Lecture 18: Question Answering

Zhizheng Wu

Agenda

- Recap
- Question answering and its applications
- Information Retrieval based Factoid QA
- Entity linking
- Knowledge-Based QA
- Other types of QA
- Benchmarks

TF-IDF

- Term frequency

$$\text{tf}_{t,d} = \text{count}(t,d)$$

Instead of using raw count, we squash a bit:

$$\text{tf}_{t,d} = \log_{10}(\text{count}(t,d)+1)$$

$$\text{tf}_{t,d} = \begin{cases} 1 + \log_{10} \text{count}(t,d) & \text{if } \text{count}(t,d) > 0 \\ 0 & \text{otherwise} \end{cases}$$

TF-IDF

- Document frequency
  - df is a term t that the number of documents it occurs in

	Collection Frequency	Document Frequency
Romeo	113	1
action	113	31

TF-IDF

- Inverse document frequency

$$\text{idf}_t = \log_{10}\left(\frac{N}{\text{df}_t}\right)$$

N is the total number of documents

Word	df	idf
Romeo	1	1.57
salad	2	1.27
Falstaff	4	0.967
forest	12	0.489
battle	21	0.246
wit	34	0.037
fool	36	0.012
good	37	0
sweet	37	0

TF-IDF

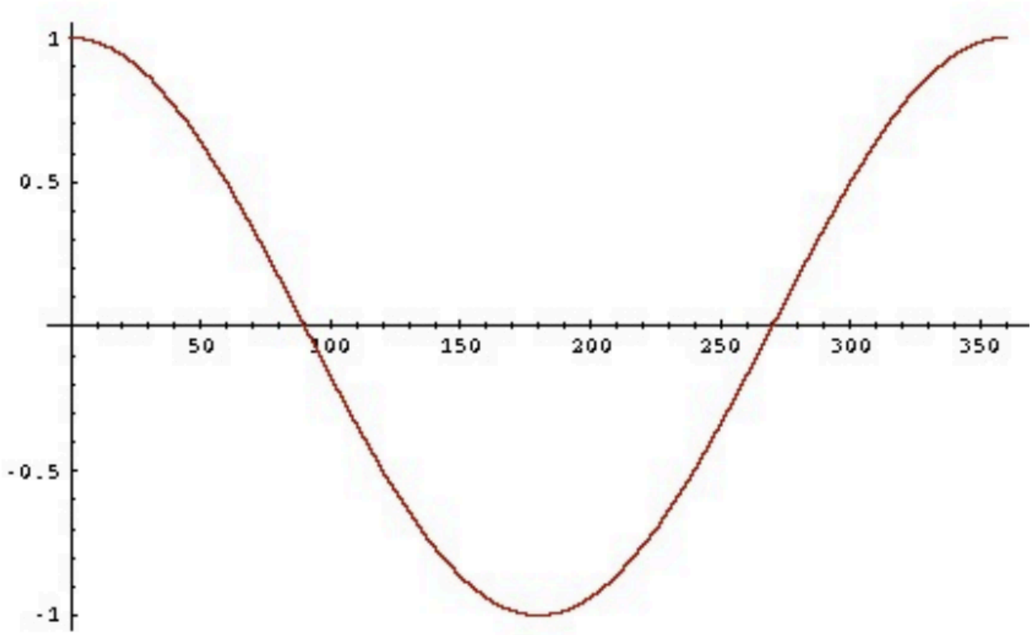
$$w_{t,d} = \text{tf}_{t,d} \times \text{idf}_t$$

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	0.074	0	0.22	0.28
good	0	0	0	0
fool	0.019	0.021	0.0036	0.0083
wit	0.049	0.044	0.018	0.022

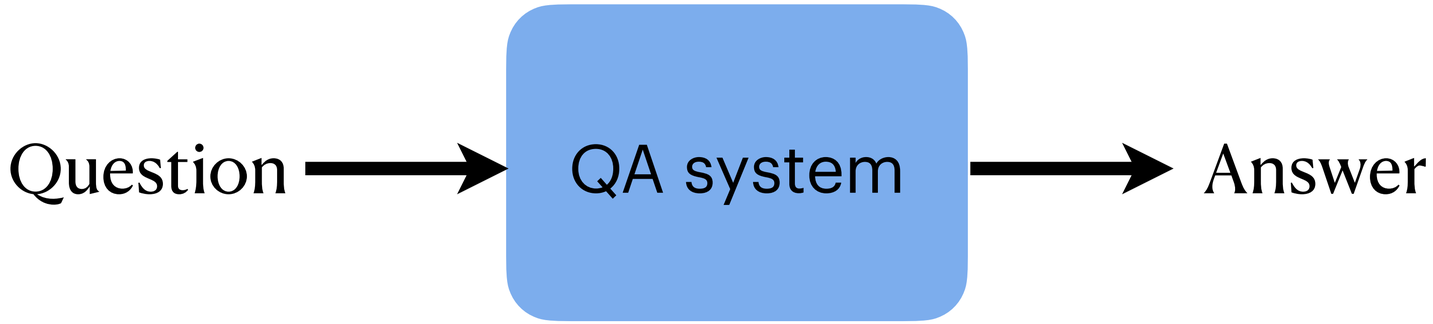
Cosine similarity

$$\text{cosine}(\vec{v}, \vec{w}) = \frac{\vec{v} \cdot \vec{w}}{|\vec{v}||\vec{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2} \sqrt{\sum_{i=1}^{N} w_i^2}}$$



What is question answering?

- The goal of question answering is to build systems that automatically answer questions posed by humans in a natural language



QA: from classification to open-ended

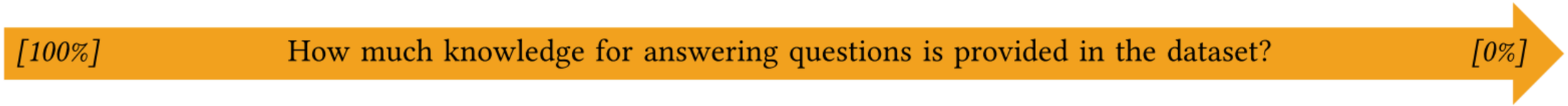
(easy)
Classification
What is thear sentiment of <STATEMENT>?

(doable)
Template-filling
When was <PERSON> born?

(difficult)
Open-ended
(too many templates and/or variables)

Rogers, Anna, Matt Gardner, and Isabelle Augenstein. "Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension." ACM Computing Surveys 55, no. 10 (2023): 1-45.

QA: sources



Single source
one document needs to be considered for answering the question

Multiple sources
evidence is provided, but it has to be ranked and found

Partial source
some evidence is provided, but it has to be combined with missing knowledge

No sources
the model has to retrieve evidence or have it memorized

Rogers, Anna, Matt Gardner, and Isabelle Augenstein. "Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension." ACM Computing Surveys 55, no. 10 (2023): 1-45.

Why do we care about this problem?

