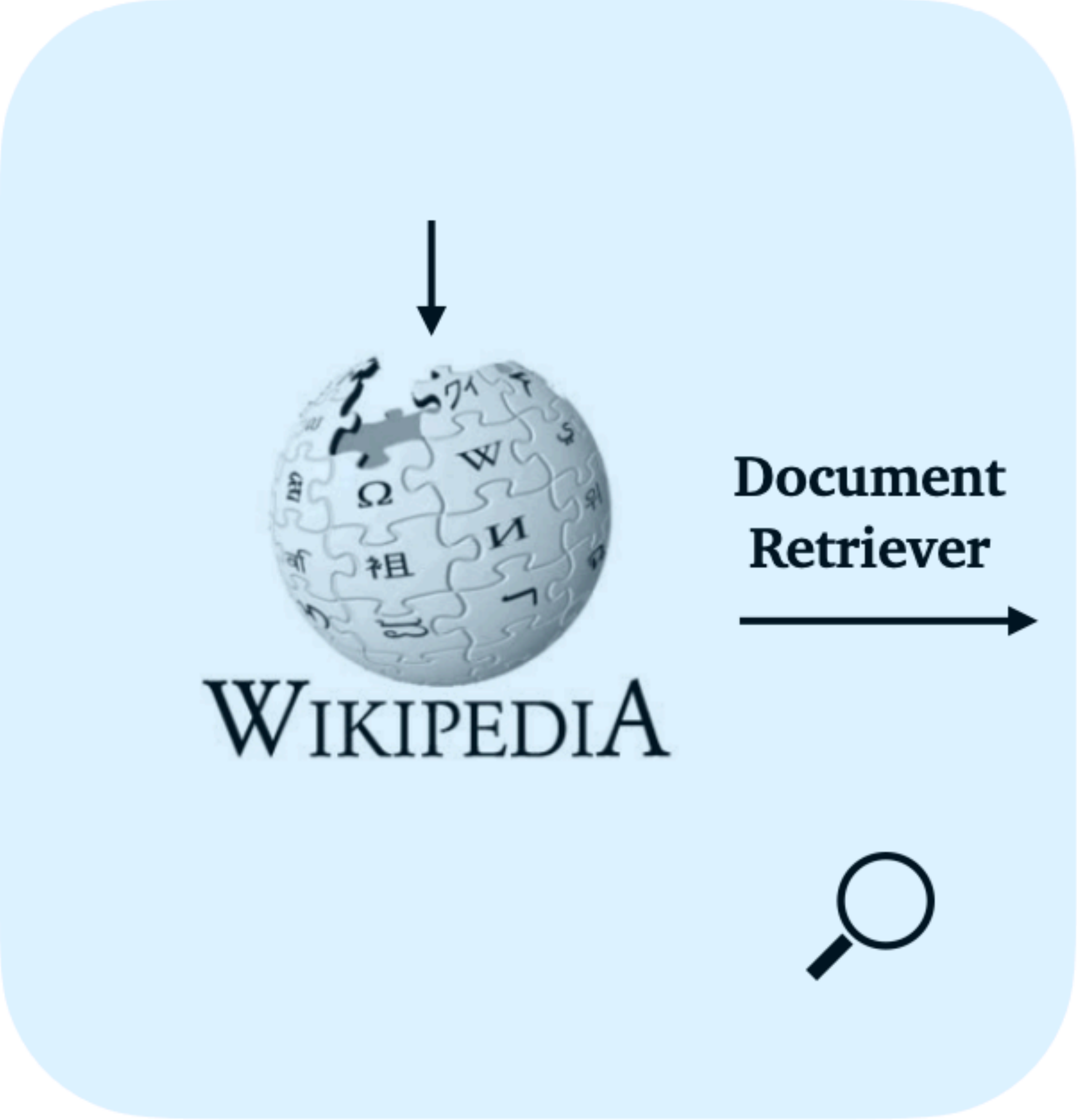
(Ribeiro et al., 2020): Beyond Accuracy: Behavioral Testing of NLP Models with CheckList

