CPSC 677 Advanced NLP Presentation

Reading Comprehension and Question Answering

Ansong Ni

Oct 15, 2020

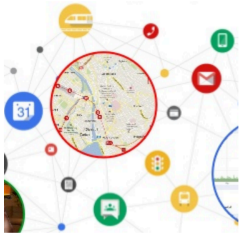
Part 1. Background

RC and QA: Motivation (1)

Question Answering:

Question (+ Context) → Answer (+ Supporting Evidence)

Different types of context:



Knowledge Graph

Databases

Category	Structure	Country	City	Height (metres)	Height (feet)
Mixed use	Burj Khalifa	United Arab Emirates	Dubai	829.8	2,722
Self-supporting tower	Tokyo Skytree	Japan	Tokyo	634	2,080
Mixed use	Shanghai Tower	China	Shanghai	632	2,073
Clock building	Abraj Al Bait Towers	Saudi Arabia	Mecca	601	1,972
Military structure	Large masts of INS Kattabomman	India	Tirunelveli	471	1,545
Mast radiator	Lualualei VLF transmitter	United States	Lualualei, Hawaii	458	1,503
Twin towers	Petronas Twin Towers	Malaysia	Kuala Lumpur	452	1,482
Residential	432 Park Avenue	United States	New York	425.5	1,396
Chimney	Ekibastuz GRES-2 Power Station	Kazakhstan	Ekibastuz	419.7	1,377
Radar	Dimona Radar Facility	Israel	Dimona	406	1,312
Lattice tower	Kiev TV Tower	Ukraine	Kiev	385	1,263
Electricity pylon	Zhoushan Island Overhead Powerline Tie	China	Zhoushan	370	1,214

Semi-structured Tables

Images

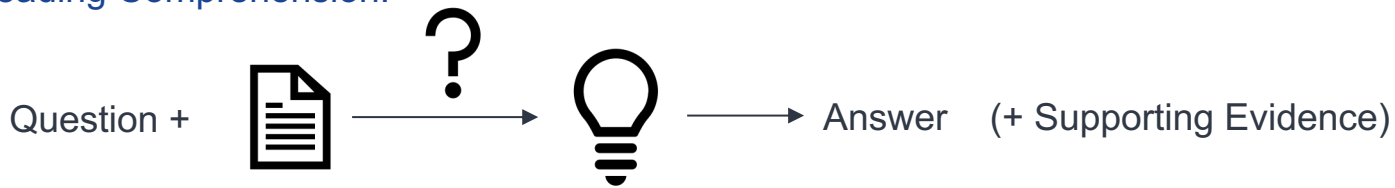
Documents

Conversation

[Danqi Chen, ACL 2020 Tutorial; PhD Thesis, 2018]

RC and QA: Motivation (2)

Reading Comprehension:



How do we know if a RC system really “understands”? -- By asking **questions.**

> *“When a person understands a story, he can demonstrate his understanding by answering questions about the story ... If a computer is said to understand a story, we must demand of the computer the same demonstration of understanding that we require of people.”* -- *Wendy Lehnert, 1977*

[Danqi Chen, PhD Thesis, 2018]

RC and QA: Motivation (3)

The ultimate goal behind RC and QA: Natural Language Understanding (NLU)

An example:

Context:
After winning re-election by defeating Republican opponent Mitt Romney, Obama was sworn in for a second term in 2013. During this term, he promoted inclusion for LGBT Americans.

Questions:
1) Which term did Obama promote inclusion for LGBT Americans?
(Named Entity Recognition, Coreference Resolution)
2) What political party is the candidate that Obama defeat in the re-election in?
(Syntactic Parsing, POS Tagging)
3) Who did Obama beat in 2008 presidential election?
(Identifying missing information)
4) What is the last name of the first lady in 2010
(Numerical reasoning; External knowledge; Commonsense)

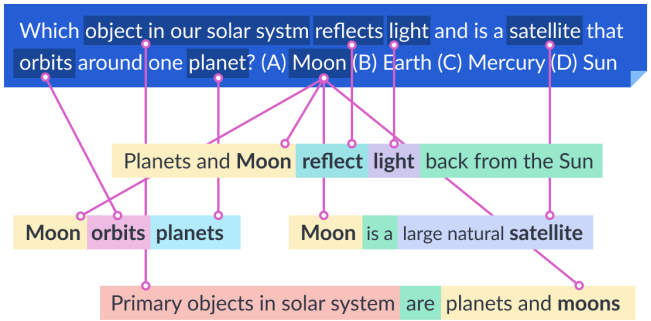
"Question answering should be considered a format which is sometimes useful for studying particular phenomena, not a phenomenon or task in itself." – [Gardner et. al., 2019]

[https://en.wikipedia.org/wiki/Barack_Obama]

RC and QA: Motivation (3)

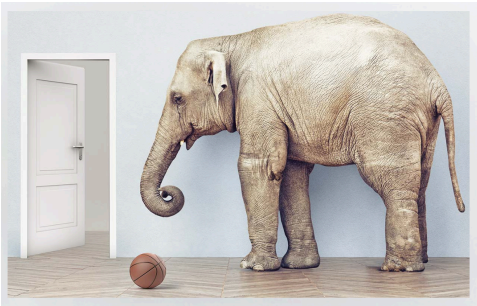
The ultimate goal behind RC and QA: Natural Language Understanding (NLU)

Some more examples:



Aristo (scientific QA)

Visual QA



Commonsense Reasoning

RC and QA: Evaluation

Evaluate by the output answer:

- F1 Score (on span): Harmonic mean of precision and recall

F1 is calculated as follows:

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

where:

$$precision = \frac{TP}{TP + FP}$$

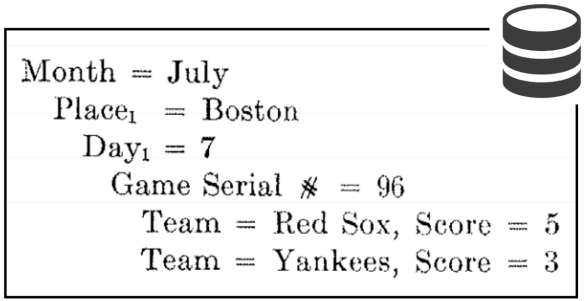
$$recall = \frac{TP}{TP + FN}$$

- Accuracy of the whole answer

RC and QA: A Brief History (1)

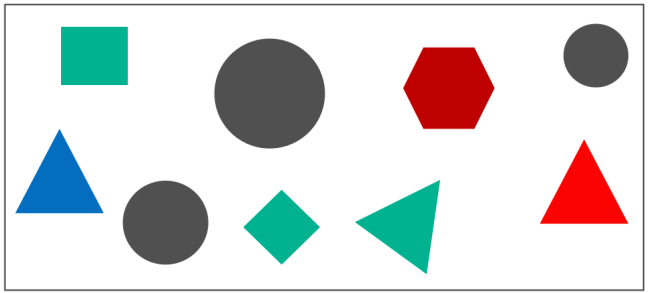
Some early systems:

How many games did the Yankees play in July?



Baseball [Green et. al., 1963]

"All circles are black circles". Is this true?



The Picture Language Machine [Krisch, 1964]

In a 1965 survey, R. F. Simmons concluded that:

- the data-base question-answerer has passed from initial research into the early developmental phase.
- The most difficult and important research questions … to be concerned with measuring meaning, dealing with ambiguities, translating into formal languages and searching large tree structures.

