ECE594: Mathematical Models of Language

Spring 2022

Lecture 10: Question Answering

Logistics

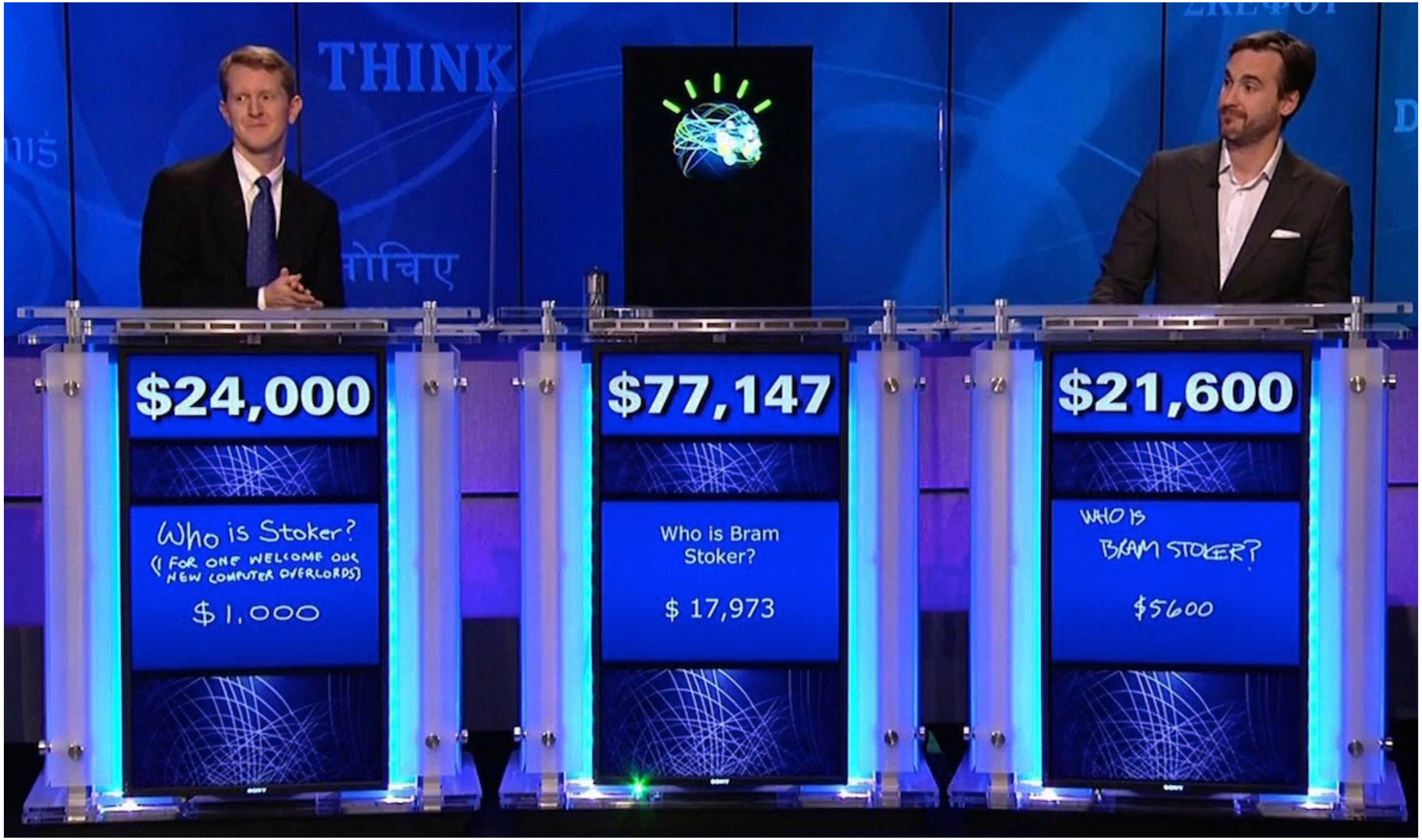
- Project progress updates 03/29
    - short presentations

UNIT 2

- Low-Resource NLP
- Summarization
- Dialog Systems
- Question Answering

Question Answering

Goal is to build systems that **automatically** answer questions posed by humans in a **natural language**



One of the oldest NLP tasks [Simmons et al. 1964]

Color Shows Mood

Cross-Curricular Focus: Visual Arts

Name: key

Artists use **color** to create patterns. Color can also show different moods. Bright colors make us feel happy and energetic. Dark colors make us feel calm or sad.

The primary colors are red, yellow, and blue. They are the colors that can be mixed together to make different colors. Mixing two primary colors makes a secondary color. The secondary colors are orange, green, and violet (purple). Orange is made by mixing yellow and red. Green is made by mixing yellow and blue. Violet is made by mixing red and blue. Intermediate colors can be made by mixing a primary and a secondary color together. Some intermediate colors are blue violet and red orange. Black, white, and gray are special colors. They are called neutral colors.

Colors have been organized into a color wheel. It shows the three primary colors, the three secondary colors, and the six intermediate colors. Artists use the the color wheel. It helps them know which colors they want to use together in their artwork.

Answer the following questions based on the reading passage. Don't forget to go back to the passage henever necessary to find or confirm your answers.

1) What kinds of colors make us feel calm?

dark colors

2) What kinds of colors make us feel like we have lots of energy?

bright colors

3) What are the primary colors?

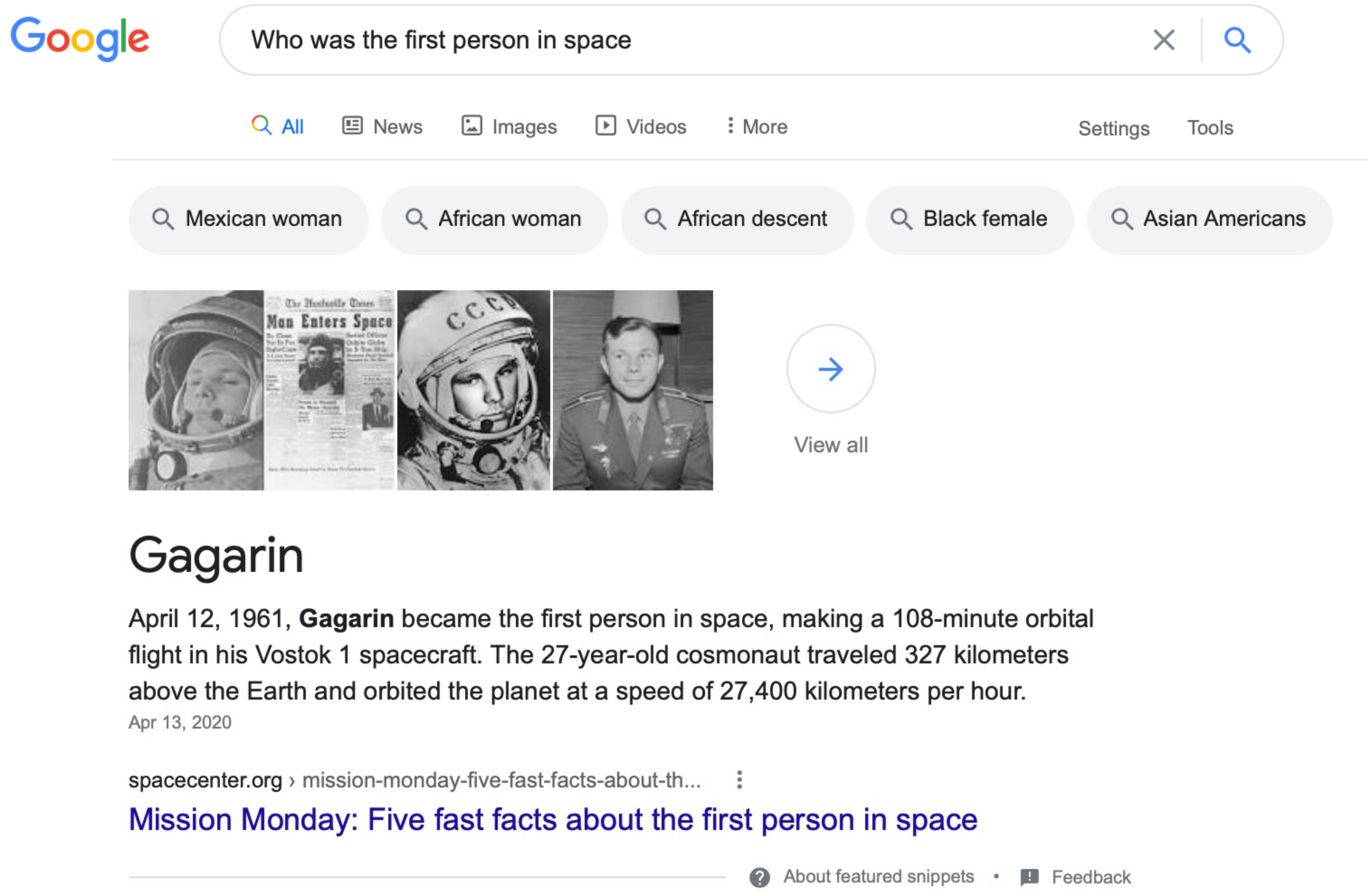
red, blue and yellow

4) What are the secondary colors?

green, orange and violet (purple)

5) What tool do artist use to organize all the colors?

a color wheel



Gagarin

April 12, 1961, **Gagarin** became the first person in space, making a 108-minute orbital flight in his Vostok 1 spacecraft. The 27-year-old cosmonaut traveled 327 kilometers above the Earth and orbited the planet at a speed of 27,400 kilometers per hour.

Apr 13, 2020

spacecenter.org › mission-monday-five-fast-facts-about-th...

Mission Monday: Five fast facts about the first person in space

- Testing language understanding
  - Understand the question
  - Relate to a position in document
  - Extract the answer

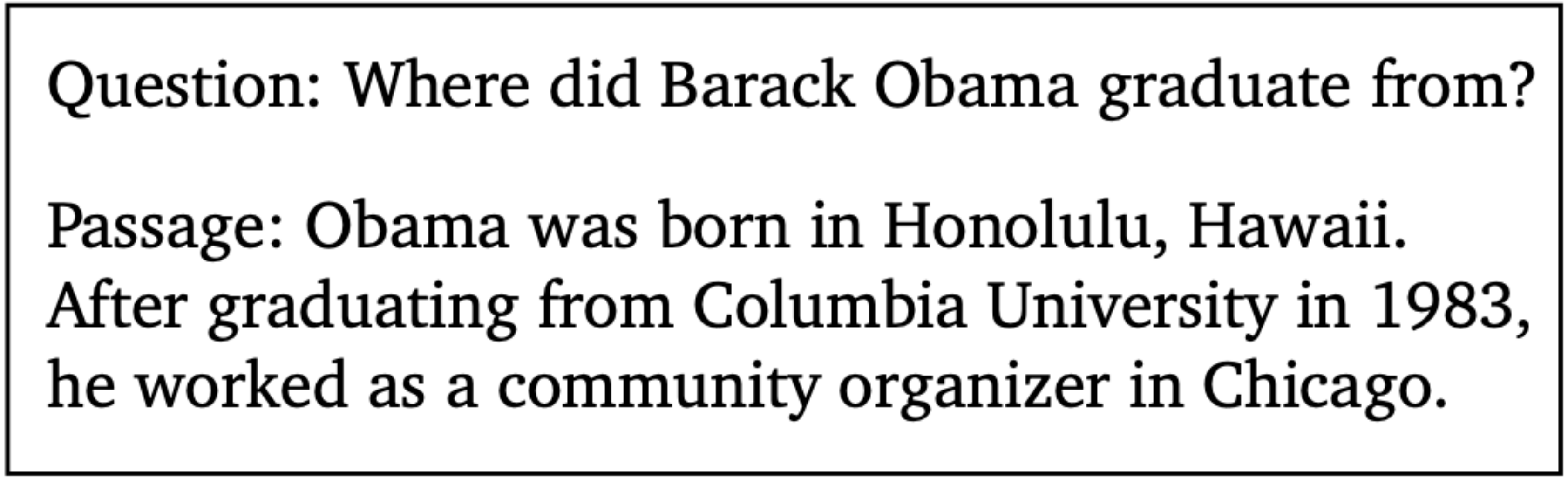
  Lehnert 1977: “Since questions can be devised to query any aspect of text comprehension, the ability to answer questions is the strongest possible demonstration of understanding.”

- Addressing human information needs
  - Talking to a virtual assistant
  - Interacting with a search engine
  - Querying a database

NLP task ⇔ QA task

- NLP tasks can be reduced to machine reading
- Given passage extract information [Levy et al. 2017]

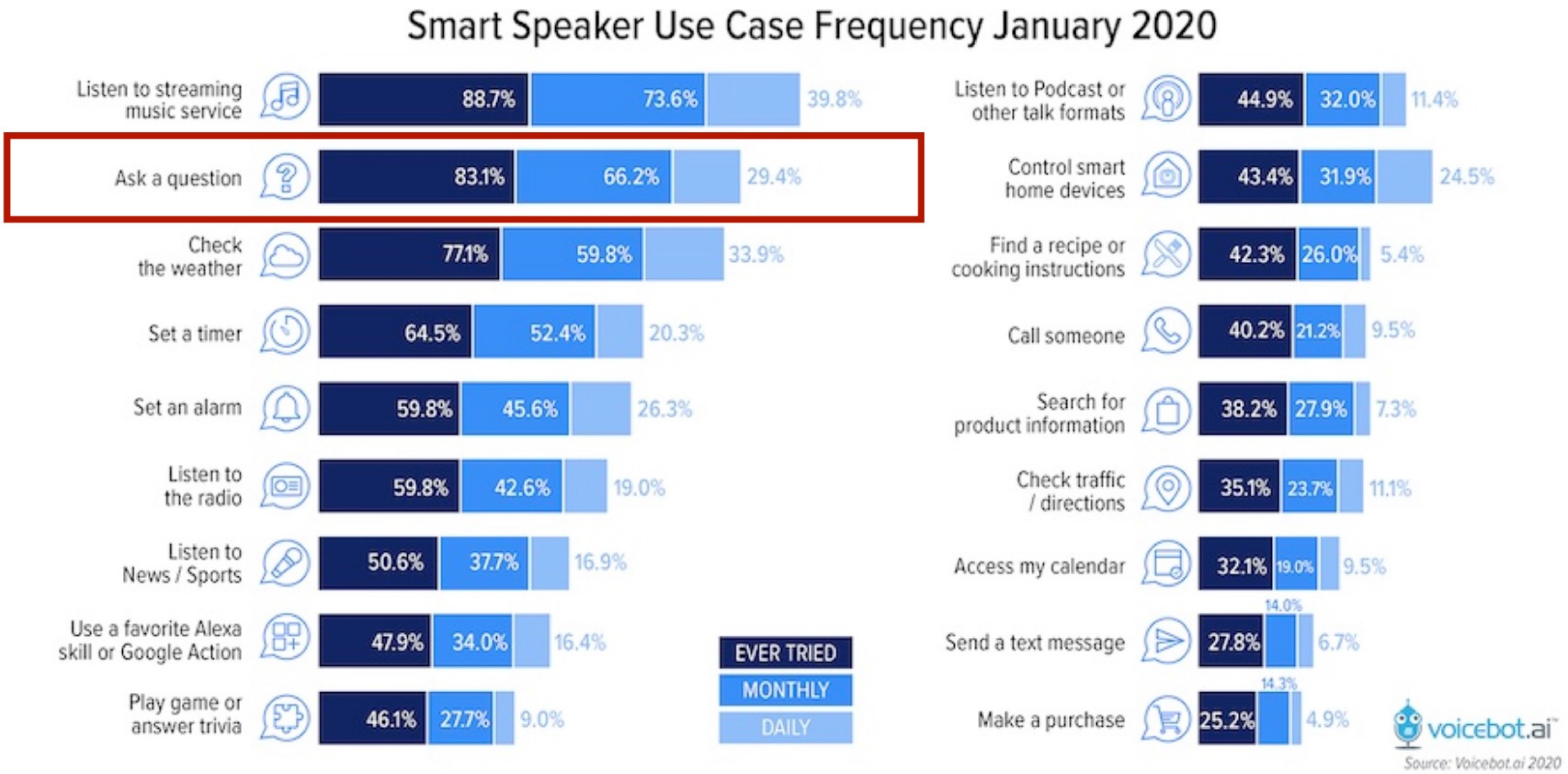
(Barack Obama, educated_at, ?)