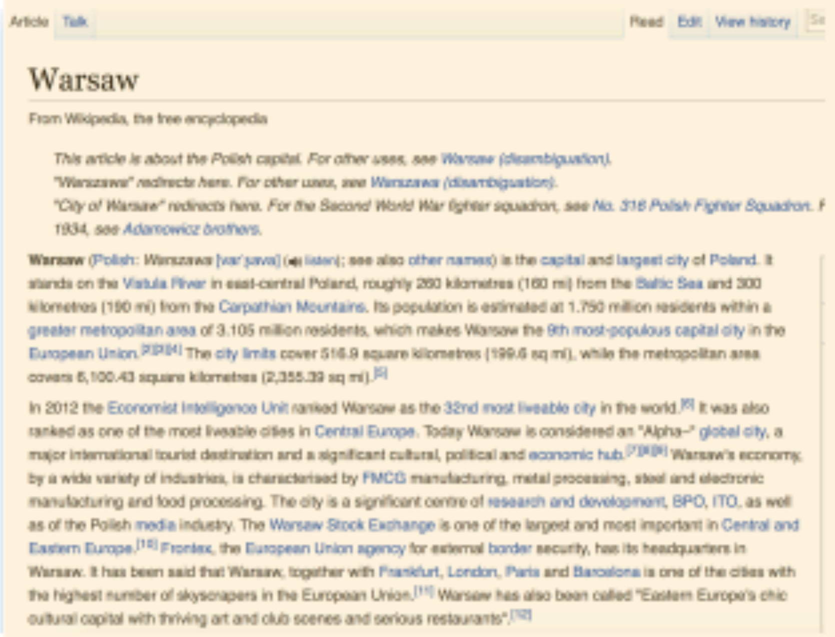
Open domain question answering



- Different from reading comprehension, we don't assume a given passage. Question (Q) Answer (A)
- Instead, we only have access to a large collection of documents (e.g., Wikipedia). We don't know where the answer is located, and the goal is to return the answer for any open-domain questions.
- Much more challenging but a more practical problem!

Retrieve and read





Article Talk Read Edit View history

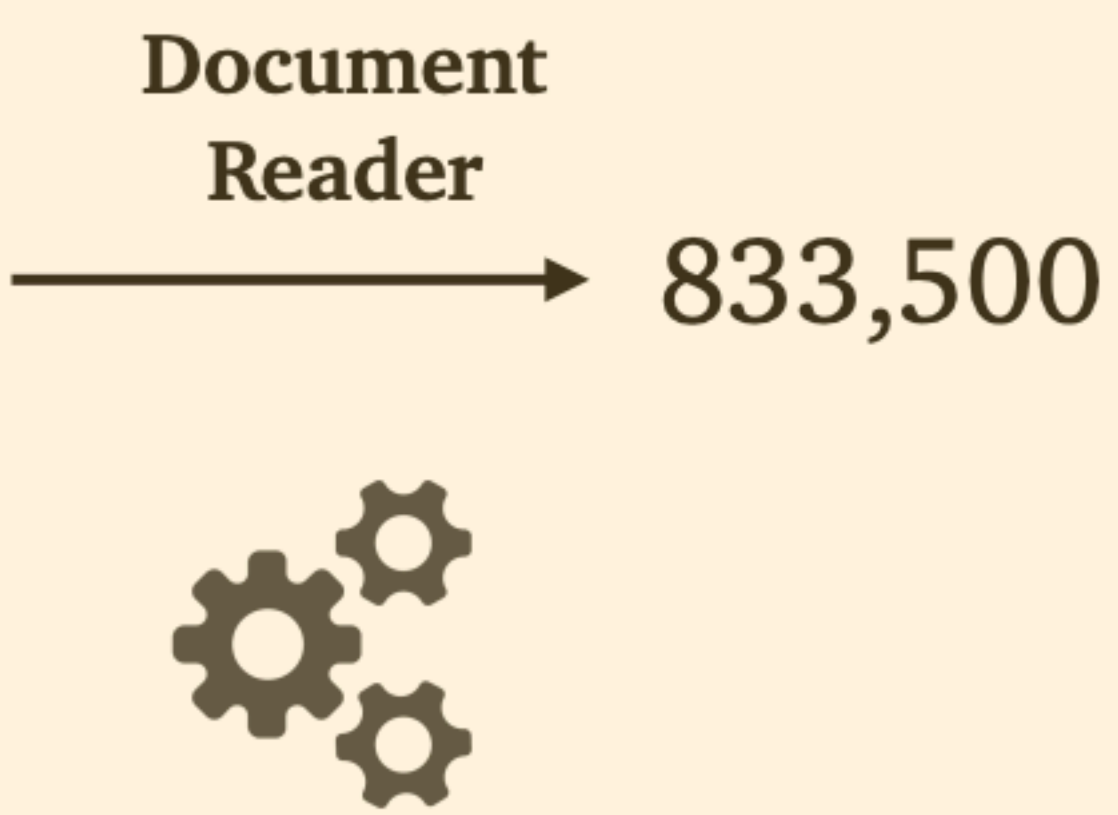
Warsaw

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]