[Scott Yih, ACL 2020 Tutorial]

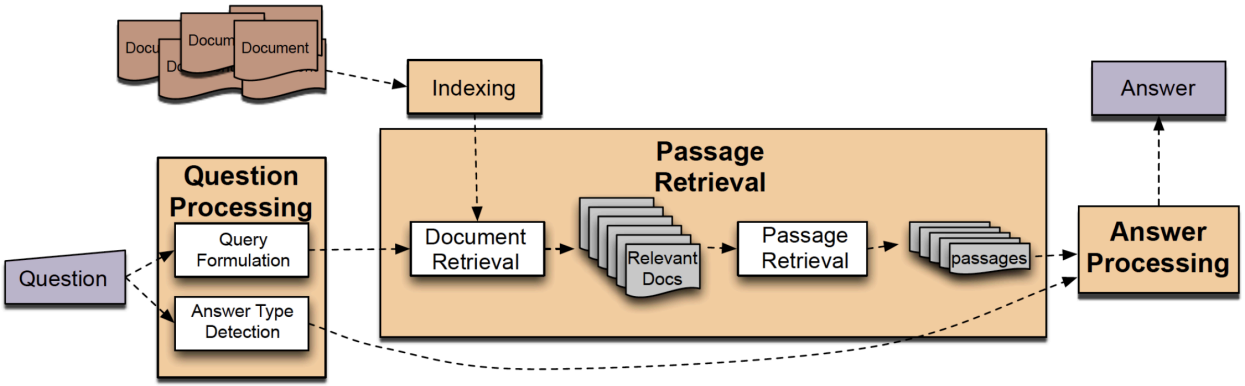
RC and QA: A Brief History (2)

Text Retrieval Conference (TREC): QA Tracks (1999-2007)

Originated from the IR community as the next generation of search:

Relevant documents ⟶ Short answers with support

A typical pipeline of the TREC-QA system:



[Scott Yih, ACL 2020 Tutorial]

RC and QA: A Brief History (3)

The *Jeopardy!* Win -- IBM Watson DeepQA project:

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

Jeopardy is a TV quiz show:

- Questions: clues in the form of answers
- Answers: phrase in the form of question

Question example:

Category: Michigan Mania
Clue: In 1894 C.W. Post created his warm cereal drink Possum in this Michigan city
Answer: Where is Battle Creek?

Many view this as an important milestone of AI.

[Scott Yih, ACL 2020 Tutorial]

Part 2. Recent Developments of the QA/RC Problem

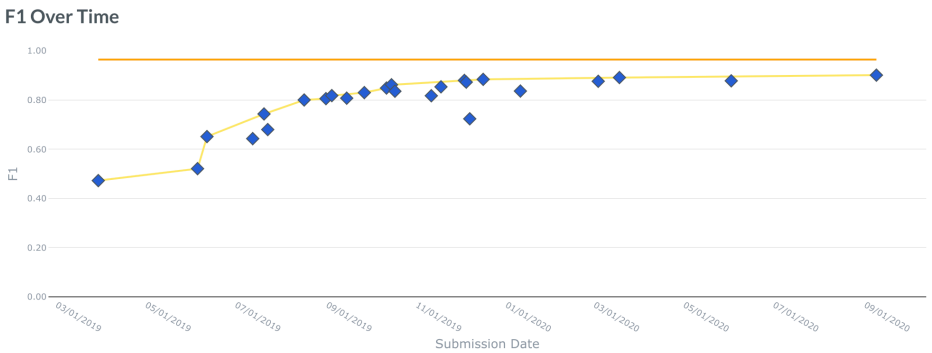
Papers included:

- SQuAD: 100,000+ Questions for Machine Comprehension of Text [Rajpurkar et. al., EMNLP 2016]
- Know What You Don't Know: Unanswerable Questions for SQuAD [Rajpurkar, Jia et. al., ACL 2018]
- **HOTPOTQA: A Dataset for Diverse, Explainable Multi-hop Question Answering [Yang, Qi, Zhang et. al., EMNLP 2018]**
- DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs [Dua et. al., NAACL 2019]

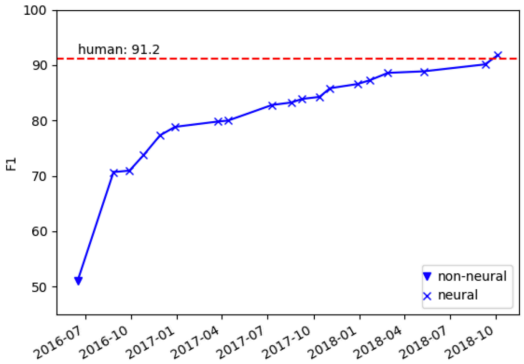
Recent Development of the RC/QA Problem

- 2020: TYDI-QA; IIRC; TORQUE; …
- 2019: ELI5; DROP; ORB; NaturalQuestions; CoQA; Quoref
- 2018: QAngaroo; NarrativeQA; HotpotQA; MultiRC; SQuAD 2.0
- 2017: TriviaQA; RACE;
- 2016: SQuAD1.1; CBT;
- 2015: CNN/Daily Mail

DROP:



SQuAD 1.1:

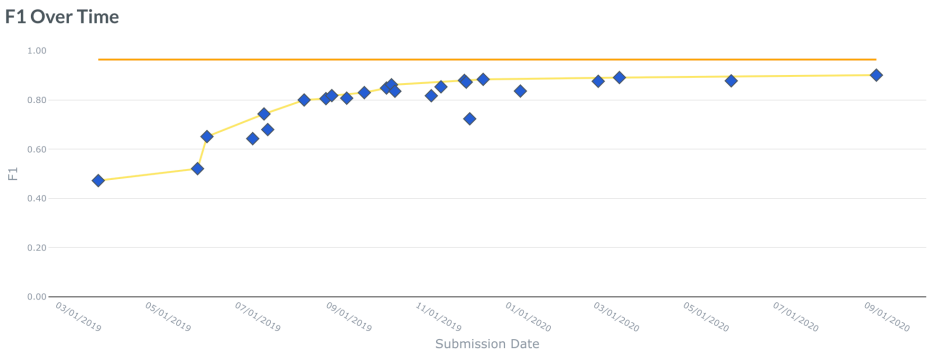


Recent Development of the RC/QA Problem

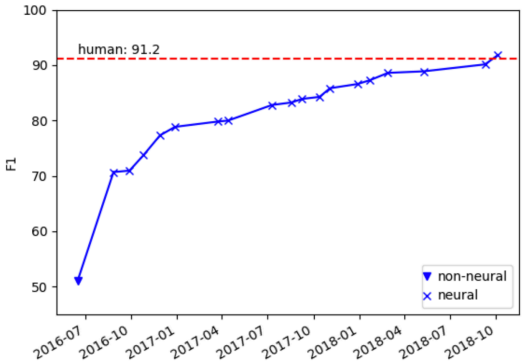
Information Seeking

- 2020: TYDI-QA; IIRC; TORQUE; …
- 2019: ELI5; DROP; ORB; NaturalQuestions; CoQA; Quoref
- 2018: QAngaroo; NarrativeQA; HotpotQA; MultiRC; SQuAD 2.0
- 2017: TriviaQA; RACE;
- 2016: SQuAD1.1; CBT;
- 2015: CNN/Daily Mail

DROP:



SQuAD 1.1:

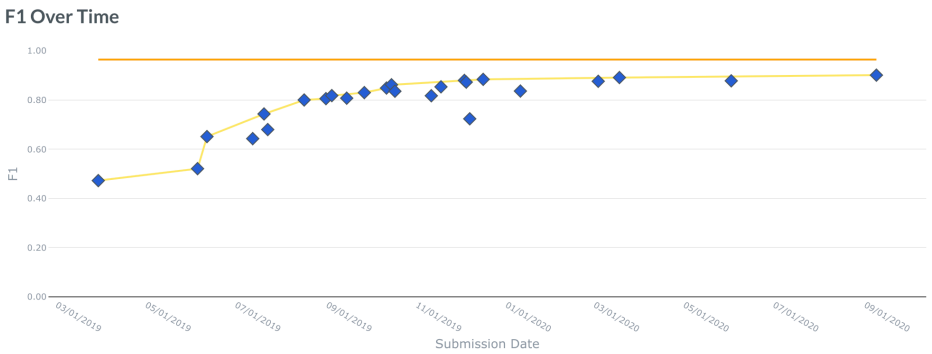


Recent Development of the RC/QA Problem

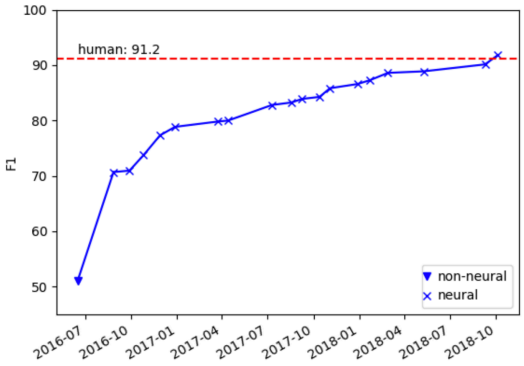
Temporal Reasoning

- 2020: TYDI-QA; IIRC; TORQUE; …
- 2019: ELI5; DROP; ORB; NaturalQuestions; CoQA; Quoref
- 2018: QAngaroo; NarrativeQA; HotpotQA; MultiRC; SQuAD 2.0
- 2017: TriviaQA; RACE;
- 2016: SQuAD1.1; CBT;
- 2015: CNN/Daily Mail

