CMPT 825: Natural Language Processing

Question Answering

Spring 2020
2020-04-02

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

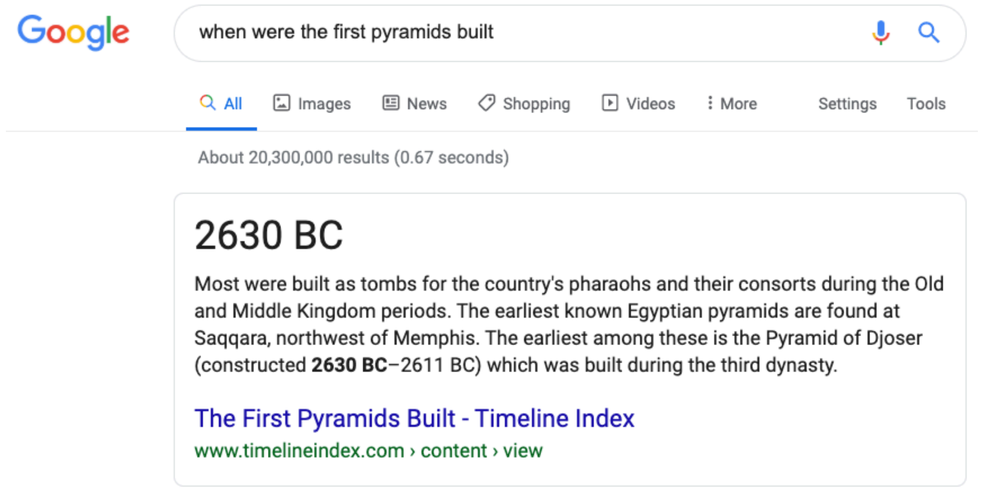
Question Answering

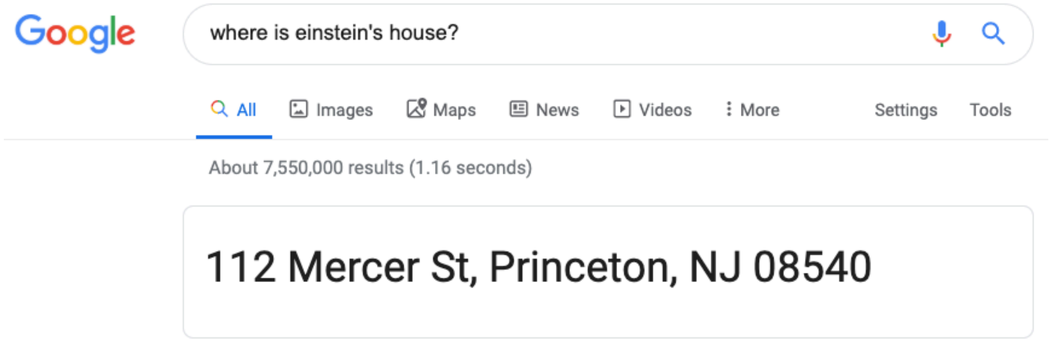
- Goal: build computer systems to answer questions

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

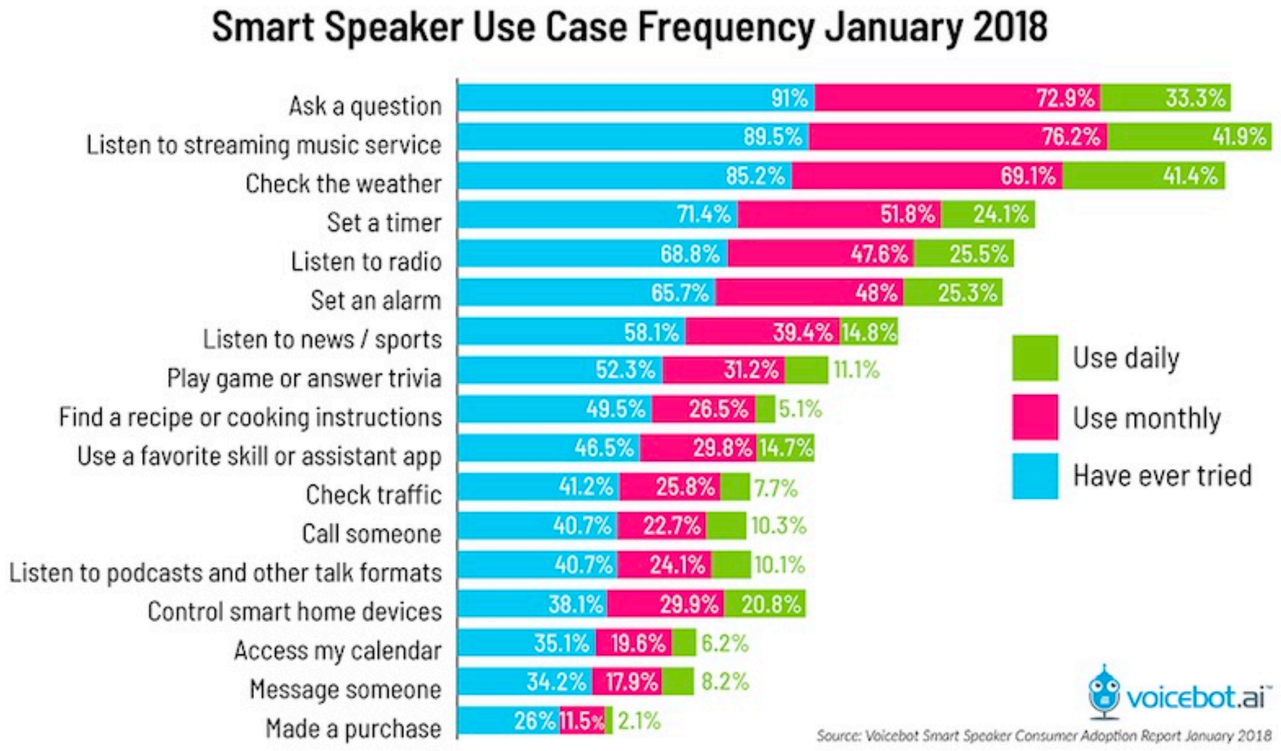
- You can easily find these answers in google today!