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.

Human Need for Information



Smart Speaker Use Case Frequency January 2020

Source: Voicebot.ai 2020

Motivation

- Search results
  - Collection of documents relevant to query, not answers
- Answers to questions
  - Specific need
  - Easily accessible to/from devices

Objective

- Tasks
- Datasets
- Models

QA Categories

- Information source
  - Text passage
  - Web documents
  - Knowledge bases
  - Tables and images

QA Categories

- Information source
- Question type
  - Factoid vs non-factoid
  - Open-domain vs closed-domain
  - Simple vs complex

Types of Questions

- Questions are very broad
  - What is 4+5?
  - How do you say [sentence] in Greek?
  - Why is ocean water salty?

- Factoid Questions
  - What is the official language of Algeria?
  - When was Marie Curie born?
    - Potentially answered using a knowledge base

QA Categories

- Information source
- Question type
- Answer type
  - Short segment
  - Paragraph
  - Yes/No
  - List

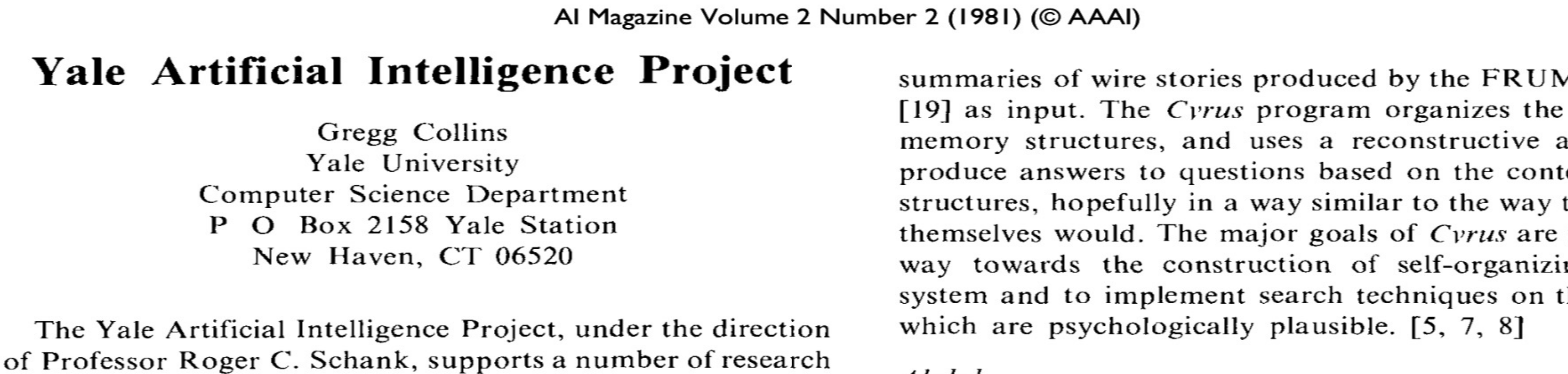
QA Tasks

- Machine reading
  - Answering questions about specific textual passages

- Open-domain question answering
  - Answering questions by collecting information from database or multiple sources

Early Work

- Simmons et al. (1964) first work on QA from text based on matching dependency parses of a question and answer
- Yale AI Project
  - Schank, Abelson, Lehnert et al. (1977)
  - Models of cognitive processes of reading comprehension



AI Magazine Volume 2 Number 2 (1981) (© AAAI)

Yale Artificial Intelligence Project

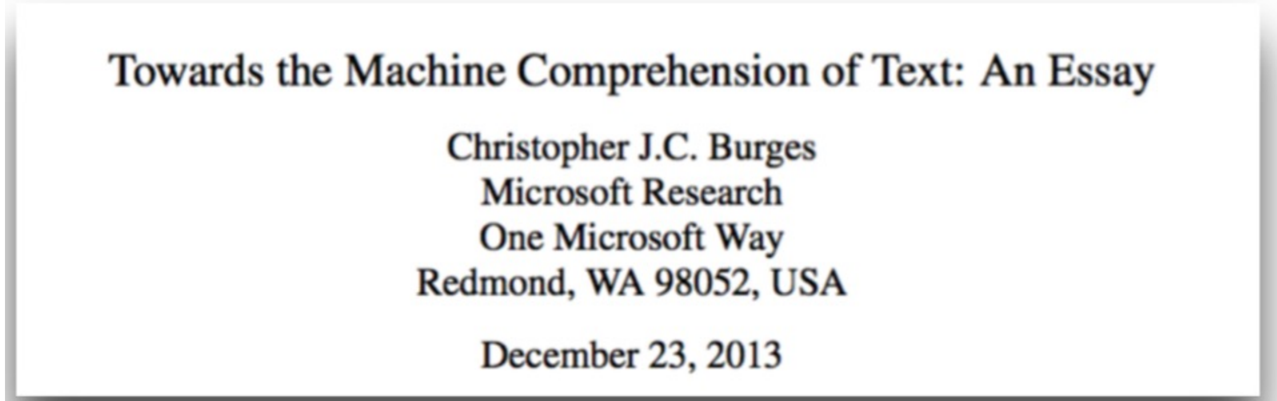
Gregg Collins
Yale University
Computer Science Department
P O Box 2158 Yale Station
New Haven, CT 06520

The Yale Artificial Intelligence Project, under the direction of Professor Roger C. Schank, supports a number of research

summaries of wire stories produced by the FRUM
[19] as input. The *Cyrus* program organizes the
memory structures, and uses a reconstructive a
produce answers to questions based on the conte
structures, hopefully in a way similar to the way t
themselves would. The major goals of *Cyrus* are
way towards the construction of self-organizi
system and to implement search techniques on t
which are psychologically plausible. [5, 7, 8]

Machine Comprehension (Burges 2013)

- "A machine **comprehends** a passage of **text** if, for any **question** regarding that text that can be **answered** correctly by a majority of native speakers, that machine can provide a string which those speakers would agree both answers that question, and does not contain information irrelevant to that question."

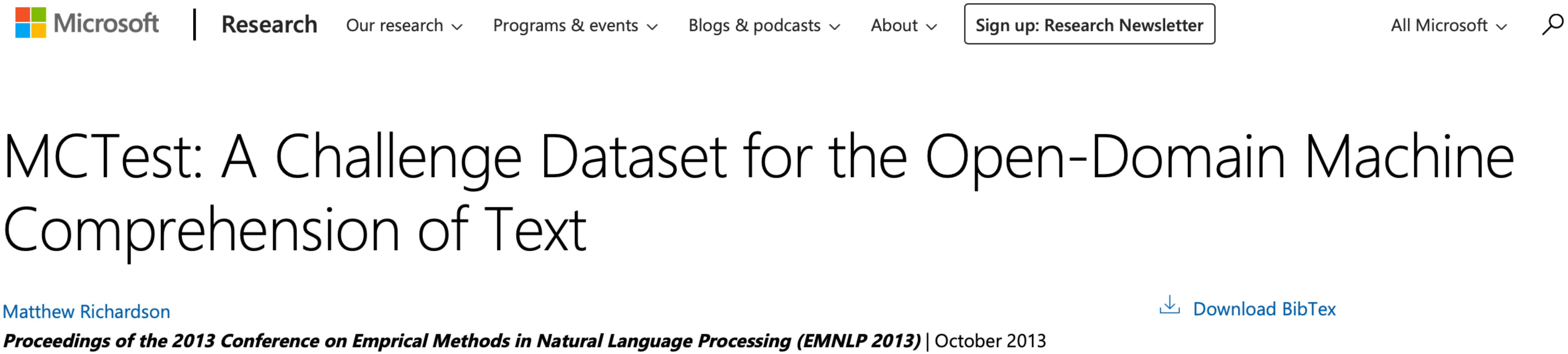


Towards the Machine Comprehension of Text: An Essay

Christopher J.C. Burges
Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA

December 23, 2013

MCTest



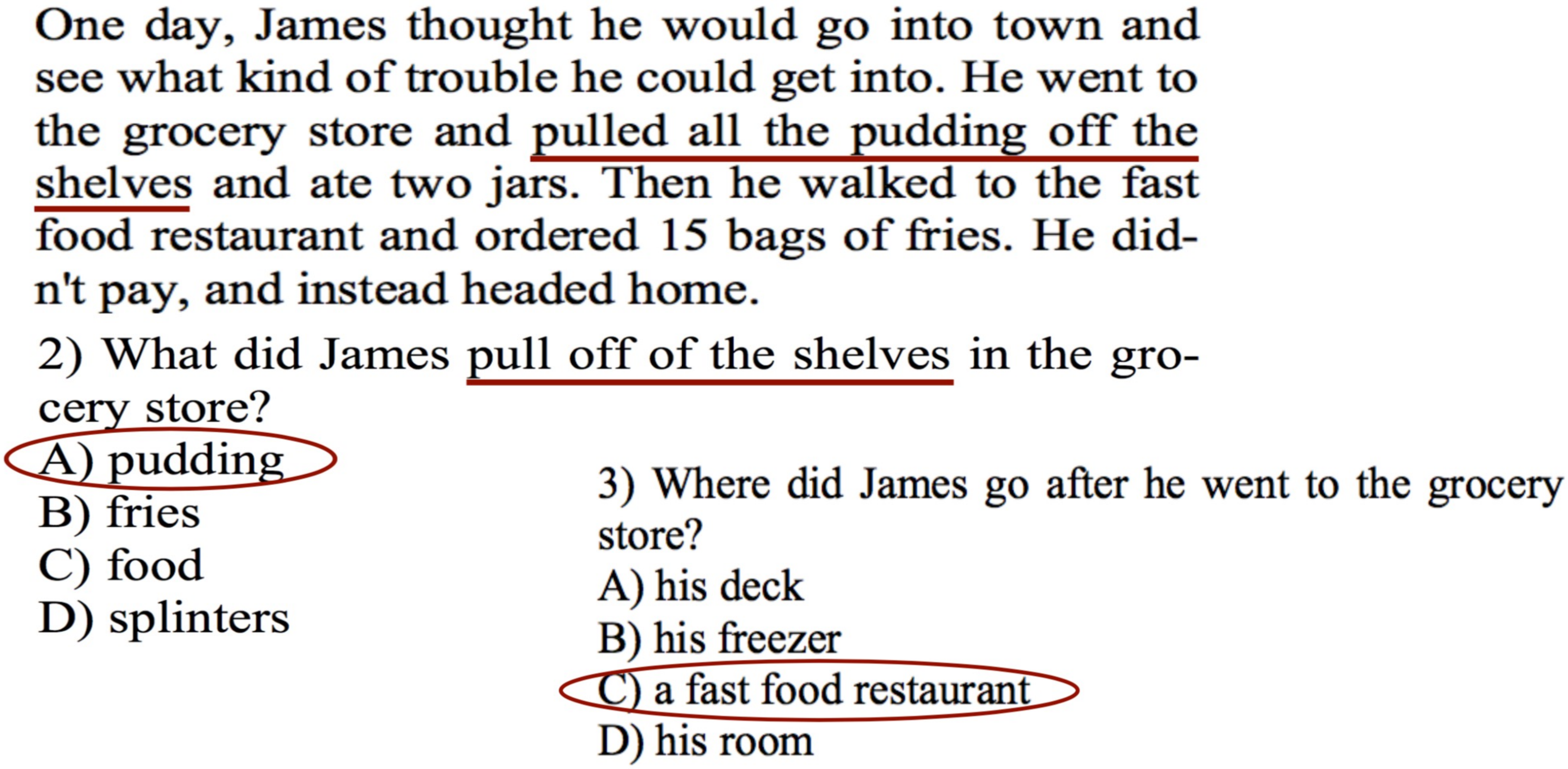
MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text

Matthew Richardson

Proceedings of the 2013 Conference on Emprical Methods in Natural Language Processing (EMNLP 2013) | October 2013

Answering questions over simple story texts

MCTest



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.

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

[✓ A) pudding]
B) fries
C) food
D) splinters

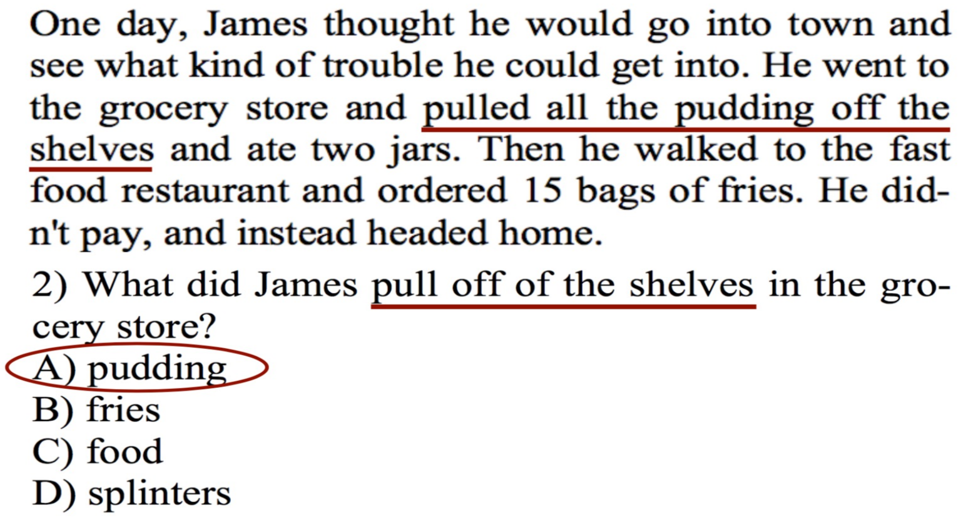
3) Where did James go after he went to the grocery store?