DROP:



SQuAD 1.1:

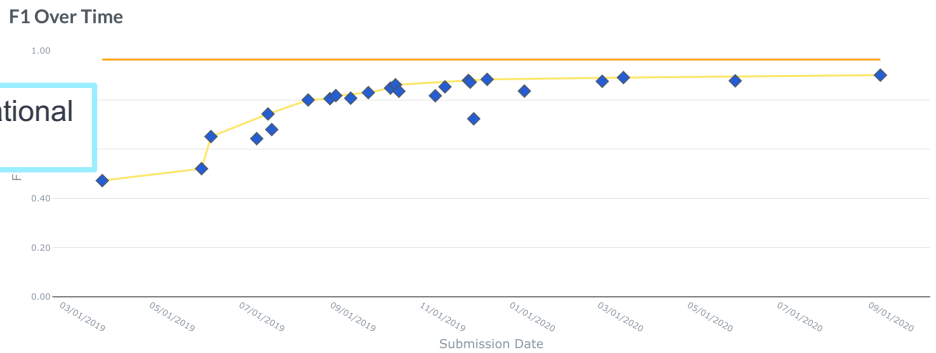


Recent Development of the RC/QA Problem

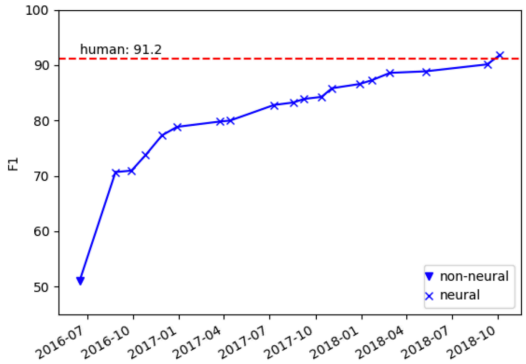
- 2020: TYDI-QA; IIRC; TORQUE; …
- 2019: ELI5; DROP; ORB; NaturalQuestions; CoQA; Quoref
- 2018: QAngaroo; NarrativeQA; HotpotQA; MultiRC; SQuAD 2.0
- 2017: TriviaQA; RACE;
- 2016: SQuAD1.1; CBT;
- 2015: CNN/Daily Mail

Conversational QA

DROP:



SQuAD 1.1:

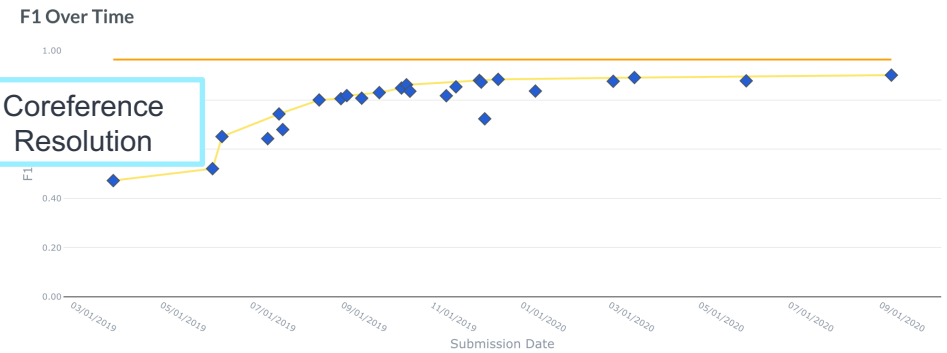


Recent Development of the RC/QA Problem

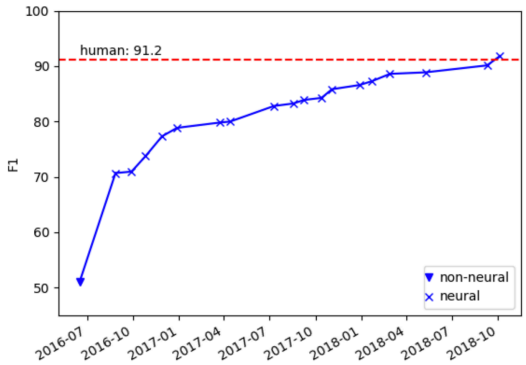
- 2020: TYDI-QA; IIRC; TORQUE; …
- 2019: ELI5; DROP; ORB; NaturalQuestions; CoQA; Quoref
- 2018: QAngaroo; NarrativeQA; HotpotQA; MultiRC; SQuAD 2.0
- 2017: TriviaQA; RACE;
- 2016: SQuAD1.1; CBT;
- 2015: CNN/Daily Mail

Coreference Resolution

DROP:



SQuAD 1.1:



SQuAD: 100,000+ Questions for Machine Comprehension of Text

An example:

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

- First large-scale (100K+) reading comprehension dataset.
- The contexts are Wikipedia article segments, questions are posed by crowd-workers, answers are spans in the context.
- Most reasoning rely on shadow semantics of the context to answer the question.
- A logistic regression based method achieves 51% F1 score.
- Estimated human performance is 86.8% F1.

[Rajpurkar et. al., EMNLP 2016]

Know What You Don't Know: Unanswerable Questions for SQuAD

An example:

Article: Endangered Species Act
Paragraph: " . . . *Other legislation followed, including the Migratory Bird Conservation Act of 1929, a 1937 treaty prohibiting the hunting of right and gray whales, and the Bald Eagle Protection Act of 1940. These later laws had a low cost to society—the species were relatively rare—and little opposition was raised.*"

Question 1: "*Which laws faced significant opposition?*"
Plausible Answer: *later laws*

Question 2: "*What was the name of the 1937 treaty?*"
Plausible Answer: *Bald Eagle Protection Act*

- Also known as "SQuAD 2.0"
- Added another 50K unanswerable questions written adversarially to look similar to original SQuAD questions.
- Strong systems achieving 86% F1 on SQuAD 1.1 only get 66% F1 on SQuAD 2.0

[Rajpurkar, Jia et. al., ACL 2018]

HOTPOTQA: A Dataset for Diverse, Explainable Multi-hop Question Answering

An example:

Paragraph A, Return to Olympus:
[1] *Return to Olympus is the only album by the alternative rock band Malfunkshun.* [2] *It was released after the band had broken up and after lead singer Andrew Wood (later of Mother Love Bone) had died of a drug overdose in 1990.* [3] Stone Gossard, of Pearl Jam, had compiled the songs and released the album on his label, Loosegroove Records.

Paragraph B, Mother Love Bone:
[4] *Mother Love Bone was an American rock band that formed in Seattle, Washington in 1987.* [5] The band was active from 1987 to 1990. [6] *Frontman Andrew Wood's personality and compositions helped to catapult the group to the top of the burgeoning late 1980s/early 1990s Seattle music scene.* [7] *Wood died only days before the scheduled release of the band's debut album, "Apple", thus ending the group's hopes of success.* [8] The album was finally released a few months later.