Question Answering

- People ask lots of questions to Digital Personal Assistants:

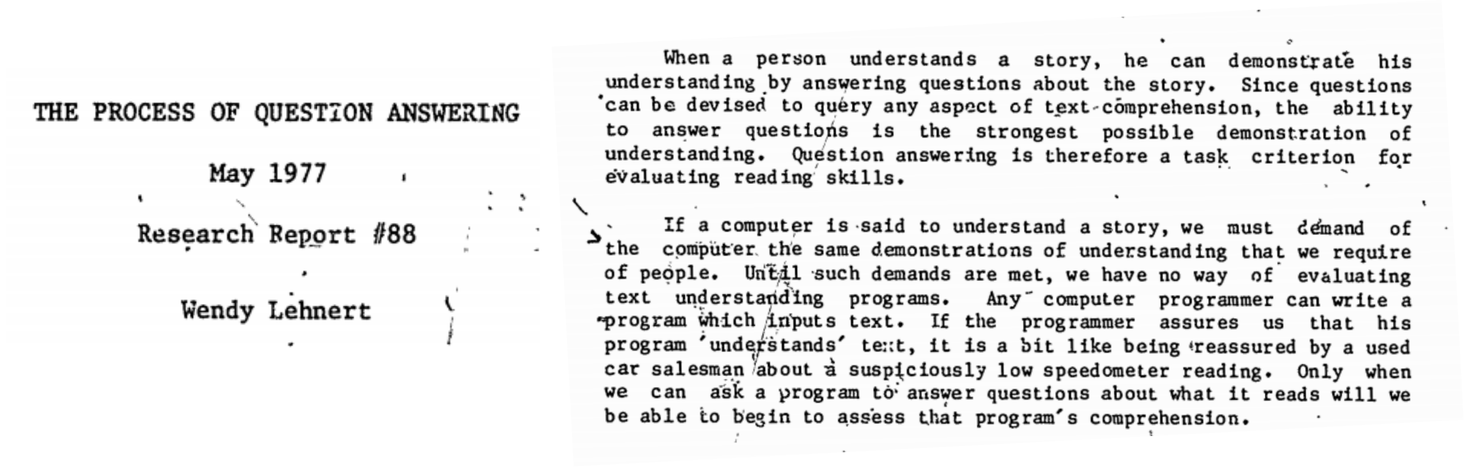


Question Answering

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

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**."

QA Taxonomy

- Factoid questions vs non-factoid questions
- Answers
  - A short span of text
  - A paragraph
  - Yes/No
  - A database entry
  - A list
- Context
  - A passage, a document, a large collection of documents
  - Knowledge base
  - Semi-structured tables
  - Images

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?