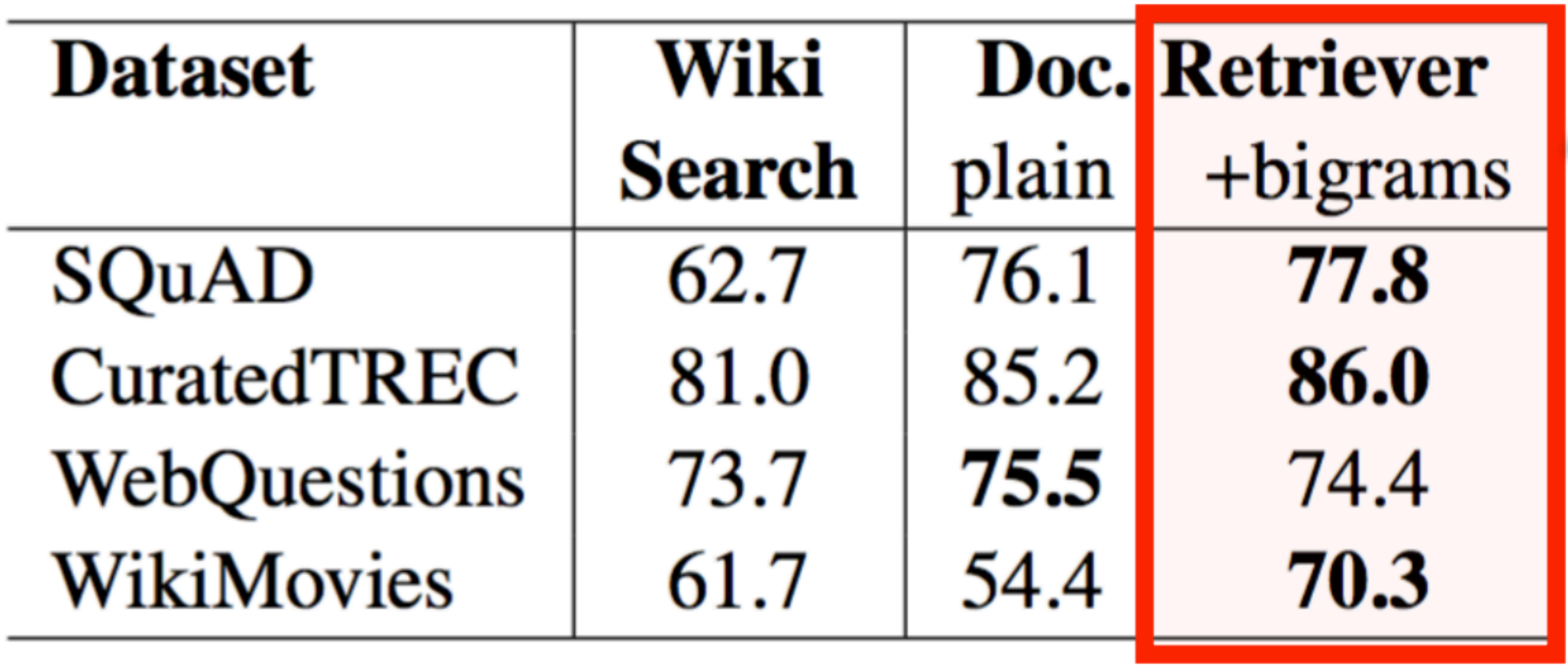
In 2012 the Economist Intelligence Unit ranked Warsaw as the 32nd most liveable city in the world.[6] It was also ranked as one of the most liveable cities in Central Europe. Today Warsaw is considered an "Alpha–" global city, a major international tourist destination and a significant cultural, political 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]