Q: What was the former band of the member of Mother Love Bone who died just before the release of "Apple"?
A: Malfunkshun
Supporting facts: 1, 2, 4, 6, 7

- One of the most popular QA/RC dataset to-date
  (over 40 submissions to the leaderboard in less than 2 years)
- 113K questions written by crowdsource workers.
- "Multi-hop reasoning": it requires the model to combine the information from multiple sentences in different paragraphs and reason about them
- Specifies two different settings for retrieval:

Distractor Setting

Full Wiki Setting

[Yang, Qi, Zhang et. al., EMNLP 2018]

DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs

Some examples:

Passage (some parts shortened)	**Question**	**Answer**
That year, his **Untitled (1981)**, a painting of a haloed, black-headed man with a bright red skeletal body, depicted amid the artists signature scrawls, was **sold by Robert Lehrman for $16.3 million, well above its $12 million high estimate**.	How many more dollars was the Untitled (1981) painting sold for than the 12 million dollar estimation?	4300000
Denver would retake the lead with kicker **Matt Prater nailing a 43-yard field goal**, yet Carolina answered as kicker **John Kasay ties the game with a 39-yard field goal**. ... Carolina closed out the half with **Kasay nailing a 44-yard field goal**. ... In the fourth quarter, Carolina sealed the win with **Kasay's 42-yard field goal**.	Which kicker kicked the most field goals?	John Kasay
In **1517, the seventeen-year-old King sailed to Castile**. There, his Flemish court .... **In May 1518, Charles traveled to Barcelona in Aragon**.	Where did Charles travel to first, Castile or Barcelona?	Castile

- The first dataset that requires **numerical reasoning** (e.g. addition/subtraction, counting, sorting) to answer the questions.
- 96K questions written by crowdsource workers based on Wikipedia paragraphs
- Tested both SOTA **semantic parsing** systems and reading comprehension systems, only achieve 32.7% F1
- Expert human performance is 96.4% F1

[Dua et. al., NAACL 2019]

Putting it together…

	SQuAD	HotpotQA (distractor)	HotpotQA (full-wiki)	DROP
Context	 (one paragraph)	 (distractors, multi-paragraph)	 (multi-paragraph, at scale)	 (one paragraph)
Question	What was the name of the 1937 treaty? (single-hot reasoning)	What was the former band of the member of Mother Love Bone who died just before the release of “Apple”? (multi-hop reasoning)	What was the former band of the member of Mother Love Bone who died just before the release of “Apple”? (multi-hop reasoning)	How many more dollars was the Untitled (1981) painting sold for than the 12-million-dollar estimation? (discrete, numerical reasoning)

Part 3. Modeling with Neural Modular Networks

Papers included:

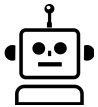
- Deep Compositional Question Answering with Neural Module Networks [Andreas et. al., CVPR 2016]
- **Neural Modular Networks for Reasoning over Text [Gupta et. al., ICLR 2020]**

Question Answering vs. Semantic Parsing (1)

An example:

argmax(count(field_goal, kicker))

Logical Form



NL Query

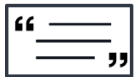
Which kicker kicked the most field goals?

Structured Knowledge

Name	Team	Goal Distance
Matt Prater	Denver	43-yard
John Kasay	Carolina	39-yard
John Kasay	Carolina	44-yard
John Kasay	Carolina	42-yard

Target

John Kasay



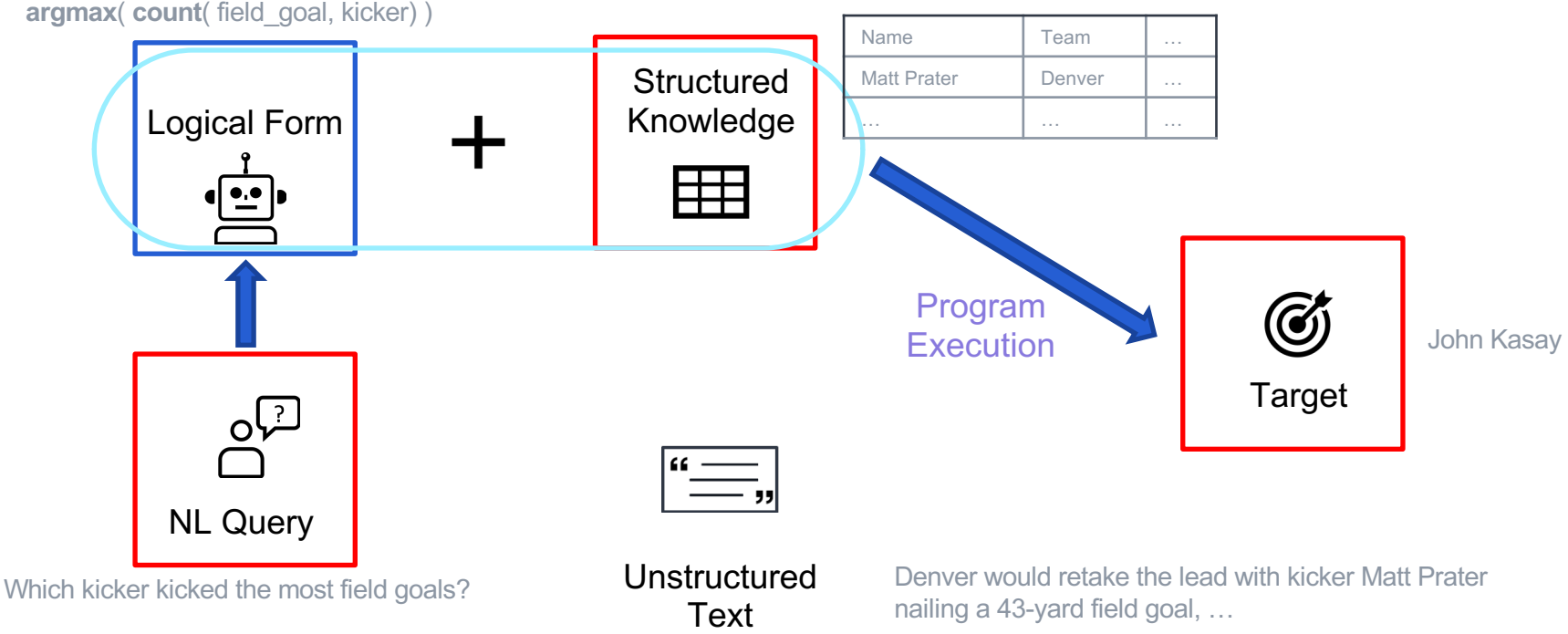
Unstructured Text

Denver would retake the lead with kicker Matt Prater nailing a 43-yard field goal, yet Carolina answered as kicker John Kasay ties the game with a 39-yard field goal. . . . Carolina closed out the half with Kasay nailing a 44-yard field goal. . . . In the fourth quarter, Carolina sealed the win with Kasay's 42-yard field goal.

Question Answering vs. Semantic Parsing (2)

Semantic Parsing:

argmax(**count**(field_goal, kicker))

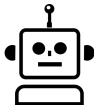


Question Answering vs. Semantic Parsing (3)

End-to-end Question Answering:

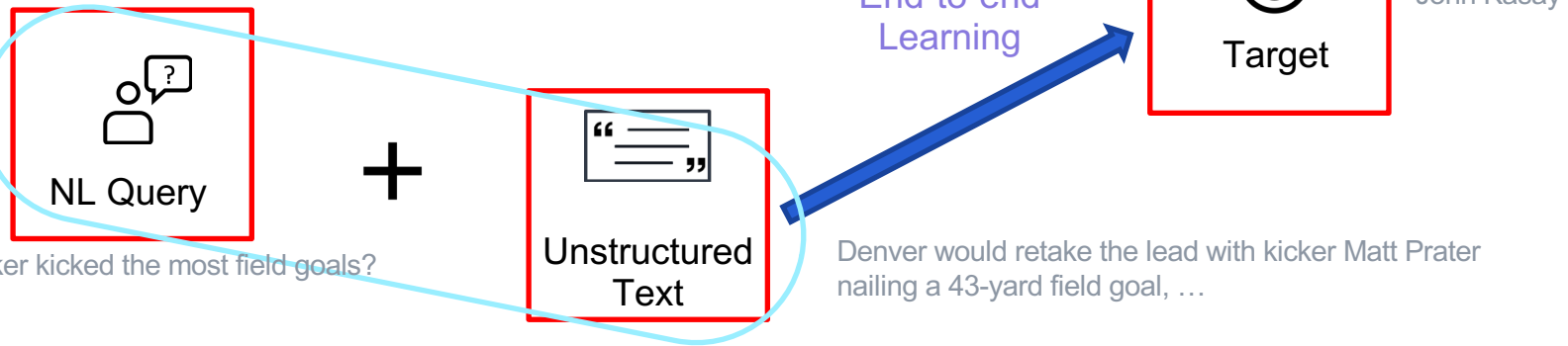
argmax(**count**(field_goal, kicker))