A) his deck
B) his freezer
[✓ C) a fast food restaurant]
D) his room

MCTest

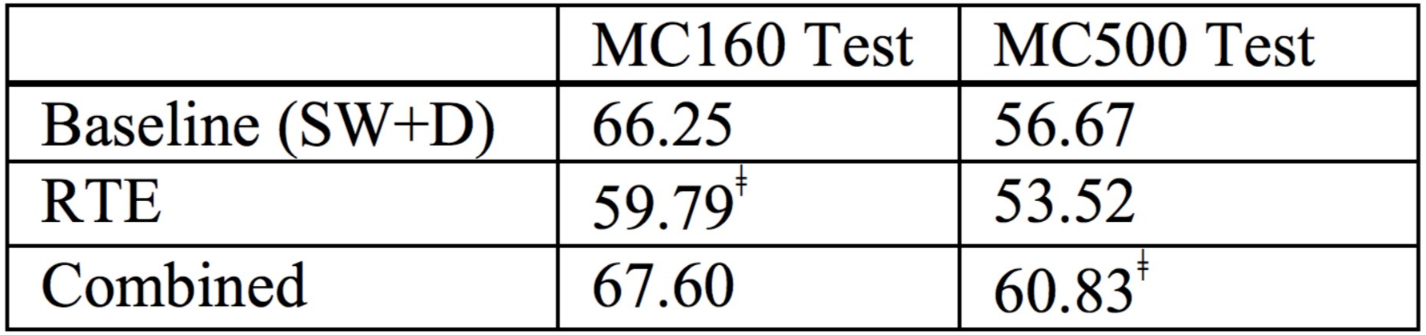
Baselines:

1. Ngram matching: append Q &A
   Return answer with highest sentence overlap
2. Dependency parse
3. Textual Entailment



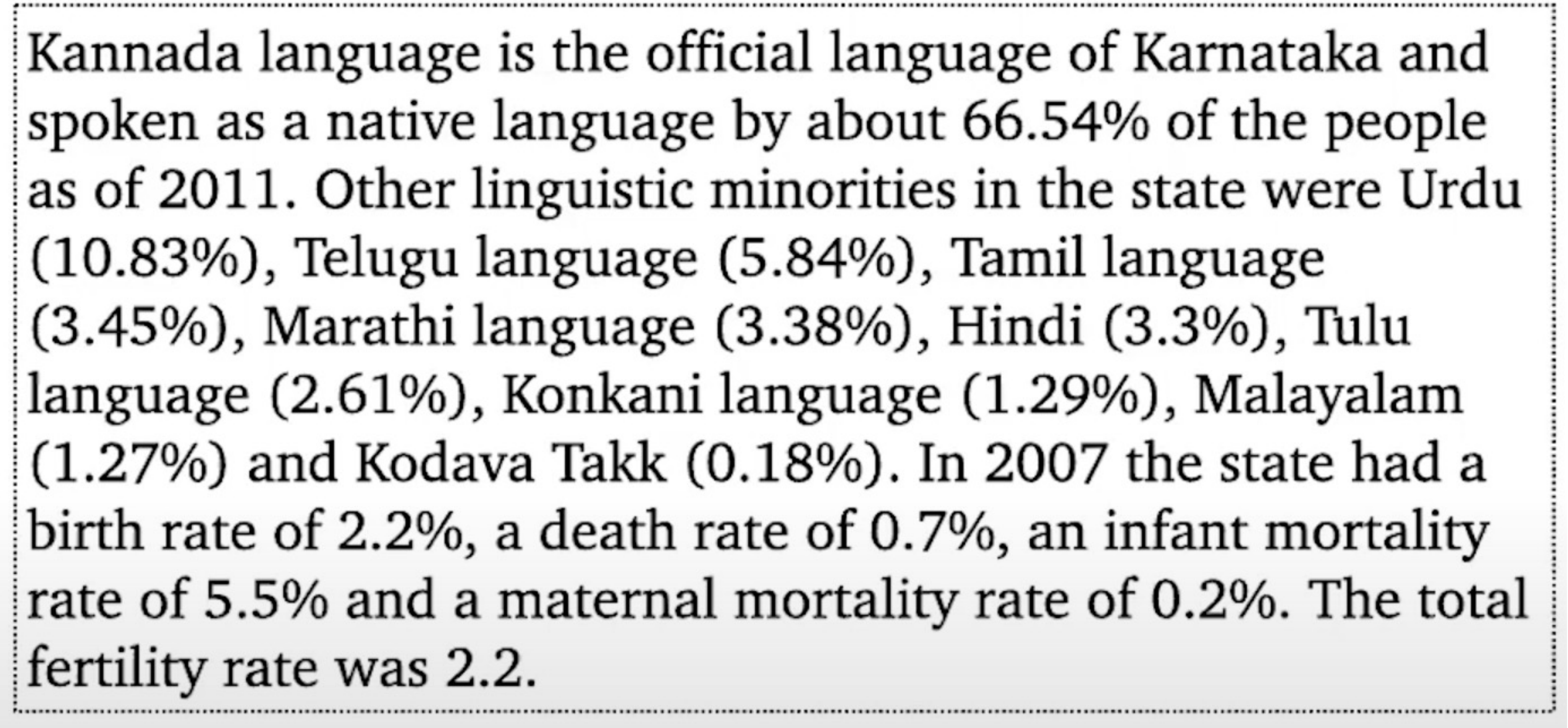
One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.

2) What did James pull off of the shelves in the grocery store?
[✓ A) pudding]
B) fries
C) food
D) splinters



	MC160 Test	MC500 Test
Baseline (SW+D)	66.25	56.67
RTE	59.79‡	53.52
Combined	67.60	60.83‡

Not enough training data



Kannada language is the official language of Karnataka and spoken as a native language by about 66.54% of the people as of 2011. Other linguistic minorities in the state were Urdu (10.83%), Telugu language (5.84%), Tamil language (3.45%), Marathi language (3.38%), Hindi (3.3%), Tulu language (2.61%), Konkani language (1.29%), Malayalam (1.27%) and Kodava Takk (0.18%). In 2007 the state had a birth rate of 2.2%, a death rate of 0.7%, an infant mortality rate of 5.5% and a maternal mortality rate of 0.2%. The total fertility rate was 2.2.

Q: Which linguistic minority is larger, Hindi or Malayalam?

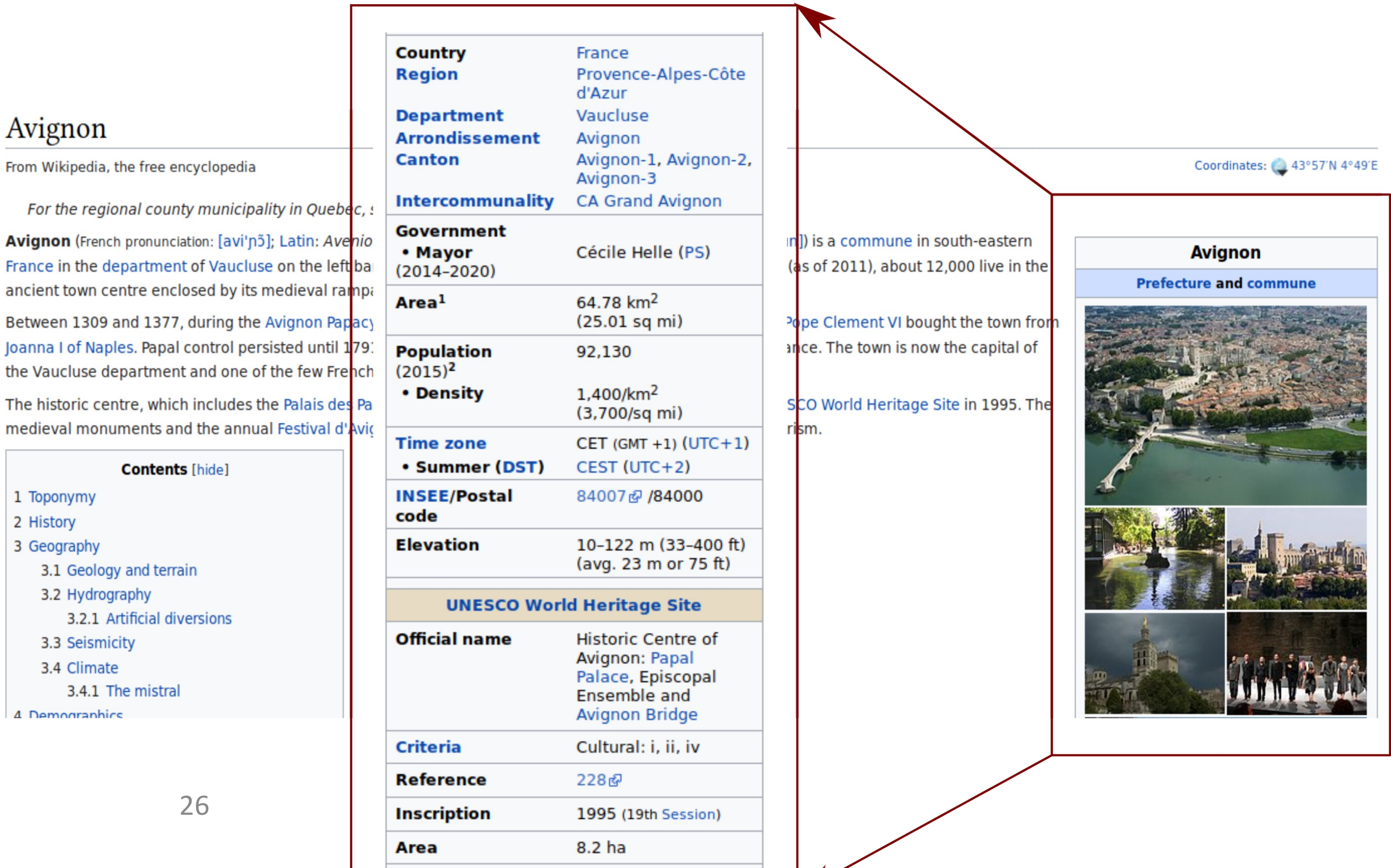
Two Paradigms of QA (factoid)

- Knowledge-base enabled
- Information-Retrieval (IR) based

Knowledge base enabled QA

- Find semantic representation of query
  - When was Ada Lovelace born? -> birth-year (Ada Lovelace, ?x)

- Query knowledge-base



Avignon

From Wikipedia, the free encyclopedia

Coordinates: 43°57′N 4°49′E

For the regional county municipality in Quebec, s

Avignon (French pronunciation: [aviˈɲɔ̃]; Latin: *Avenio* n]) is a commune in south-eastern France in the department of Vaucluse on the left ba (as of 2011), about 12,000 live in the ancient town centre enclosed by its medieval rampa

Between 1309 and 1377, during the Avignon Papacy Pope Clement VI bought the town from Joanna I of Naples. Papal control persisted until 179 nce. The town is now the capital of the Vaucluse department and one of the few French

The historic centre, which includes the Palais des Pa SCO World Heritage Site in 1995. The medieval monuments and the annual Festival d'Avig rism.

Contents [hide]

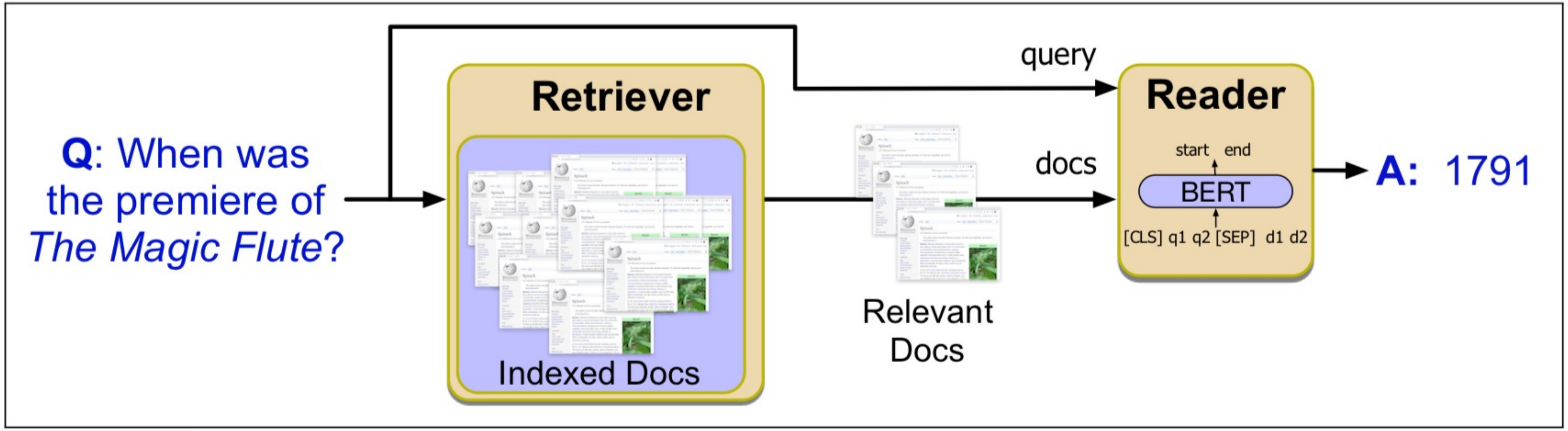
Country	France
Region	Provence-Alpes-Côte d'Azur
Department	Vaucluse
Arrondissement	Avignon
Canton	Avignon-1, Avignon-2, Avignon-3
Intercommunality	CA Grand Avignon
Government	
• **Mayor** (2014–2020)	Cécile Helle (PS)
Area[1]	64.78 km^2 (25.01 sq mi)
Population (2015)[2]	92,130
• **Density**	$1{,}400/\text{km}^2$ (3,700/sq mi)
Time zone	CET (GMT +1) (UTC+1)
• **Summer (DST)**	CEST (UTC+2)
INSEE/Postal code	84007 /84000
Elevation	10–122 m (33–400 ft) (avg. 23 m or 75 ft)
UNESCO World Heritage Site	
Official name	Historic Centre of Avignon: Papal Palace, Episcopal Ensemble and Avignon Bridge
Criteria	Cultural: i, ii, iv
Reference	228
Inscription	1995 (19th Session)
Area	8.2 ha

Avignon

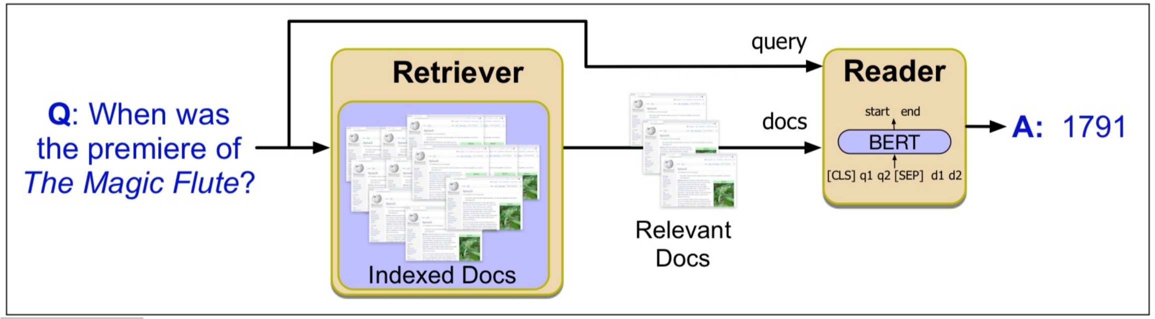
Prefecture and commune

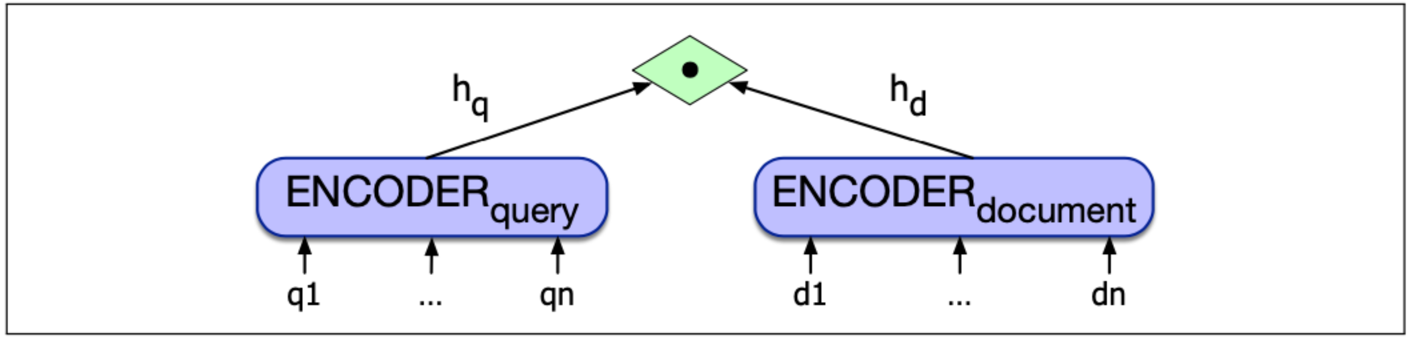
Information-Retrieval based QA

- Information-Retrieval based (aka open-domain QA)
  - Use IR to find documents with answer
  - Use machine reading comprehension to locate answer
    - Retrieve answer from a passage/document



