

Natural Language Processing



Question Answering

Dan Klein – UC Berkeley

The following slides are largely from Greg Durrett and Chris Manning, including many slides originally from Sanda Harabagiu, ISI, and Nicholas Kushmerick.



QA is Very Broad

- ▶ Factoid QA: *what states border Mississippi?, when was Barack Obama born?*
 - ▶ Lots of this could be handled by QA from a knowledge base, if we had a big enough knowledge base
- ▶ “Question answering” as a term is so broad as to be meaningless
 - ▶ *What is the meaning of life?*
 - ▶ *What is 4+5?*
 - ▶ *What is the translation of [sentence] into French?* [McCann et al., 2018]



QA Limits

- ▶ Focus on questions where the answer might appear in text — still hard!
 - ▶ *What were the main causes of World War II?* — requires summarization
 - ▶ *Can you get the flu from a flu shot?* — want IR to provide an explanation of the answer, not just yes/no
 - ▶ *How long should I soak dry pinto beans?* — could be written down in a KB but probably isn't
- ▶ Today: QA when it requires retrieving the answer from a passage



Related: Reading Comprehension

- ▶ “AI challenge problem”:
answer question given
context
- ▶ Recognizing Textual
Entailment (2006)
- ▶ MCTest (2013): 500
passages, 4 questions
per passage
- ▶ Two questions per
passage explicitly require
cross-sentence reasoning

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.

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

Richardson (2013)



Related: Reading Comprehension

- ▶ “AI challenge answer questions in context”
- ▶ Recognizing Textual Entailment (2010)
- ▶ MCTest (2013) 10 passages, 4 questions per passage
- ▶ Two question types per passage explicit cross-sentence

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 went home.

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

James went to the fast food restaurant after the grocery store.

ld go into town and get into. He went to 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 went to the grocery

Richardson (2013)



A Brief (Academic) History

- Question answering is not a new research area
- Question answering systems can be found in many areas of NLP:
 - Natural language / database systems
 - A lot of early NLP work on these
 - Conversational / assistant systems
 - Currently very active and commercially relevant
- The focus on open-domain QA is (relatively) new
 - TREC QA competition: 1999+
 - Modern large-scale factoid QA, eg SQuAD: 2016+
 - Search increasingly includes question answering
- General approach (across all eras): retrieval + entailment



Classic Question Answering

- ▶ Form semantic representation from semantic parsing, execute against structured knowledge base

Q: *where was Barack Obama born*

$\lambda x. \text{type}(x, \text{Location}) \wedge \text{born_in}(\text{Barack_Obama}, x)$

(also Prolog / GeoQuery, etc.)

- ▶ How to deal with open-domain data/relations? Need data to learn how to ground every predicate or need to be able to produce predicates in a zero-shot way



Famous QA Example: Watson (2011)

"a camel is a horse designed by"

About Wiktionary: a multilingual free encyclopedia

Wiktionary [ˈwɪkʃənəri] *n.*, a wiki-based Open Content dictionary

Log in / create account

Entry Discussion Read Edit History Search

a camel is a horse designed by a committee

Contents [hide]

- 1 English
- 1.1 Alternative forms
- 1.2 Proverb

The Phrase Finder

Discussion Forum

Google Custom Search Search

A camel is a horse designed by committee

Posted by Ruben P. Mendez on April 16, 2004

Does anyone know the origin of this maxim? I heard it way back at the United Nations, which is chockfull of committees. It may have originated there, but I'd like an authoritative explanation. Thanks

- [Re: A camel is a horse designed by committee](#) SR 16/April/04
 - [Re: A camel is a horse designed by committee](#) Henry 18/April/04



Jeopardy...

Category: General Science

Clue: When hit by electrons, a phosphor gives off electromagnetic energy in this form.

Answer: Light (or Photons)

Category: Lincoln Blogs

Clue: Secretary Chase just submitted this to me for the third time; guess what, pal. This time I'm accepting it.

Answer: his resignation

Category: Head North

Clue: They're the two states you could be reentering if you're crossing Florida's northern border.

Answer: Georgia and Alabama

Category: Decorating

Clue: Though it sounds "harsh," it's just embroidery, often in a floral pattern, done with yarn on cotton cloth.

Answer: crewel

Category: "Rap" Sheet

Clue: This archaic term for a mischievous or annoying child can also mean a rogue or scamp.

Subclue 1: This archaic term for a mischievous or annoying child.

Subclue 2: This term can also mean a rogue or scamp.

Answer: Rapscaillon

Category: Before and After Goes to the Movies

Clue: Film of a typical day in the life of the Beatles, which includes running from bloodthirsty zombie fans in a Romero classic.

Subclue 2: Film of a typical day in the life of the Beatles.

Answer 1: (*A Hard Day's Night*)