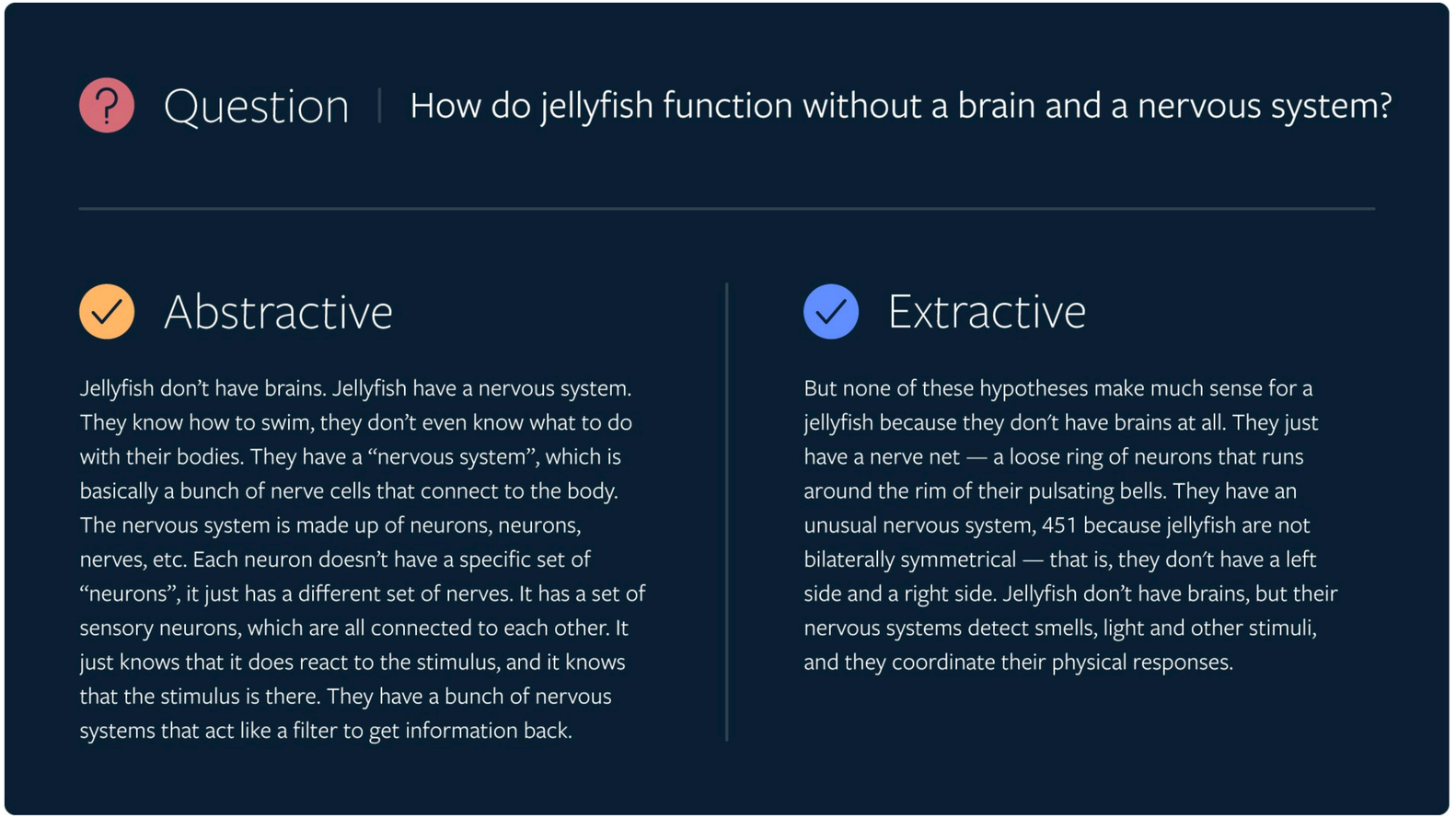
A: Republican

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



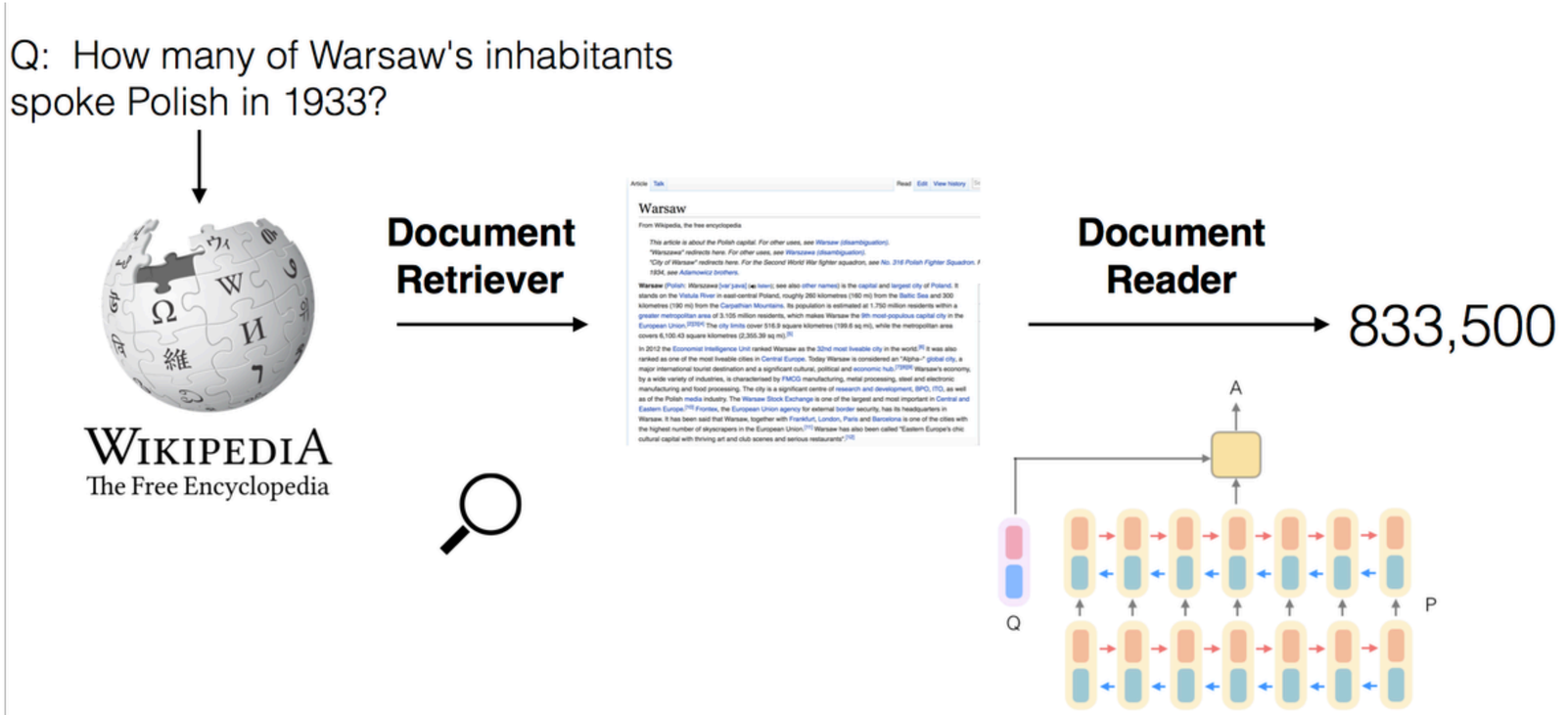
Extractive:
Select excerpts (extracts) and concatenate them to form the answer.

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)

