COS 484: Natural Language Processing

# Question Answering

Fall 2019

# Announcements

- Final project presentation: January 13, 10am-12pm
- Revised project proposal: November 22
  - Come meet us during OHs!

# Course planning

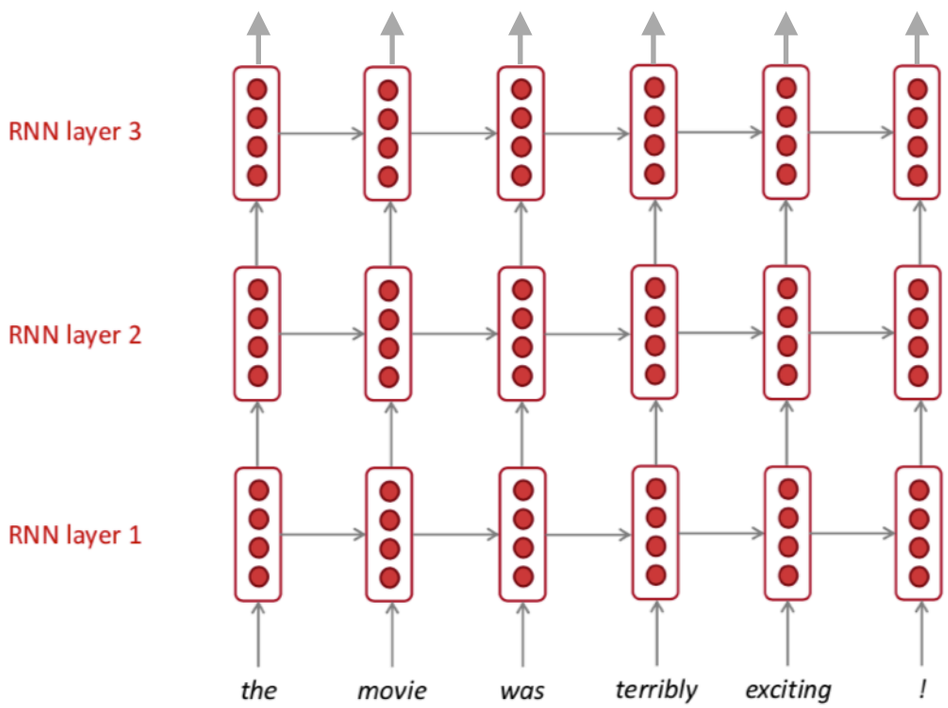
| | |
|---|---|
| Nov 19 | Question answering |
| Nov 21 | Coreference resolution |
| Nov 26 | Guest lecture: Jason Weston |
| Nov 28 | NO CLASS (Thanksgiving) |
| Dec 3 | Guest lecture: Tom Kwiatkowski |
| Dec 5 | Information extraction |
| Dec 10 | Grounding |
| Dec 12 | Project advice |