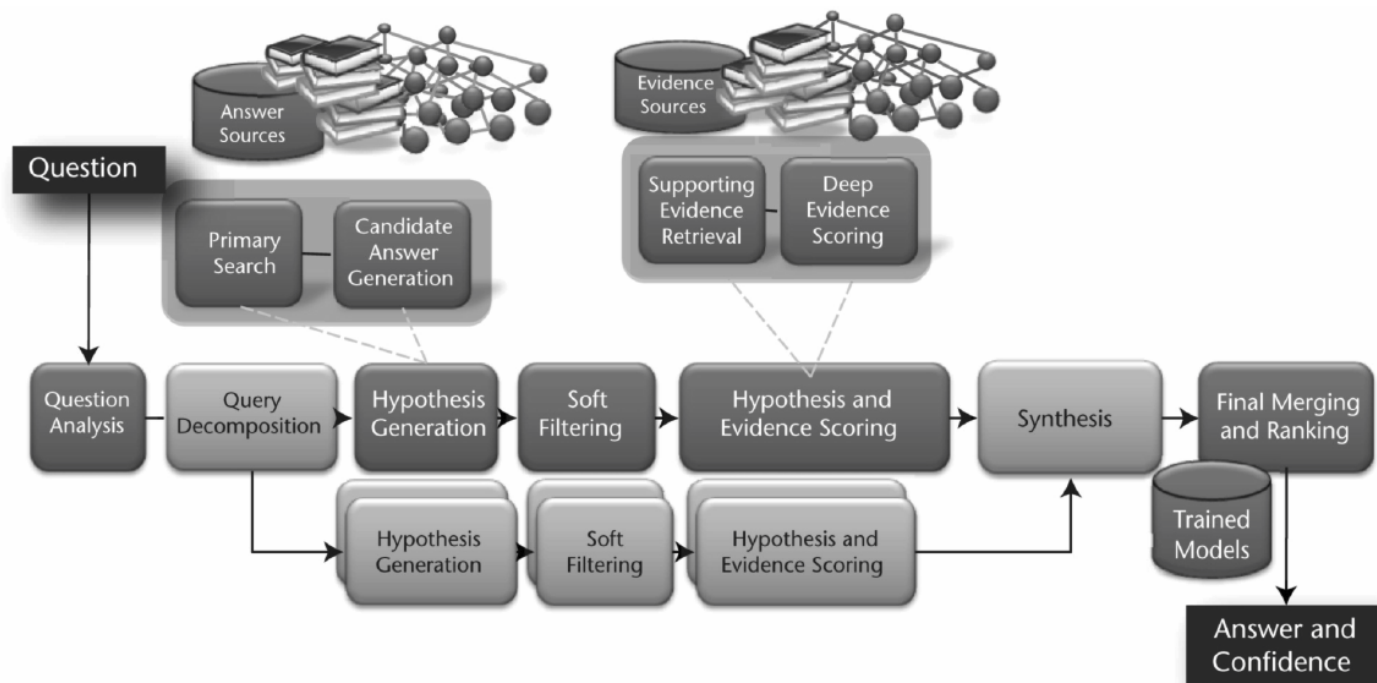
Subclue 2: Running from bloodthirsty zombie fans in a Romero classic.

Answer 2: (*Night of the Living Dead*)

Answer: *A Hard Day's Night of the Living Dead*



Architecture



Retrieval



Querying Documents with Keywords

- Goal: take a query and find relevant documents
- Example: songs in a database
- Constraints: want documents that contain query words, but not exact match on query...
- Idea: rank documents by how much of the query they contain?

Query:

we wait on trains

New Romantics

Fift

We're all bored

You

We're all so tired of everything

He

And

***We wait for trains** that just aren't coming*

Drive

It's

We show off our different scarlet letters

I th

You

Trust me, mine is better

No

aw

But

⋮



Term Frequency

- Idea: Score of document by summing over query words:

$$score(q, d) = \sum_{w \in q} score(w, d) \quad score(w, d) = (w \in d)$$

- Problem: Many documents could contain all query words (e.g., once)
- Solution: Term frequency: $score(q_i, d) = tf(w, d) := C(w, d)$

	<i>Out of the Woods</i>	<i>Speak Now</i>	<i>Fifteen</i>	<i>New Romantics</i>	<i>Tim McGraw</i>	<i>CMUDICT</i>
"we"	73		2	12	3	1
"wait"		4		1		1
"on"	1	3	2	4	10	1
"trains"				1		1



Saturation and Normalization

- Problem: One query term being repeated many times can outweigh other terms being entirely absent
- Solution: Give counts of any one term diminishing returns
- Example: (Log) Normalization: $tf(w, d) := \log(C(w, d) + 1)$

	<i>Out of the Woods</i>	<i>Speak Now</i>	<i>Fifteen</i>	<i>New Romantics</i>	<i>Tim McGraw</i>
"we"	4.30		1.01	2.56	1.39
"wait"		1.61		0.69	
"on"	0.69	1.39	1.01	1.61	2.40
"trains"				0.69	



Weighting Terms with TF-IDF

- Problem: Common words will have high counts but carry little information
- Solution: Downweight words that occur in many documents
- Inverse document frequency (IDF):
 - Basic document frequency of w is fraction of documents with the w : $\frac{D(w)}{N}$
- TF-IDF: Classic IR baseline
 - Normalized term frequency: $\text{tf}(w, d) := \log(C(w, d) + 1)$
 - Normalized inverse document frequency: $\text{idf}(w) := \log\left(\frac{N}{D(w)}\right)$
 - $\text{TF-IDF}(w, d) = \text{tf}(w, d) \times \text{idf}(w)$



Weighting Terms with TF-IDF

- Term frequency: $\text{tf}(w, d) := \log(C(w, d) + 1)$
- Inverse document frequency: $\text{idf}(w) := \log\left(\frac{N}{D(w)}\right)$
- $\text{TF-IDF}(w, d) = \text{tf}(w, d) \times \text{idf}(w)$

	<i>Out of the Woods</i>	<i>Speak Now</i>	<i>Fifteen</i>	<i>New Romantics</i>	<i>Tim McGraw</i>
"we"	4.30		1.01	2.56	1.39
"wait"		1.61		0.69	
"on"	0.69	1.39	1.01	1.61	2.40
"trains"				0.69	



Weighting Terms with TF-IDF

- Term frequency: $\text{tf}(w, d) := \log(C(w, d) + 1)$
- Inverse document frequency: $\text{idf}(w) := \log\left(\frac{N}{D(w)}\right)$
- $\text{TF-IDF}(w, d) = \text{tf}(w, d) \times \text{idf}(w)$

	<i>Out of the Woods</i>	<i>Speak Now</i>	<i>Fifteen</i>	<i>New Romantics</i>	<i>Tim McGraw</i>	IDF
"we"	2.47		0.63	1.47	0.79	0.57
"wait"		3.78		1.63		2.35
"on"	0.22	0.45	0.36	0.52	0.78	0.32
"trains"				3.15		4.54
Sum of terms	2.69	4.22	0.99	6.77	1.57	



BM25: Addressing Issues with TF-IDF

Some remaining issues with TF-IDF:

- Handling of term saturation
- Varying document length

Still a strong zero-shot retrieval baseline as of December 2021

Given a query $w = \{w_1, w_2, \dots, w_n\}$:

$$\text{BM25}(w, d) := \sum_{i=1}^n \text{idf}(w_i) \cdot \frac{\text{tf}(w_i, d) \cdot \overset{\text{hyperparameter}}{\underset{|}{k+1}}}{\text{tf}(w_i, d) + \underset{\text{hyperparameter}}{\underset{|}{(1-b)}} + \underset{\text{average document length}}{\underset{|}{b \cdot |d|/L}}}$$