Information-Retrieval based QA



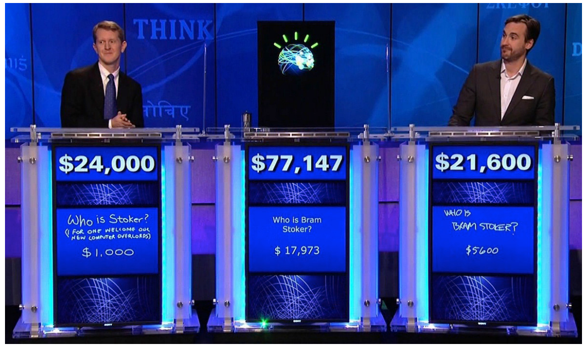


$$h_q = \mathrm{BERT}_Q(\mathrm{q})\,[\mathrm{CLS}]$$

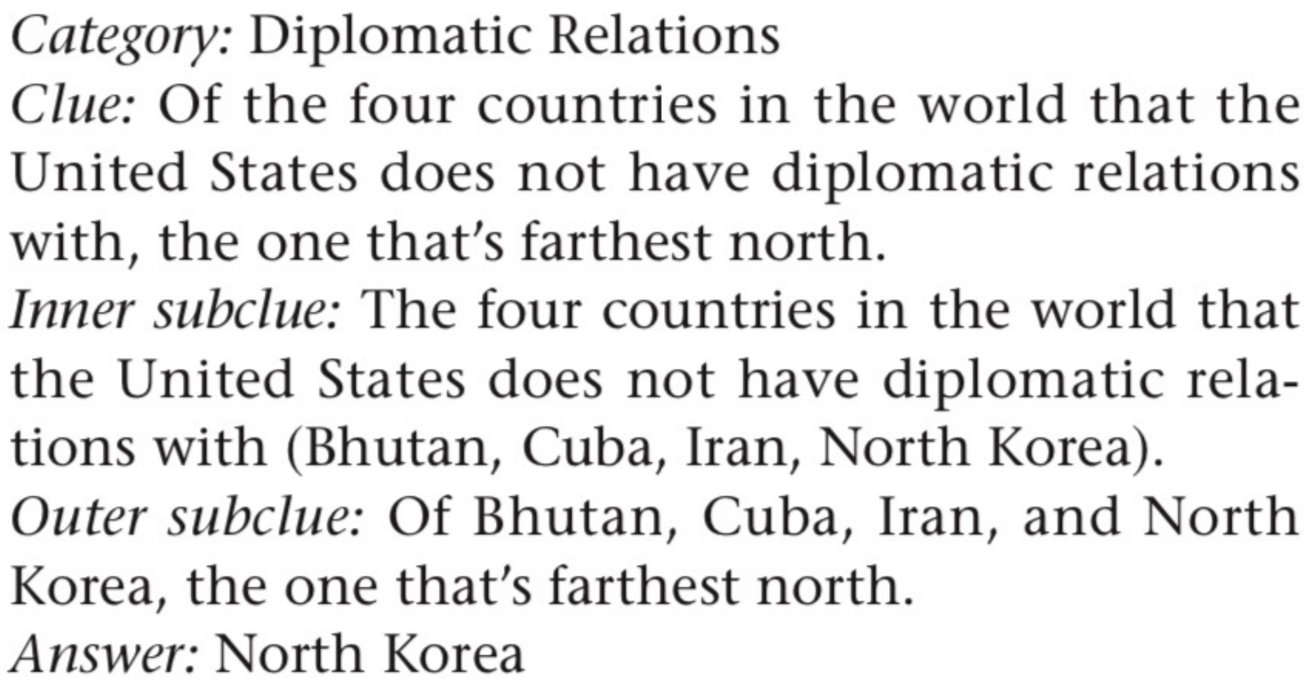
$$h_d = \mathrm{BERT}_D(\mathrm{d})\,[\mathrm{CLS}]$$

$$\mathrm{score}(d,q) = h_q \cdot h_d$$

IBM Watson



- Questions full of subtlety, puns and wordplay
  - Clue: “Colorful fourteenth century plague that became a hit play by Arthur Miller.”
  - Response: “What is
  - Exact answer not a
    - Putting together sources, becaus written anywhere



Category: Diplomatic Relations

Clue: Of the four countries in the world that the United States does not have diplomatic relations with, the one that’s farthest north.

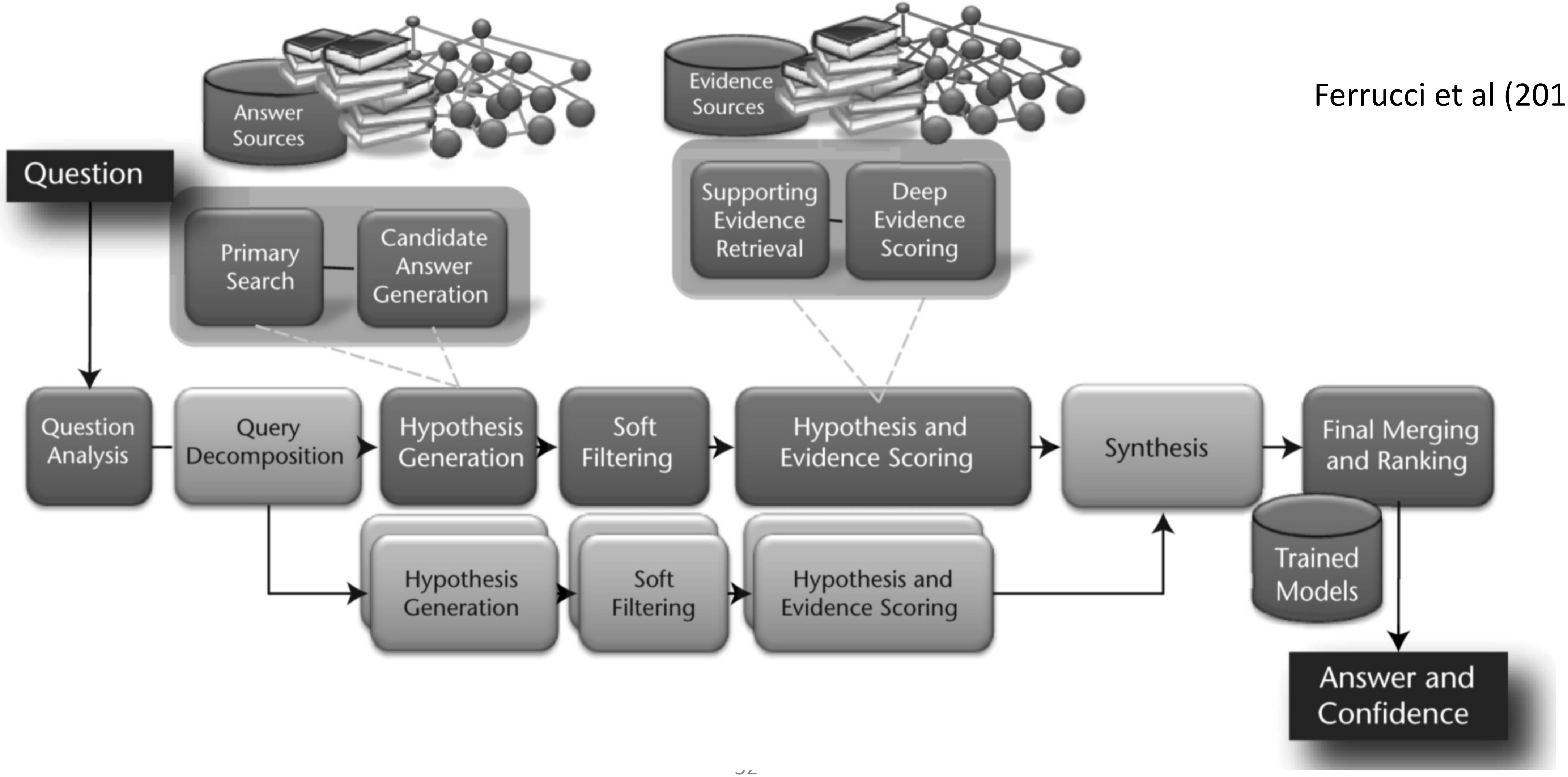
Inner subclue: The four countries in the world that the United States does not have diplomatic relations with (Bhutan, Cuba, Iran, North Korea).

Outer subclue: Of Bhutan, Cuba, Iran, and North Korea, the one that’s farthest north.

Answer: North Korea

DeepQA

Ferrucci et al (2010)



DeepQA

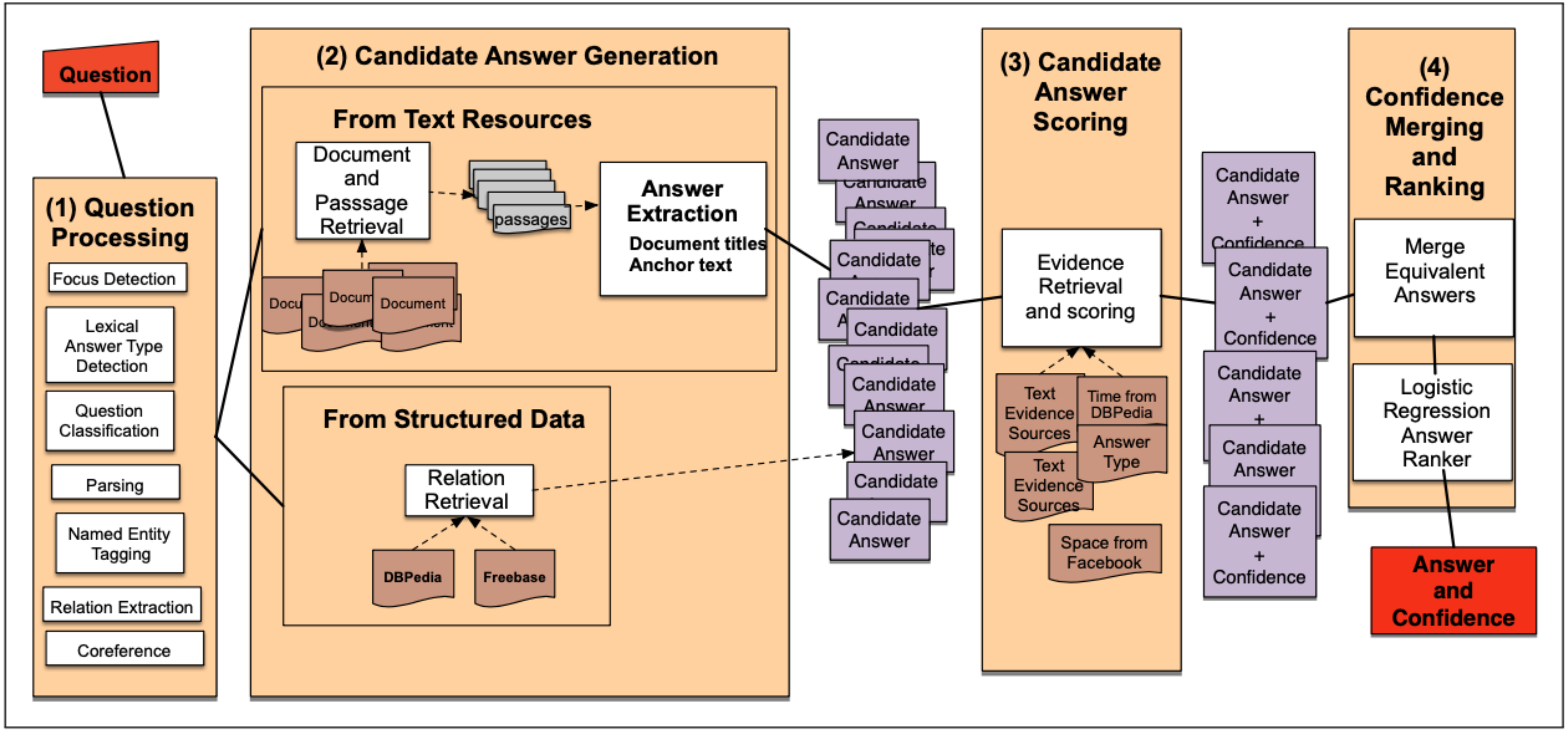


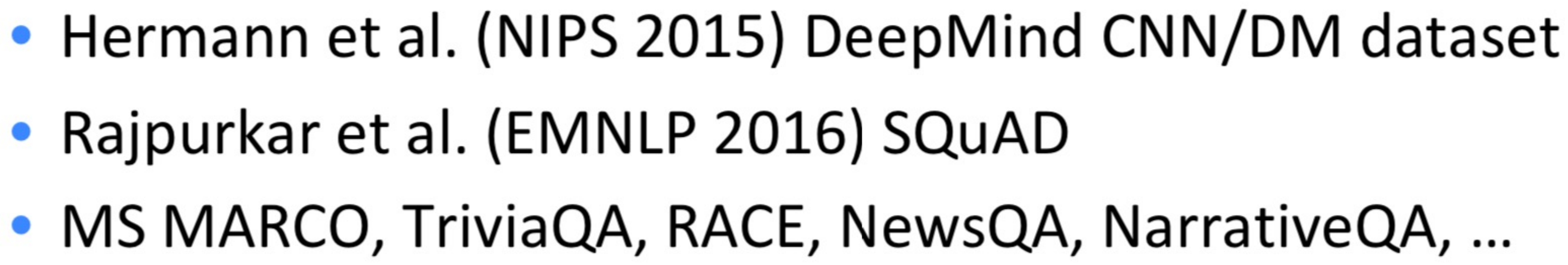
Figure 23.17 The 4 broad stages of Watson QA: (1) Question Processing, (2) Candidate Answer Generation, (3) Candidate Answer Scoring, and (4) Answer Merging and Confidence Scoring.

QA tasks

- Question answering: questions are in natural language
  - Answers: multiple choice, require picking from the passage, or generate freeform answer (last is pretty rare)
  - Require human annotation
- "Cloze" task: word (often an entity) is removed from a sentence
  - Answers: multiple choice, pick from passage, or pick from vocabulary
  - Can be created automatically from things that aren't questions

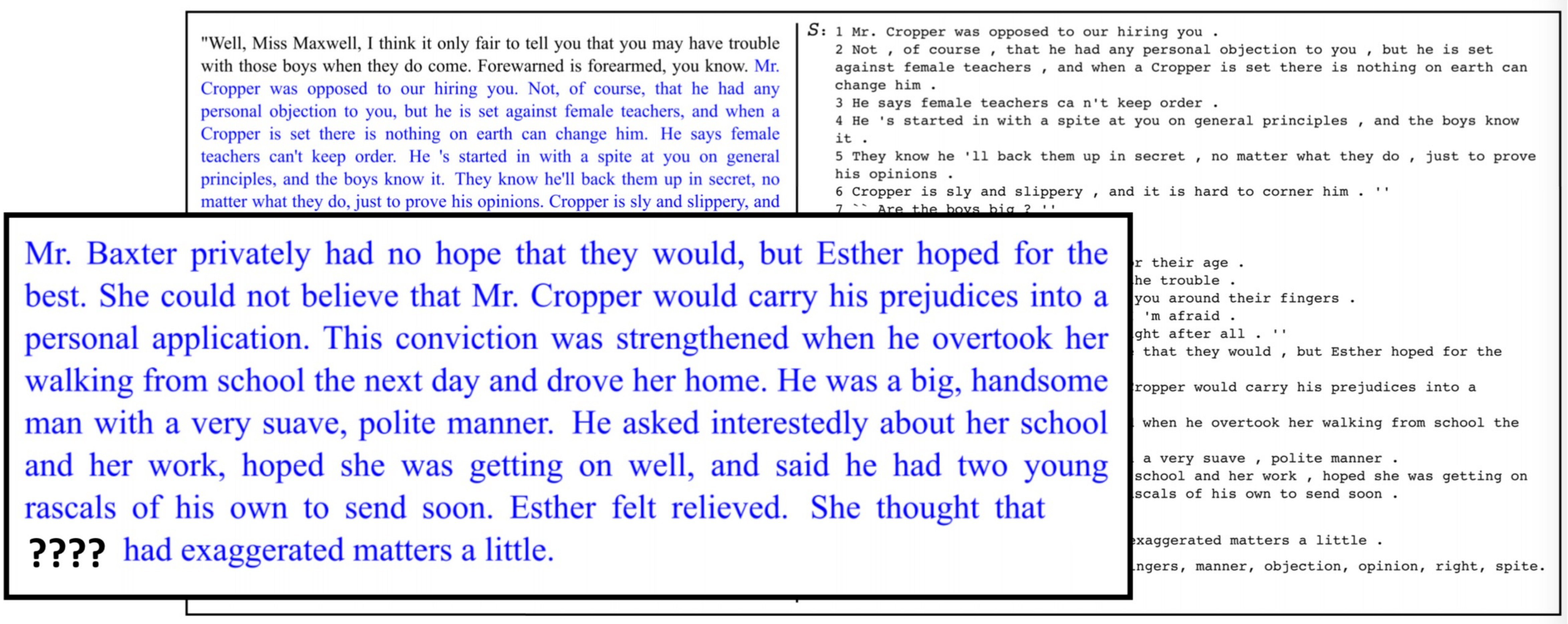
Datasets

- With rapid progress in deep learning
  - Renewed interest shown in multitude of datasets (30+)
  - Train supervised approaches