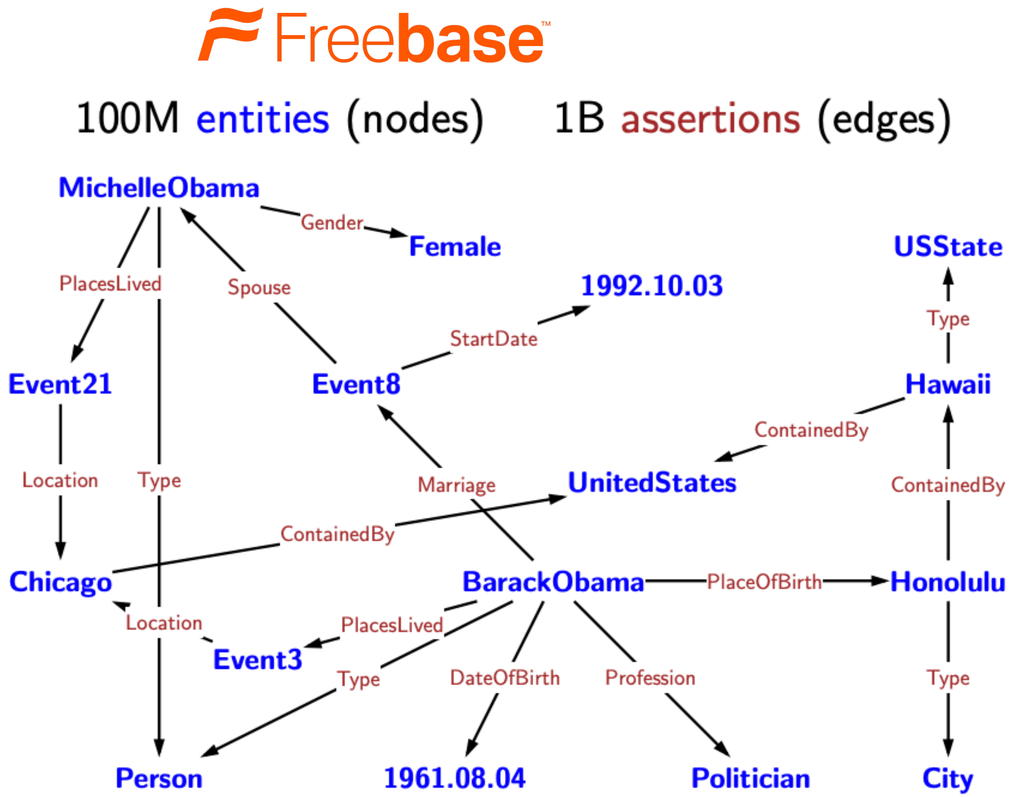



```
>>> process('What is the answer to life, the universe, and everything?')

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



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

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, 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.

3 gold answers are collected for each question

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

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.

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.

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

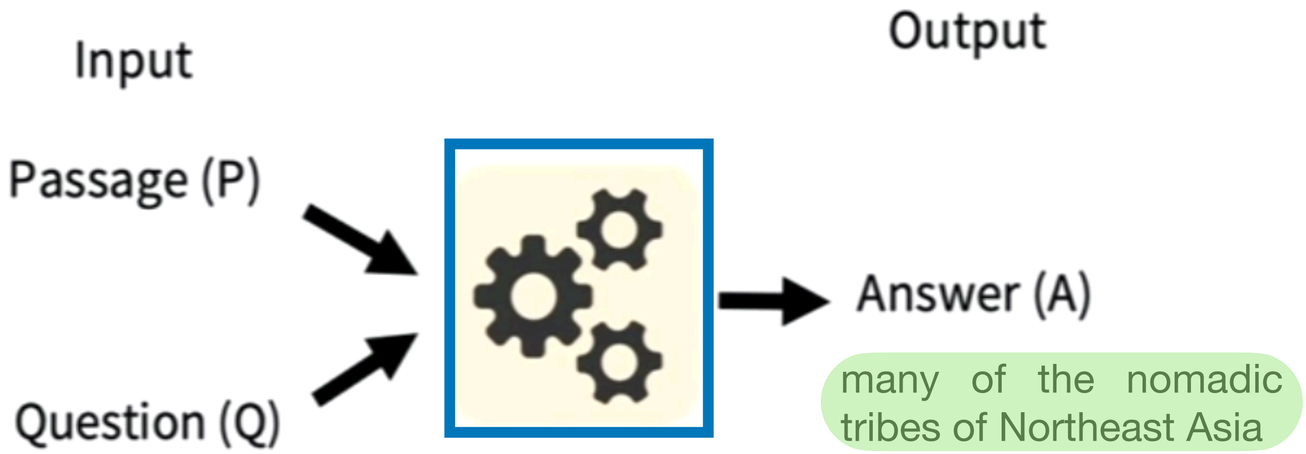
A: {tuition, charging their students tuition, tuition}

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

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

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

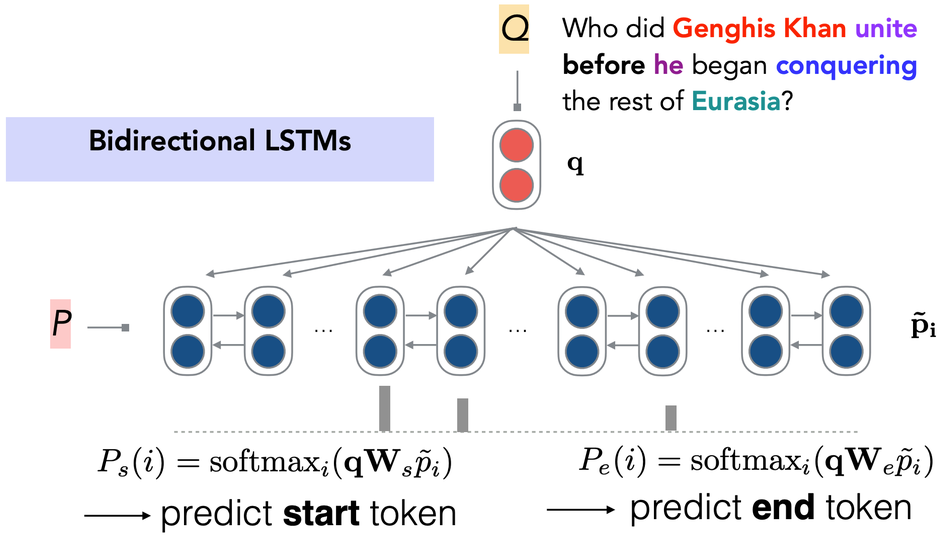
- Define a feature vector $\phi(p, q, a_i) \in \mathbb{R}^d$:
  - Word/bigram frequencies
  - Parse tree matches
  - Dependency labels, length, part-of-speech tags

- Apply a (multi-class) logistic regression model