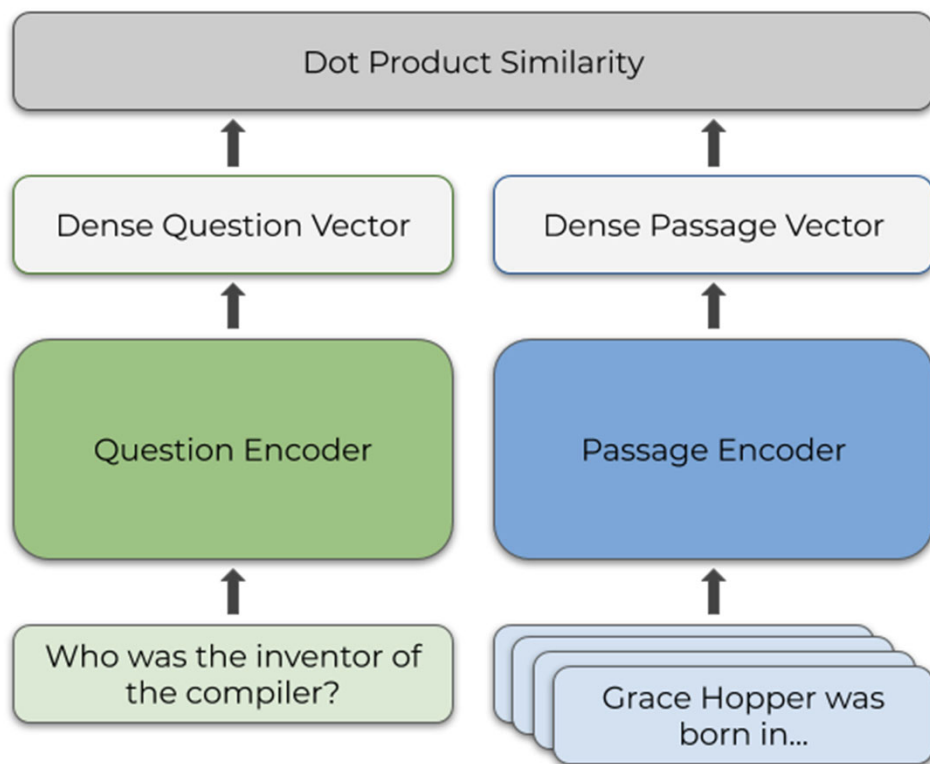
Beyond Term Scoring

- Term scoring suited classic inverted index construction very well!
- Modern IR systems include other term factors
 - Contiguous match (e.g. n-grams)
 - Positional information (e.g. titles)
 - Related word match (e.g. synonyms)
- ... And some of the most important features aren't term-derived at all
 - Link analysis (e.g. PageRank)
 - User behavior (e.g. clickstream analysis)

Neural Retrieval



Dense Passage Retrieval



Contrastive loss function:

$$-\log \frac{\exp(\text{sim}(w_i, d_i^+))}{\exp(\text{sim}(w_i, d_i^+)) + \sum_{j=1}^n \exp(\text{sim}(w_i, d_{i,j}^-))}$$

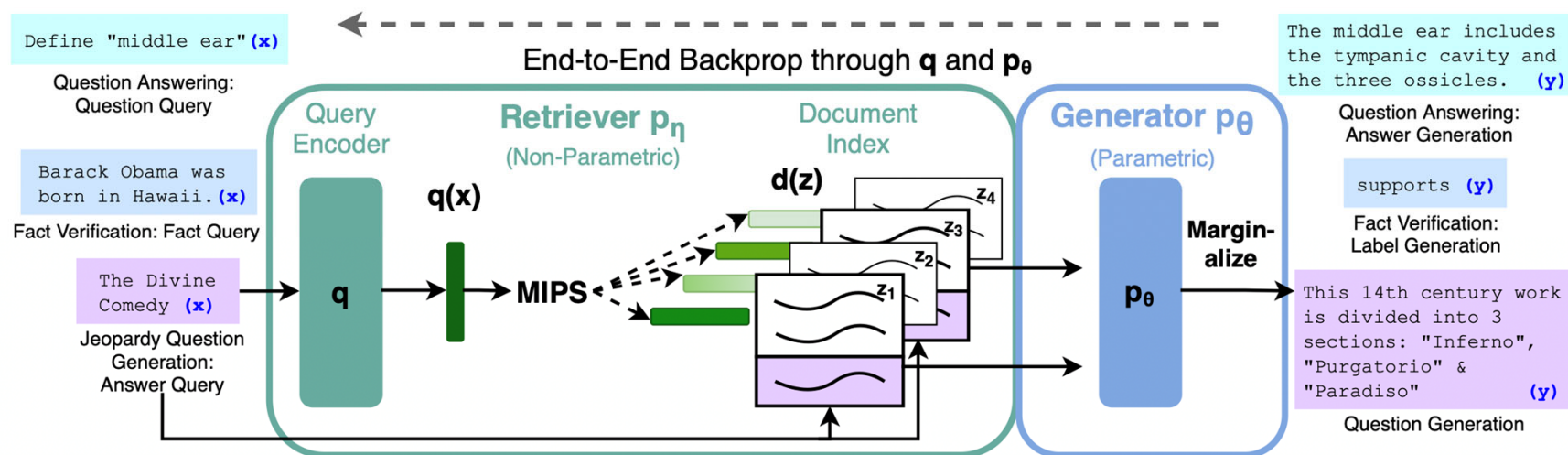
where $d_{i,j}^-$ are negatives and sim is vector similarity.