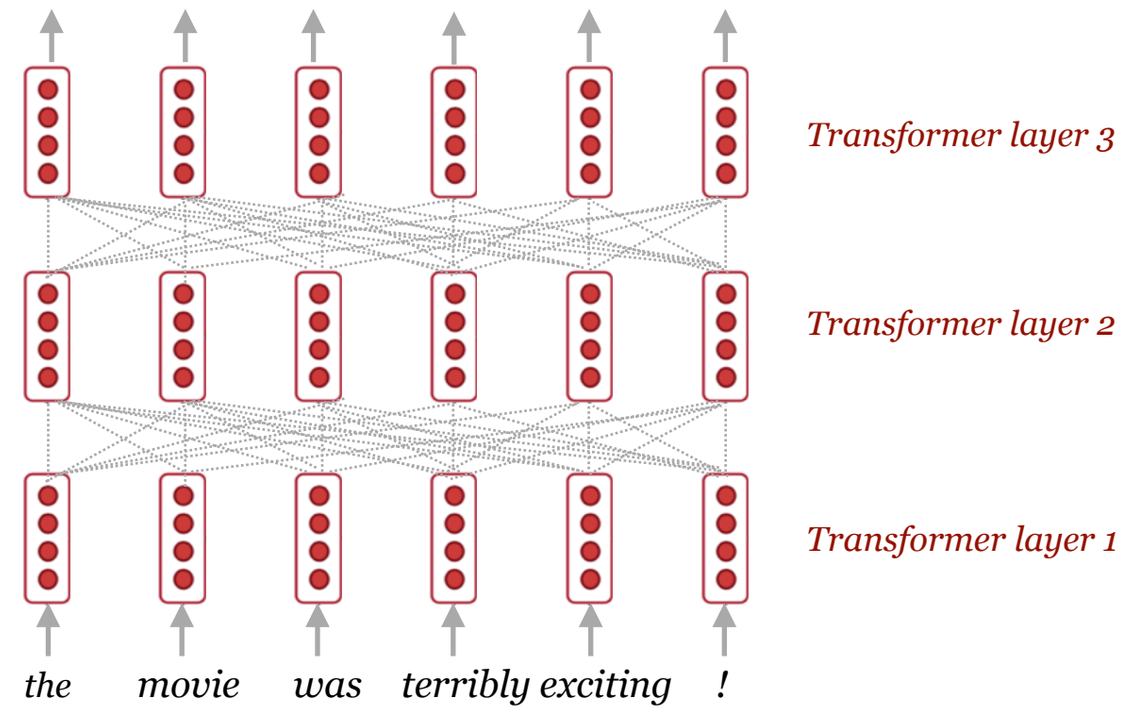
Transformers +
Question Answering

Dialogue

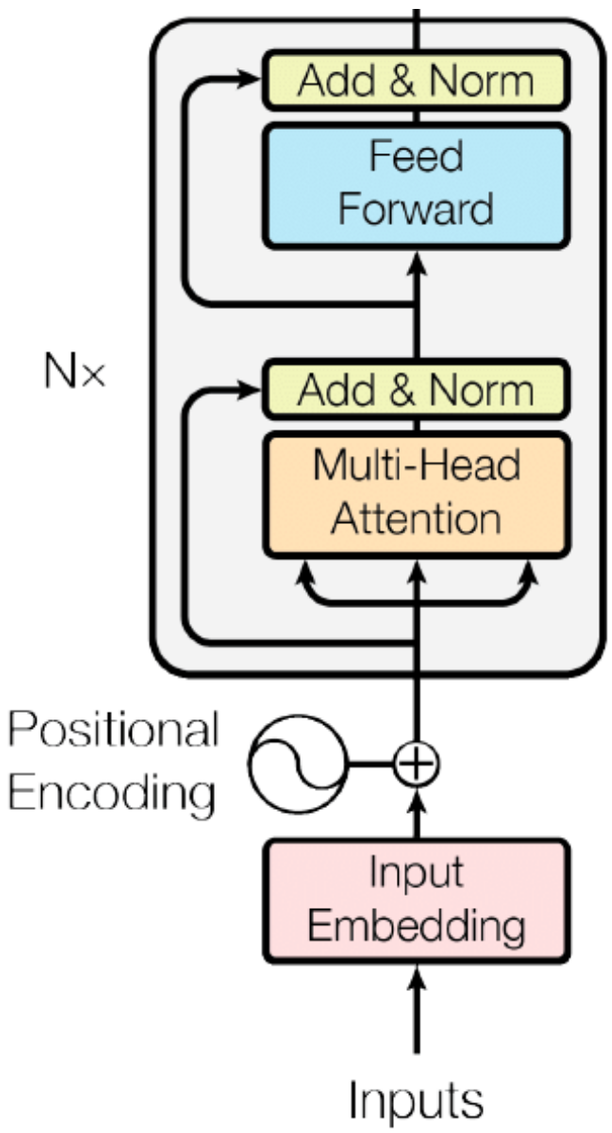
Advanced topics
in QA and others

# RNNs vs Transformers



# Transformers

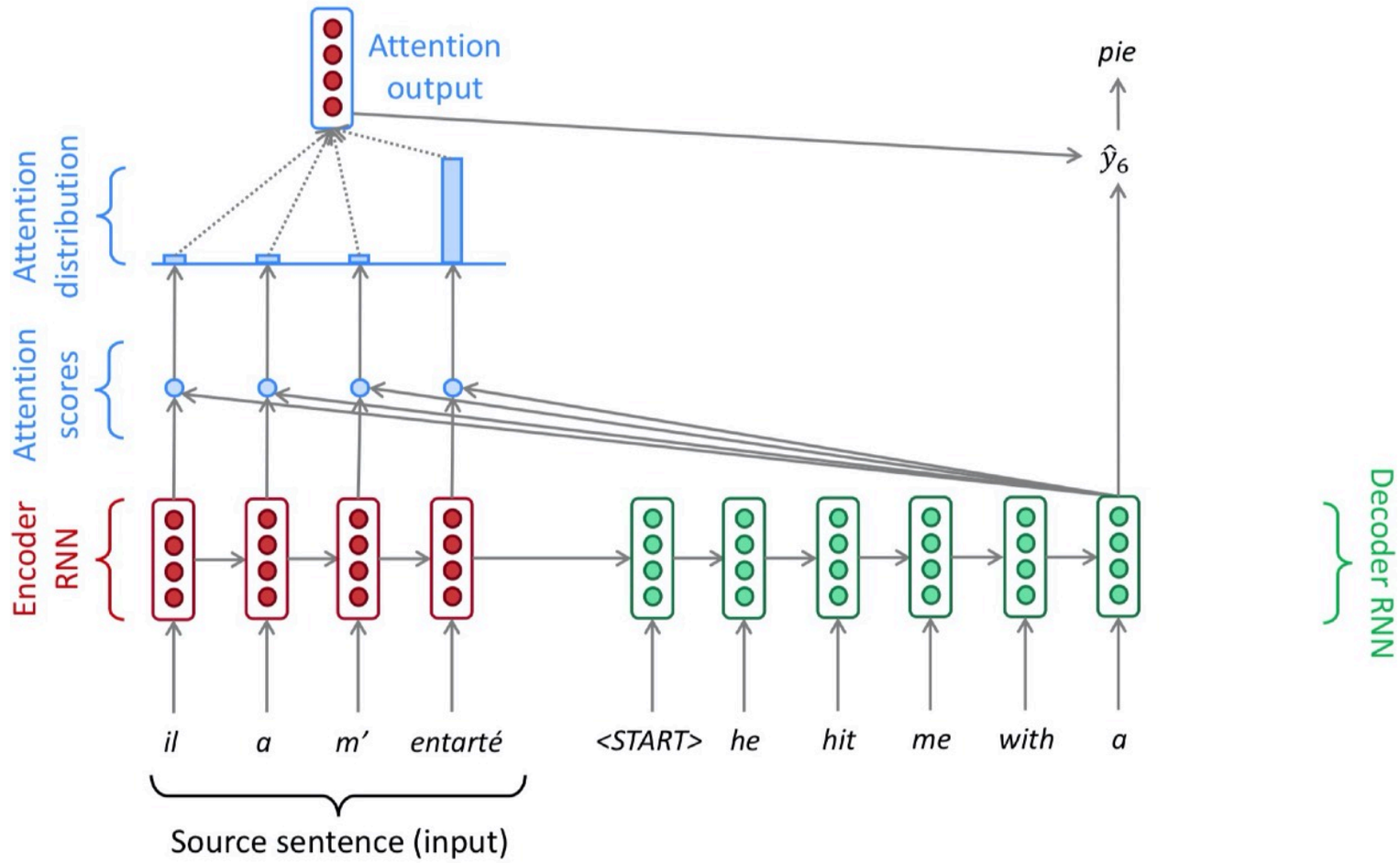



**Key concepts:**

- (scaled) dot-product attention
- Self-attention
- Multi-head self-attention

(Vaswani et al, 2017): Attention is all you need

# Recap: seq2seq with attention




key

$$e_i^t = g(\mathbf{h}_i^{\text{enc}}, \mathbf{h}_t^{\text{dec}})$$

query

$$\alpha_i^t = \frac{\exp(e_i^t)}{\sum_{j=1}^{n} \exp(e_j^t)}$$

$$\mathbf{a}^t = \sum_{i=1}^{n} \alpha_i^t \mathbf{h}_i^{\text{enc}}$$

value

# Generalized Attention

- A query $q$ and a set of key-value $(k_i, v_i)$ pairs to an output

- Dot-product attention:

$$A(q, \{k_i, v_i\}) = \sum_i \frac{e^{q \cdot k_i}}{\sum_j e^{q \cdot k_j}} v_i$$

$$k_i, v_i, q \in \mathbb{R}^d$$

- If we have multiple queries:

$$A(Q, K, V) = \text{softmax}(QK^\intercal)V$$

$$Q \in \mathbb{R}^{n_Q \times d}, K, V \in \mathbb{R}^{n \times d}$$

- Scaled dot-product attention:

$$A(Q, K, V) = \text{softmax}(\frac{QK^\intercal}{\sqrt{d}})V$$

# Self-attention

- Input: $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathbb{R}^{d_{in}}$
- Output: $\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_n \in \mathbb{R}^{d}$
- Key idea: let's use each word as query and compute the attention with all the other words
- Input: $X \in \mathbb{R}^{n \times d_{in}}$

$$A(XW^Q, XW^K, XW^V) \in \mathbb{R}^{n \times d}$$

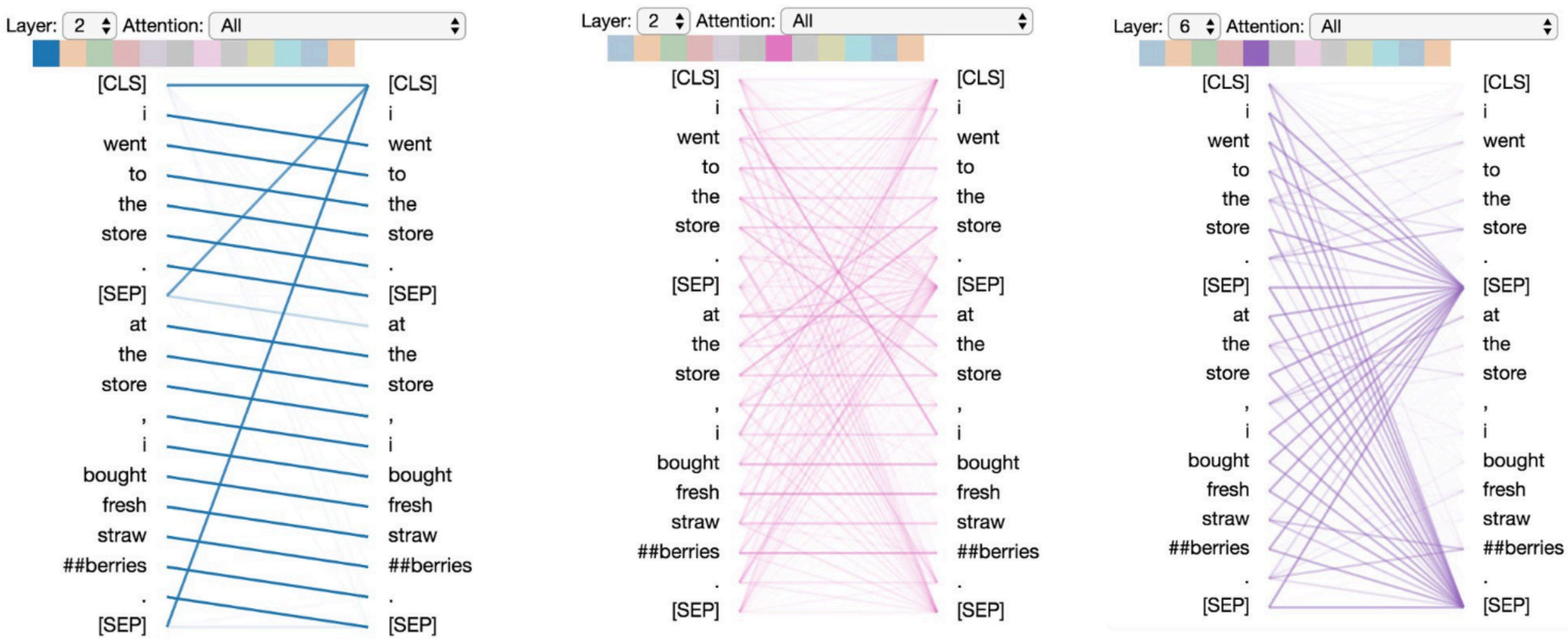
$$W^Q, W^K, W^V \in \mathbb{R}^{d_{in} \times d}$$