- Hermann et al. (NIPS 2015) DeepMind CNN/DM dataset
- Rajpurkar et al. (EMNLP 2016) SQuAD
- MS MARCO, TriviaQA, RACE, NewsQA, NarrativeQA, …

MR as Cloze Task



"Well, Miss Maxwell, I think it only fair to tell you that you may have trouble with those boys when they do come. Forewarned is forearmed, you know. Mr. Cropper was opposed to our hiring you. Not, of course, that he had any personal objection to you, but he is set against female teachers, and when a Cropper is set there is nothing on earth can change him. He says female teachers can't keep order. He 's started in with a spite at you on general principles, and the boys know it. They know he'll back them up in secret, no matter what they do, just to prove his opinions. Cropper is sly and slippery, and

```
S: 1 Mr. Cropper was opposed to our hiring you .
   2 Not , of course , that he had any personal objection to you , but he is set
   against female teachers , and when a Cropper is set there is nothing on earth can
   change him .
   3 He says female teachers ca n't keep order .
   4 He 's started in with a spite at you on general principles , and the boys know
   it .
   5 They know he 'll back them up in secret , no matter what they do , just to prove
   his opinions .
   6 Cropper is sly and slippery , and it is hard to corner him . ''
   7 `` Are the boys big ? ''
```

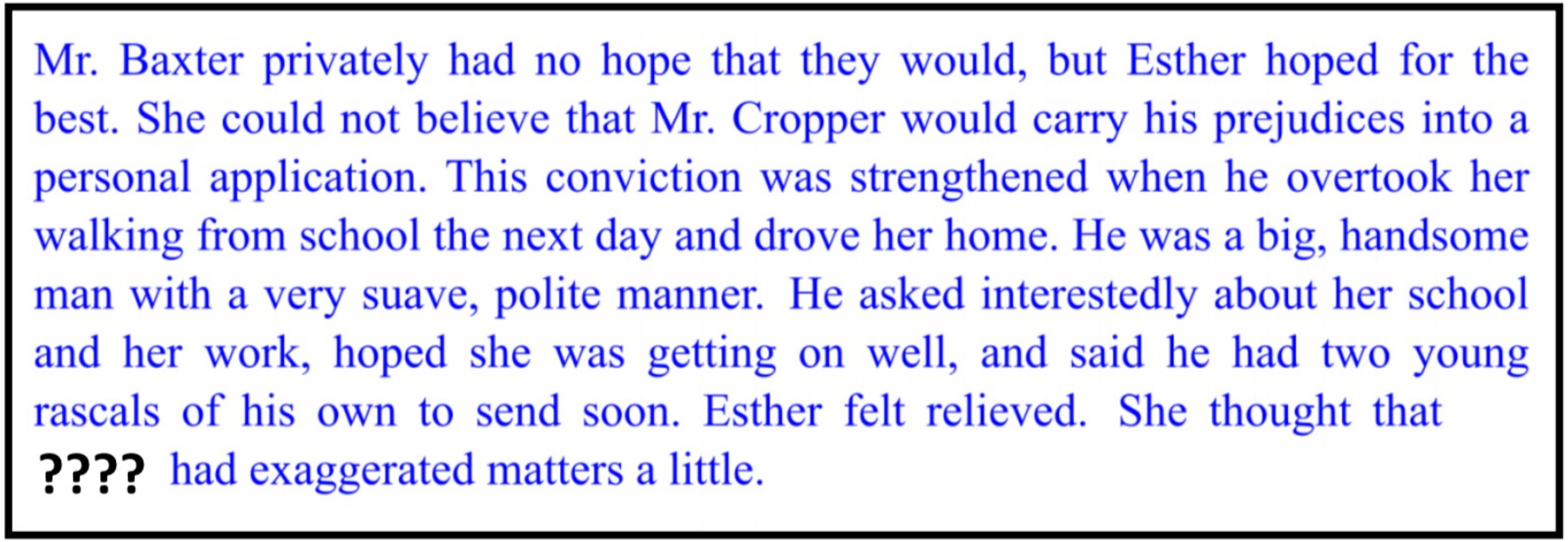
Mr. Baxter privately had no hope that they would, but Esther hoped for the best. She could not believe that Mr. Cropper would carry his prejudices into a personal application. This conviction was strengthened when he overtook her walking from school the next day and drove her home. He was a big, handsome man with a very suave, polite manner. He asked interestedly about her school and her work, hoped she was getting on well, and said he had two young rascals of his own to send soon. Esther felt relieved. She thought that

???? had exaggerated matters a little.

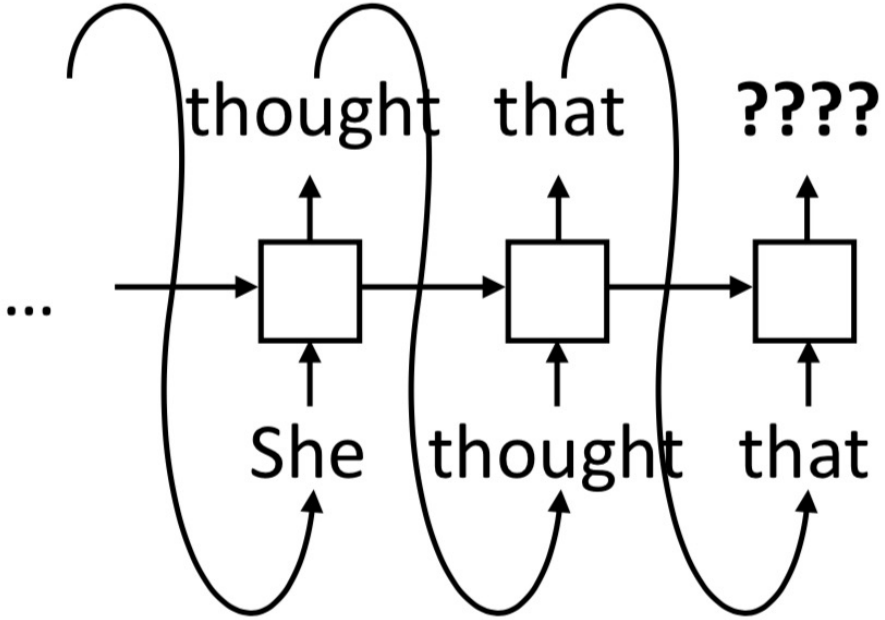
```
r their age .
he trouble .
you around their fingers .
 'm afraid .
ght after all . ''
 that they would , but Esther hoped for the

ropper would carry his prejudices into a

 when he overtook her walking from school the

 a very suave , polite manner .
school and her work , hoped she was getting on
scals of his own to send soon .

xaggerated matters a little .
ngers, manner, objection, opinion, right, spite.
```

- Children's Book Test: take a section of a children's story, block out an entity and predict it (one-doc multi-sentence cloze task)

Hill et al. (2015)



Mr. Baxter privately had no hope that they would, but Esther hoped for the best. She could not believe that Mr. Cropper would carry his prejudices into a personal application. This conviction was strengthened when he overtook her walking from school the next day and drove her home. He was a big, handsome man with a very suave, polite manner. He asked interestedly about her school and her work, hoped she was getting on well, and said he had two young rascals of his own to send soon. Esther felt relieved. She thought that **????** had exaggerated matters a little.



- Predict next word with LSTM LM
- Context: either just the current sentence (query) or the whole document up to this point (query+context)

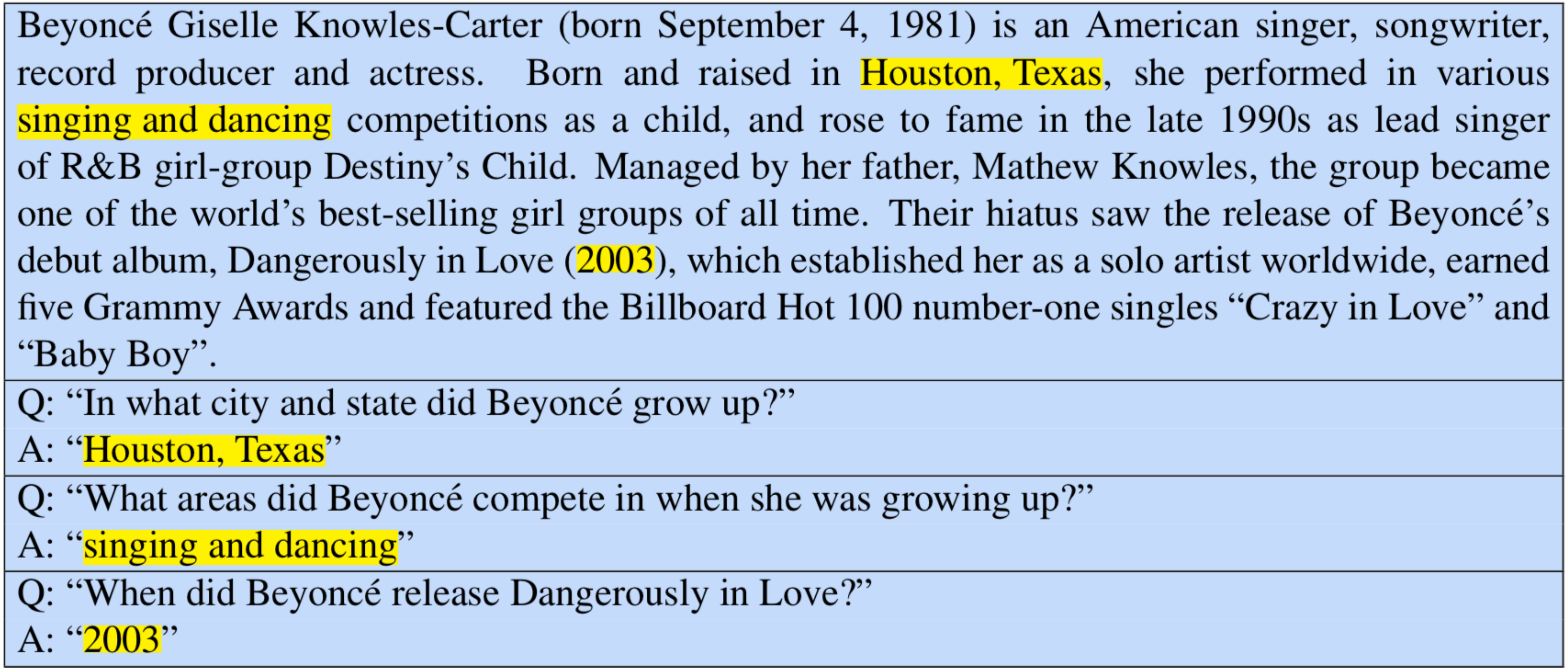
Hill et al. (2015)

IR-Based QA

- Stage 1: Retrieve using IR

- Stage 2: Machine Reading
  - Currently most systems do extractive QA
    - Answer is a span of text in the passage
    - Modeled by span labeling: identifying in the passage a span that constitutes an answer
    - Given a question *q* of *n* tokens $q_1,...,q_n$ and a passage *p* of *m* tokens p_1 , ..., p_m compute the probability $P(a|q, p)$ that each possible span *a* is the answer

SQuAD 1.0 and 2.0 (Rajpurkar et al. 2016)



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"

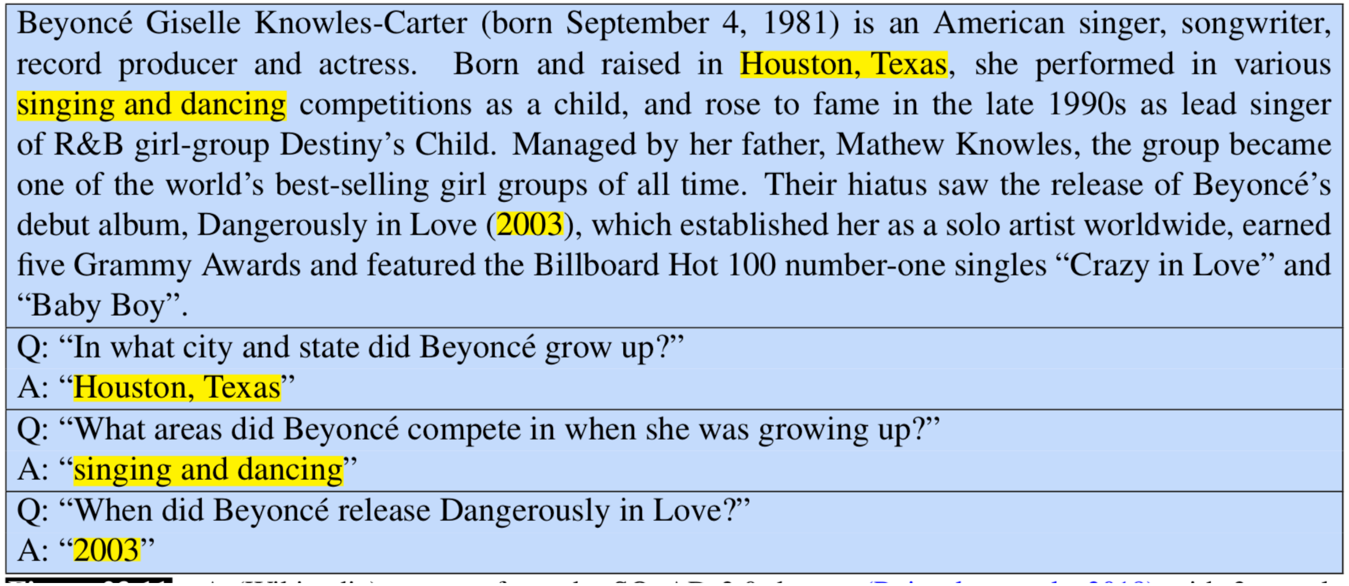
Q: "What areas did Beyoncé compete in when she was growing up?"

A: "singing and dancing"

Q: "When did Beyoncé release Dangerously in Love?"

A: "2003"

SQuAD 1.0 and 2.0



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"

- Annotators given Wikipedia article
  - Create questions based on article
  - Problem?