Obtain “hard negatives” using incorrect answers from a BM25 baseline model



Retrieval-Augmented Models

- Generation models (including QA) have a bottleneck where parameters must capture all information from the training data.
- Retrieval-augmented models let a system look directly at source data.
- Similar to how attention let encoder-decoders look directly at the input
- Also allows a system to dynamically respond to new data after training

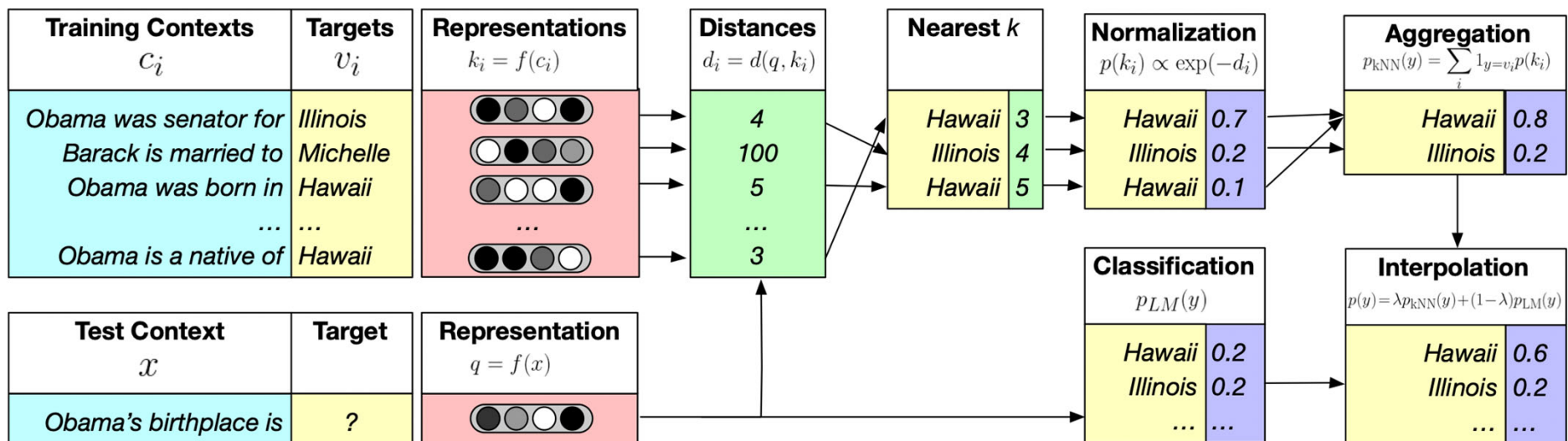


[Lewis et al, 2020]



Nearest Neighbor Language Modeling

- Nonparametrics: condition generation on retrieved documents
- Slightly improves perplexity at the cost of inference speed and storage, but can update knowledge without retraining



[Image from Khandelwal, et al. (2019): Nearest Neighbor Language Models]

Question Answering



AskMSR

- **Web Question Answering: Is More Always Better?**
 - [Dumais, Banko, Brill, Lin, Ng 2002]
- **Q: “Where is the Louvre located?”**
 - Want “Paris” or “France”, etc
 - These answers are often all over the documents returned by a web search
 - Idea: the answer is probably a frequent n-gram in the search results

The screenshot shows a Google search interface with the query "Where is the Louvre museum located?". The search results include several entries:

- PDF: An Analysis of the AskMSR Question-Answering System**
File Format: PDF/Adobe Acrobat - [View as HTML](#)
... Page 2. Question Rewrite Query <Search Engine> Collect Summaries, Mine N-grams Filter N-Grams Tile N-Grams N-Best Answers Where is the **Louvre Museum located?** ...
research.microsoft.com/~sdumais/EMNLP_Final.pdf - [Similar pages](#)
- hotel montpensier - located near louvre museum, opera house, ...**
Located in the heart of Paris, Hotel Montpensier offers 43 rooms, incl. ... The hotel is at walking distance from the **Louvre museum**, the Opera House, Champs ...
www.away-to-paris.com/Hotels/MONTPENSIER/MainNS.htm - 2k - [Cached](#) - [Similar pages](#)
- hotel montpensier - located near louvre museum, opera house, ...**
Located in the heart of Paris, Hotel Montpensier offers 43 rooms, incl. 35 with bath or shower, direct-line telephone, TV set and hair dryer. The hotel is ...
www.away-to-paris.com/Hotels/MONTPENSIER/TheHotel2.htm - 2k - [Cached](#) - [Similar pages](#)
[[More results from www.away-to-paris.com](#)]
- PDF: AskMSR: Question Answering Using the Worldwide Web**
File Format: PDF/Adobe Acrobat - [View as HTML](#)
... 49.2 40 Question Rewrite Query <Search Engine> Collect Summaries, Mine N-grams Filter N-Grams Tile N-Grams N-Best Answers Where is the **Louvre Museum located?** ...
www.ai.mit.edu/people/jimmylin/publications/Banko-etal-AAAI02.pdf - [Similar pages](#)
- Louvre Museum Official Website: Publications**
... Médiathèque" Located on the first floor of the area "Accueil des groupes", the "Médiathèque" is accessible for ... The Bookshop at the **Louvre Museum** ...
www.louvre.fr/anglais/publications/lieux.htm - 21k - 29 Sep 2002 - [Cached](#) - [Similar pages](#)

At the bottom, there is a link for [Louvre Museum Official Website](#).



AskMSR: Shallow approach

- *In what year did Abraham Lincoln die?*
- Ignore hard documents and find easy ones

Abraham Lincoln, 1809-1865

***LINCOLN, ABRAHAM** was born near Hodgenville, Kentucky, on February 12, 1809. In 1816, the Lincoln family moved to Pigeon Creek in Perry (now Spencer) County. Two years later, Abraham Lincoln's mother died and his father married a woman his "angel" mother. Lincoln attended a formal school for only a few months but acquired knowledge through the reading of books in Illinois, in 1830 where he obtained a job as a store clerk and the local postmaster. He served without distinction in the Black Hawk War, lost his attempt at the state legislature, but two years later he tried again, was successful, and Lincoln was admitted to the bar and became noteworthy as a witty, honest, competent circuit lawyer. He served a one-year term in the U.S. House in 1846, at which time he opposed the war with Mexico. By 1858, Lincoln had gained national attention for his series of debates with Stephen A. Douglas. After losing the election he became a significant figure in his party. On the day of his inauguration on March 4, seven southern states had seceded from the Union. Lincoln called for 75,000 volunteers (approximately 40,000 were accepted) to form the United States Army. Lincoln called for 75,000 volunteers (approximately 40,000 were accepted) to form the United States Army. Lincoln called for 75,000 volunteers (approximately 40,000 were accepted) to form the United States Army. Lincoln called for 75,000 volunteers (approximately 40,000 were accepted) to form the United States Army.