# Multi-head self-attention

One head is not expressive enough. Let's have multiple heads!

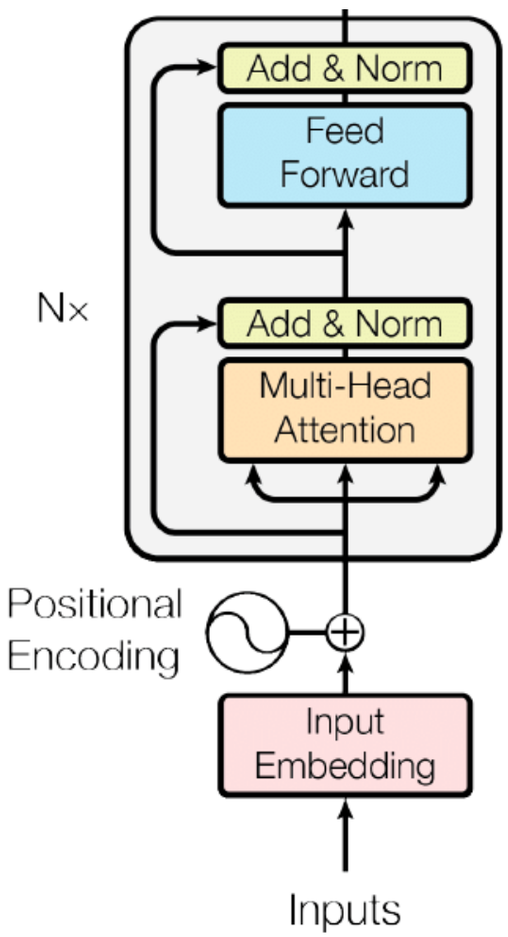
$$A(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O$$
$$\text{head}_i = A(XW_i^Q, XW_i^K, XW_i^V)$$

In practice, $h = 8$,
$d = d_{out}/h$, $W^O \in \mathbb{R}^{d_{out} \times d_{out}}$




https://github.com/jessevig/bertviz

# Putting it all together




- Each Transformer block has two sub-layers
  - Multi-head attention
  - 2-layer feedforward NN (with ReLU)
- Each sublayer has a residual connection and a layer normalization

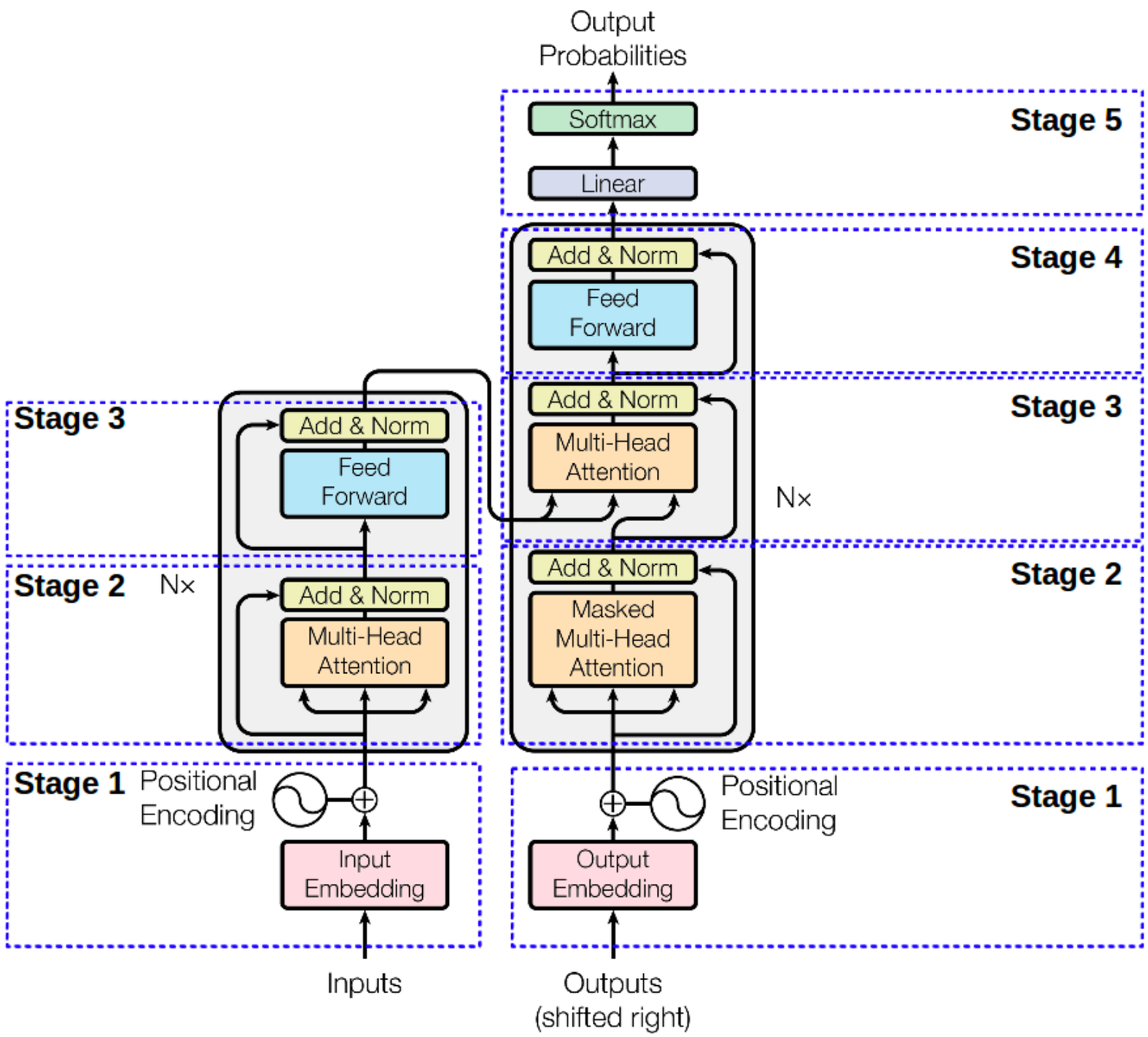
$$\text{LayerNorm}(x + \text{SubLayer}(x))$$

- Input layer has a positional encoding

- BERT_base: 12 layers, 12 heads, hidden size = 768, 110M parameters
- BERT_large: 24 layers, 16 heads, hidden size = 1024, 340M parameters