  - TriviaQA meant to address that

Neural Models for MR

- Problem formulation
  - Input: $C = (c_1, c_2, \ldots, c_N), Q = (q_1, q_2, \ldots, q_M), c_i, q_i \in V$
  - Output: $1 \leq$ start $\leq$ end $\leq N$

N~100, M ~15

answer is a span in the passage

- A family of LSTM-based models with attention (2016–2018)

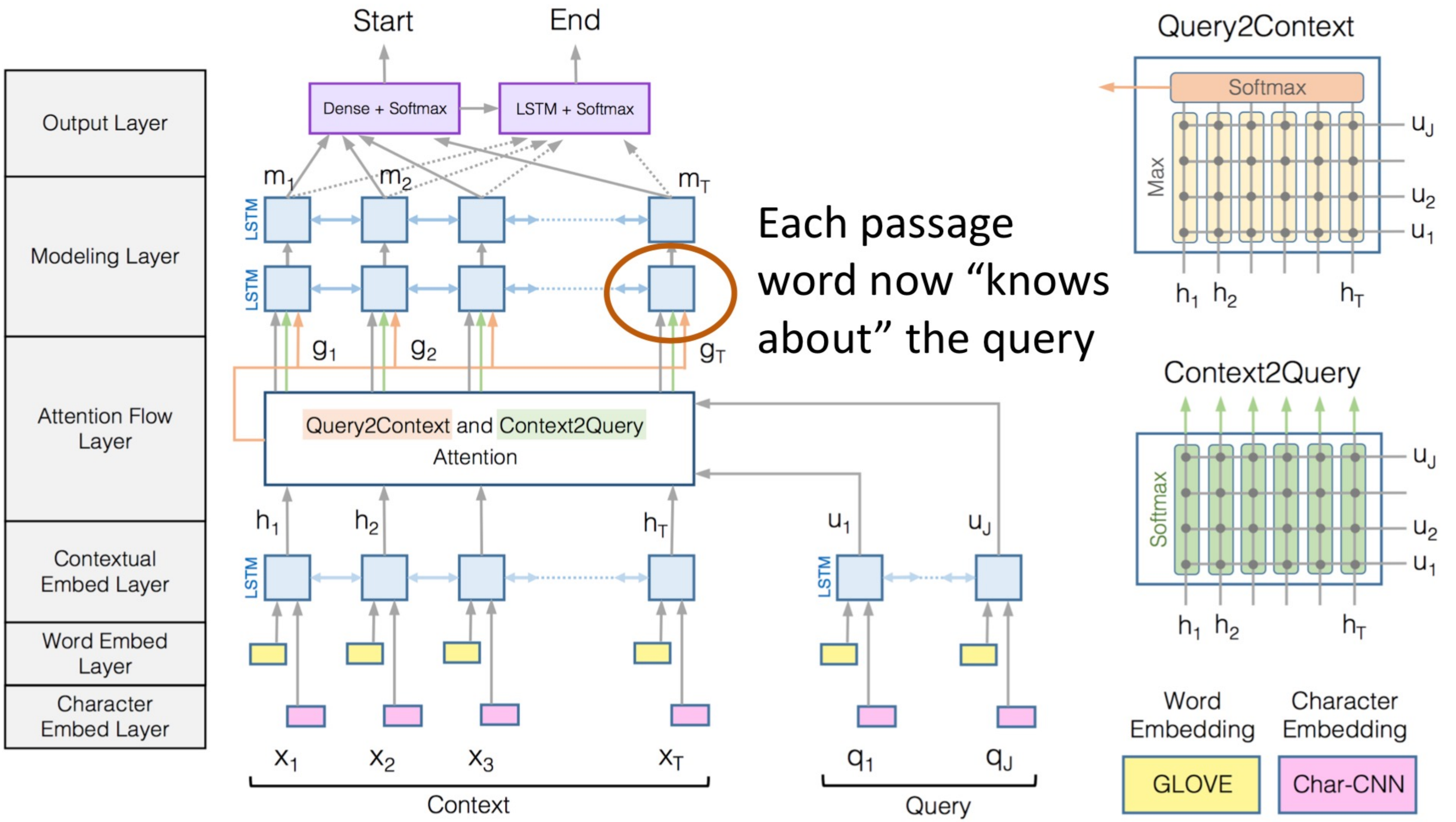
  Attentive Reader (Hermann et al., 2015), Stanford Attentive Reader (Chen et al., 2016), Match-LSTM (Wang et al., 2017), BiDAF (Seo et al., 2017), Dynamic coattention network (Xiong et al., 2017), DrQA (Chen et al., 2017), R-Net (Wang et al., 2017), ReasoNet (Shen et al., 2017)..

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

Key idea: Need to model which words in the passage are most relevant to the question (and question words)

Slides adapted from: Danqi Chen

Bidirectional Attention Flow (BiDAF)



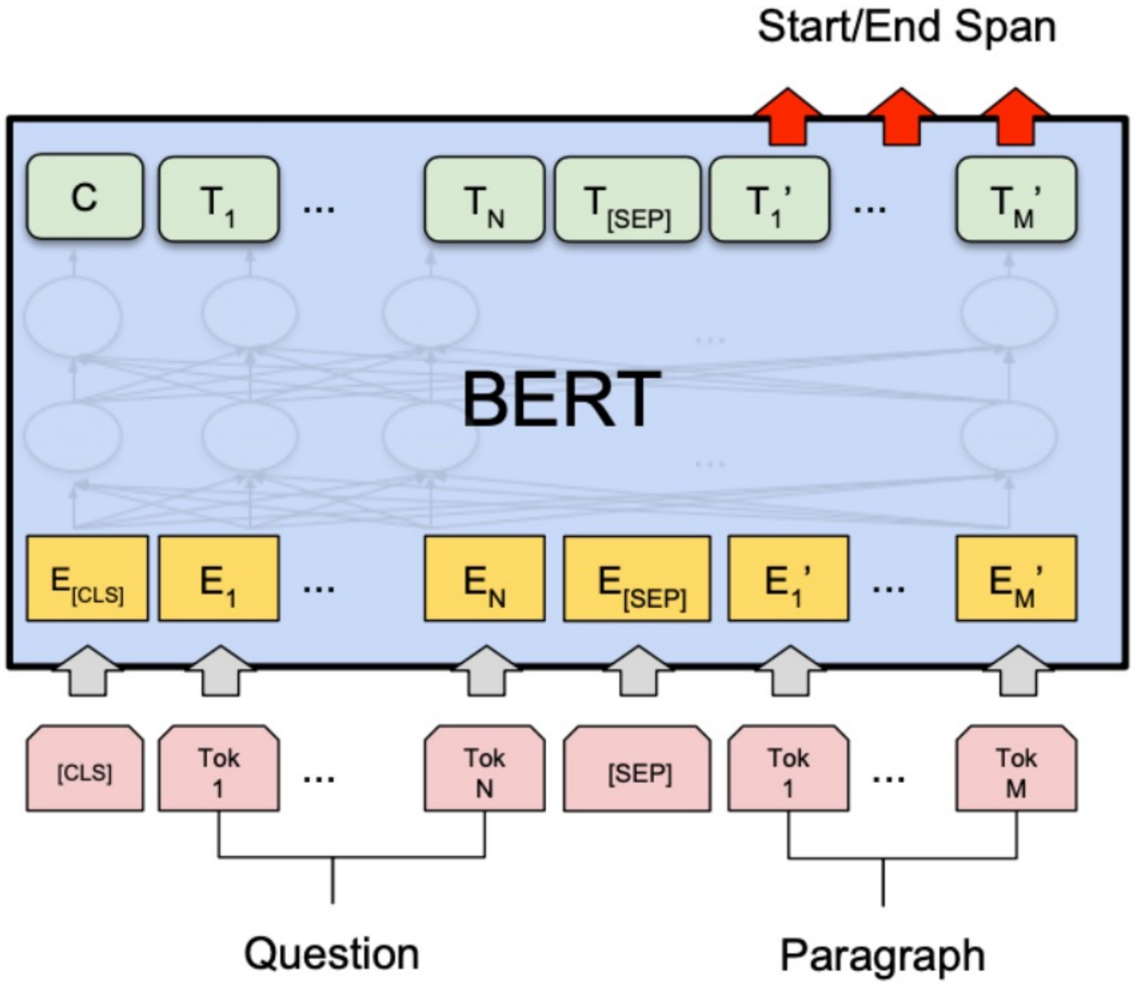
Each passage word now "knows about" the query

Seo et al. (2016)

Bidirectional Attention Flow (BiDAF)

- **Attention Flow layer**
- **Idea:** attention should flow both ways – from the context to the question and from the question to the context

Pretraining + Finetuning



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

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

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

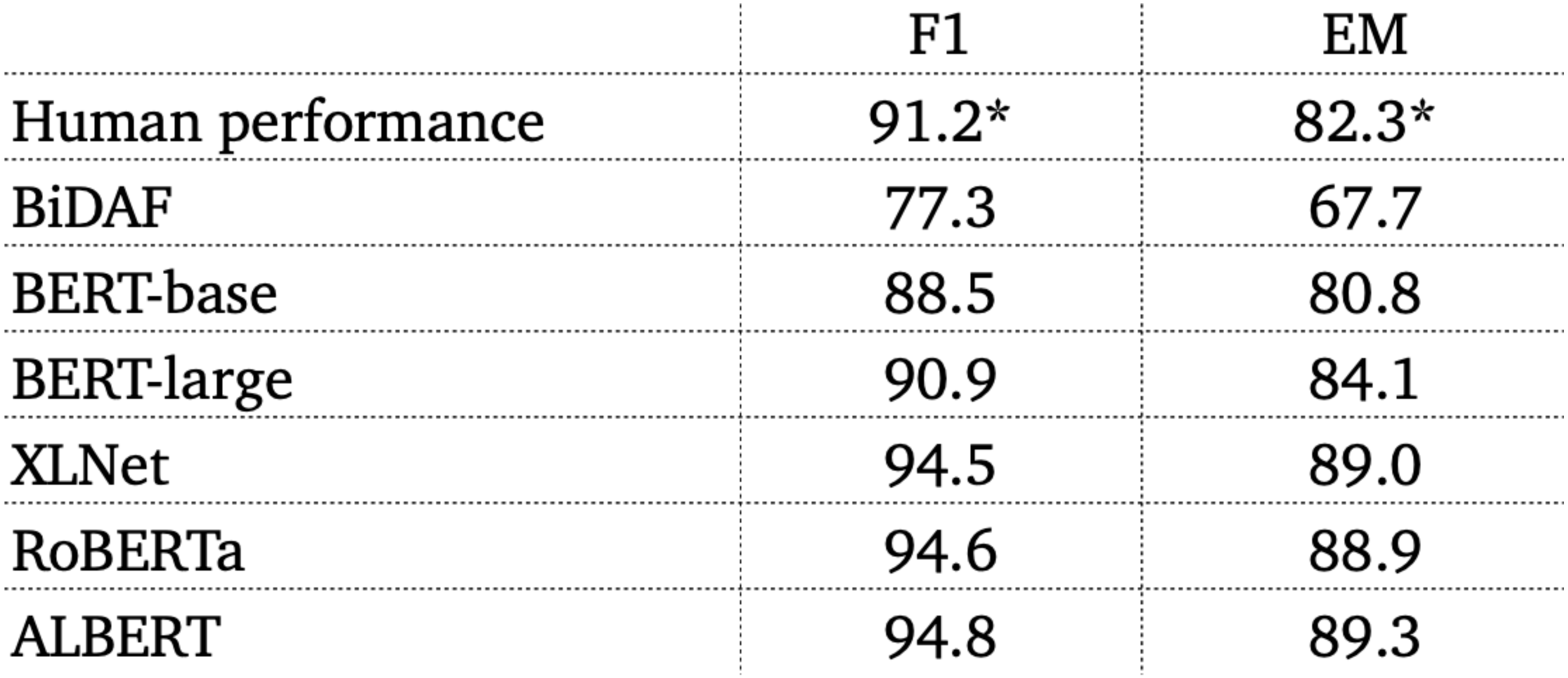
where $\mathbf{h}_i$ is the hidden vector of c_i, returned by BERT

All BERT parameters (e.g., 110M) including newly introduced parameters (w's) optimized for L

Slide credits: Danqi Chen

Pretraining + Finetuning

Evaluation: exact match (0 or 1) and F1 (partial credit)



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

Pretraining + Finetuning

- For development and testing sets, 3 gold answers are collected, because there could be multiple plausible answers.
- We compare the predicted answer to *each* gold answer (a, an, the, punctuations are removed) and take max scores. Finally, we take the average of all the examples for both exact match and F1.
- Estimated human performance: EM = 82.3, F1 = 91.2

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

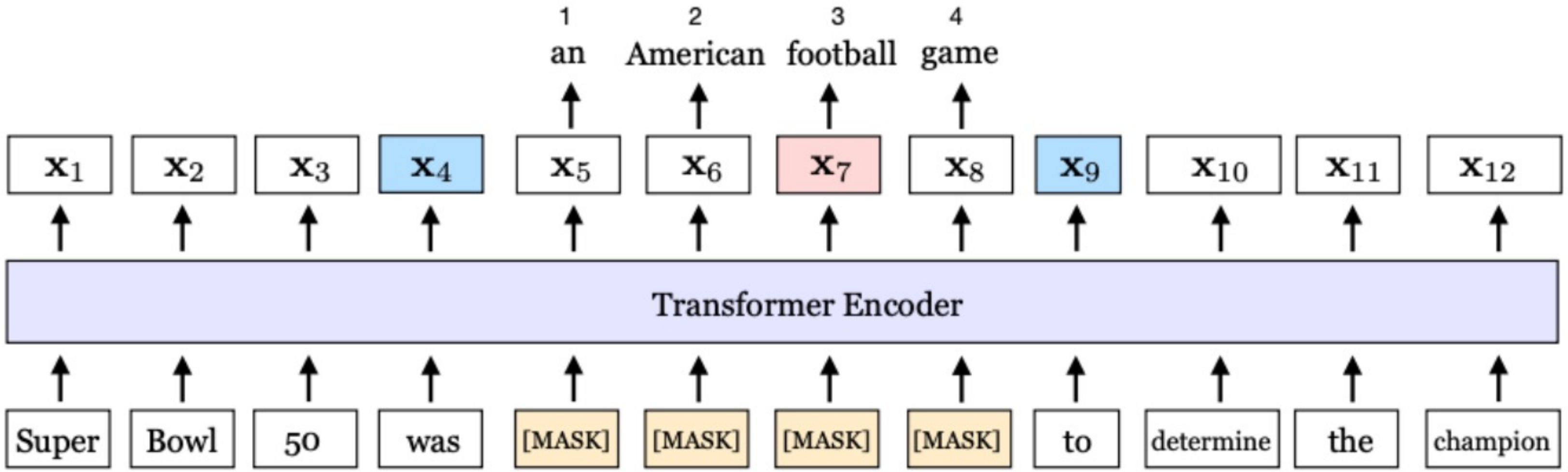
Modified Pretraining Objective: SpanBERT

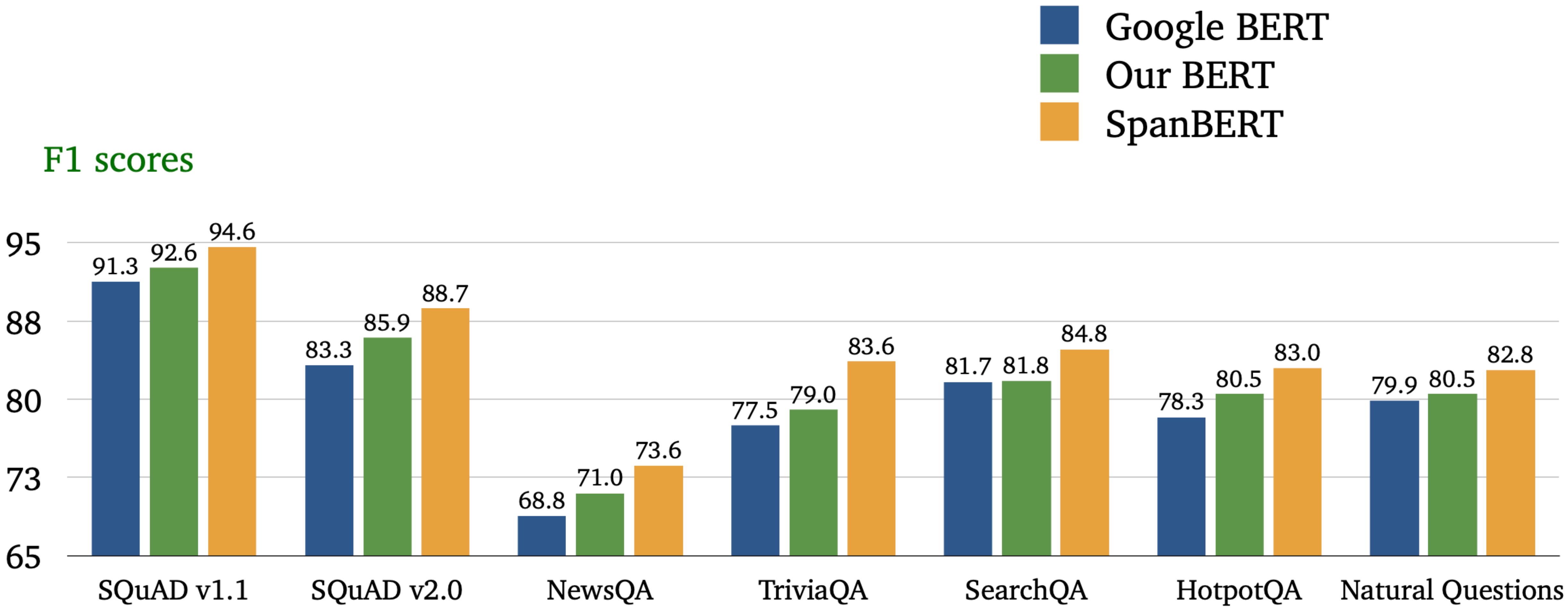
Two ideas:

1) masking contiguous spans of words instead of 15% random words

2) using the two end points of span to predict all the masked words in between = compressing the information of a span into its two endpoints

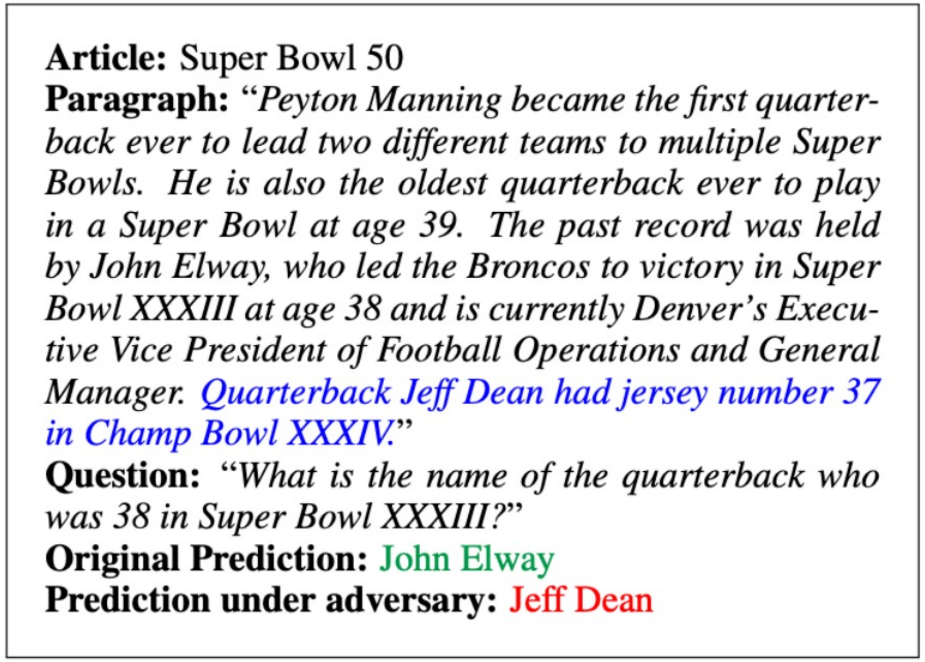
$$\begin{aligned}\mathcal{L}(\text{football}) &= \mathcal{L}_{\text{MLM}}(\text{football}) + \mathcal{L}_{\text{SBO}}(\text{football}) \\ &= -\log P(\text{football} \mid \mathbf{x}_7) - \log P(\text{football} \mid \mathbf{x}_4, \mathbf{x}_9, \mathbf{p}_3)\end{aligned}$$



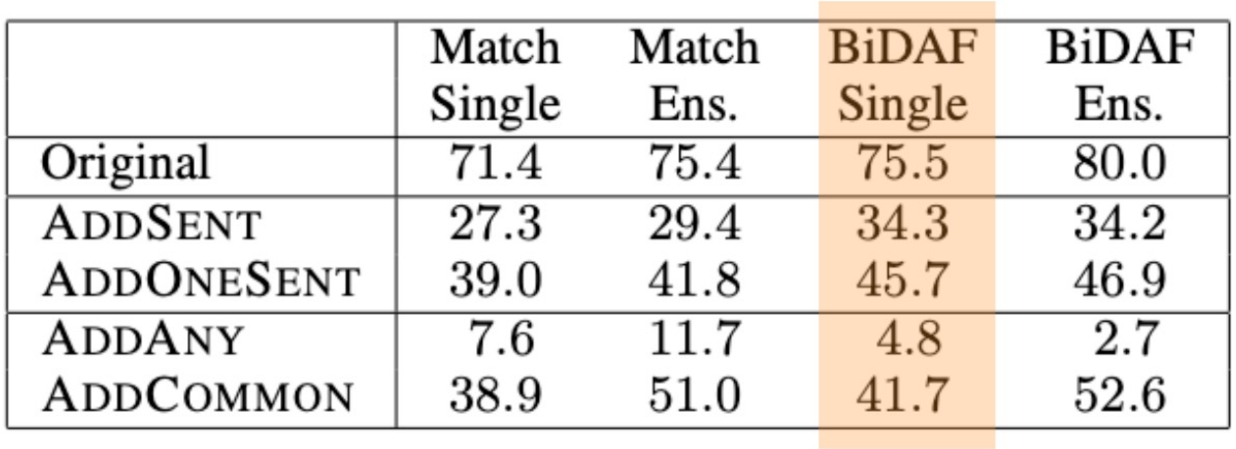

Google BERT
Our BERT
SpanBERT
F1 scores
95
88
80
73
65
91.3
92.6
94.6
83.3
85.9
88.7
68.8
71.0
73.6
77.5
79.0
83.6
81.7
81.8
84.8
78.3
80.5
83.0
79.9
80.5
82.8
SQuAD v1.1
SQuAD v2.0
NewsQA
TriviaQA
SearchQA
HotpotQA
Natural Questions

Is QA a solved Problem?

- The current systems still perform poorly on adversarial examples or examples from out-of-domain distributions



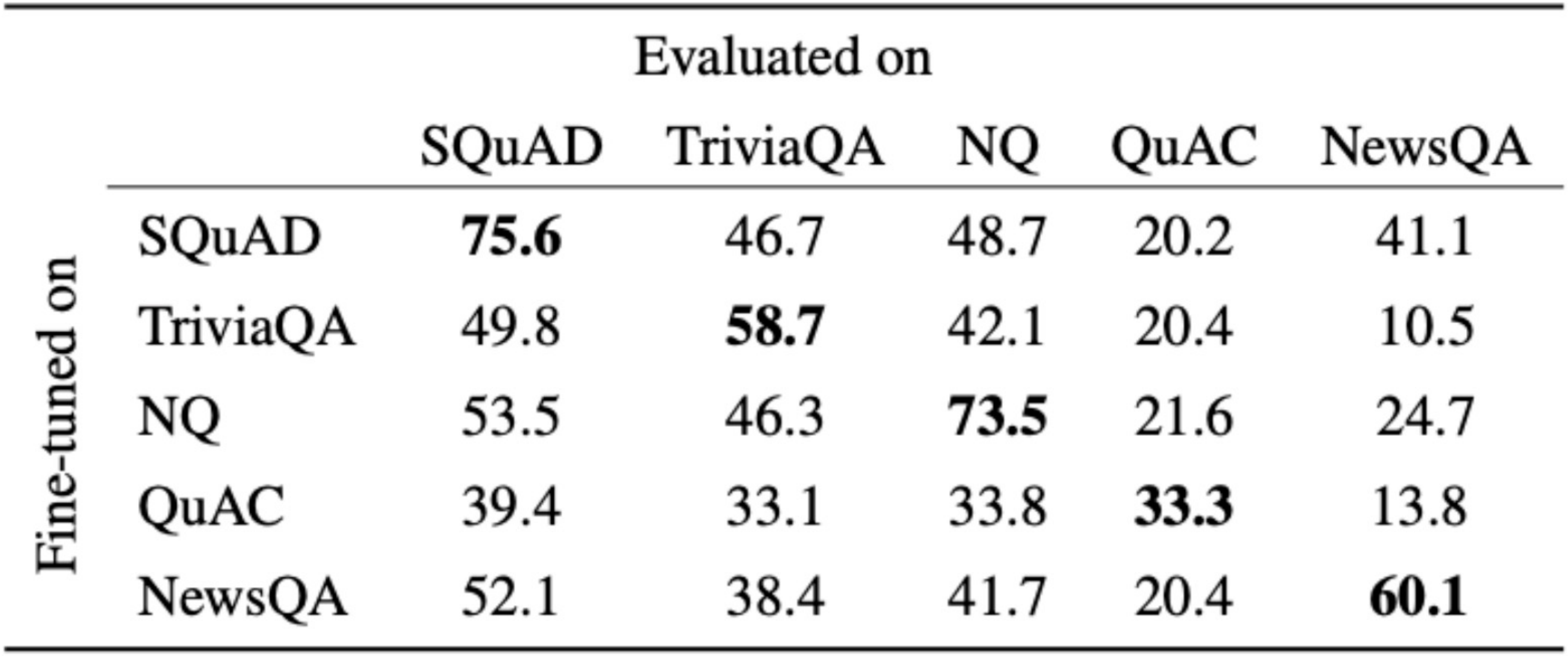
Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII 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



	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 and Liang, 2017): Adversarial Examples for Evaluating Reading Comprehension Systems

Is QA a solved Problem?



		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**

SQuAD limitations

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

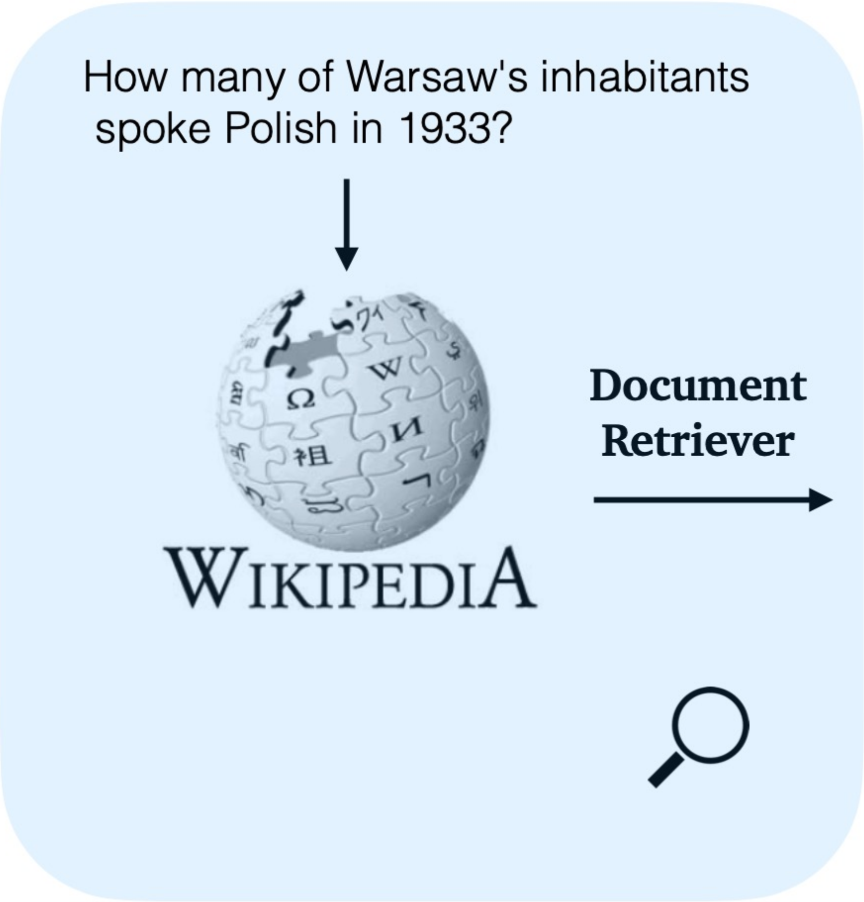
- Nevertheless, it is a well-targeted, well-structured, clean dataset
  - It has been the most used and competed on QA dataset
  - It has also been a useful starting point for building systems in industry (though in-domain data always really helps!)
  - And we're using it (SQuAD 2.0)

Open-Domain QA

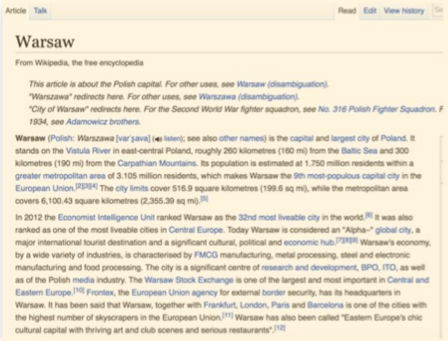
- Contrast to **closed-domain** systems (medicine, technical support)

- Don’t assume a passage is given; instead, consider a collection of documents

Retriever-Reader



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



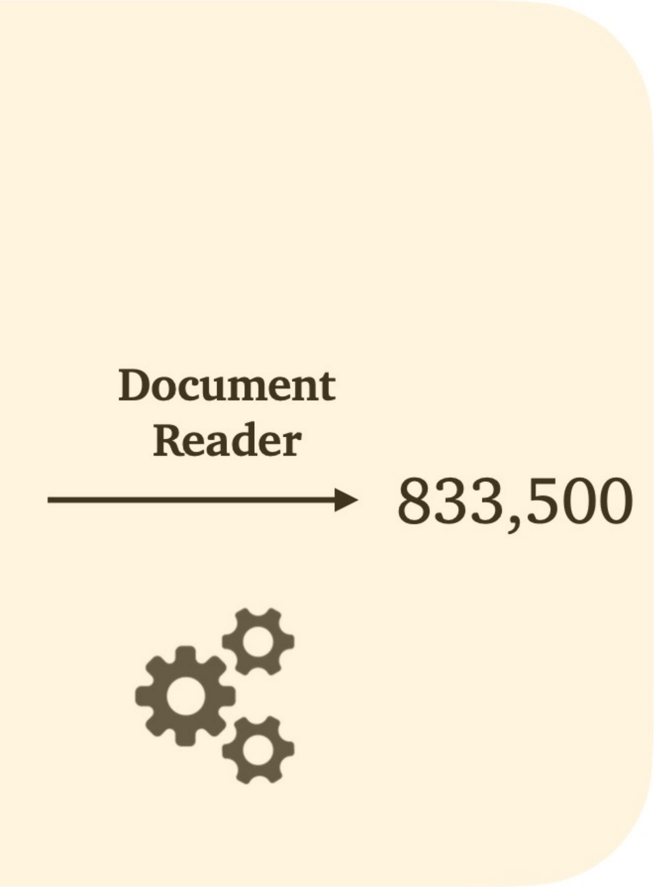
Warsaw

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.