Sixteenth President
1861-1865
Married to Mary Todd Lincoln

ABRAHAM LINCOLN



Sixteenth President of the United States

Born in 1809 - Died in 1865

Abraham Lincoln

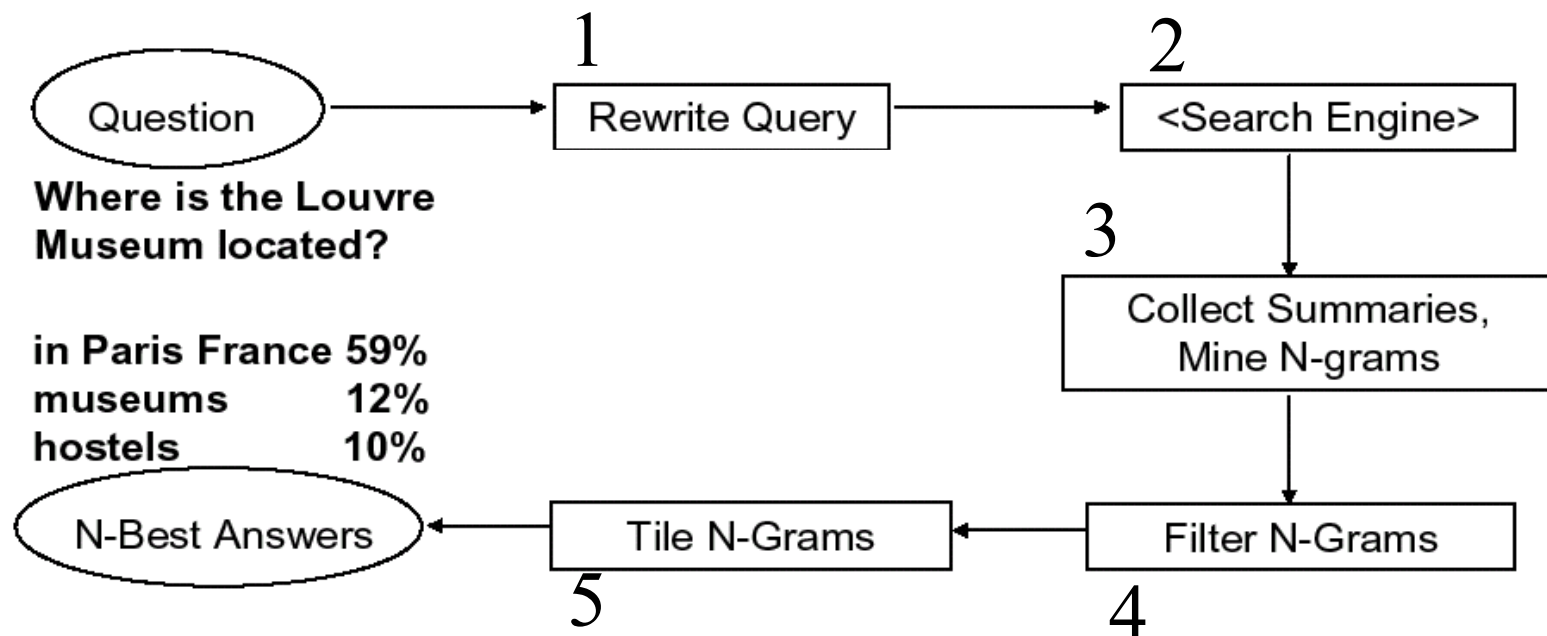
16th President of the United States (March 4, 1861 to April 15, 1865)
Born: February 12, 1809, in Hardin County, Kentucky
Died: April 15, 1865, at Petersen's Boarding House in Washington, D.C.

"I was born February 12, 1809, in Hardin County, Kentucky. My parents were both born in Virginia, of undistinguished families, perhaps I should say. My mother, who died in my tenth year, was of a family of the name of Hanks."