- Useful for many practical applications
- Viewed as an important testbed for evaluating how well computer systems understand human language
- Many other NLP tasks can be reduced to a reading comprehension problem

Applications

who is the president of the united states

All News Images Books More Tools

About 1,410,000,000 results (0.80 seconds)

United States / President

Joe Biden

People also search for


Tucker Carlson
Trending


Jill Biden
Trending


Kamala Harris


Donald Trump


Vladi… Putin


Joe Manc…
Trending


Nikki Haley
Trending

Joe Biden

46th U.S. President

Joseph Robinette Biden Jr. is an American politician who is the 46th and current president of the United States. A member of the Democratic Party, he previously served as the 47th vice president from 2009 to 2017 under President Barack Obama and represented Delaware in the United States Senate from 1973 to 2009. Wikipedia

Born: November 20, 1942 (age 80 years), Scranton, PA

Marriage location: New York, NY

Edited works: Halting the Spread of HIV/AIDS: Future Efforts in the U. S. Bilateral and Multilateral Response: Congressional Hearings, MORE

Applications

how old is Dilraba

All Images News Videos More Tools

About 689,000 results (0.52 seconds)

Dilraba Dilmurat / Age

30 years

June 3, 1992

People also search for

Yang Yang
31 years

Yang Mi
36 years

Zhao Lusi
24 years

Feedback

Dilraba Dilmurat

Chinese actress

Dilraba Dilmurat, known in Chinese as Dilireba, is a Chinese actress, host, dancer, singer and model of Uyghur ethnicity. Wikipedia