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

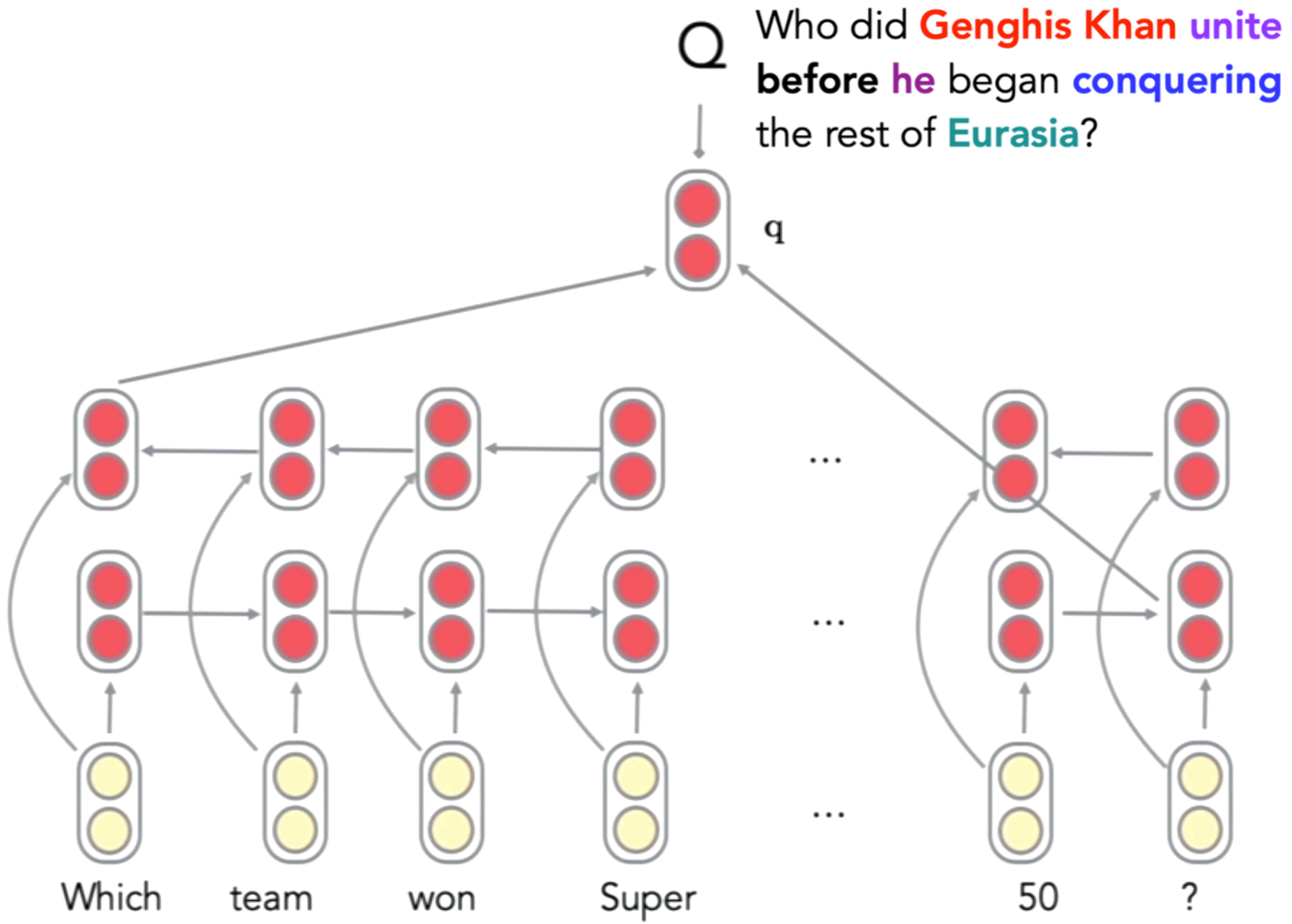
Stanford Attentive Reader (Chen, Bolten, and Manning, 2016)

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



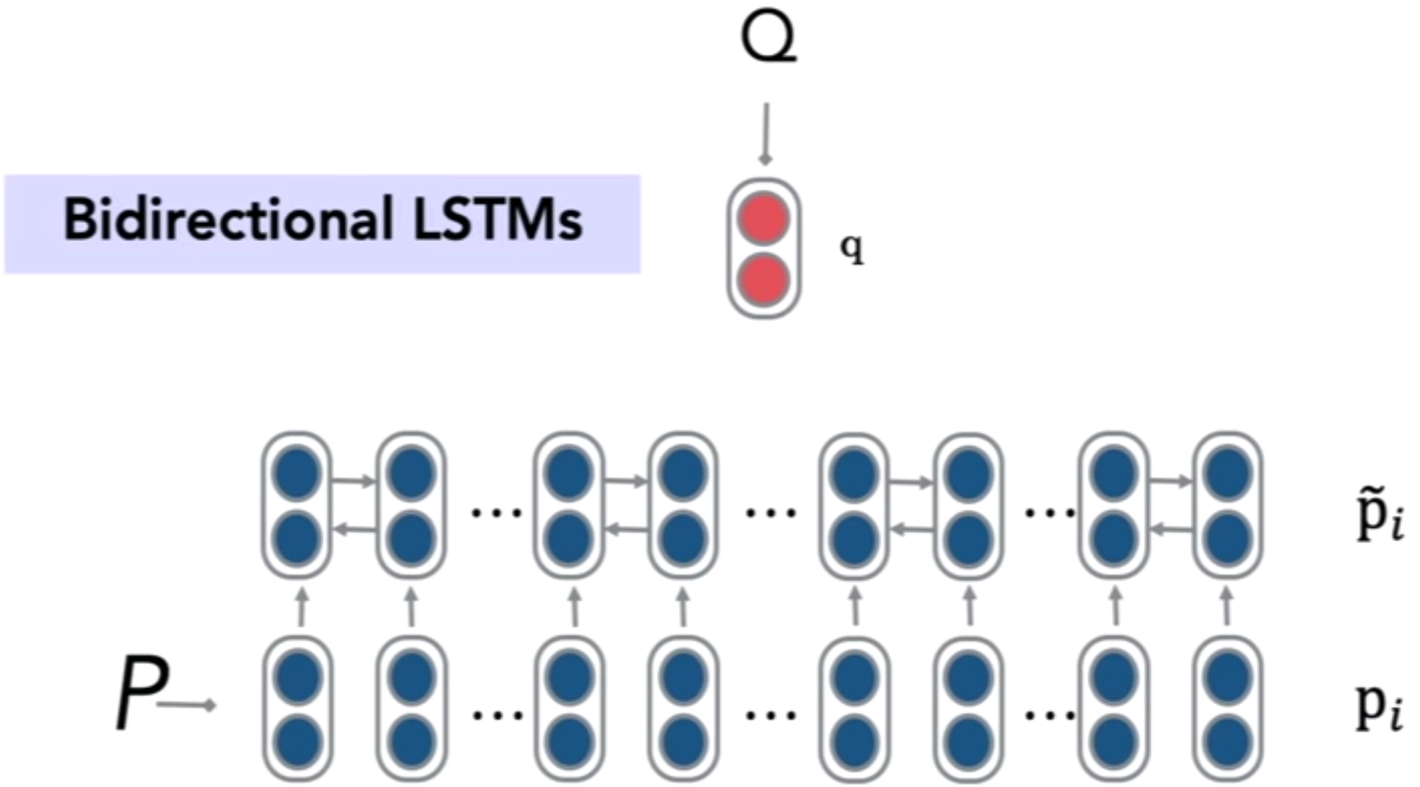
Also used in DrQA (Chen et al, 2017)

Stanford Attentive Reader Question Encoder



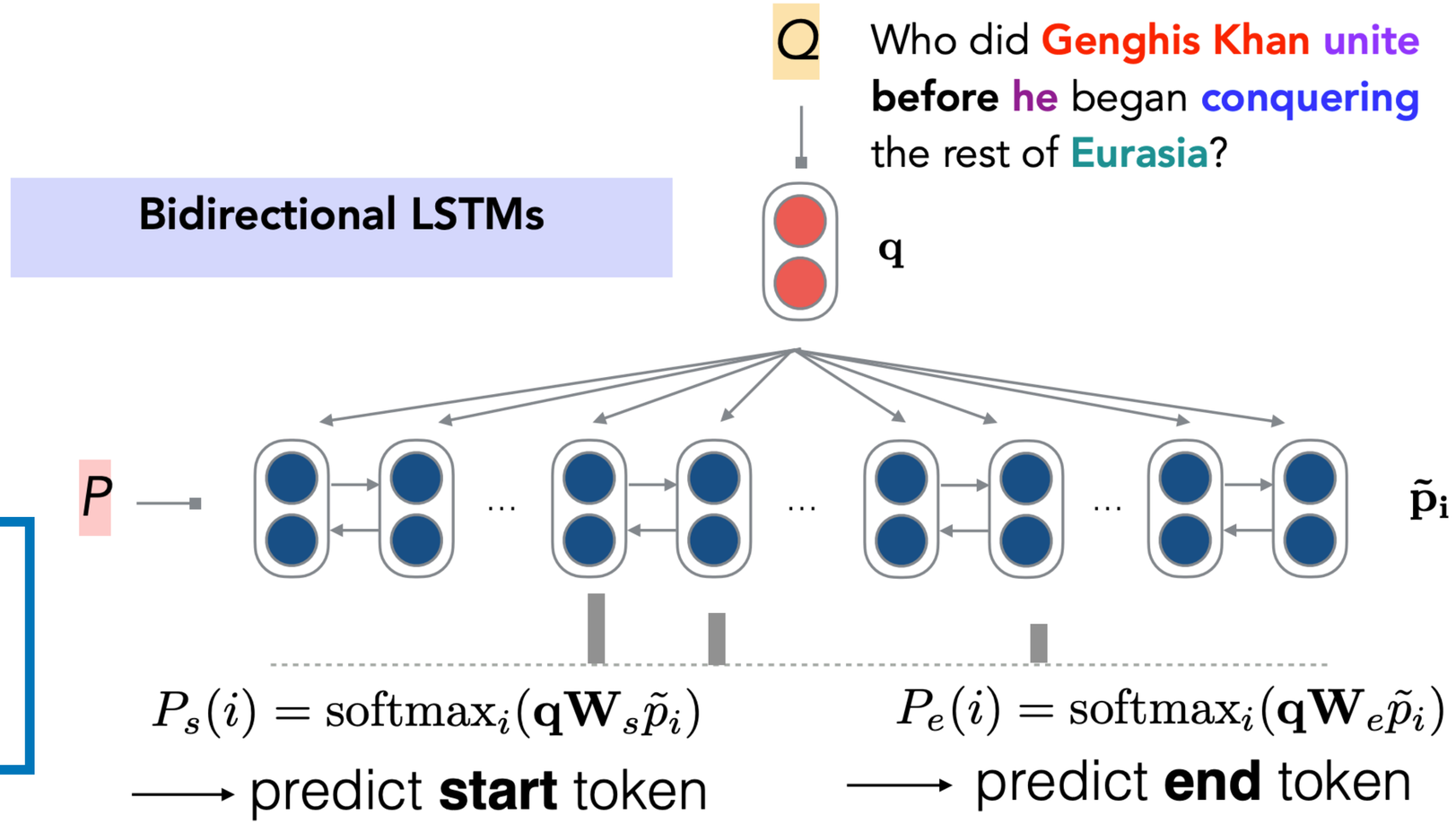
Slide credit: Chris Manning

Stanford Attentive Reader
Passage encoder



Slide credit: Chris Manning

Stanford Attentive Reader



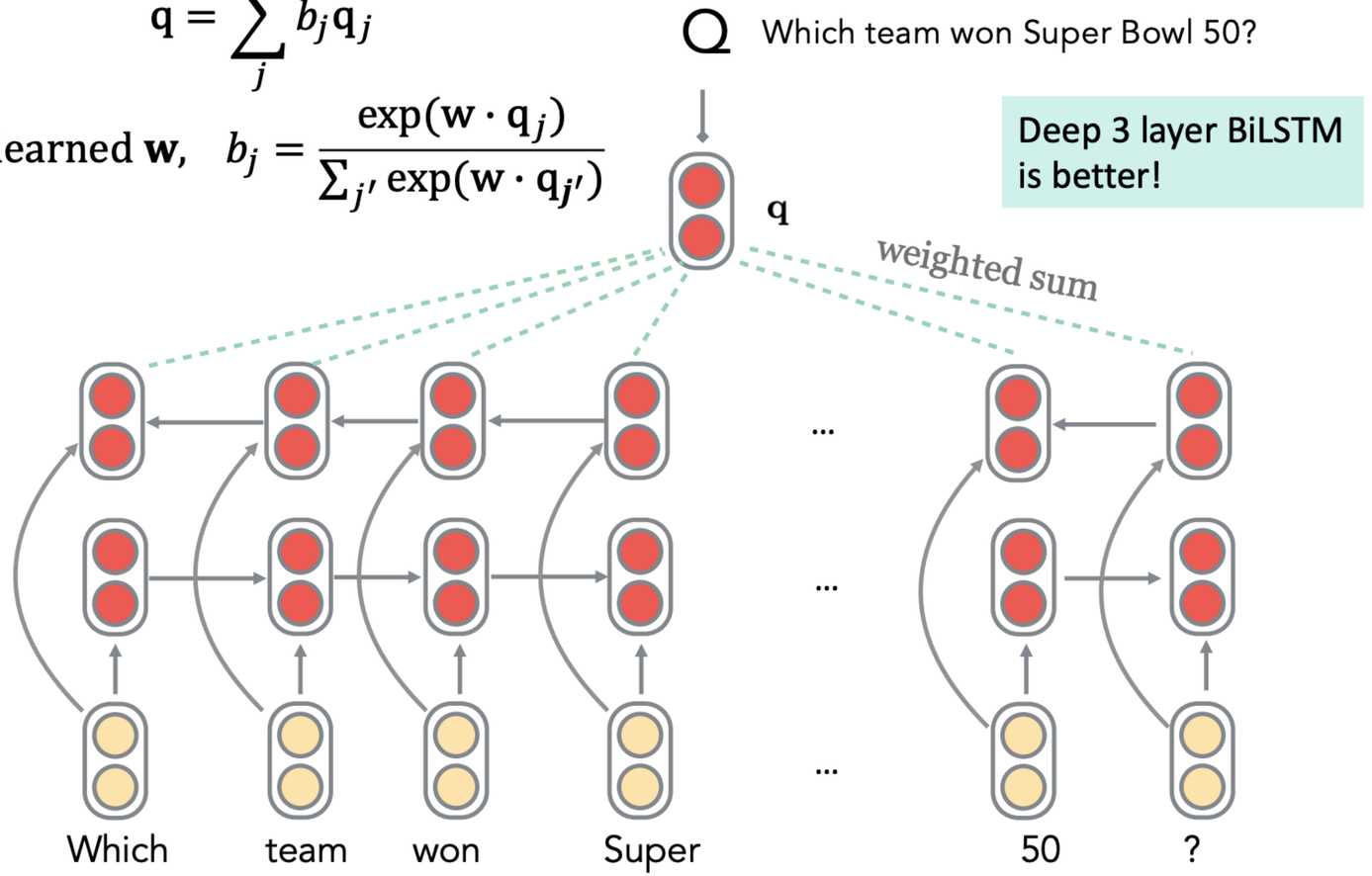
Use **attention** to predict span

Stanford Attentive Reader++

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

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

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



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")

Improved passage word/position representations

Matching of words in the question to words in the passage

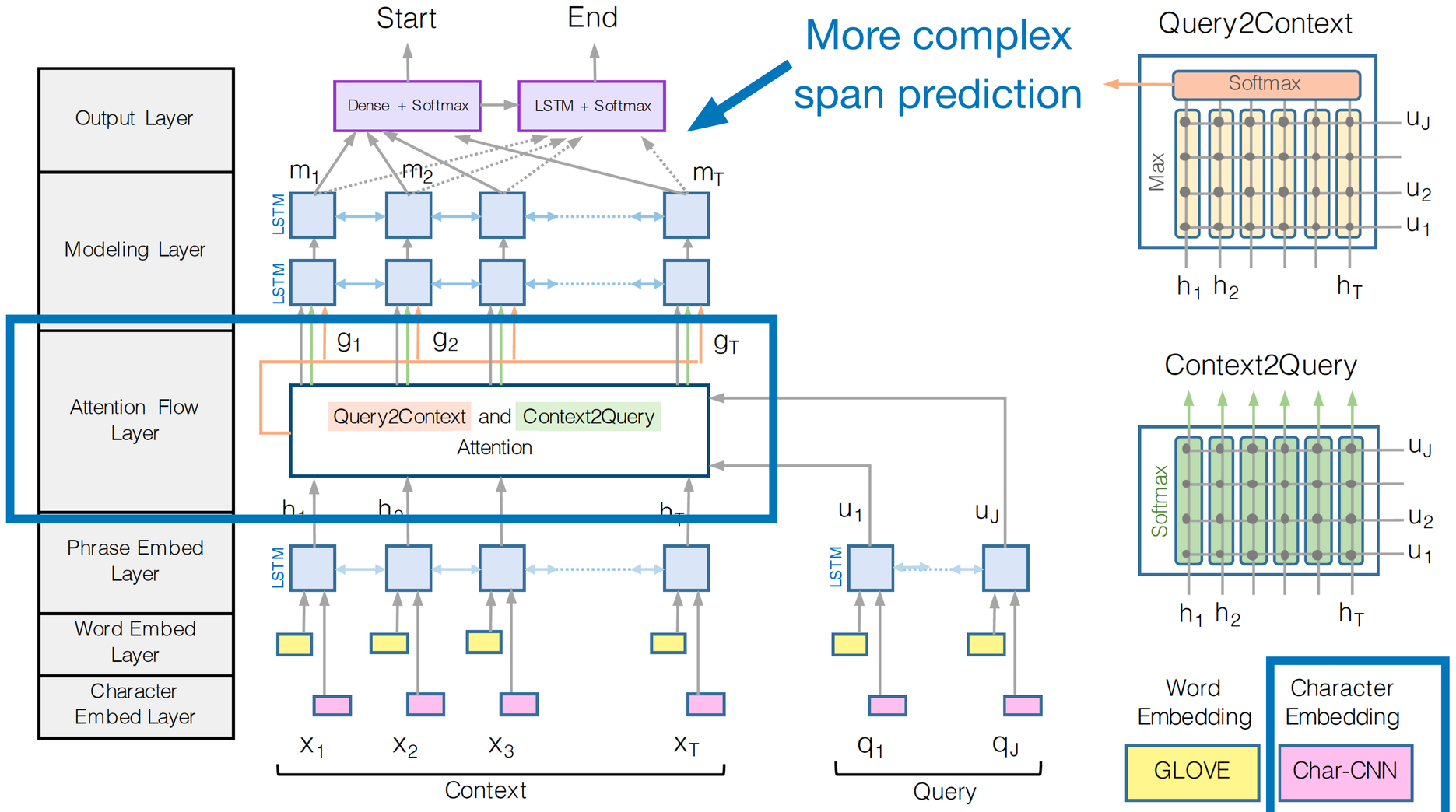
$$f_{align}(p_i) = \sum_j a_{i,j}\mathbf{E}(q_j) \qquad q_{i,j} = \frac{\exp(\alpha(\mathbf{E}(p_i)) \cdot \alpha(\mathbf{E}(q_j)))}{\sum_{j'} \exp(\alpha(\mathbf{E}(p_i)) \cdot \alpha(\mathbf{E}(q'_j)))}$$