(Ba et al, 2016): Layer Normalization

# Encoder-decoder architecture




(Vaswani et al, 2017): Attention is all you need

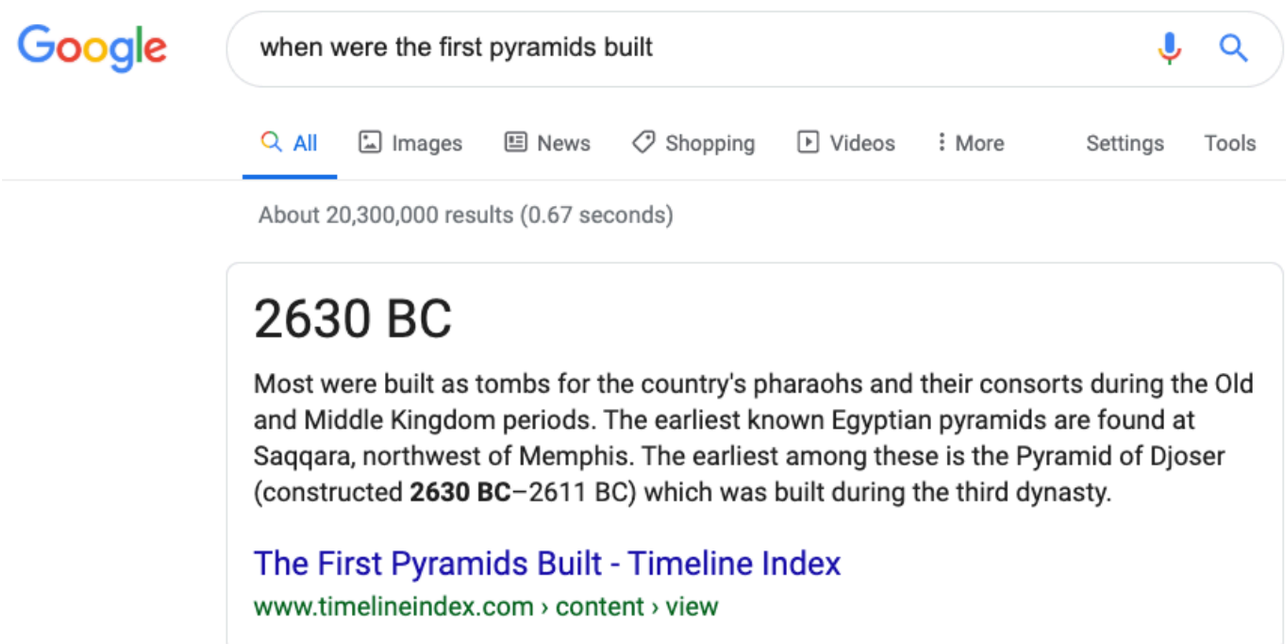
# Question Answering

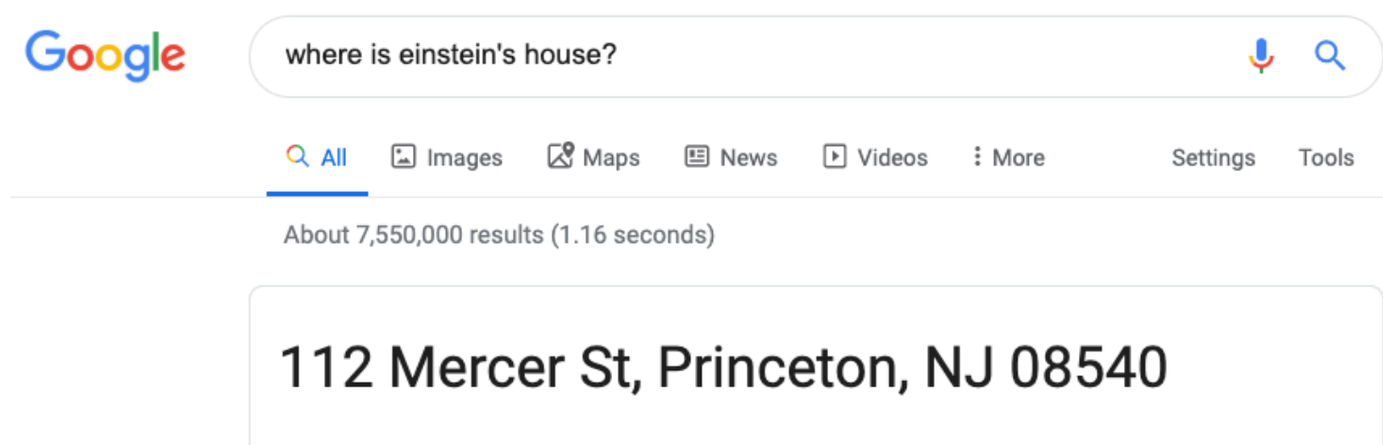
- Goal: build computer systems to answer questions

| Question | Answer |
| --- | --- |
| When were the first pyramids built? | 2630 BC |
| What's the weather like in Princeton? | 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

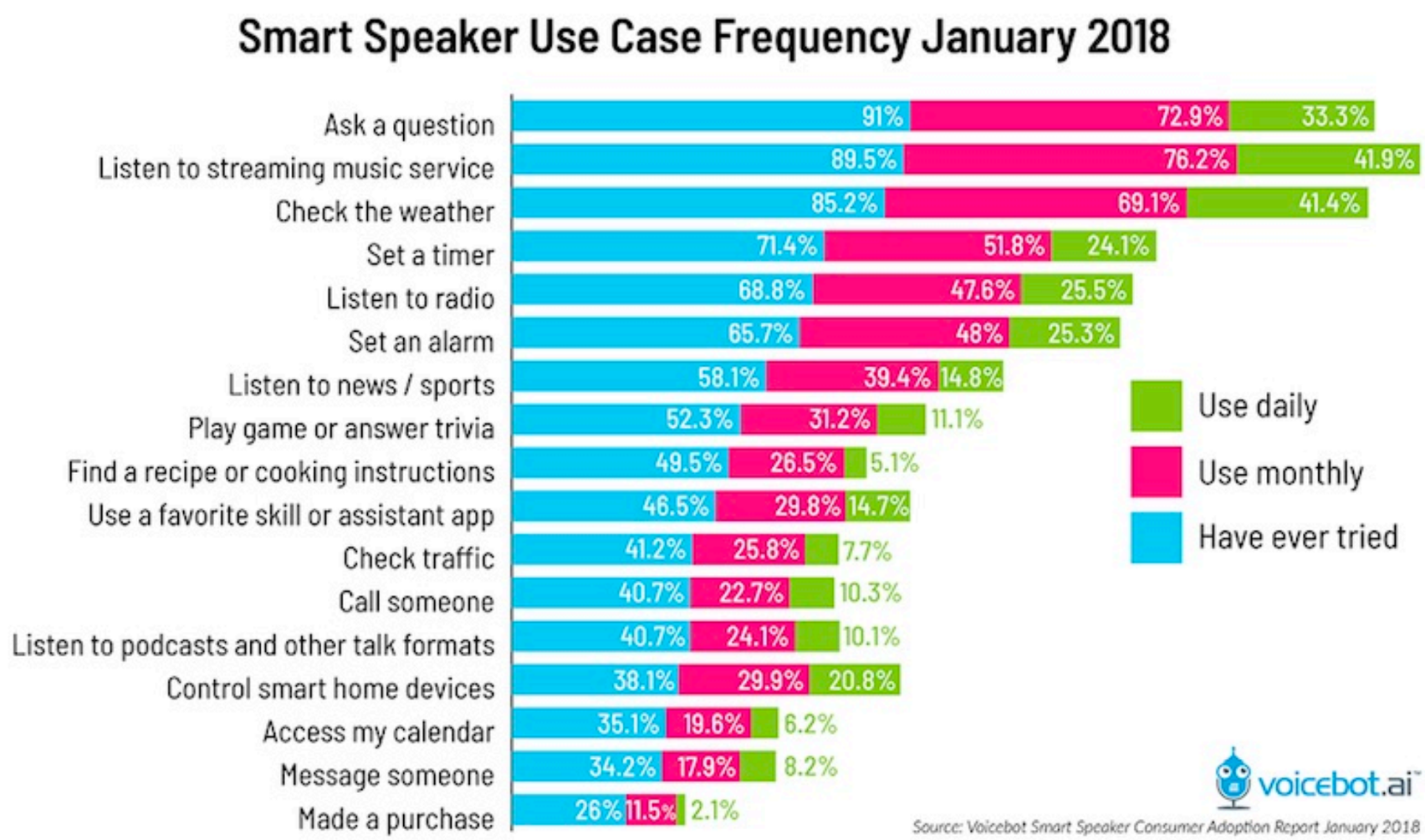
- You can easily find these answers in google today!