Logical Form

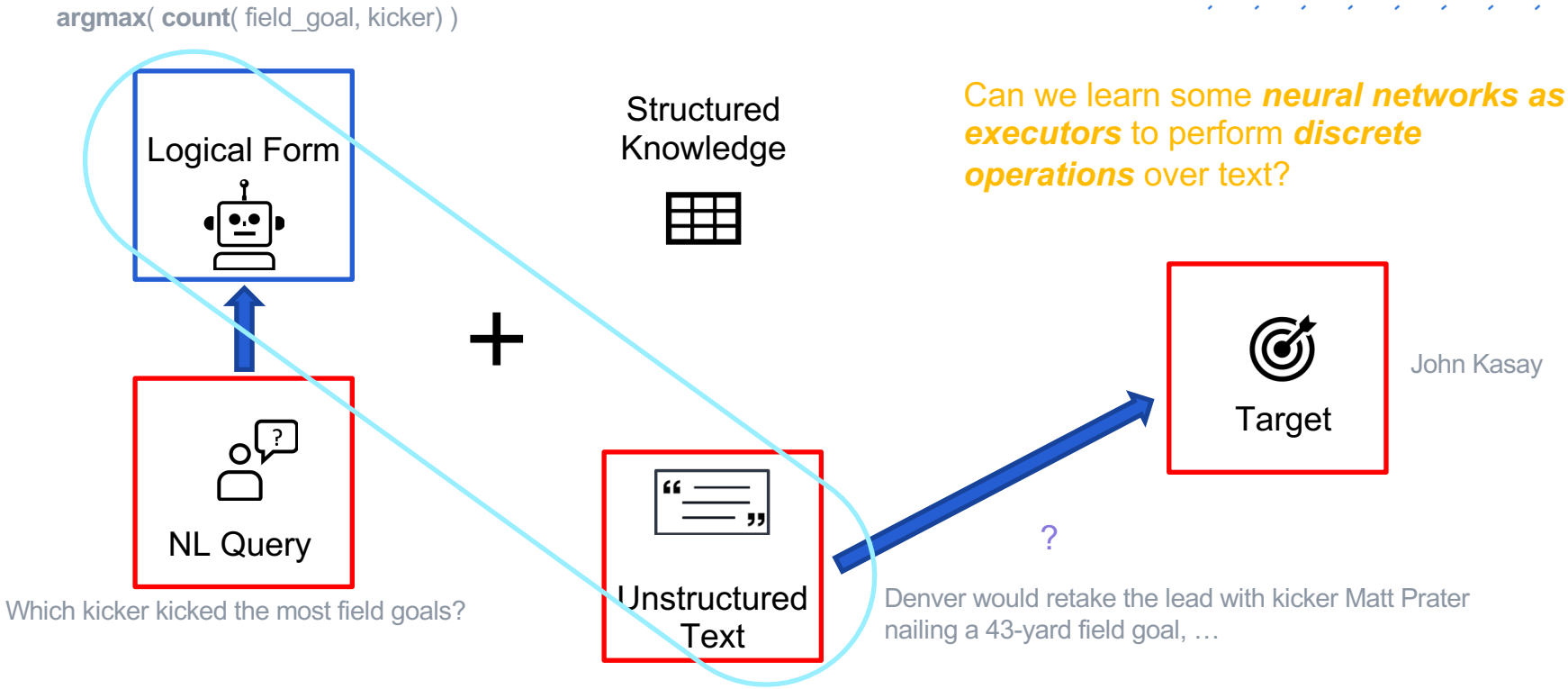


Structured Knowledge

Name	Team	…
Matt Prater	Denver	…
…	…	…



Question Answering vs. Semantic Parsing (3)



Deep Compositional Question Answering with Neural Module Networks (1)

Visual Question Answering (VQA):

Question (+ Context) →  → Answer (+ Supporting Evidence)

What color is the tie?

Yellow

[Andreas et. al., CVPR 2016]

Deep Compositional Question Answering with Neural Module Networks (2)

"jointly trained neural *modules,* dynamically composed into deep networks based on linguistic structure"

Neural Modules

- Attn()
- Re-Attn()
- Measure()
- Combine()
- Classify()

[Andreas et. al., CVPR 2016]

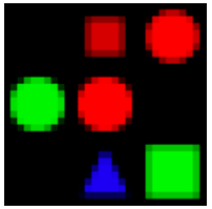
[Photo from: https://blog.se.com/datacenter/2014/09/15/new-way-think-data-center-design-optimizing-data-center-like-box-legos/]

Deep Compositional Question Answering with Neural Module Networks (3)

Some examples:

What color is the tie?

How many lights are there?

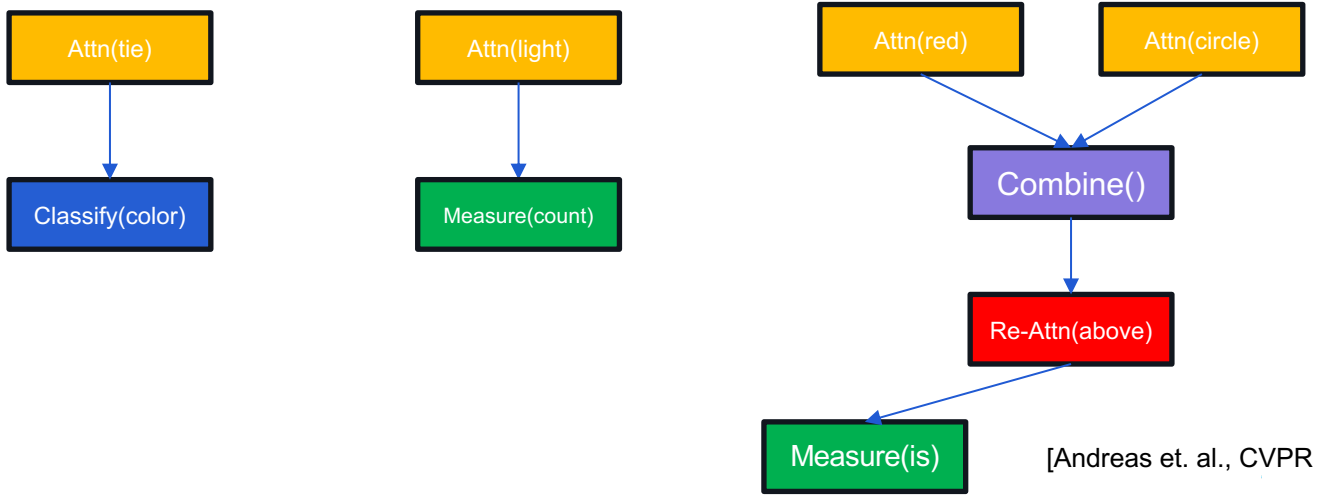


Is there a red shape above a circle?