Born: June 3, 1992 (age 30 years), Ürümqi, China

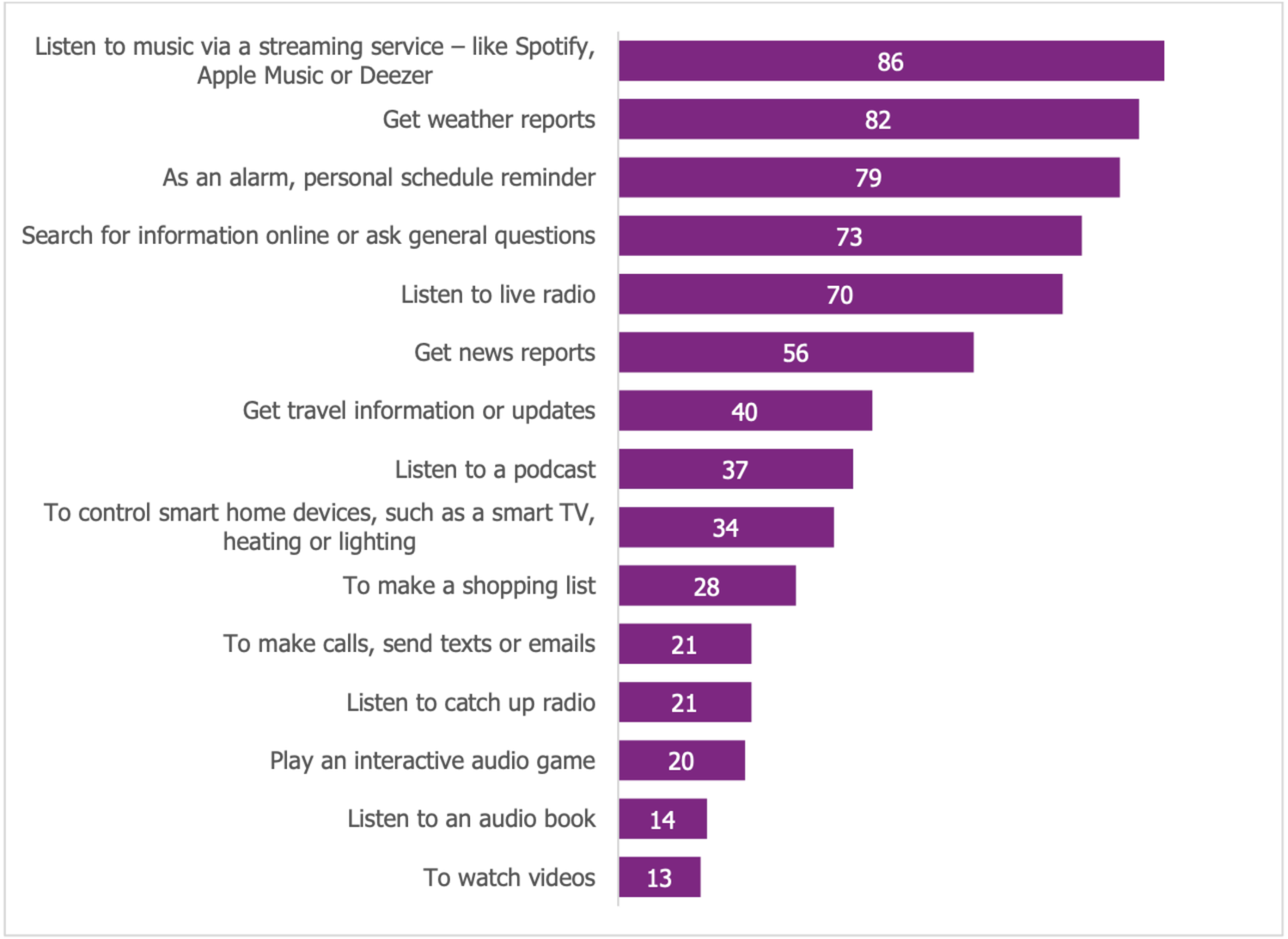
Height: 5′ 7″

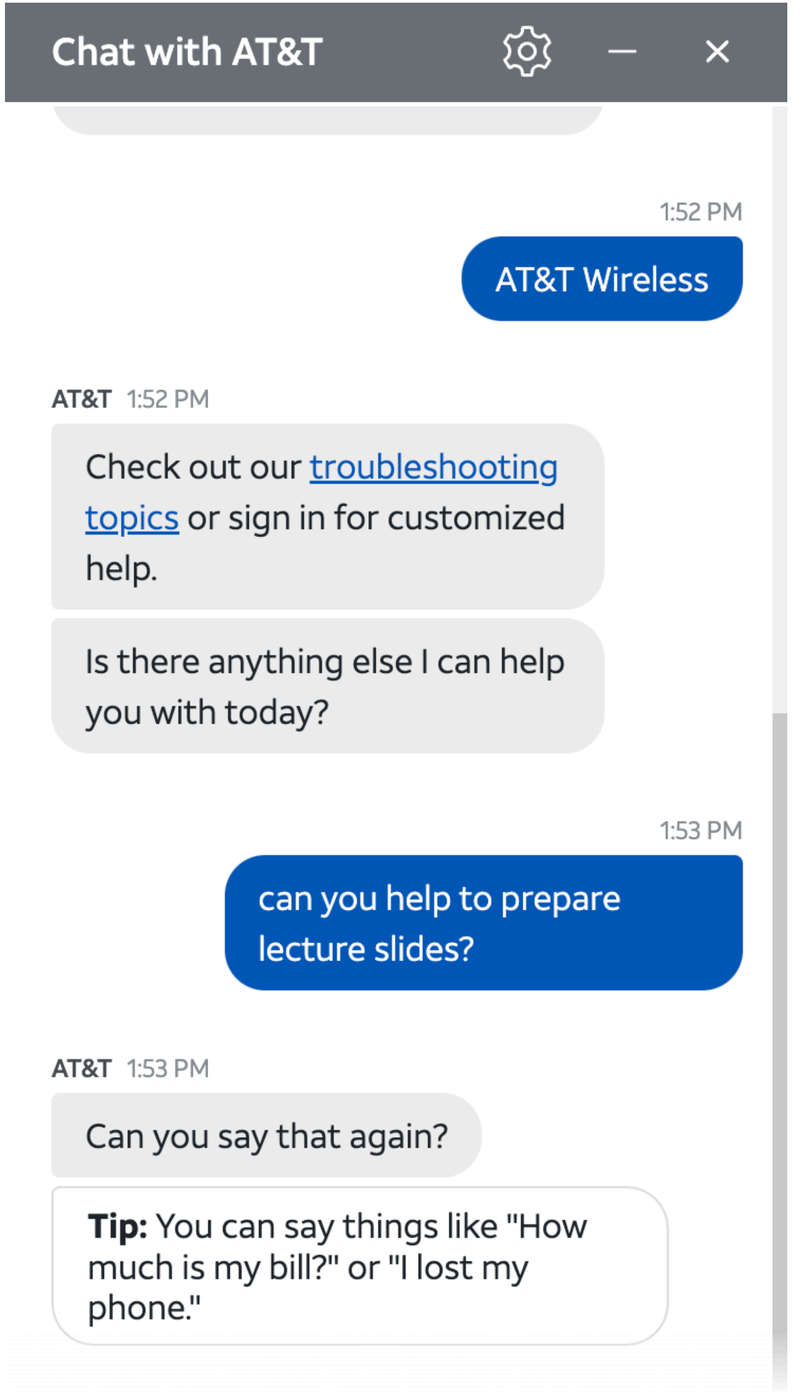
Record label: 宏揚國際有限公司

Albums: 《漂亮的李慧珍》電視劇原聲帶

Genre: Pop, Mandarin pop

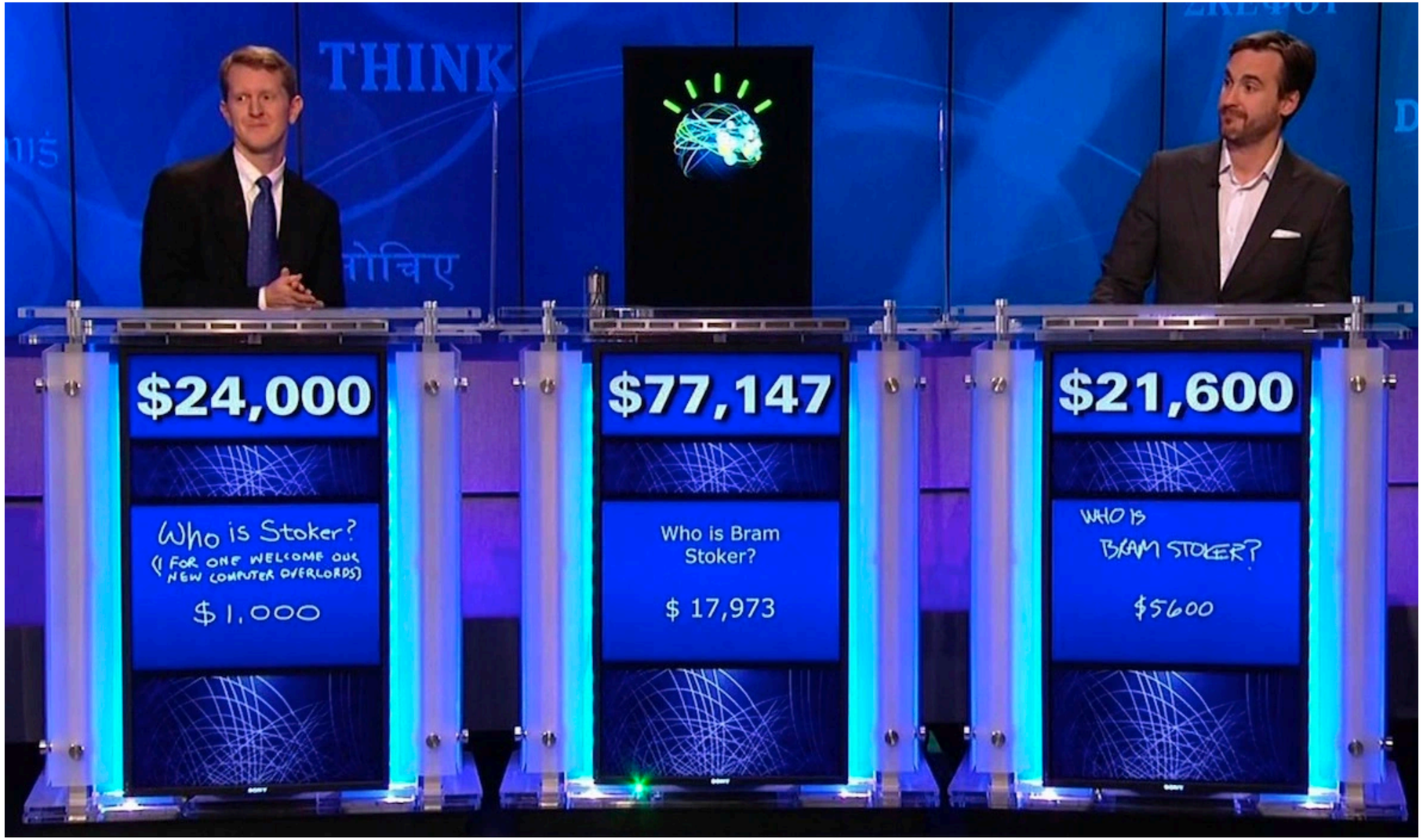
Applications



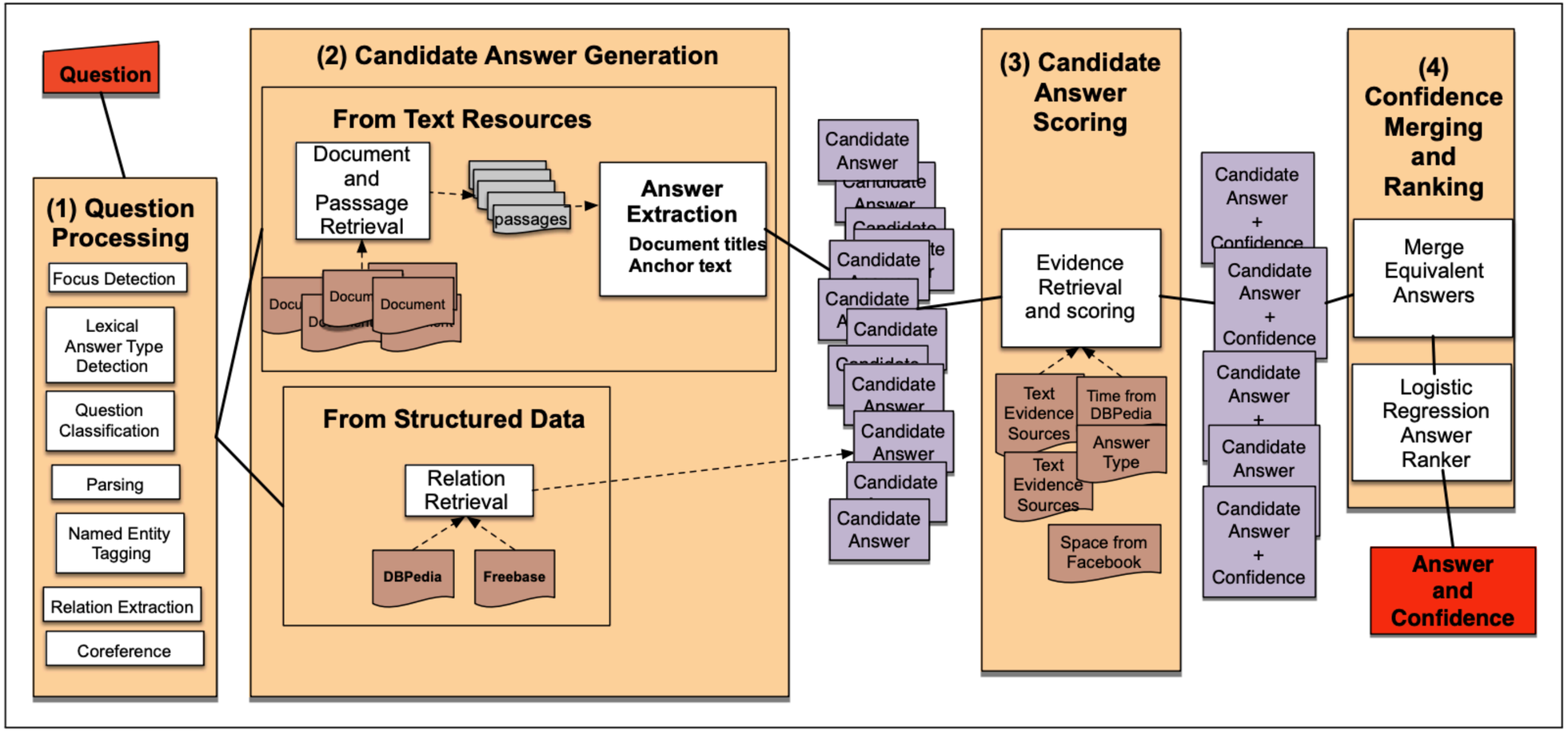


Smart speakers research with the public Research report December 2022

IBM Watson beated Jeopardy champions



Four stages of Watson QA



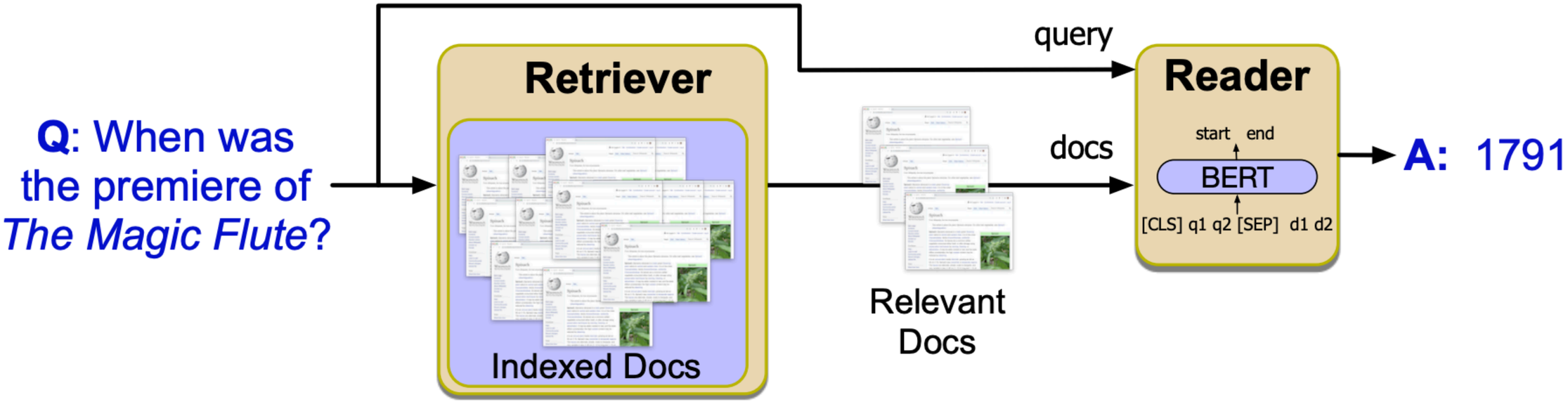
Information Retrieval based Factoid QA

- Also called open domain QA
- answer a user's question by finding short text segments from the web or some other large collection of documents

Question	Answer
Where is the Louvre Museum located?	in Paris, France