# Question Answering

- People ask lots of questions to Digital Personal Assistants:

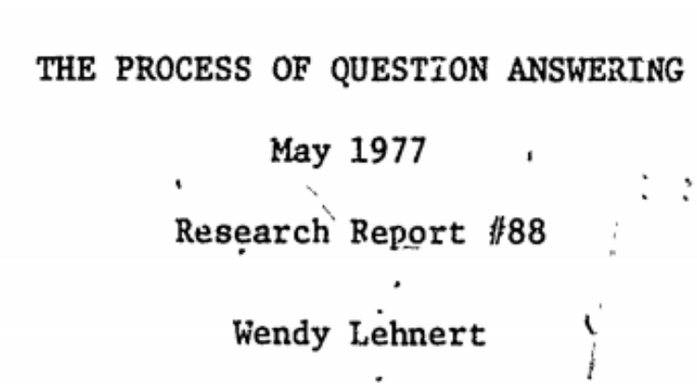


# Question Answering


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

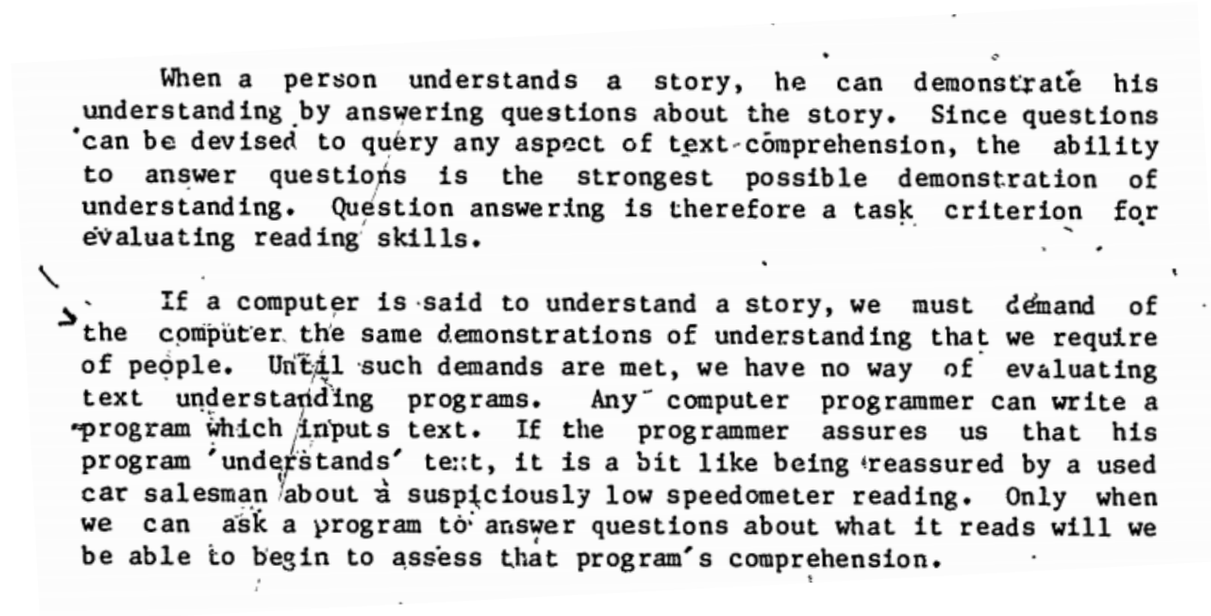


THE PROCESS OF QUESTION ANSWERING

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

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.

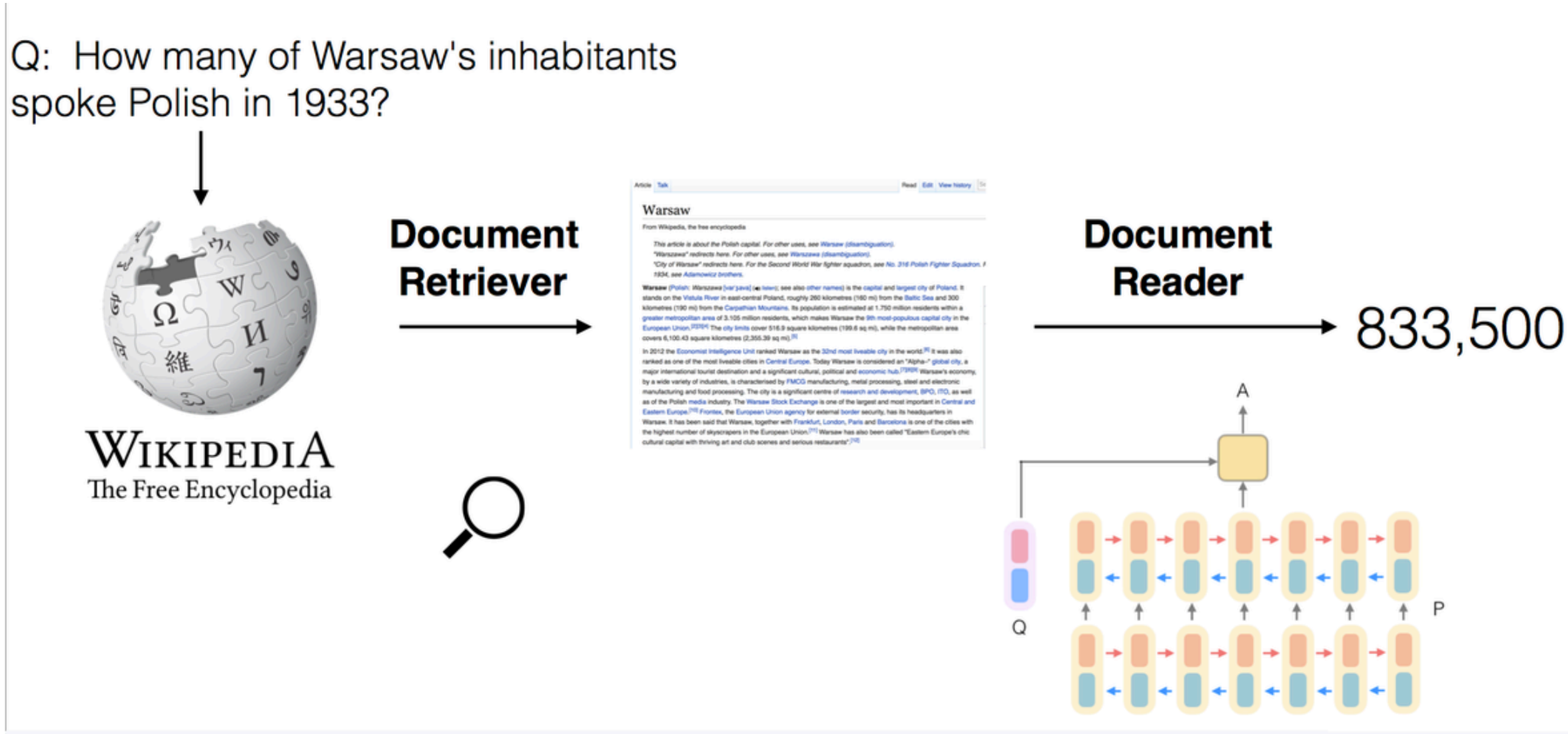
**Extractive**

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.