https://github.com/facebookresearch/DrQA

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

DrQA: Document Retrieval



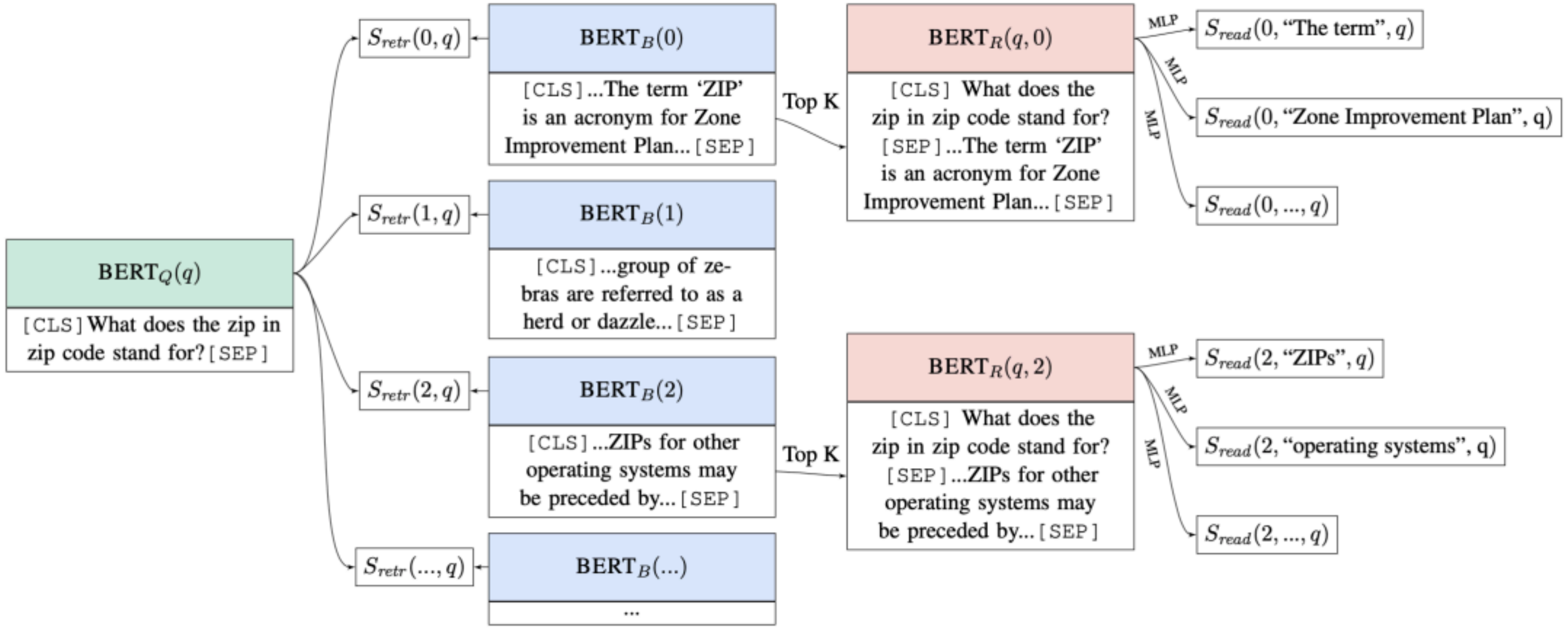
Dataset	**Wiki Search**	**Doc. Retriever** plain	+bigrams
SQuAD	62.7	76.1	**77.8**
CuratedTREC	81.0	85.2	**86.0**
WebQuestions	73.7	**75.5**	74.4
WikiMovies	61.7	54.4	**70.3**