Where α is a simple one layer FFNN

Slide credit: Chris Manning

BiDAF

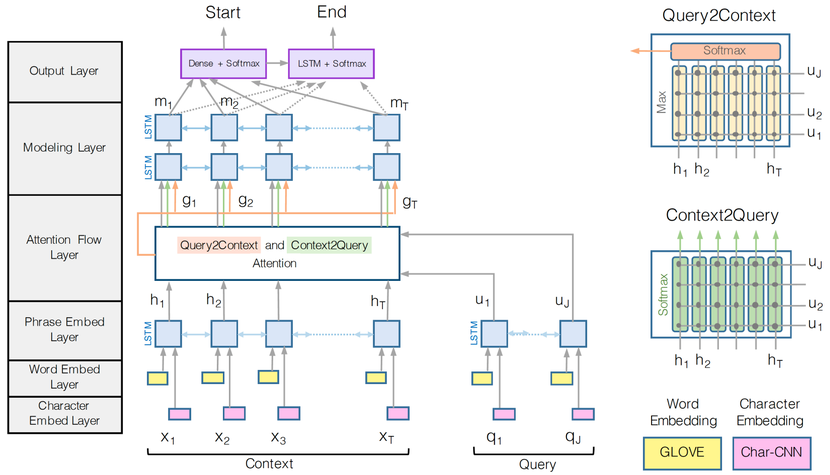
Attention flowing between question (query) and passage (context)



(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

- There are variants of and improvements to the BiDAF architecture over the years, but the central idea is **the Attention Flow layer**
- **Idea:** attention should flow both ways – from the context to the question and from the question to the context
- Make similarity matrix (with **w** of dimension 6*d*):

$$\boldsymbol{S}_{ij} = \boldsymbol{w}_{\text{sim}}^T[\boldsymbol{c}_i; \boldsymbol{q}_j; \boldsymbol{c}_i \circ \boldsymbol{q}_j] \in \mathbb{R}$$

- Context-to-Question (C2Q) attention:
  (which query words are most relevant to each context word)

$$\alpha^i = \text{softmax}(\boldsymbol{S}_{i,:}) \in \mathbb{R}^M \quad \forall i \in \{1, \ldots, N\}$$

$$\boldsymbol{a}_i = \sum_{j=1}^{M} \alpha_j^i \boldsymbol{q}_j \in \mathbb{R}^{2h} \quad \forall i \in \{1, \ldots, N\}$$

c_i = passage word

q_j = question word

Each are of dimension $2d$