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

# Open-domain Question Answering

DrQA

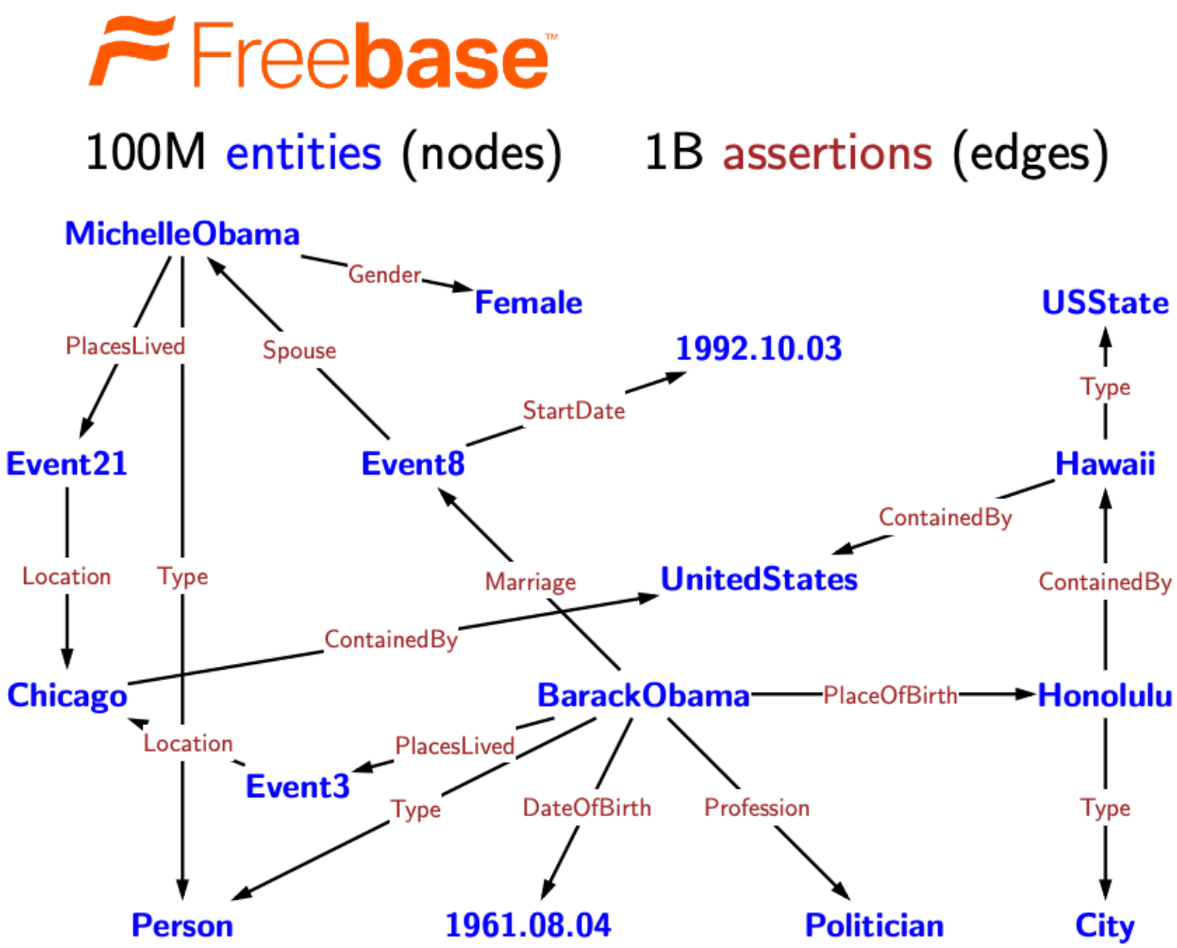



```
>>> process('What is the answer to life, the universe, and everything?')

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




*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

(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

# 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

- (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)