Traditional tf.idf inverted index + efficient bigram hash

For **70–86%** of questions, the answer segment appears in the top 5 articles

Slide credit: Chris Manning

Joint training of retriever and reader

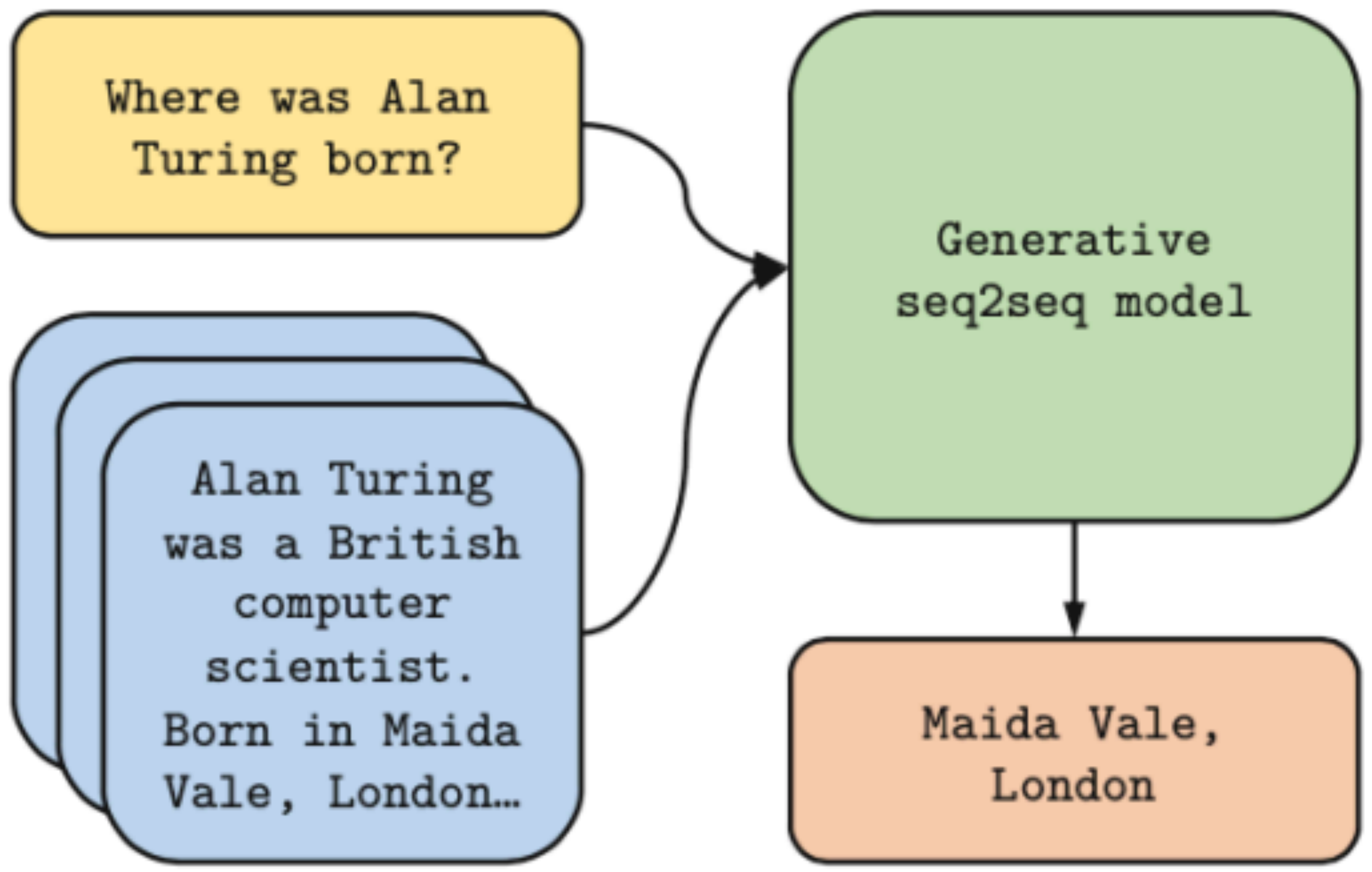

$\text{BERT}_Q(q)$
[CLS] What does the zip in zip code stand for? [SEP]
$S_{retr}(0, q)$
$\text{BERT}_B(0)$
[CLS]...The term 'ZIP' is an acronym for Zone Improvement Plan... [SEP]
$S_{retr}(1, q)$
$\text{BERT}_B(1)$
[CLS]...group of zebras are referred to as a herd or dazzle... [SEP]
$S_{retr}(2, q)$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