AskMSR: Details





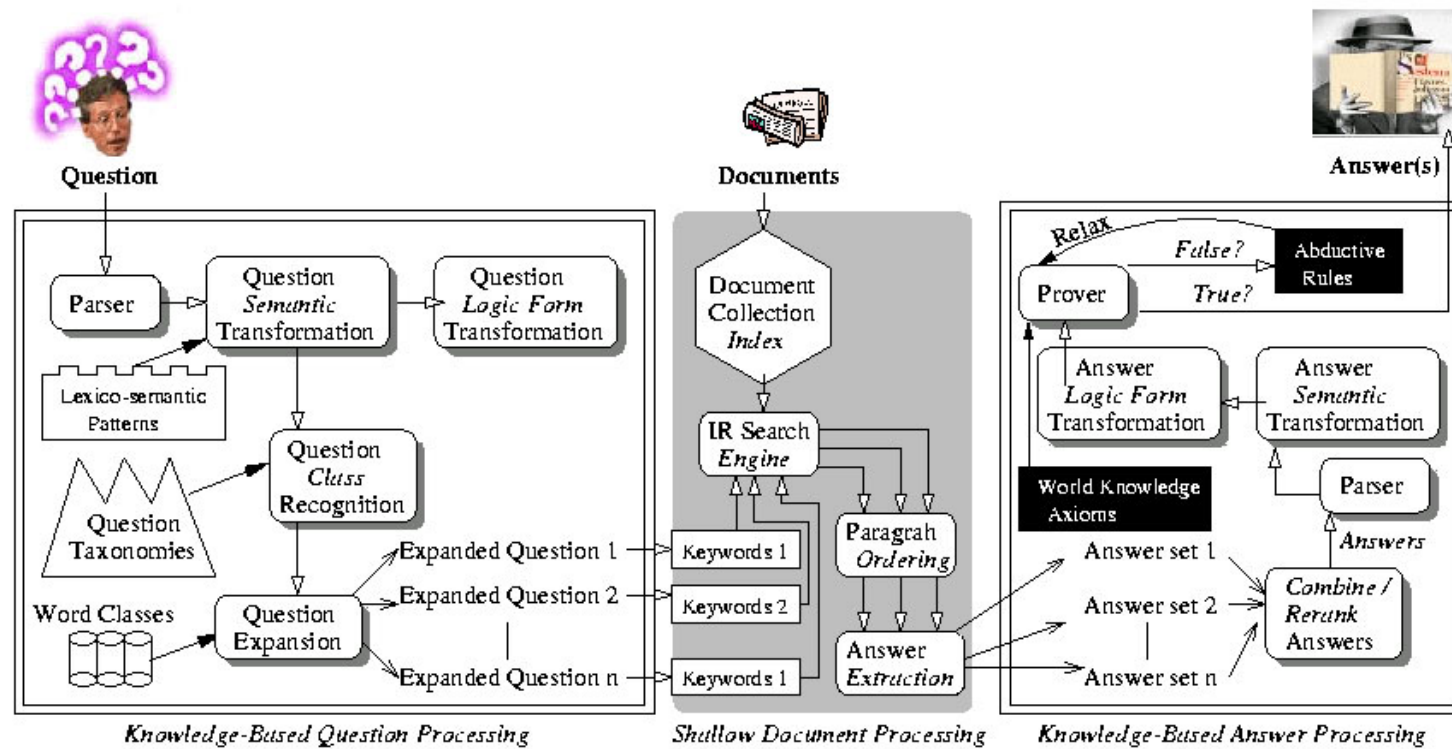
Results

- Standard TREC contest test-bed:
~1M documents; 900 questions
- Technique doesn't do too well (though would have placed in top 9 of ~30 participants!)
 - MRR = 0.262 (ie, right answer ranked about #4-#5 on average)
 - Why? Because it relies on the redundancy of the Web
- Using the Web as a whole, not just TREC's 1M documents... MRR = 0.42 (ie, on average, right answer is ranked about #2-#3)

Abduction: LCC



LCC: Harabagiu, Moldovan et al.





Abductive inference

- System attempts inference to justify an answer (often following lexical chains)
 - This inference is a kind of middle ground between logic and pattern matching
 - ... but it can be effective: 30% improvement at the time
- Example:
 - *Q: When was the internal combustion engine invented?*
 - *A: The first internal-combustion engine was built in 1867.*
 - invent -> create_mentally -> create -> build



Question Answering Example

- How hot does the inside of an active volcano get?
- “lava fragments belched out of the mountain were as hot as 300 degrees Fahrenheit”
 - volcano ISA mountain
 - lava ISPARTOF volcano ■ lava IN volcano
 - fragments of lava HAVEPROPERTIESOF lava
- The needed semantic information is in WordNet definitions, and was successfully translated into a form that was used for rough ‘proofs’

Span-Based QA



SQuAD

- ▶ Single-document question-answering task where the answer is always a substring of the passage (= a paragraph from Wikipedia)
- ▶ Predict start and end indices of the answer in the passage

One of the most famous people born in Warsaw was Maria Skłodowska-Curie, who achieved international recognition for her research on radioactivity and was the first female recipient of the Nobel Prize. Famous musicians include Władysław Szpilman and Frédéric Chopin. Though Chopin was born in the village of Żelazowa Wola, about 60 km (37 mi) from Warsaw, he moved to the city with his family when he was seven months old. Casimir Pulaski, a Polish general and hero of the American Revolutionary War, was born here in 1745.

What was Maria Curie the first female recipient of?

Ground Truth Answers: Nobel Prize Nobel Prize Nobel Prize

What year was Casimir Pulaski born in Warsaw?

Ground Truth Answers: 1745 1745 1745

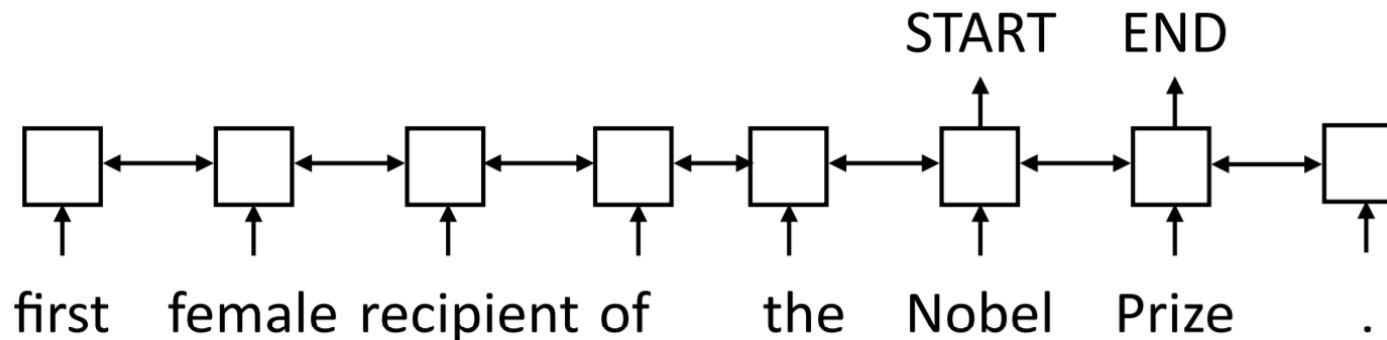
Who was one of the most famous people born in Warsaw?

Ground Truth Answers: Maria Skłodowska-Curie Maria Skłodowska-Curie Maria Skłodowska-Curie



Just Seq2Seq?

What was Marie Curie the first female recipient of?



- ▶ Like a tagging problem over the sentence (not multiclass classification), but we need some way of attending to the query