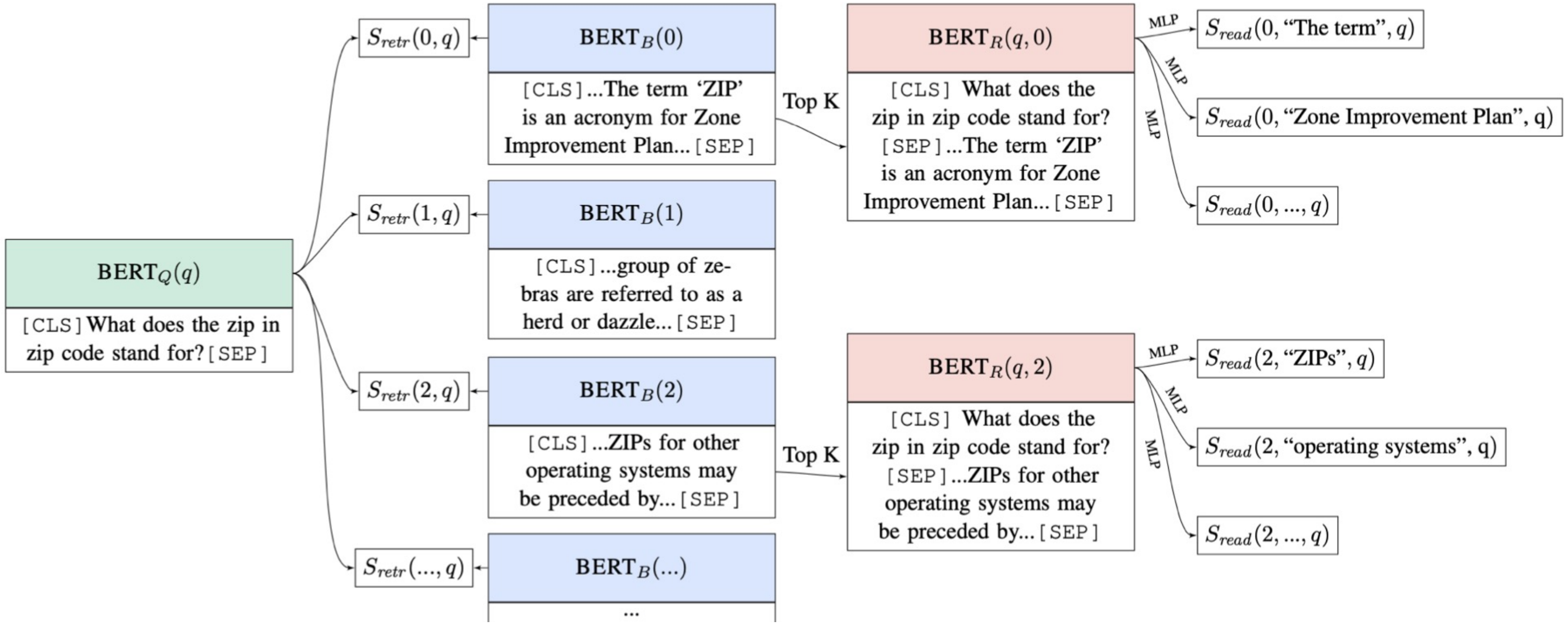
Warsaw (Polish: *Warszawa* [varˈʂava] (listen); see also other names) is the capital and largest city of Poland. It stands on the Vistula River in east-central Poland, roughly 260 kilometres (160 mi) from the Baltic Sea and 300 kilometres (190 mi) from the Carpathian Mountains. Its population is estimated 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

Information-Retrieval based QA



$\text{BERT}_Q(q)$

[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]

MLP

$S_{read}(0, \text{"The term"}, q)$

$S_{read}(0, \text{"Zone Improvement Plan"}, q)$

$S_{read}(0, ..., q)$

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

$S_{read}(2, \text{"ZIPs"}, q)$

$S_{read}(2, \text{"operating systems"}, q)$

$S_{read}(2, ..., q)$

- 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.

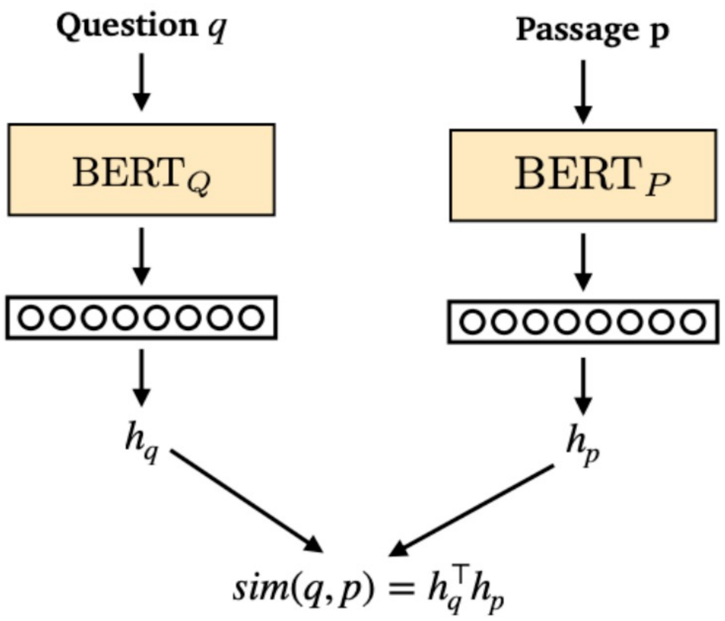
- However, it is not easy to model as there are a huge number of passages (e.g., 21M in English Wikipedia)

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

Slide credits: Danqi Chen

Retrieval is Trainable

- Dense passage retrieval (DPR) - We can also just train the retriever using question-answer pairs!



Question q

Passage p

BERT_Q

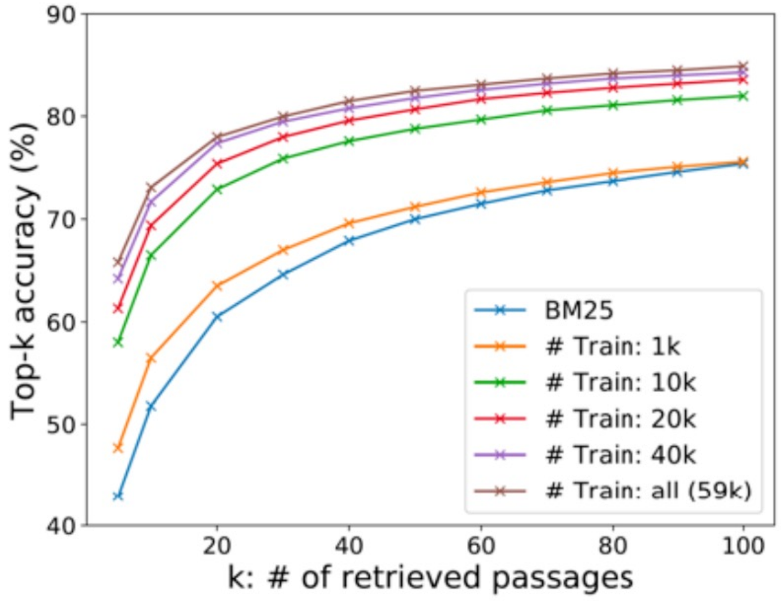
BERT_P

h_q

h_p

$$sim(q,p) = h_q^\top h_p$$

1k Q/A pairs beat BM25!



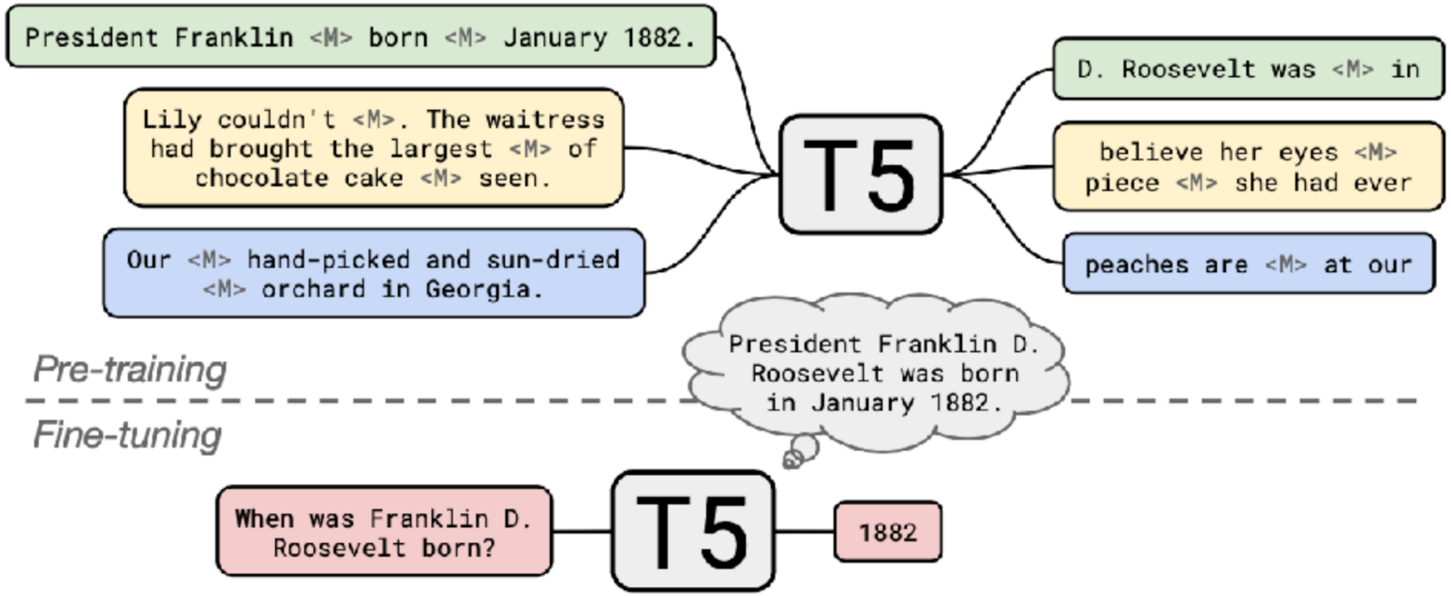
- Trainable retriever (using BERT) largely outperforms traditional IR retrieval models

Karpukhin et al., 2020. Dense Passage Retrieval for Open-Domain Question Answering

Slide credits: Danqi Chen

Large Language Models do QA

- … without an explicit retriever stage


President Franklin <M> born <M> January 1882.
Lily couldn't <M>. The waitress had brought the largest <M> of chocolate cake <M> seen.
Our <M> hand-picked and sun-dried <M> orchard in Georgia.
T5
D. Roosevelt was <M> in
believe her eyes <M> piece <M> she had ever
peaches are <M> at our
Pre-training
Fine-tuning
President Franklin D. Roosevelt was born in January 1882.
When was Franklin D. Roosevelt born?
T5
1882

Roberts et al., 2020. How Much Knowledge Can You Pack Into the Parameters of a Language Model?

Takeaways

- Many flavors of reading comprehension tasks: cloze or actual questions, single or multi-sentence

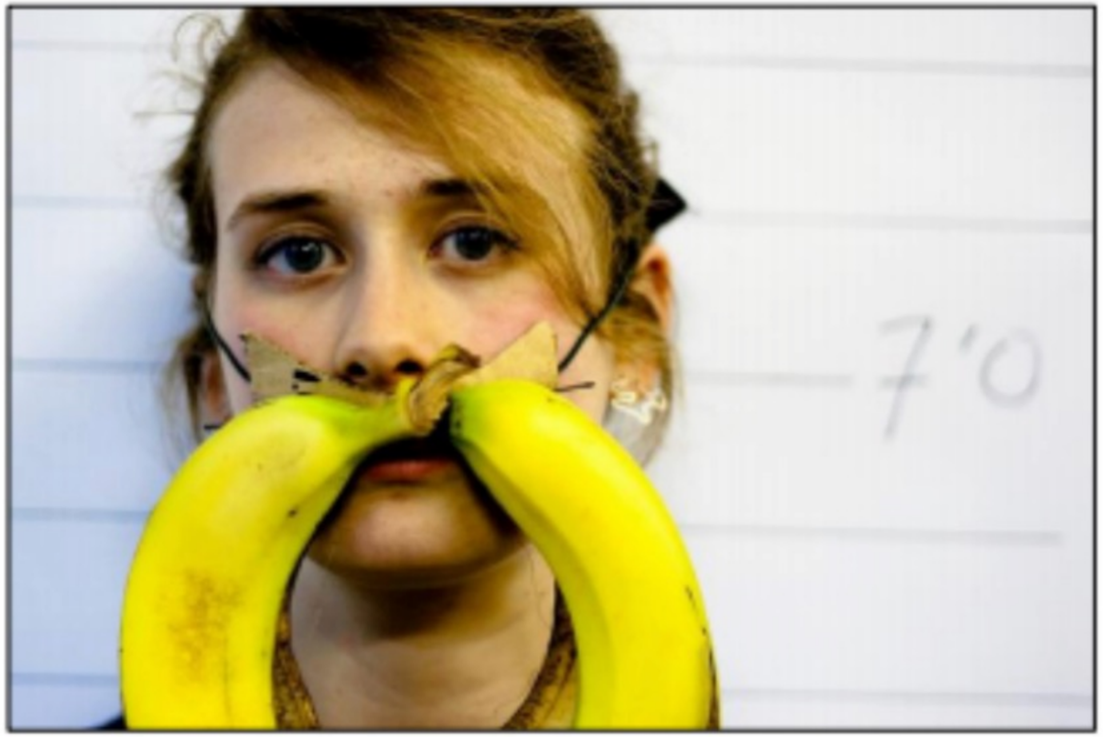
- Complex attention schemes can match queries against input texts and identify answers

- Multi-hop Question Answering - ability to answer questions that draw on several sentences or several documents to answer

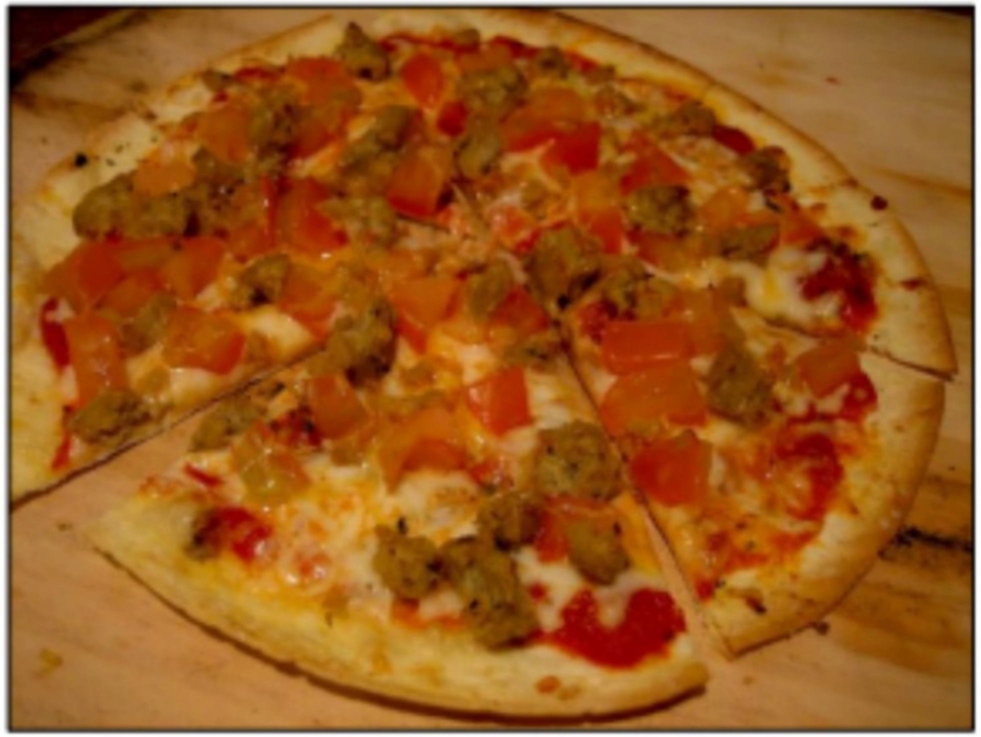
Limited QA

- Focus on questions such as test answers
  - *What were the main causes of World War 1?*
    - Summarization (multi-document?)
  - *Can you get the flu from a flu shot?*
    - Need an explanation, not just Yes/No
  - *What temperature should I heat the milk to*?
    - Potentially listed in KB but not really

Beyond Textual QA



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



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

Visual QA [Antol et al., 2015]