What are the names of Odin's ravens?	Huginn and Muninn
What kind of nuts are used in marzipan?	almonds
What instrument did Max Roach play?	drums
What's the official language of Algeria?	Arabic

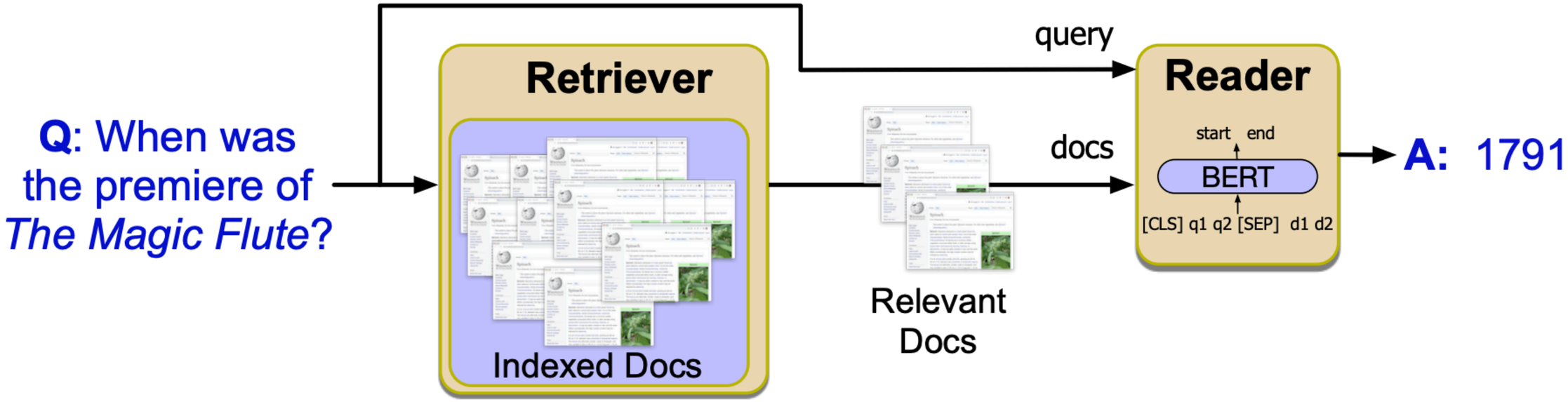
Retriever-reader framework

- Assumption
  - We have access to a large collection of documents that we have processed in advanced (“indexed documents”)
  - The question can be answered by returning a snippet of text (“span”) from one (or more) of these documents



Retriever-reader framework

- Procedure
  - Identify a (small) subset of documents that are relevant to the question
  - Identify (and return) the most likely answer span



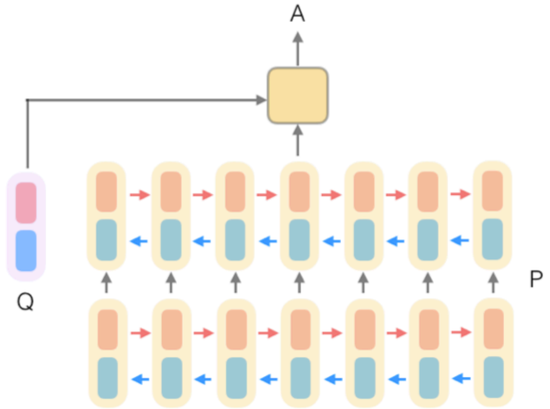
Retriever-reader framework

Q: How many of Warsaw's inhabitants spoke Polish in 1933?