A Simple Neural Architecture

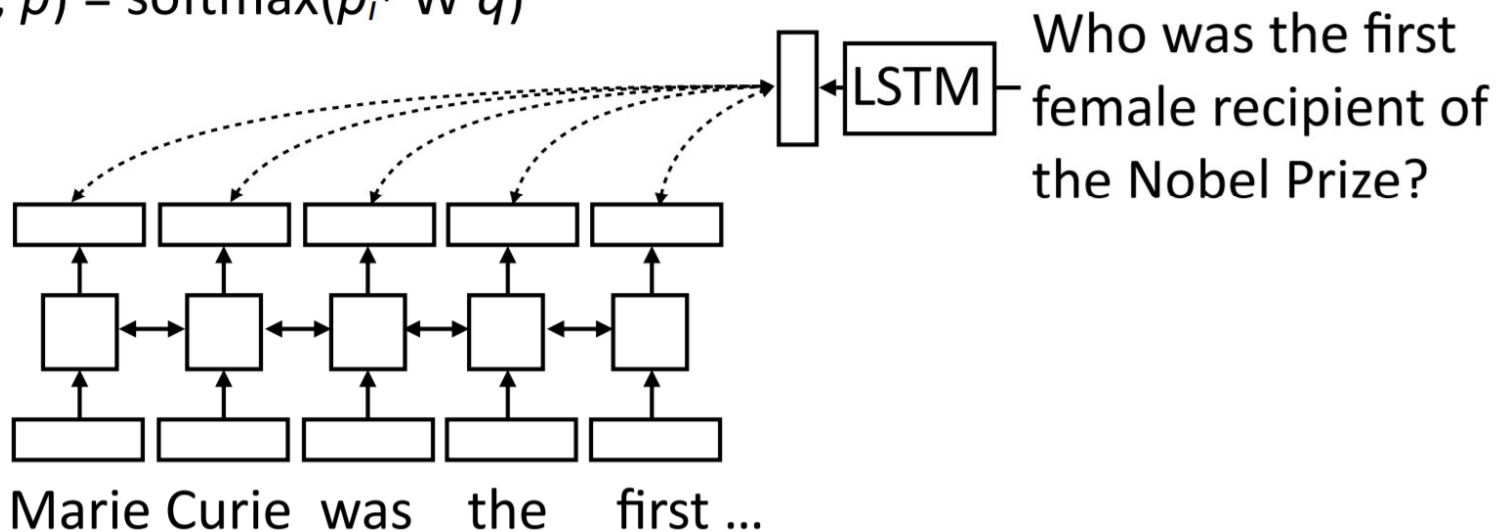
- ▶ Predict a distributions over start and end points of the answer

$P(\text{end} \mid q, p)$ computed similarly

$$P(\text{start} = i \mid q, p) = \text{softmax}(p_i^T W q)$$

encoding
of passage

BiLSTM
encoder





Training

- ▶ Train on labeled data with start and end points, maximize likelihood of correct decisions: $\log \sum_{i \in \text{gold starts}} p(\text{start} = i | p, q) + \log \sum_{i \in \text{gold ends}} p(\text{end} = i | p, q)$

In September 1958, Bank of America launched a new product called **BankAmericard** in Fresno. After a troubled gestation during which its creator resigned, **BankAmericard** went on to become the first **successful credit card**; that is, a financial instrument that was usable across a large number of merchants and also allowed **cardholders** to revolve a balance (earlier financial products could do one or the other but not both). In 1976, **BankAmericard** was renamed and spun off into a separate company known today as Visa Inc.

What was the name of the first successful credit card?

- ▶ Inference: maximize $P(\text{start}) + P(\text{end})$ with the constraint that (start, end) isn't too big a span



Some Outputs

Question: who caught a 16-yard pass on this drive ?

Answer: devin funchess

START

there would be no more scoring in the third quarter , but early in the fourth , the broncos drove to the panthers 41-yard line . on the next play , ealy knocked the ball out of manning 's hand as he was winding up for a pass , and then recovered it for carolina on the 50-yard line . a 16-yard reception by devin funchess and a 12-yard run by stewart then set up gano 's 39-yard field goal , cutting the panthers deficit to one score at 16â€"10 . the next three drives of the game would end in punts .

END

there would be no more scoring in the third quarter , but early in the fourth , the broncos drove to the panthers 41-yard line . on the next play , ealy knocked the ball out of manning 's hand as he was winding up for a pass , and then recovered it for carolina on the 50-yard line . a 16-yard reception by devin funchess and a 12-yard run by stewart then set up gano 's 39-yard field goal , cutting the panthers deficit to one score at 16â€"10 . the next three drives of the game would end in punts .



Some Outputs

Question: how many victorians are non - religious ?

Answer: 20 %

START

about 61.1 % of victorians describe themselves as christian . roman catholics form the single largest religious group in the state with 26.7 % of the victorian population , followed by anglicans and members of the uniting church . buddhism is the state 's largest non - christian religion , with 168,637 members as of the most recent census . victoria is also home of 152,775 muslims and 45,150 jews . hinduism is the fastest growing religion . around 20 % of victorians claim no religion . amongst those who declare a religious affiliation , church attendance is low .

END

about 61.1 % of victorians describe themselves as christian . roman catholics form the single largest religious group in the state with 26.7 % of the victorian population , followed by anglicans and members of the uniting church . buddhism is the state 's largest non - christian religion , with 168,637 members as of the most recent census . victoria is also home of 152,775 muslims and 45,150 jews . hinduism is the fastest growing religion . around 20 % of victorians claim no religion . amongst those who declare a religious affiliation , church attendance is low .