(from the bidirectional LSTM)

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

Slide credit: Chris Manning

BiDAF

- **Attention Flow Idea:** attention should flow both ways – from the context to the question and from the question to the context
- Question-to-Context (Q2C) attention:
  (the weighted sum of the most important words in the context with respect to the query – slight asymmetry through max)

$$\boldsymbol{m}_i = \max_j \boldsymbol{S}_{ij} \in \mathbb{R} \quad \forall i \in \{1, \ldots, N\}$$

$$\beta = \text{softmax}(\boldsymbol{m}) \in \mathbb{R}^N$$

$$\boldsymbol{c}' = \sum_{i=1}^{N} \beta_i \boldsymbol{c}_i \in \mathbb{R}^{2h}$$

- For each passage position, output of BiDAF layer is:

$$\boldsymbol{b}_i = [\boldsymbol{c}_i; \boldsymbol{a}_i; \boldsymbol{c}_i \circ \boldsymbol{a}_i; \boldsymbol{c}_i \circ \boldsymbol{c}'] \in \mathbb{R}^{8h} \quad \forall i \in \{1, \ldots, N\}$$

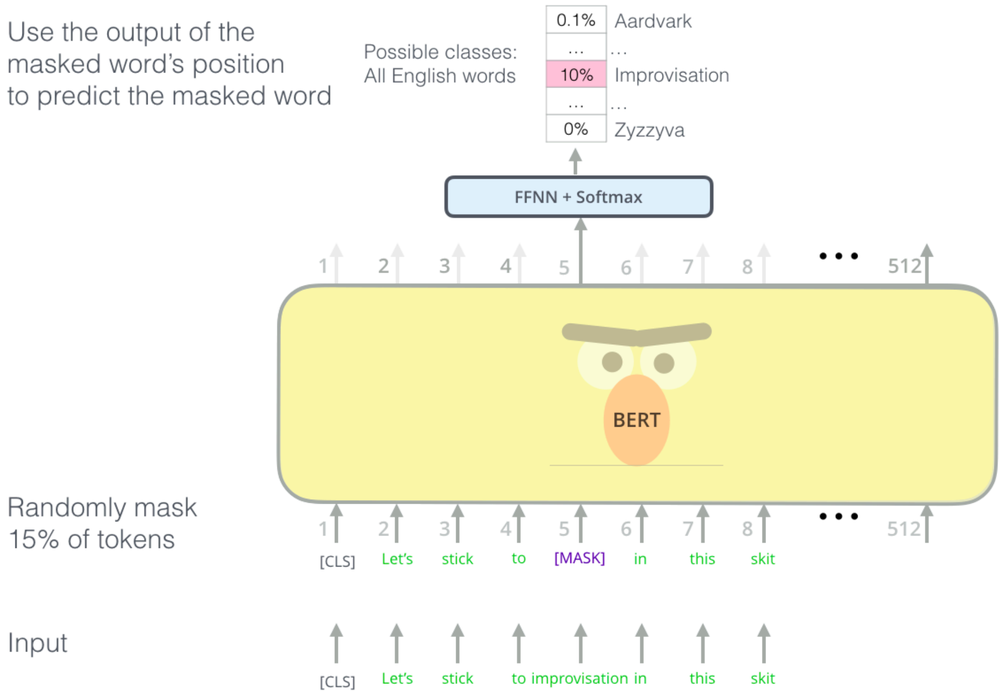
(Seo et al, 2017): Bidirectional Attention Flow for Machine Comprehension

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

BERT-based models

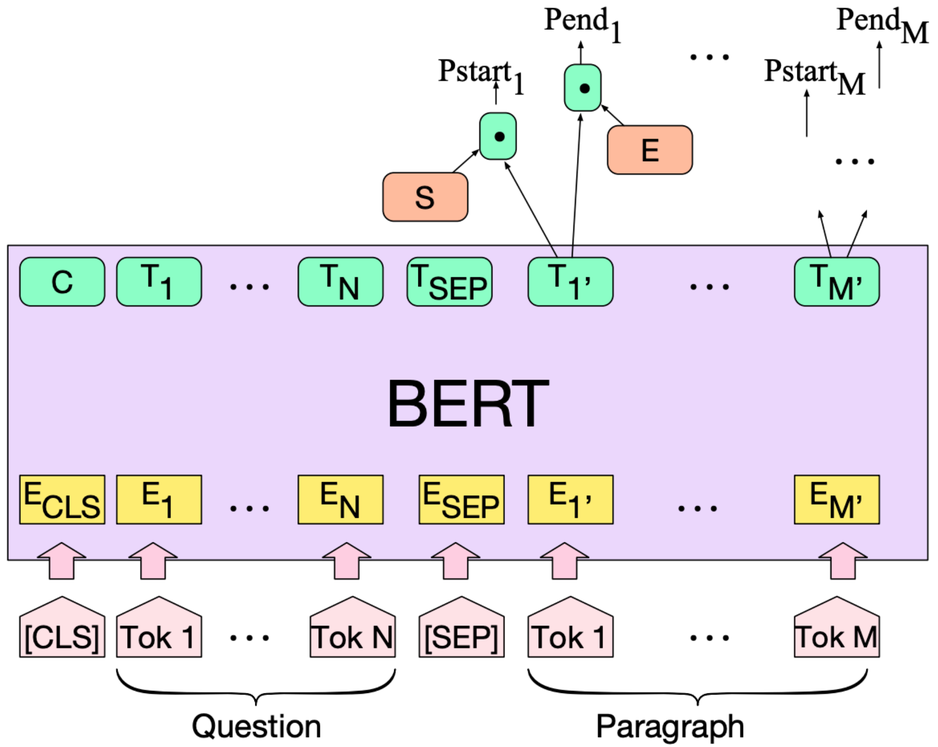