Document Retriever

Document Reader

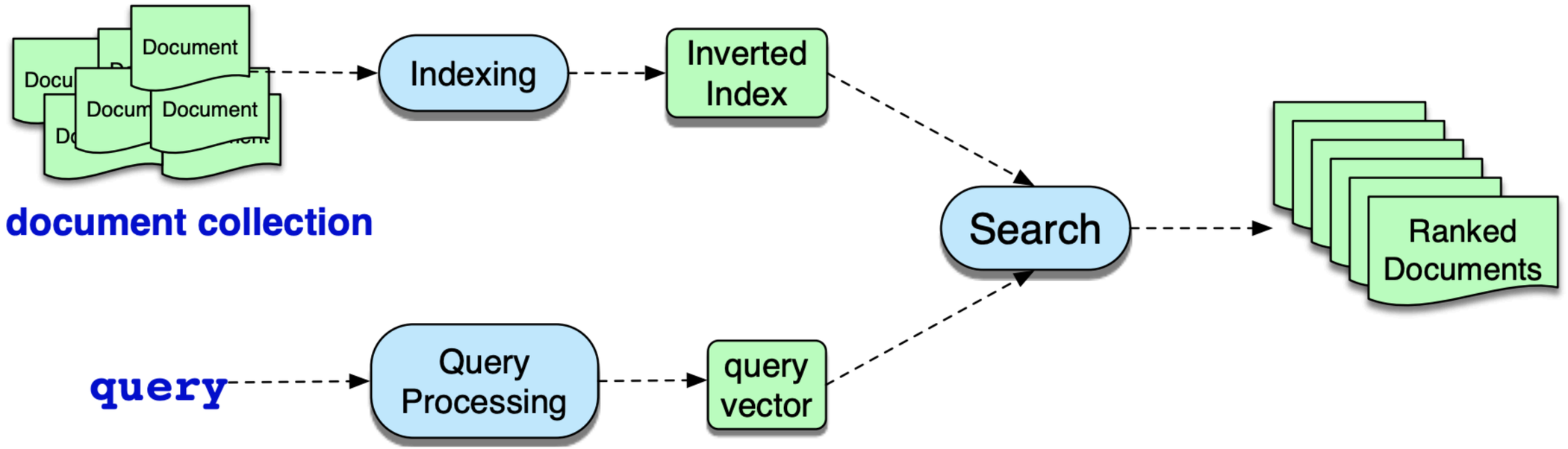
833,500



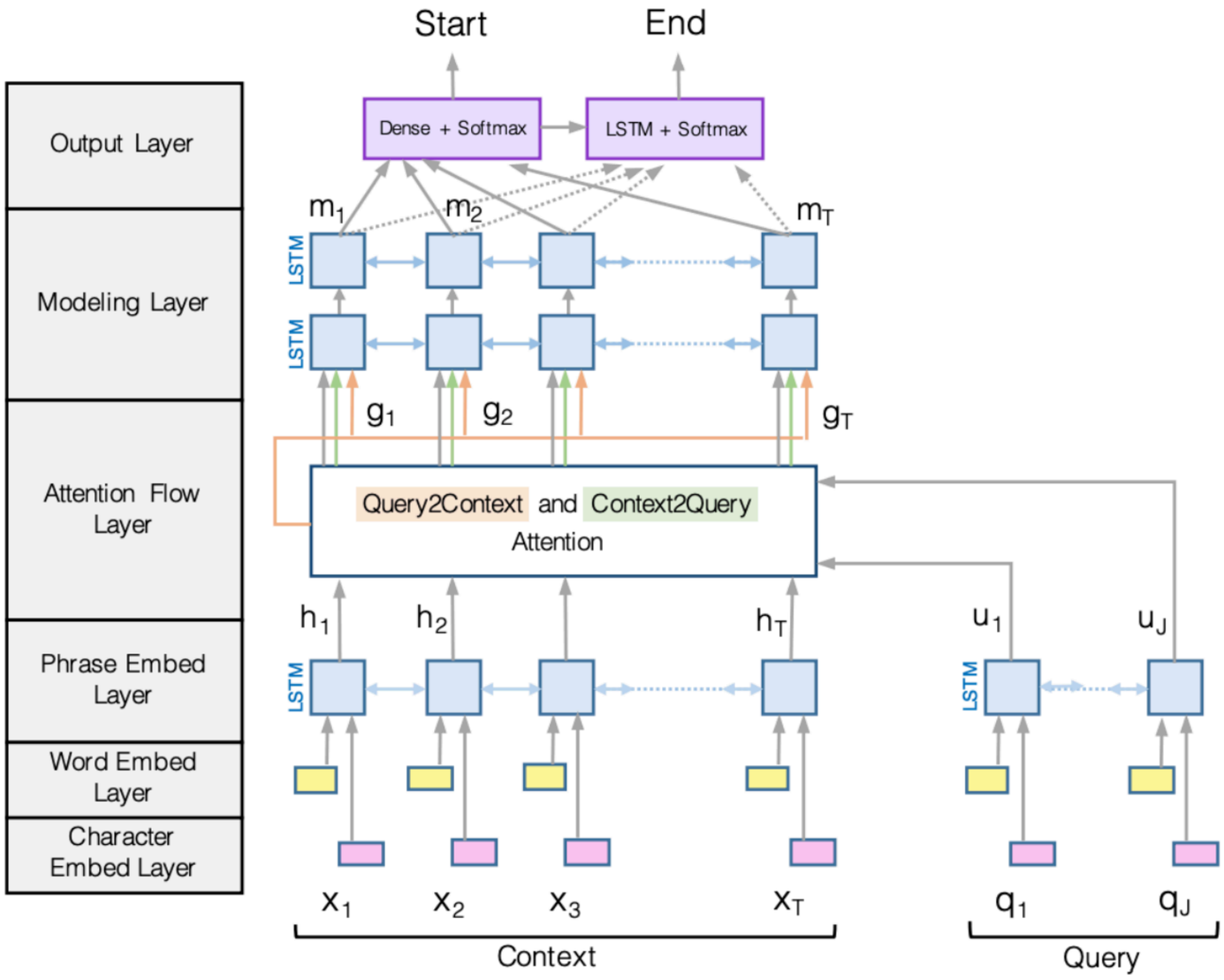
https://github.com/facebookresearch/DrQA