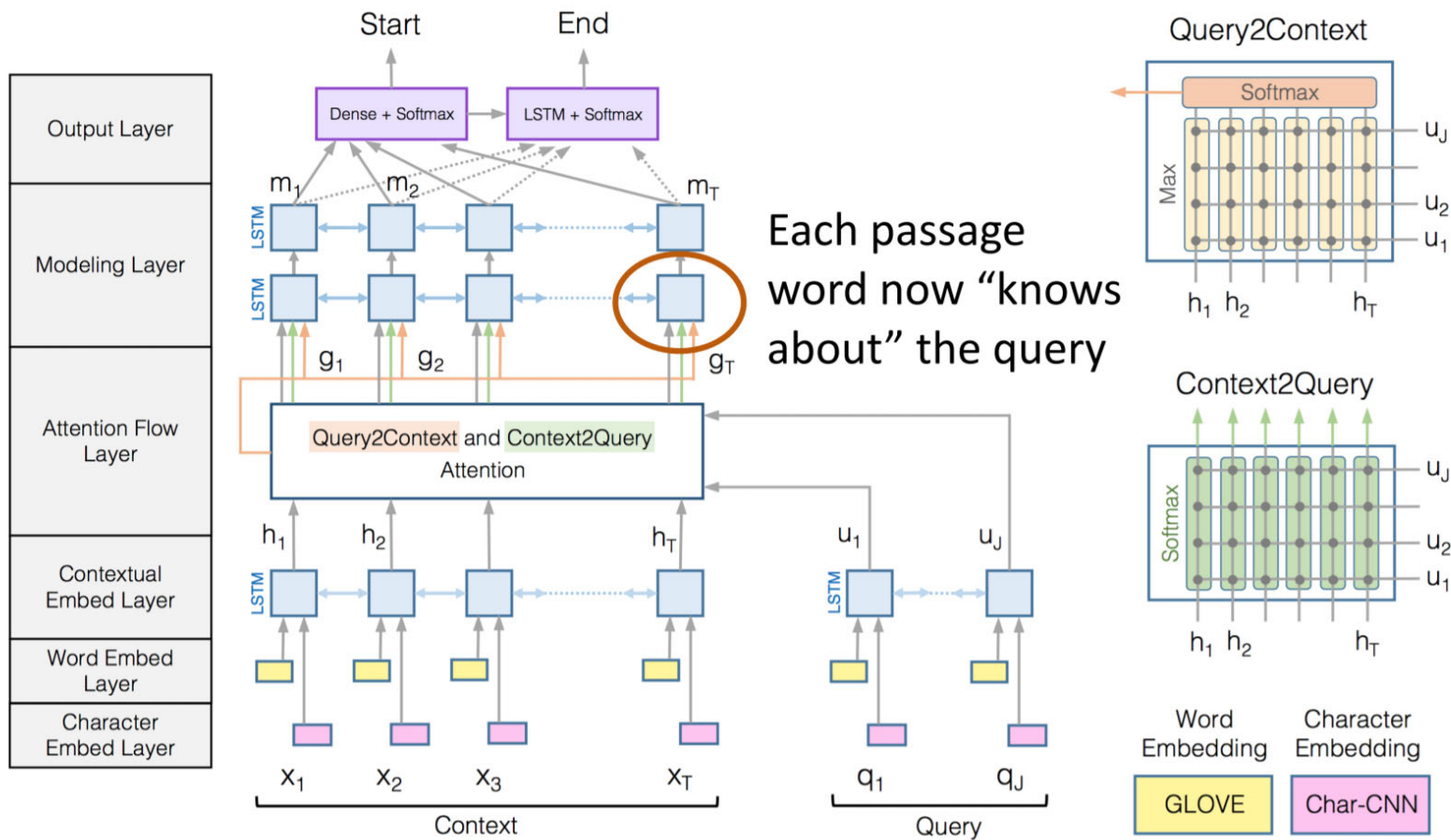


Why did SQuAD Take Off?

- ▶ SQuAD was **big**: >100,000 questions at a time when deep learning was exploding
- ▶ SQuAD was **pretty easy**: year-over-year progress for a few years until the dataset was essentially solved
- ▶ SQuAD had **room to improve**: ~50% performance from a logistic regression baseline (classifier with 180M features over constituents)



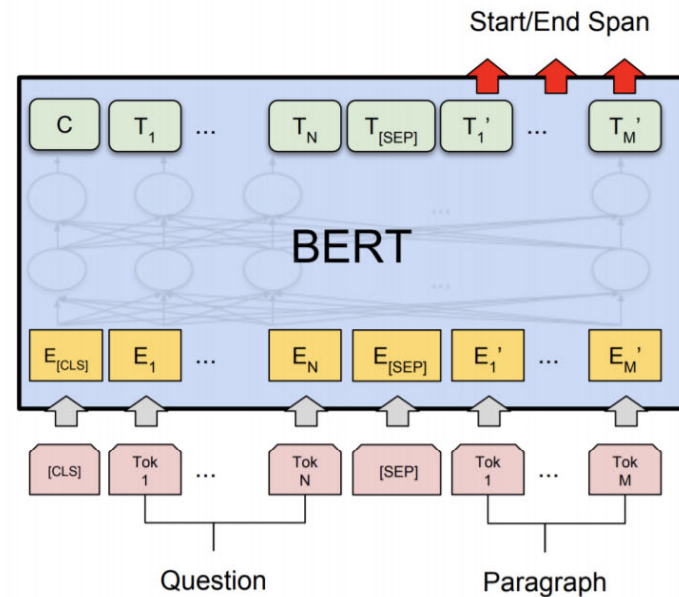
Example: Richer Model Structures



Seo et al. (2016)



Pre-Training



What was Marie Curie the first female recipient of ? $[SEP]$ Marie Curie was the first female recipient of ...

- ▶ Predict start and end positions in passage
- ▶ No need for cross-attention mechanisms!

Devlin et al. (2019)



Leaderboards

2018

Rank	Model	EM	F1
	Human Performance <i>Stanford University</i> (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) <i>Google AI Language</i> https://arxiv.org/abs/1810.04805	87.433	93.160
2 Oct 05, 2018	BERT (single model) <i>Google AI Language</i> https://arxiv.org/abs/1810.04805	85.083	91.835
2 Sep 09, 2018	nlnet (ensemble) <i>Microsoft Research Asia</i>	85.356	91.202
2 Sep 26, 2018	nlnet (ensemble) <i>Microsoft Research Asia</i>	85.954	91.677
3 Jul 11, 2018	QANet (ensemble) <i>Google Brain & CMU</i>	84.454	90.490
4 Jul 08, 2018	r-net (ensemble) <i>Microsoft Research Asia</i>	84.003	90.147
5 Mar 19, 2018	QANet (ensemble) <i>Google Brain & CMU</i>	83.877	89.737

2021

Rank	Model	EM	F1
	Human Performance <i>Stanford University</i> (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Feb 21, 2021	FPNet (ensemble) <i>Ant Service Intelligence Team</i>	90.871	93.183
2 Feb 24, 2021	IE-Net (ensemble) <i>RICOH_SRCB_DML</i>	90.758	93.044
3 Apr 06, 2020	SA-Net on Albert (ensemble) <i>QIANXIN</i>	90.724	93.011
4 May 05, 2020	SA-Net-V2 (ensemble) <i>QIANXIN</i>	90.679	92.948
4 Apr 05, 2020	Retro-Reader (ensemble) <i>Shanghai Jiao Tong University</i> http://arxiv.org/abs/2001.09694	90.578	92.978
4 Feb 05, 2021	FPNet (ensemble) <i>YuYang</i>	90.600	92.899
5 Apr 18, 2021	TransNets + SFVerifier + SFEnsembler (ensemble) <i>Senseforth AI Research</i> https://www.senseforth.ai/	90.487	92.894