**SQuAD 1.1 evaluation:**

- 3 gold answers are collected for each answer
- 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

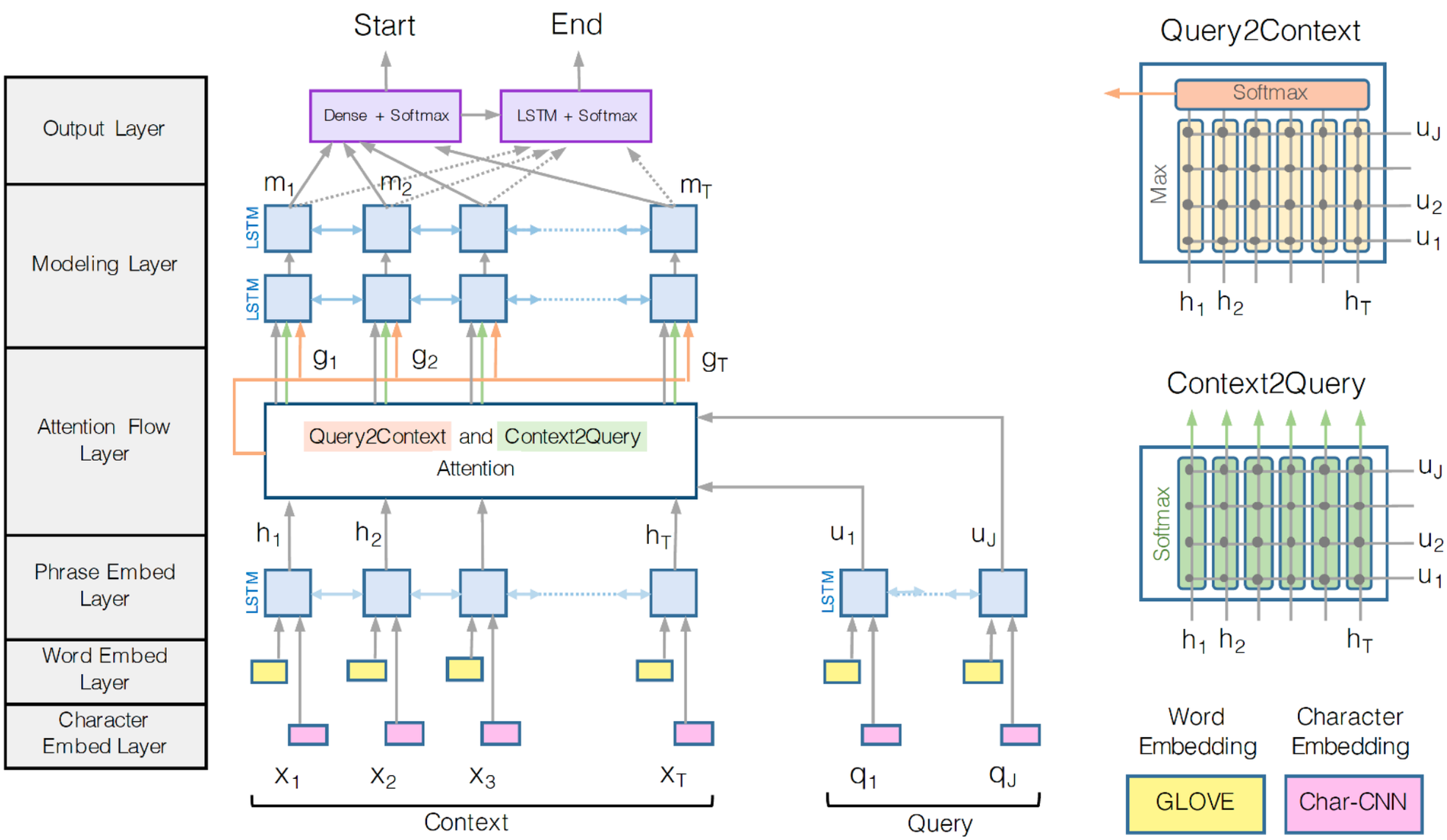
# Feature-based models

- 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

# BiLSTM-based models

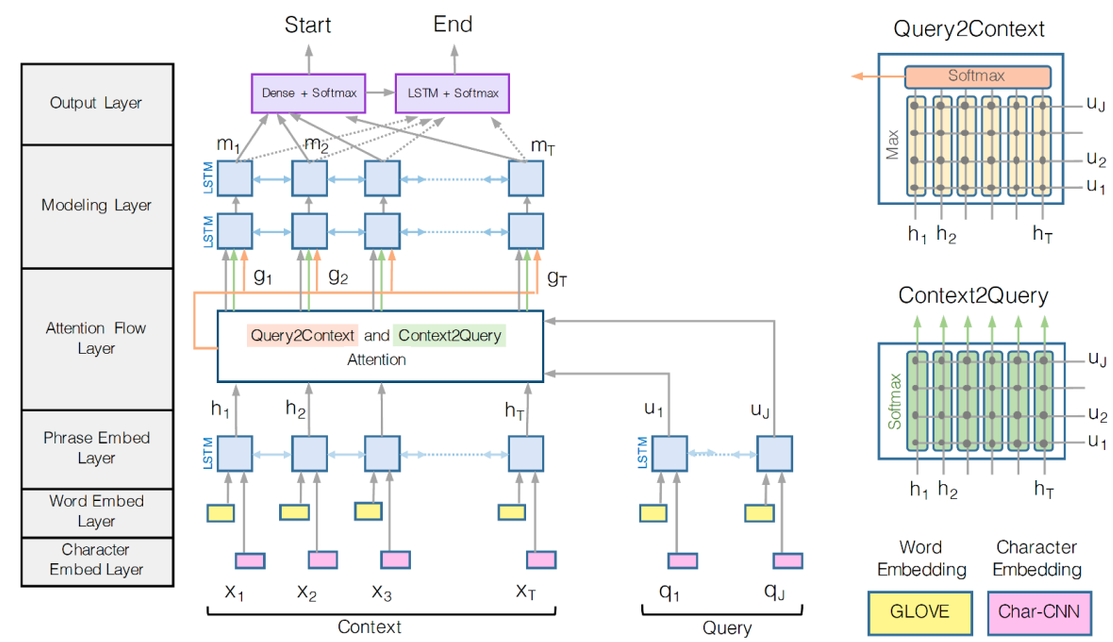
BiDAF




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

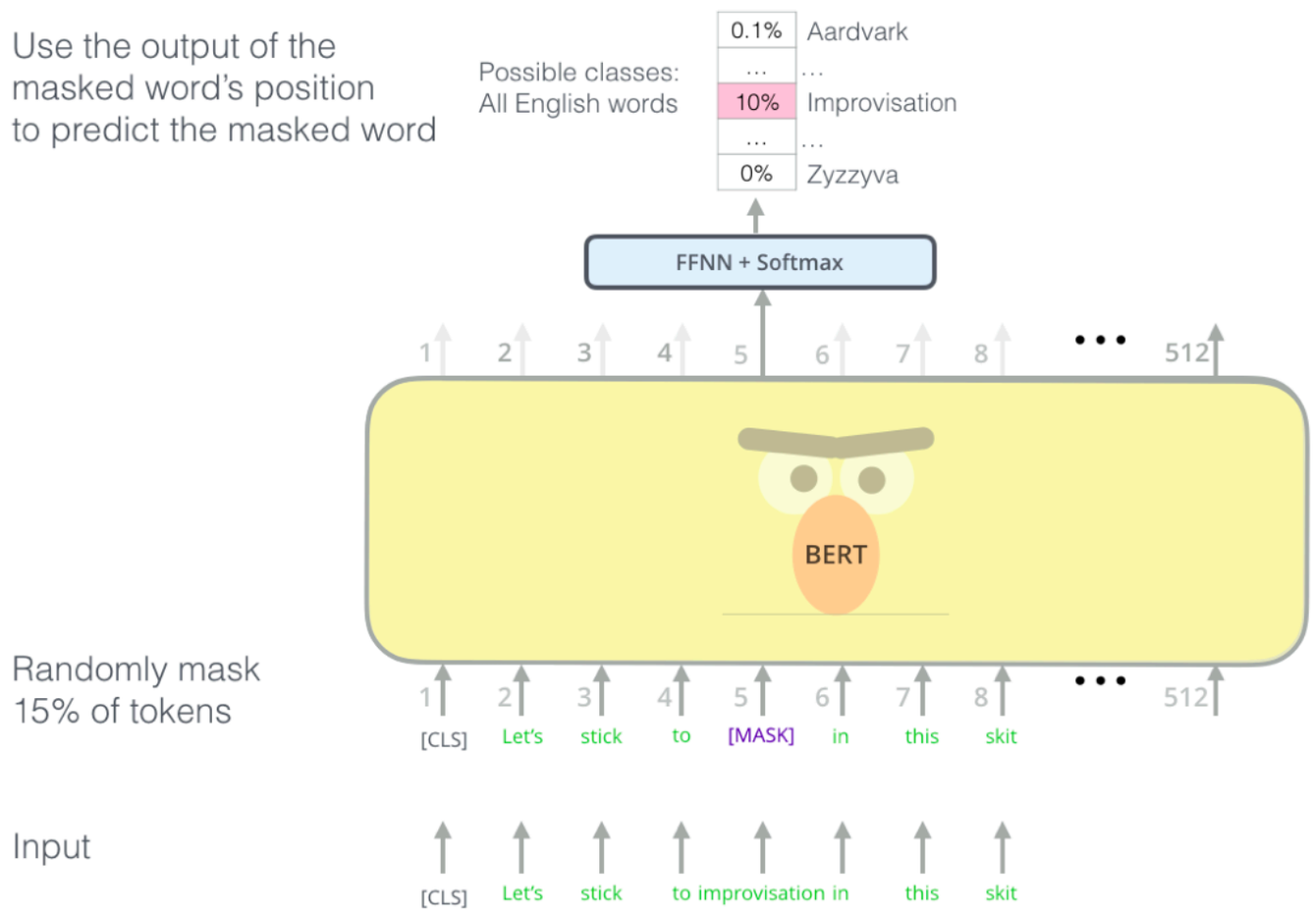
# BiLSTM-based models

- 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

# BERT-based models

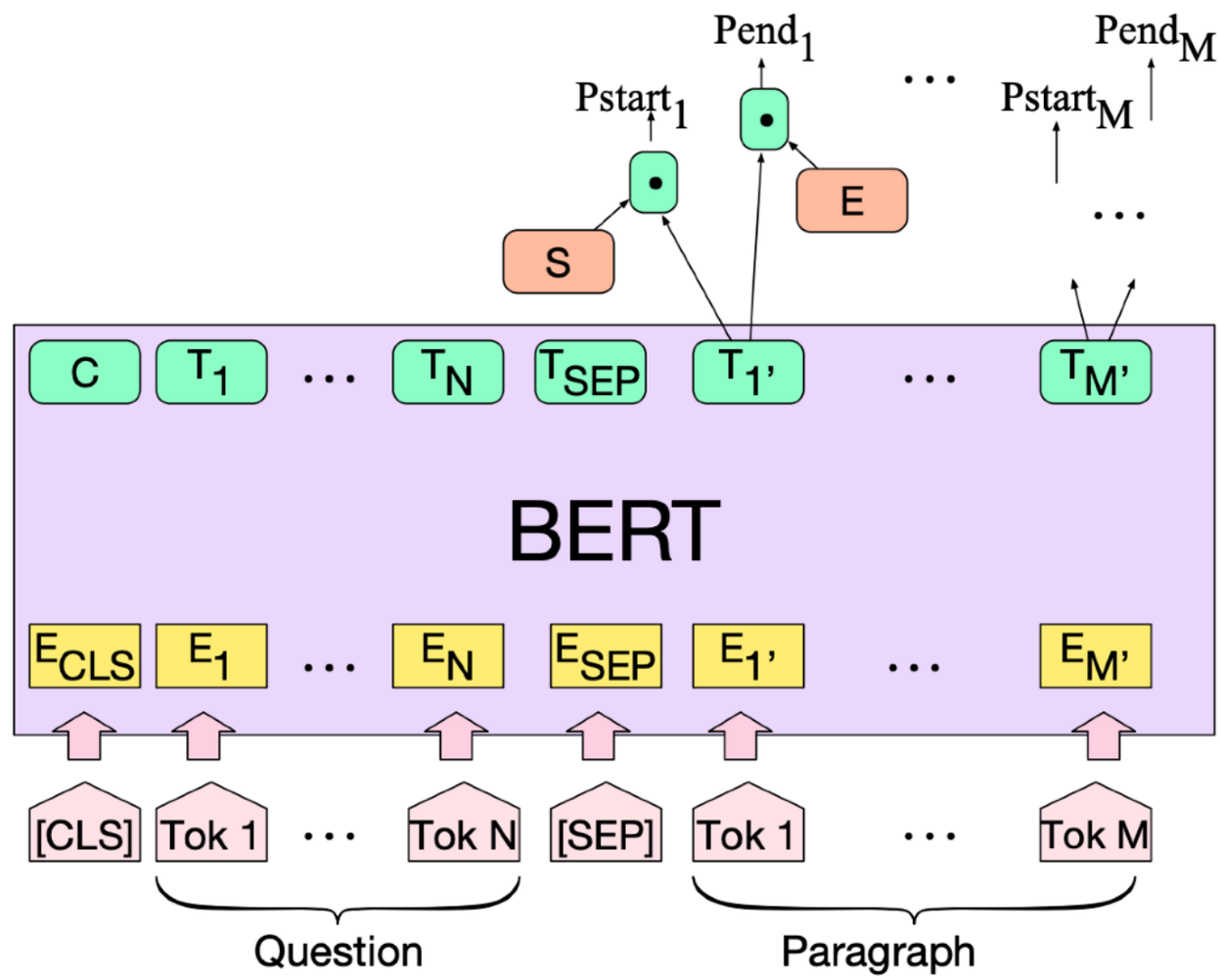