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

IR architecture

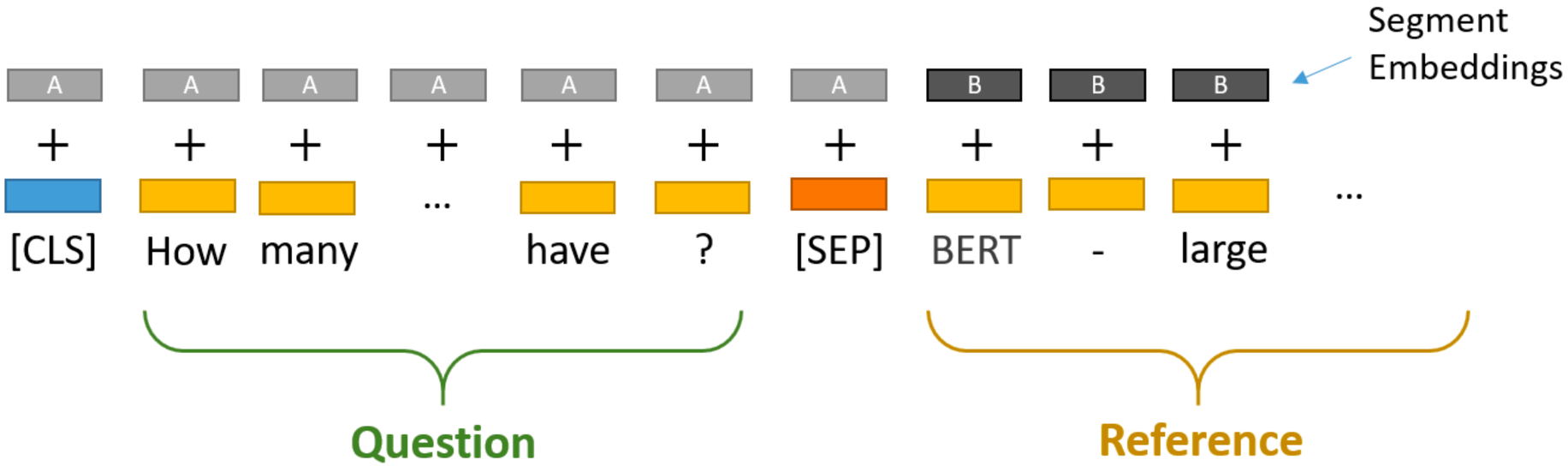


Reading comprehension: LSTM-based models with attention



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

Reading comprehension: BERT

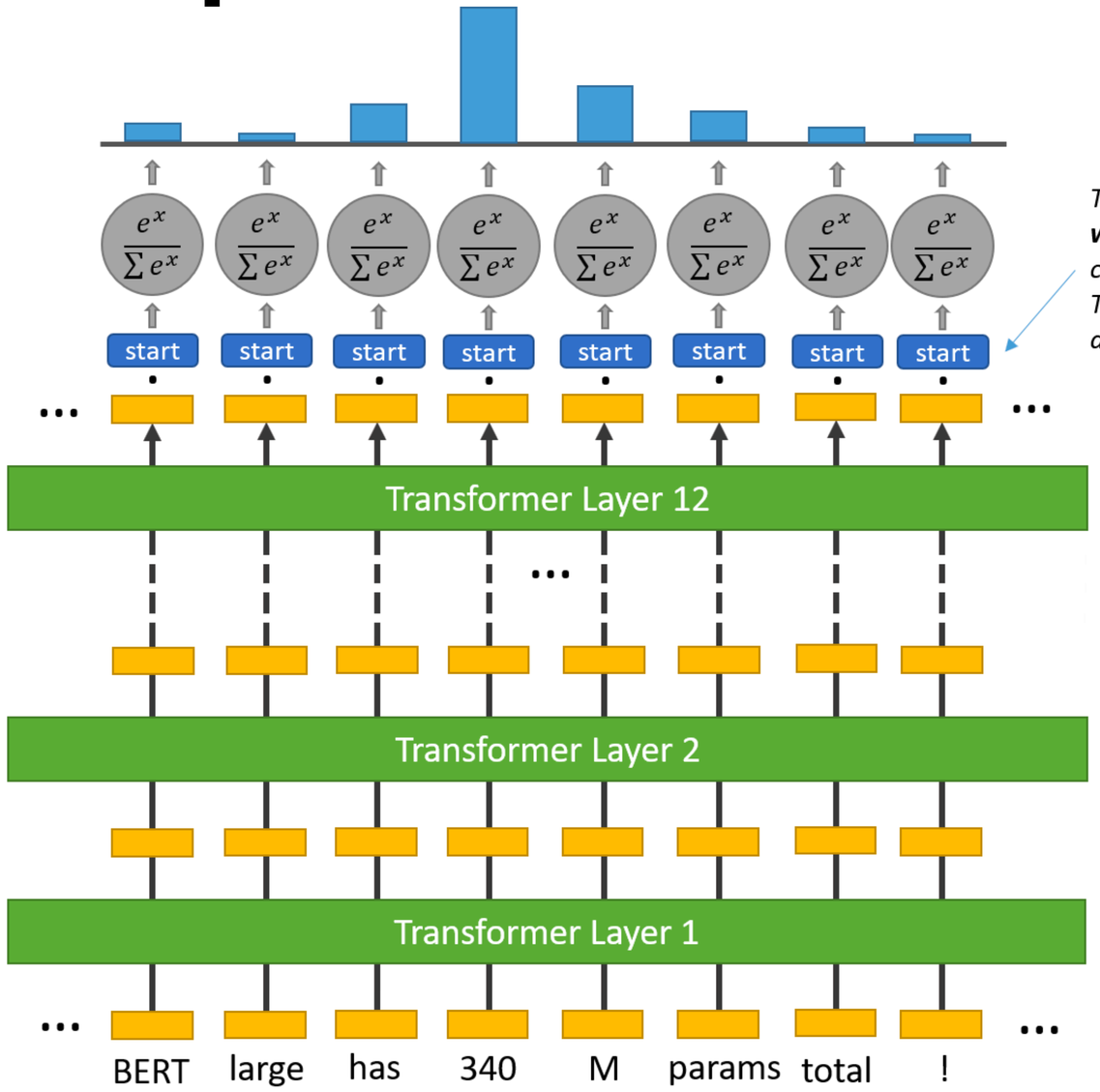


Question: How many parameters does BERT-large have?

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

Credit: https://mccormickml.com/2020/03/10/question-answering-with-a-fine-tuned-BERT/

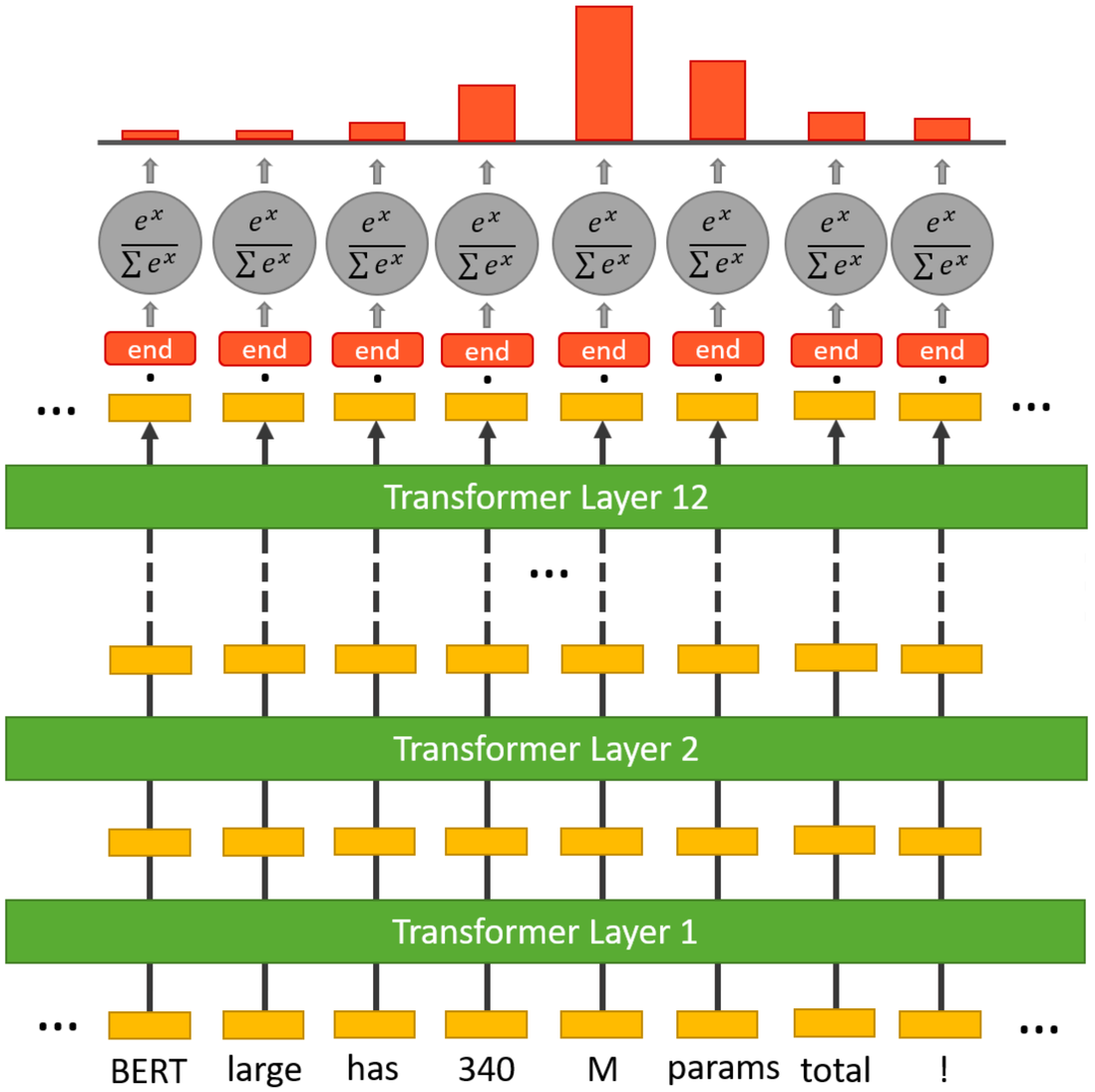
Reading comprehension: BERT



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

Credit: https://mccormickml.com/2020/03/10/question-answering-with-a-fine-tuned-BERT/