Neural Modules

- Attn()
- Measure()
- Classify()
- Re-Attn()
- Combine()

Semantic Parsing



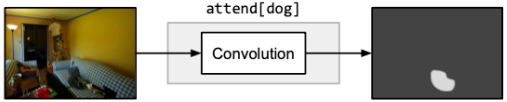
[Andreas et. al., CVPR 2016]

Deep Compositional Question Answering with Neural Module Networks (4)

What these neural modules really look like:

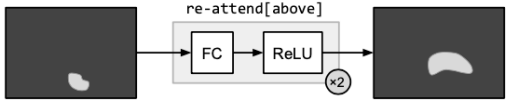
Attention

attend : *Image* → *Attention*



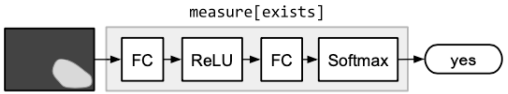
Re-attention

re-attend : *Attention* → *Attention*



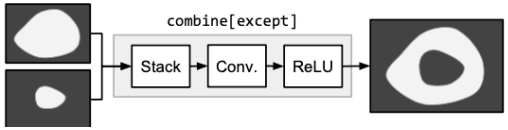
Measurement

measure : *Attention* → *Label*



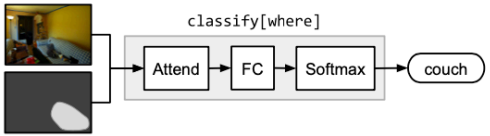
Combination

combine : *Attention* × *Attention* → *Attention*



Classification

classify : *Image* × *Attention* → *Label*



A classification module `classify[c]` takes an attention and the input image and maps them to a distribution over labels. For example, `classify[color]` should return a distribution over colors in the region attended to.

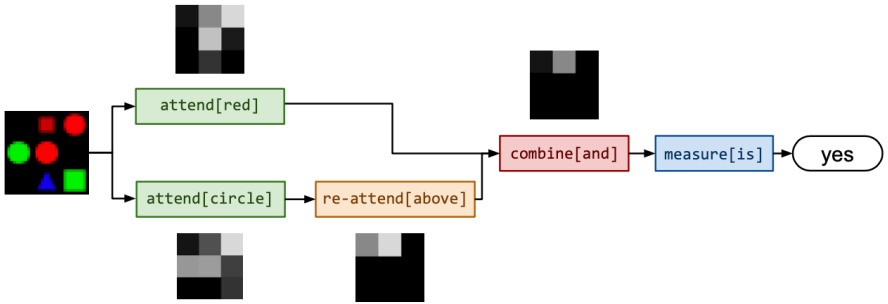
Use attention map as an intermediate representation:

- Interpretable
- Continuous and differentiable
- All intermediate representations are from the same space

[Andreas et. al., CVPR 2016]

Deep Compositional Question Answering with Neural Module Networks (4)

A concrete example:



Pros:

- Interpretability
- Better generalization

Cons:

- Hard to train
- Scalability
- Faithfulness (talk about later)

[Andreas et. al., CVPR 2016]

Neural Modular Networks for Reasoning over Text (1)

Dataset: DROP (partial)

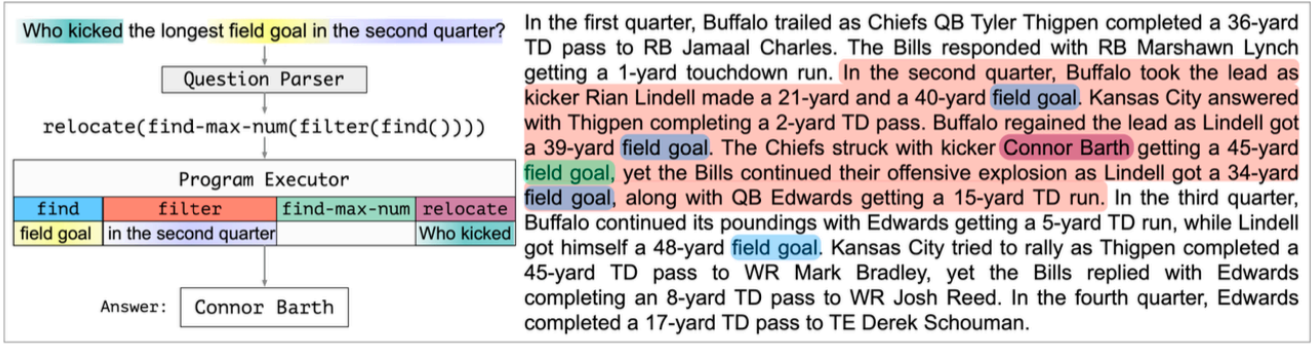
Passage (some parts shortened)	**Question**	**Answer**
That year, his **Untitled (1981)**, a painting of a haloed, black-headed man with a bright red skeletal body, depicted amid the artists signature scrawls, was **sold by Robert Lehrman for $16.3 million, well above its $12 million high estimate**.	How many more dollars was the Untitled (1981) painting sold for than the 12 million dollar estimation?	4300000
Denver would retake the lead with kicker **Matt Prater nailing a 43-yard field goal**, yet Carolina answered as kicker **John Kasay ties the game with a 39-yard field goal**. ... Carolina closed out the half with **Kasay nailing a 44-yard field goal**. ... In the fourth quarter, Carolina sealed the win with **Kasay's 42-yard field goal**.	Which kicker kicked the most field goals?	John Kasay
In **1517, the seventeen-year-old King sailed to Castile**. There, his Flemish court .... **In May 1518, Charles traveled to Barcelona in Aragon**.	Where did Charles travel to first, Castile or Barcelona?	Castile

- Defined 10 different neural modules
- Defined 5 types for the intermediate/final execution result
- Jointly learn the question parser and neural modules
- Train and tested on a subset of DROP dataset
- Other tricks include: heuristic for question parsing/decomposition, auxiliary loss for some neural modules, curriculum learning, etc

[Gupta et. al., ICLR 2020]

Neural Modular Networks for Reasoning over Text (2)

A concrete example:

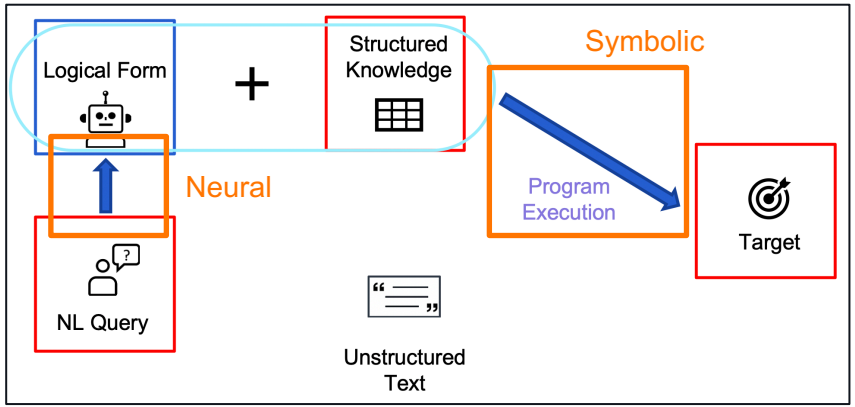


AllenNLP Demo: https://demo.allennlp.org/reading-comprehension

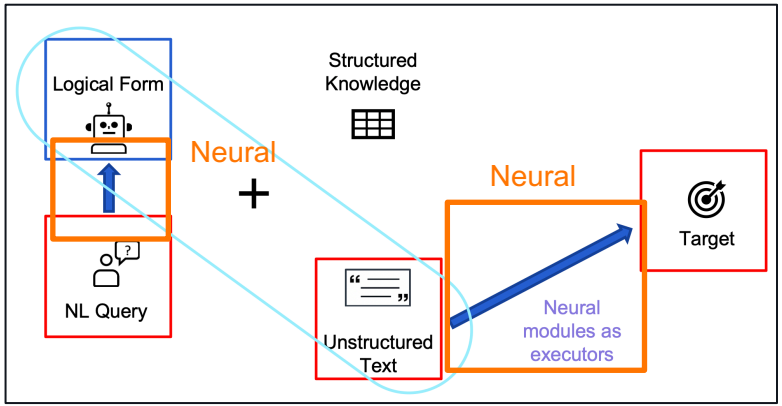
[Gupta et. al., ICLR 2020]

Neural Modular Networks for Reasoning over Text (3)

Joint learning of semantic parser and the neural modules:



QA with Semantic Parsing



QA with Neural Module Network

[Gupta et. al., ICLR 2020]

Part 4. Context Retrieval for QA/RC

Papers included:

- **Learning to Retrieve Reasoning Paths over Wikipedia Graphs for Question Answering [Asai et. al., ICLR 2020]**

Why retrieval?