Pre-training

# BERT-based models

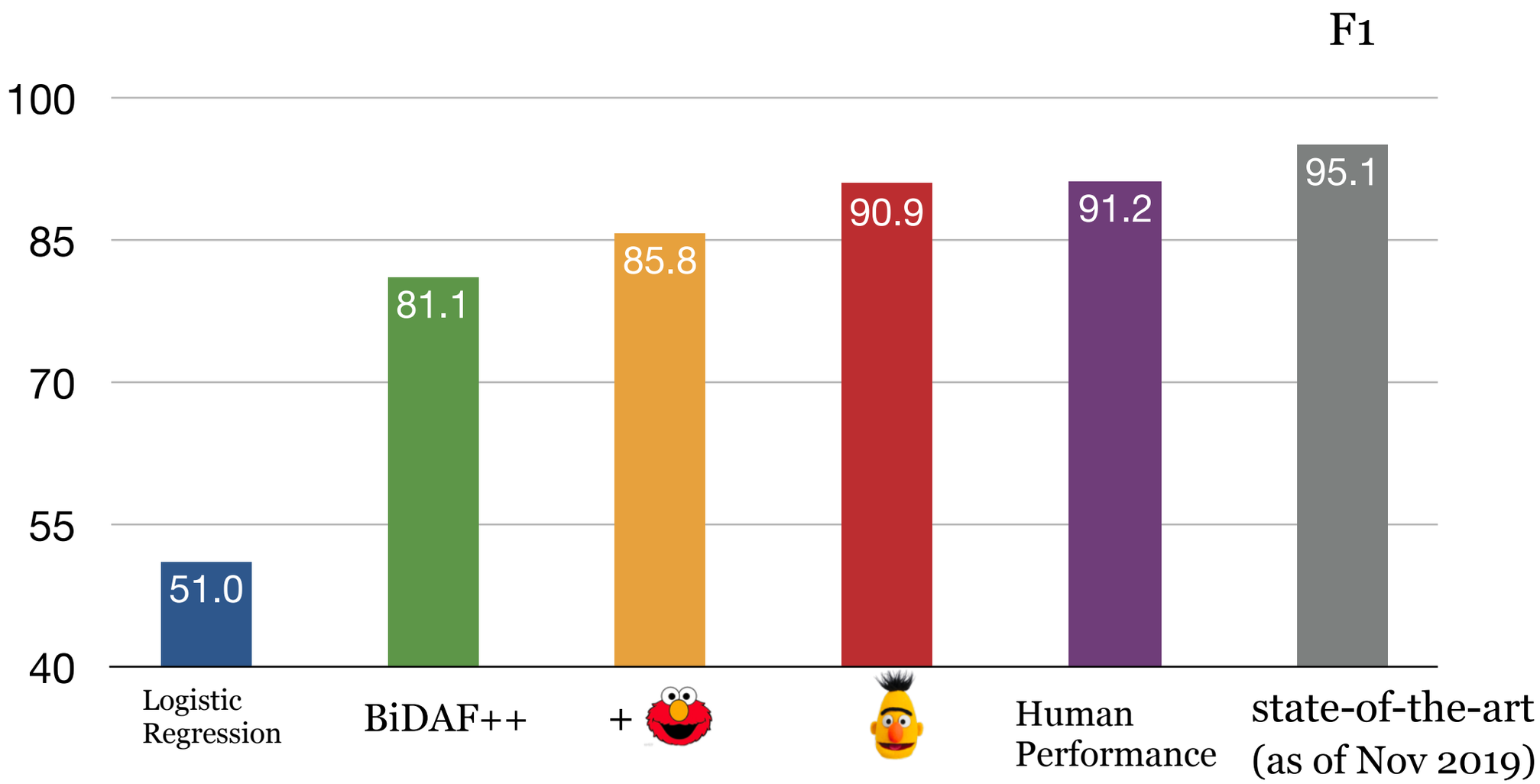
$$Pstart_i = \frac{e^{S \cdot T_i}}{\sum_j e^{S \cdot T_j}}$$

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




- 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

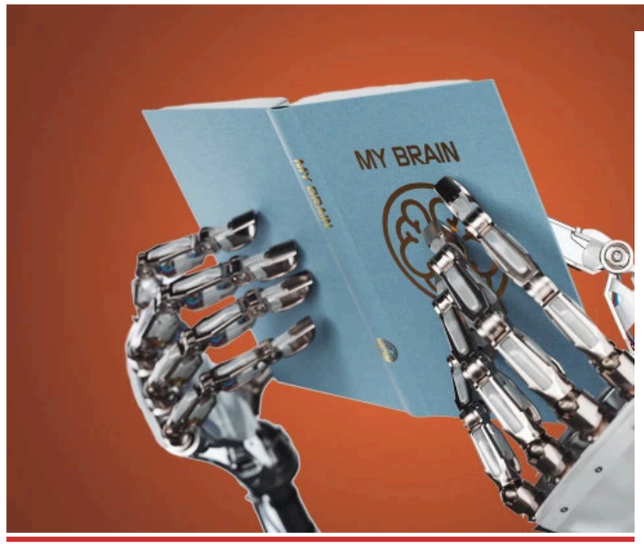



*: single model only

# 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.

# 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 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