Reading comprehension: BERT

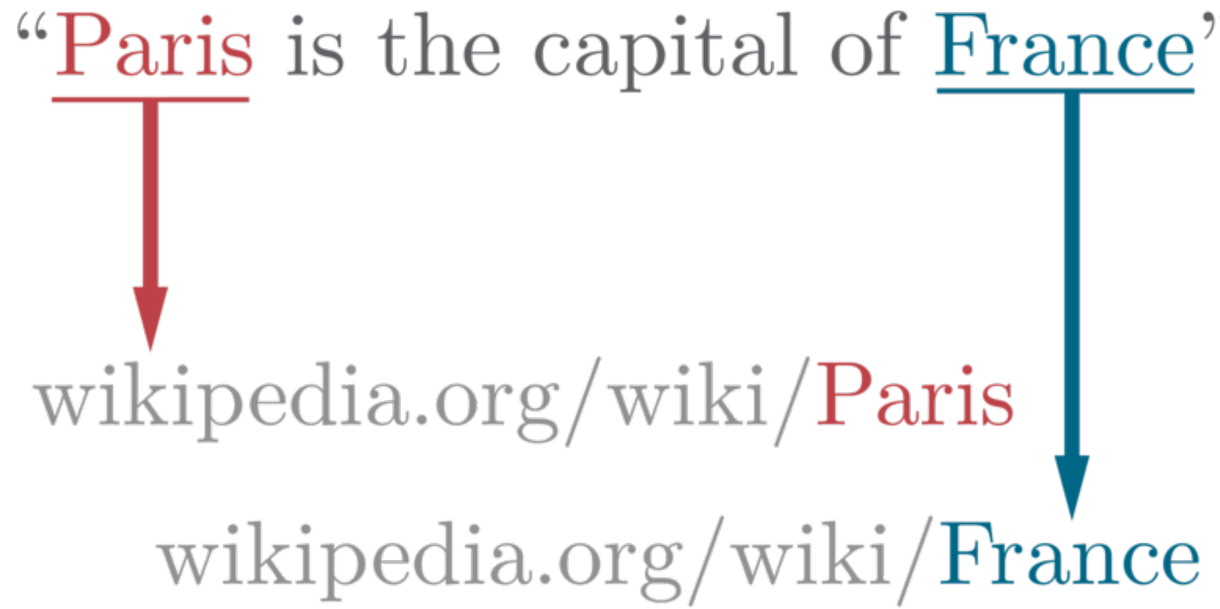


Credit: https://mccormickml.com/2020/03/10/question-answering-with-a-fine-tuned-BERT/

Entity linking

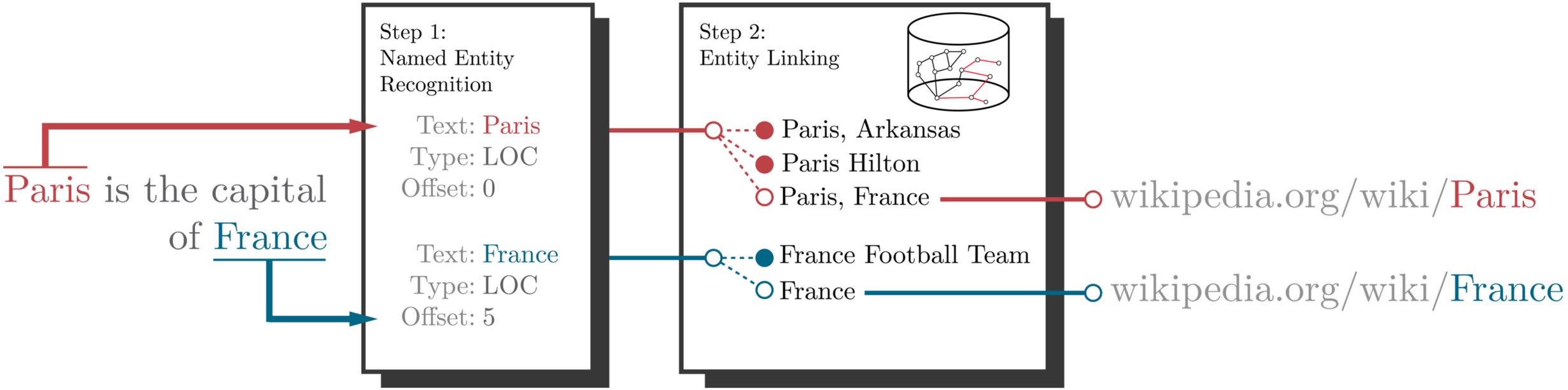
- A task of associating a mention in text with the representation of some real-world entity in an ontology

- The most common ontology for factoid question-answering is Wikipedia



Entity linking

- Done in (roughly) two stages: mention detection and mention disambiguation



Knowledge-Based QA

- Answering a natural language question by mapping it to a query over a structured database

- Two Paradigms
  - Graph-based QA
  - QA by semantic parsing

Graph-based QA

- Assumes we have a knowledge based of “facts” (facts = RDF triplets: predicate with two arguments, can also be expressed as a knowledge graph):

```
Ada Lovelace    birth-year   1815
Claude Shannon  birth-year   1916
William Shakespeare author   Hamlet
…                    …       ….
```

[data sets: SimpleQuestions, FreebaseQA, WebQA etc.]

When was Ada Lovelace born?