Most of the time, QA is not like this:

What causes precipitation to fall?
gravity

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

It's more like this…

Retrieval is important, and hard… (1)

Retrieval is important for QA:

- We want all relevant information to be retrieved (Recall)
- We don’t want to retrieve extraneous context to introduce noise (Precision)

An illustration:

In what year was Isaac Newton born?

What you see:

> Isaac Newton was born on Christmas Day, 25 December 1642

What neural QA models (probably) see:

> Isaac Newton, xxx xxx xxxx 1642.

With some noise:

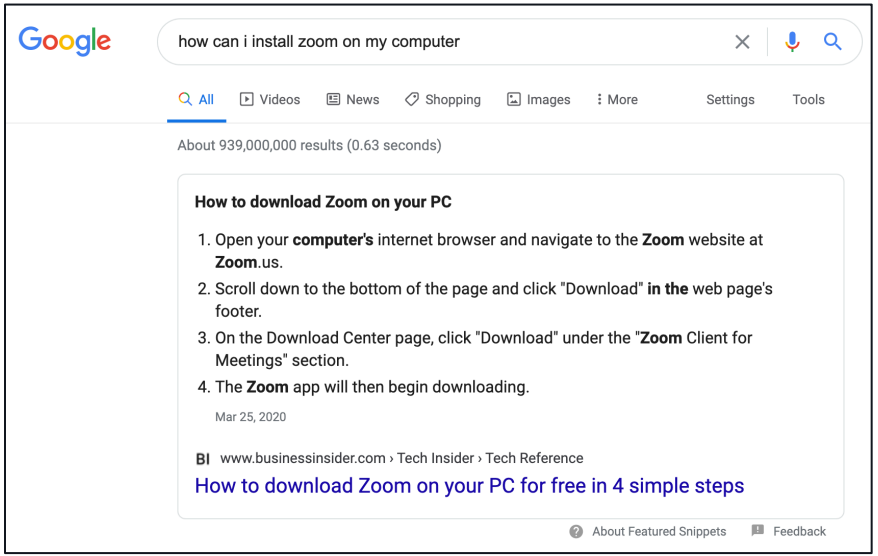
> xx 1703, xxx xxx 1727 xxx xxx, xxxxx xxxxxxxx xxxxxxxxxxx 1643 xxxxx. Isaac Newton, xxxx xxx 1642. xxx 1699 xxxxxx

Retrieval is important, and hard… (2)

Retrieval for QA is also very hard:

- The search space is usually very large (*e.g.* Wikipedia, the whole internet)
- For multi-hop questions, usually more than one piece of evidences need to be retrieved
- For complex questions, knowing what to retrieve would already require some reasoning ability

Previously, the retrieval mostly relies on lexical overlapping/similarities:

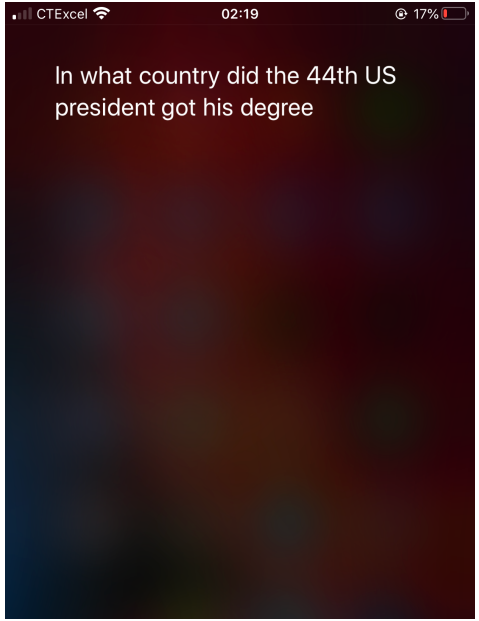


Retrieval is important, and hard… (3)

Retrieval for QA is also very hard:

- The search space is usually very large (*e.g.* Wikipedia, the whole internet)
- For multi-hop questions, usually more than one piece of evidences need to be retrieved
- For complex questions, knowing what to retrieve would already require some reasoning ability

Now, we need some notion of “semantic” retrieval:



Learning to Retrieve Reasoning Paths over Wikipedia Graphs for Question Answering (1)

Dataset: HotpotQA (fullwiki setting)

Paragraph A, Return to Olympus:
[1] *Return to Olympus is the only album by the alternative rock band Malfunkshun.* [2] *It was released after the band had broken up and after lead singer Andrew Wood (later of Mother Love Bone) had died of a drug overdose in 1990.* [3] Stone Gossard, of Pearl Jam, had compiled the songs and released the album on his label, Loosegroove Records.

Paragraph B, Mother Love Bone:
[4] *Mother Love Bone was an American rock band that formed in Seattle, Washington in 1987.* [5] The band was active from 1987 to 1990. [6] *Frontman Andrew Wood's personality and compositions helped to catapult the group to the top of the burgeoning late 1980s/early 1990s Seattle music scene.* [7] *Wood died only days before the scheduled release of the band's debut album, "Apple", thus ending the group's hopes of success.* [8] The album was finally released a few months later.

Q: What was the former band of the member of Mother Love Bone who died just before the release of "Apple"?
A: Malfunkshun
Supporting facts: 1, 2, 4, 6, 7

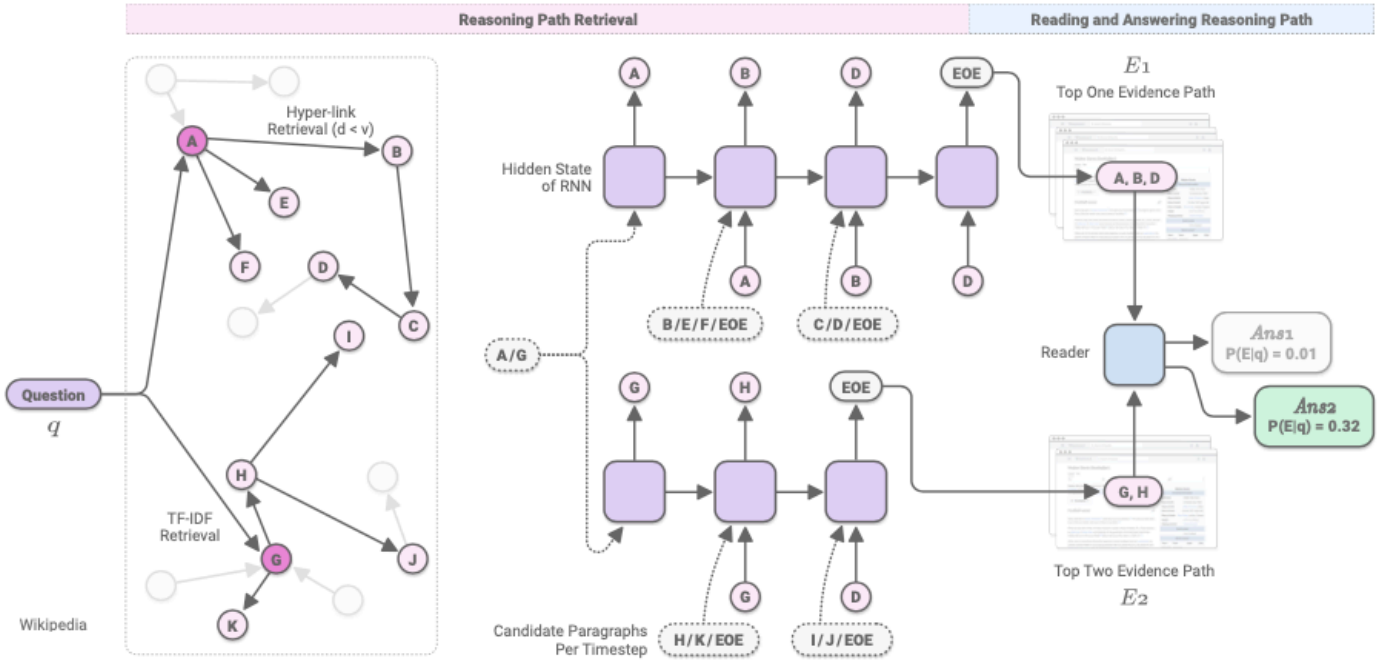
Use Wikipedia as a graph connected by hyperlinks:

[photo from: https://rethinkingvis.com/visualizations/217]

[Asai et. al., ICLR 2020]

Learning to Retrieve Reasoning Paths over Wikipedia Graphs for Question Answering (2)

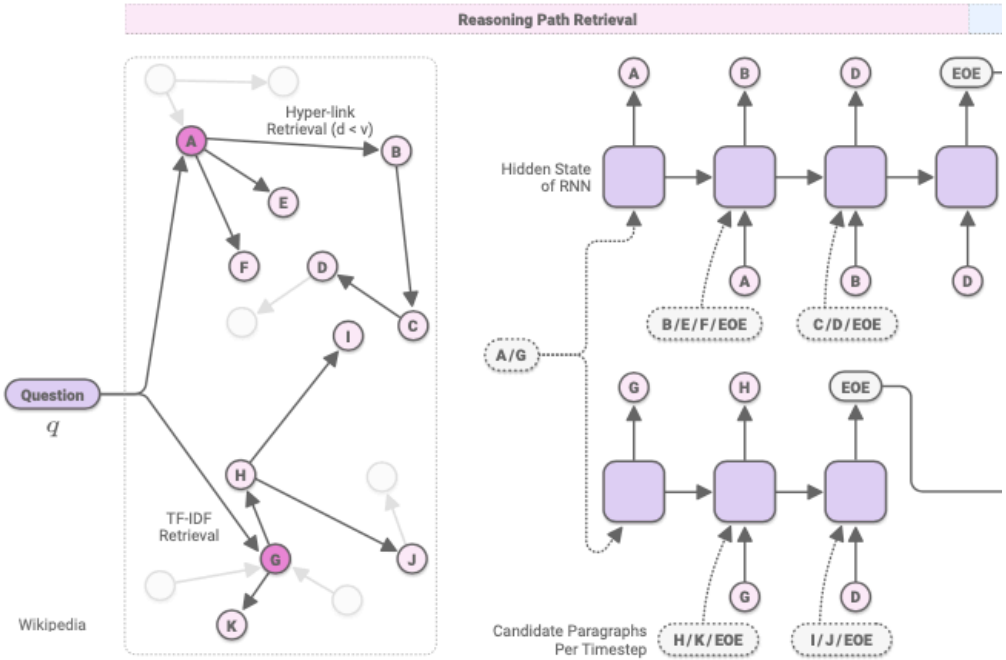
An overview:



[Asai et. al., ICLR 2020]

Learning to Retrieve Reasoning Paths over Wikipedia Graphs for Question Answering (3)

Reasoning path retrieved in detail:



Data Augmentation:

Augment with additional reasoning paths by adding a paragraph linked to the first paragraph (e.g. A/G) and have high TF-IDF score.

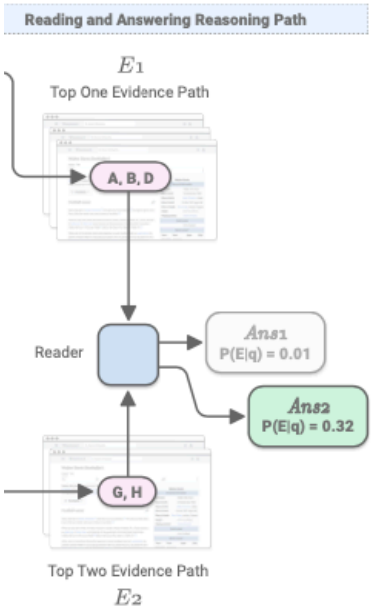
Negative Example for Robustness:

1) TF-IDF based negative examples
2) Hyperlink based negative examples

[Asai et. al., ICLR 2020]

Learning to Retrieve Reasoning Paths over Wikipedia Graphs for Question Answering (4)

Reading and answering in detail:



Model the reader as multi-task learning:

1) Reading comprehension (span extraction)
2) Reasoning path re-ranking

The reader itself is a BERT model.

[Asai et. al., ICLR 2020]

Part 5. Findings and Discussion

Papers included:

- **Compositional Questions Do NOT Necessitate Multi-hop Reasoning [Min, Wallace et. al., ACL 2019]**
- Obtaining Faithful Interpretations from Compositional Neural Networks [Subramanian, Bogin, Gupta et. al., ACL 2020]

Compositional Questions Do NOT Necessitate Multi-hop Reasoning

Argument: *"Even highly compositional questions can be answered with a single hop if they target on specific entity types or the facts needed to answer them are redundant."*

Question: What is the former name of the animal whose habitat the Réserve Naturelle Lomako Yokokala was established to protect?
Paragraph 5: The Lomako Forest Reserve is found in Democratic Republic of the Congo. It was established in 1991 especially to protect the habitat of the Bonobo apes.
Paragraph 1: The bonobo ("Pan paniscus"), formerly called the **pygmy chimpanzee** and less often, the dwarf or gracile chimpanzee, is an endangered great ape and one of the two species making up the genus "Pan".

They also found that:

1) A single-hop BERT-based model achieves 67 F1 – comparable to SOTA
2) Humans can answer 80% of the question without all the necessary paragraphs being shown to them

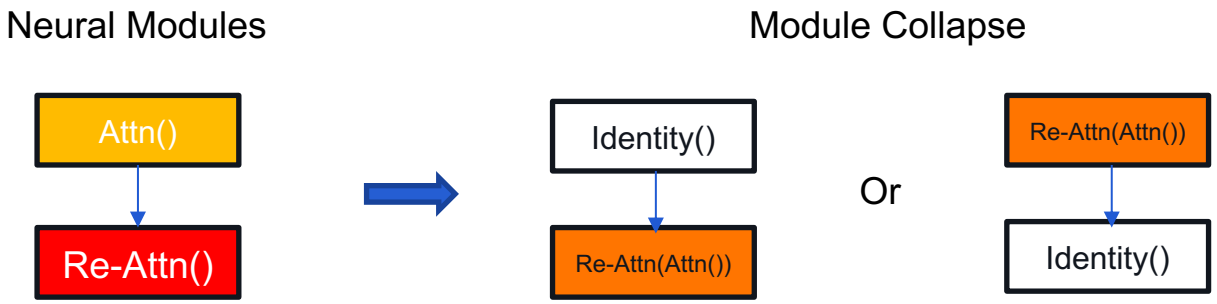
Discussion question 1:

- How does this finding affects your view on formatting RC as a QA problem?

[Min, Wallace et. al., ACL 2019]

Obtaining Faithful Interpretations from Compositional Neural Networks

Argument: *"Prior work assumes that all modules perform their intended behavior… we illustrate that the network structure does not provide a faithful explanation of the model behavior"*



Discussion question 2:

- Improving faithfulness/interpretability is usually at the cost of performance, what is your take for this trade-off? Is it not enough to get good results on the test set?

[Subramanian, Bogin, Gupta et. al., ACL 2020]

Some other discussion questions…

Discussion question 3:

- We observed great success with QA & RC in recent years: **~95** F1 for SQuAD 1.1; **~93** F1 for SQuAD 2.0; **~83** F1 for Hotpot (distractor); **~79** F1 for Hotpot (fullwiki); **~90** F1 for DROP… Do you think RC/QA is a solved problem? Or do you think the settings we have are too easy?

Discussion question 4:

- We have seen what current QA/RC models can do, but can you think of a question that you think will be hard for the current QA systems to answer, or have no hope to answer at all?

Thanks!