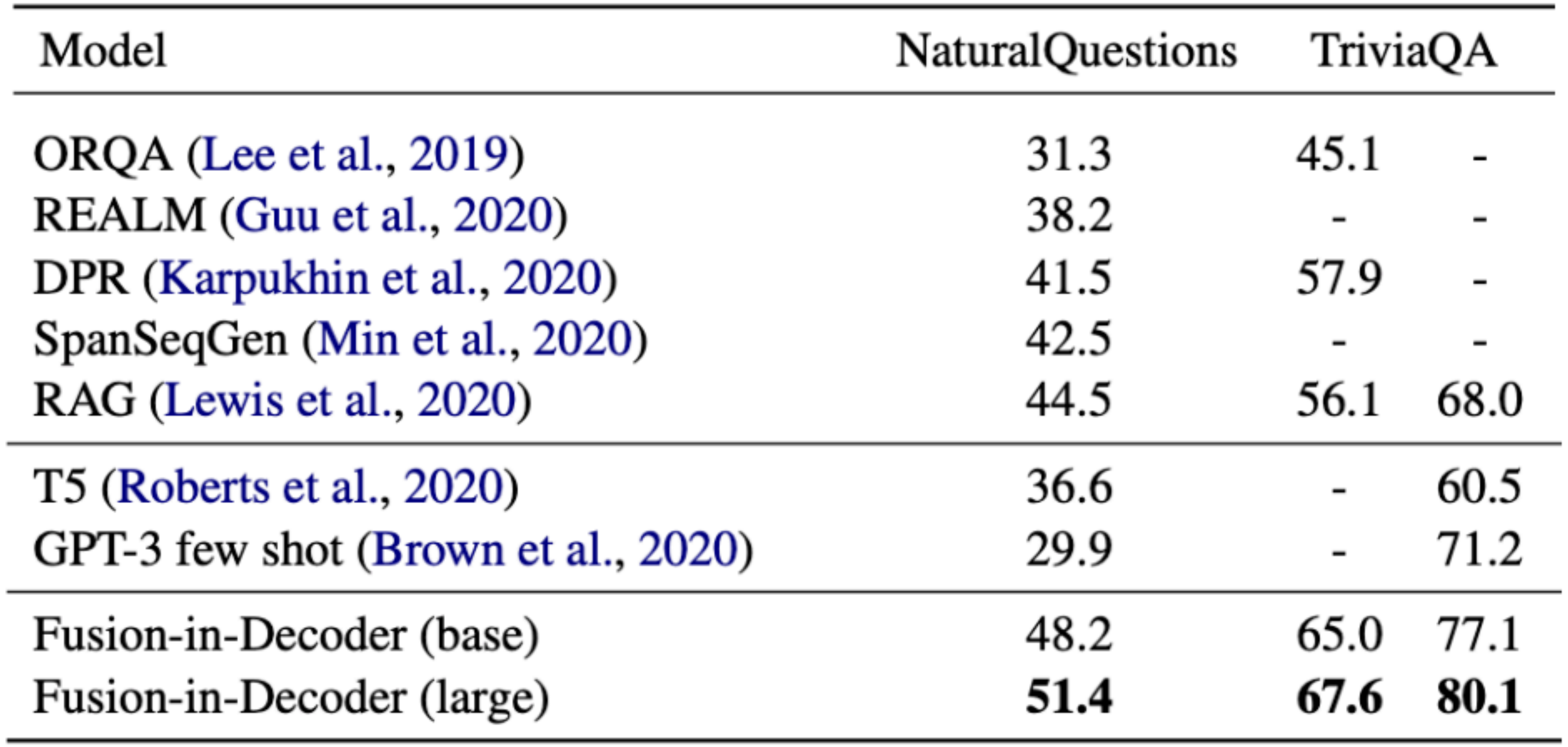
- Each text passage can be encoded as a vector using BERT and the retriever score can be measured as the dot product between the question representation and passage representation.
- However, it is not easy to model as there are a huge number of passages (e.g., 21M in English Wikipedia)

Lee et al., 2019. Latent Retrieval for Weakly Supervised Open Domain Question Answering

Dense retrieval + generate answers

Fusion-in-decoder (FID) = DPR + T5


Where was Alan Turing born?
Alan Turing was a British computer scientist. Born in Maida Vale, London…
Generative seq2seq model
Maida Vale, London



Model	NaturalQuestions	TriviaQA	
ORQA (Lee et al., 2019)	31.3	45.1	-
REALM (Guu et al., 2020)	38.2	-	-
DPR (Karpukhin et al., 2020)	41.5	57.9	-
SpanSeqGen (Min et al., 2020)	42.5	-	-
RAG (Lewis et al., 2020)	44.5	56.1	68.0
T5 (Roberts et al., 2020)	36.6	-	60.5
GPT-3 few shot (Brown et al., 2020)	29.9	-	71.2
Fusion-in-Decoder (base)	48.2	65.0	77.1
Fusion-in-Decoder (large)	**51.4**	**67.6**	**80.1**

Izacard and Grave 2020. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering

Summary

- Many different types of question answering
- Reading comprehension
  - Given passage + question, come up with answer
  - SQuAD: answer is span of text in passage
  - Train classifier to predict span
- Reading comprehension is not solved!
- Lots of ongoing work on QA!