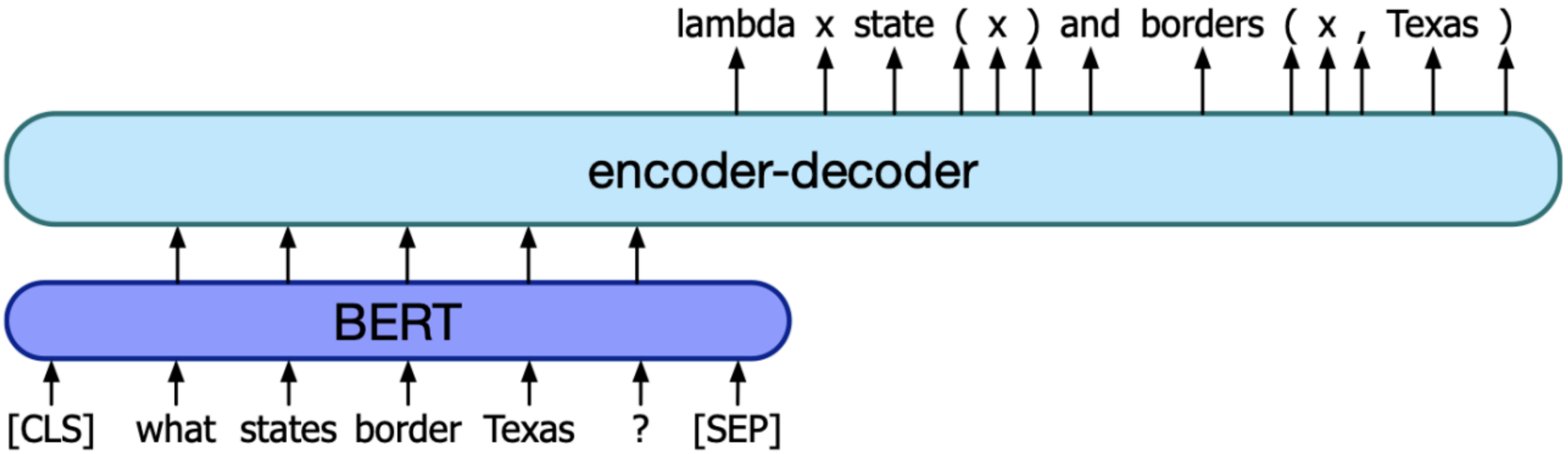
QA by semantic parsing

- map the question to a structured program to produce an answer

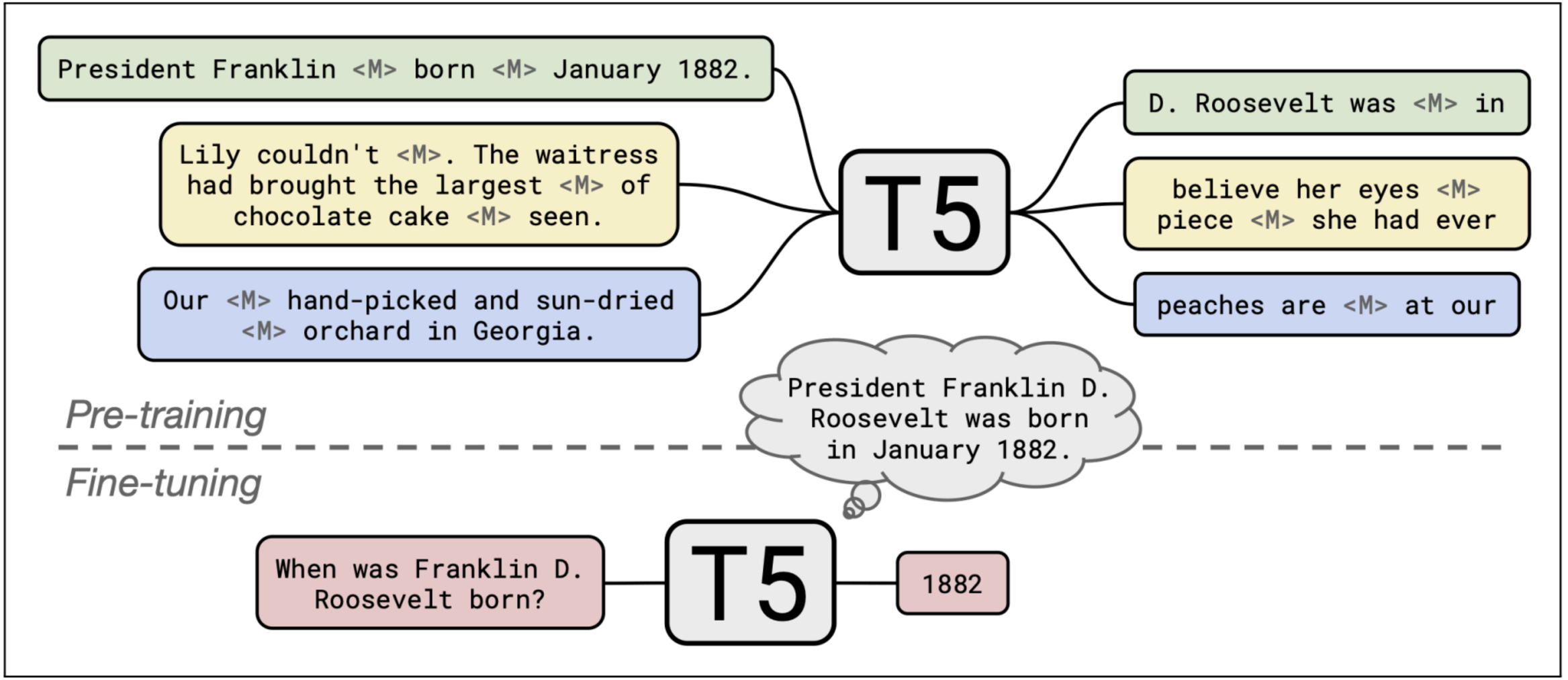
Question	Logical form
What states border Texas?	$\lambda x.\text{state}(x) \wedge \text{borders}(x, \text{texas})$
What is the largest state?	$\text{argmax}(\lambda x.\text{state}(x), \lambda x.\text{size}(x))$
I'd like to book a flight from San Diego to Toronto	SELECT DISTINCT f1.flight_id FROM flight f1, airport_service a1, city c1, airport_service a2, city c2 WHERE f1.from_airport=a1.airport_code AND a1.city_code=c1.city_code AND c1.city_name= 'san diego' AND f1.to_airport=a2.airport_code AND a2.city_code=c2.city_code AND c2.city_name= 'toronto'
How many people survived the sinking of the Titanic?	`(count (!fb:event.disaster.survivors fb:en.sinking_of_the_titanic))`
How many yards longer was Johnson's longest touchdown compared to his shortest touchdown of the first quarter?	ARITHMETIC diff(SELECT num(ARGMAX(SELECT)) SELECT num(ARGMIN(FILTER(SELECT))))

Semantic Parsing for QA

- sequence-to-sequence model



Large Language Models for QA



Visual QA

- Answer questions about an image

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

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

Antol, Stanislaw, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. “VQA: Visual question answering." In *Proceedings of the IEEE international conference on computer vision*, pp. 2425-2433. 2015.

Datasets: SQuAD

- Stanford question answering dataset (SQuAD)
  - 100k annotated (passage, question, answer) triples
  - Passages are selected from English Wikipedia, usually 100~150 words
  - Questions are crowd-sourced
  - Each answer is a short segment of text (or span) in the passage

Beyoncé Giselle Knowles-Carter (born September 4, 1981) is an American singer, songwriter, record producer and actress. Born and raised in Houston, Texas, she performed in various singing and dancing competitions as a child, and rose to fame in the late 1990s as lead singer of R&B girl-group Destiny's Child. Managed by her father, Mathew Knowles, the group became one of the world's best-selling girl groups of all time. Their hiatus saw the release of Beyoncé's debut album, Dangerously in Love (2003), which established her as a solo artist worldwide, earned five Grammy Awards and featured the Billboard Hot 100 number-one singles "Crazy in Love" and "Baby Boy".
Q: "In what city and state did Beyoncé grow up?" A: "Houston, Texas"
Q: "What areas did Beyoncé compete in when she was growing up?" A: "singing and dancing"
Q: "When did Beyoncé release Dangerously in Love?" A: "2003"

Benchmarks

Leaderboard

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

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jun 04, 2021	IE-Net (ensemble) *RICOH_SRCB_DML*	**90.939**	**93.214**
2 Feb 21, 2021	FPNet (ensemble) *Ant Service Intelligence Team*	90.871	93.183
3 May 16, 2021	IE-NetV2 (ensemble) *RICOH_SRCB_DML*	90.860	93.100
4 Apr 06, 2020	SA-Net on Albert (ensemble) *QIANXIN*	90.724	93.011
5 May 05, 2020	SA-Net-V2 (ensemble) *QIANXIN*	90.679	92.948

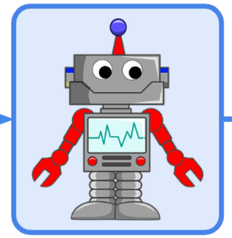
https://rajpurkar.github.io/SQuAD-explorer/

Benchmark: ScienceQA

Question: Which type of force from the baby's hand opens the cabinet door?

Options: (A) pull (B) push

Context: A baby wants to know what is inside of a cabinet. Her hand applies a force to the door, and the door opens.



Answer: The answer is A.

BECAUSE:

Lecture: A force is a push or a pull that one object applies to a second object. The direction of a push is away from the object that is pushing. The direction of a pull is toward the object that is pulling.

Explanation: The baby's hand applies a force to the cabinet door. This force causes the door to open. The direction of this force is toward the baby's hand. This force is a pull.

https://scienceqa.github.io/

Benchmark: ScienceQA

Biology
Genes to traits
Classification
Adaptations
Traits and heredity
Ecosystems
Classification
Scientific names
Heredity
Ecological interactions
Cells
Plants
Animals
Plant reproduction

Earth Science
Weather and climate
Rocks and minerals
Astronomy
Fossils
Earth events
Plate tectonics

Physics
Materials
Magnets
Velocity and forces
Force and motion
Particle motion and energy
Heat and thermal energy
States of matter
Kinetic and potential energy
Mixture

Chemistry
Solutions
Physical and chemical change
Atoms and molecules
Chemical reactions

Engineering
Designing experiments
Engineering practices

Units and Measurement
Weather and climate

Geography
State capitals
Geography
Maps
Oceania: geography
Physical Geography
The Americas: geography
Oceans and continents
Cities
States

History
Colonial America
English colonies in North America
The American Revolution

World History
Greece
Ancient Mesopotamia
World religions
American history
Medieval Asia

Civics
Social skills
Government
The Constitution

Economics
Basic economic principles
Supply and demand
Banking and finance

Global Studies
Society and environment

Writing Strategies
Supporting arguments
Sentences, fragments, and run-ons
Word usage and nuance
Creative techniques
Audience, purpose, and tone
Pronouns and antecedents
Persuasive strategies
Editing and revising
Visual elements
Opinion writing

Vocabulary
Categories
Shades of meaning
Comprehension strategies
Context clues

Grammar
Sentences and fragments
Phrases and clauses

Figurative Language
Literary devices

Verbs
Verb tense

Capitalization
Formatting

Punctuation
Fragments

Phonology
Rhyming

Reference
Research skills

https://scienceqa.github.io/

Summary

- Question answering and its applications
- Information Retrieval based Factoid QA
- Entity linking
- Knowledge-Based QA
- Other types of QA
- Benchmarks

Readings

- Chapter 14: Question Answering and Information Retrieval
  - https://web.stanford.edu/~jurafsky/slp3/14.pdf