Pre-training

BERT-based models

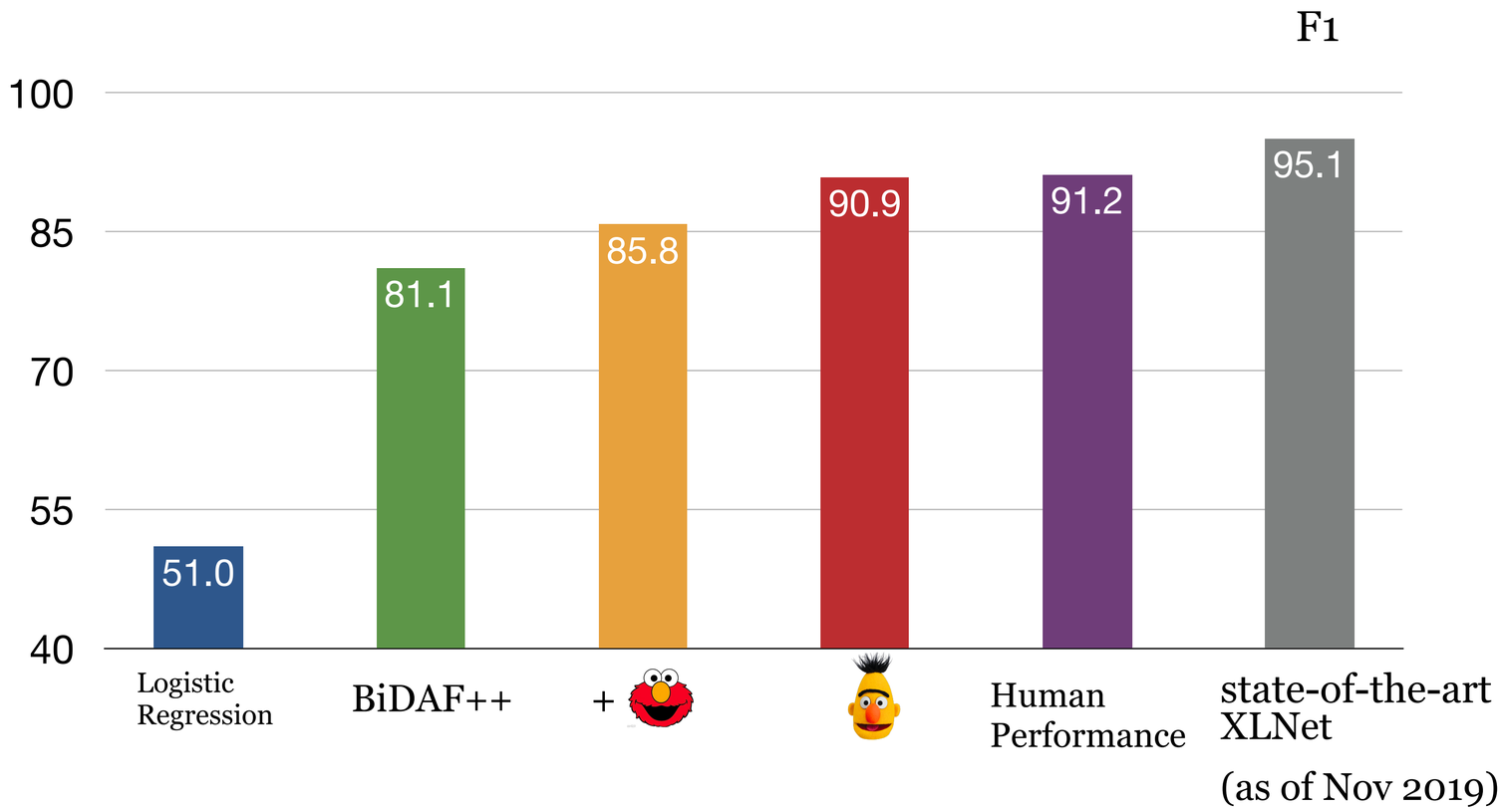
$$Pstart_i = \frac{e^{S \cdot T_i}}{\sum_j e^{S \cdot T_j}}$$

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



- 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

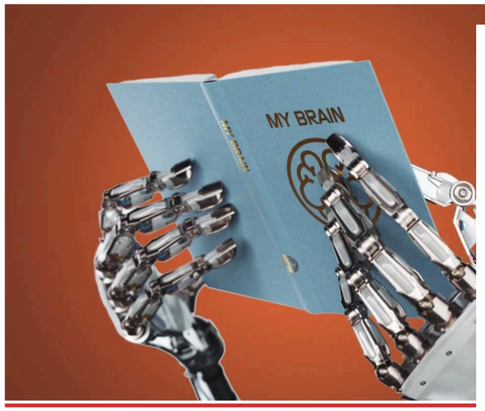


*: single model only

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

(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

DrQA: Document Retrieval

Dataset	Wiki Search	Doc. plain	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**

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

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

General Questions

Combined with **Web search**, DrQA can answer **57.5%** of **trivia questions** correctly

bing

Q: The Dodecanese Campaign of WWII that was an attempt by the Allied forces to capture islands in the Aegean Sea was the inspiration for which acclaimed 1961 commando film?

A: The Guns of Navarone

Q: American Callan Pinckney's eponymously named system became a best-selling (1980s-2000s) book/video franchise in what genre?

A: Fitness

Slide credit: Chris Manning