CS 371N: Natural Language Processing
Guest Lecture: Question Answering

TEXAS The University of Texas at Austin Eunsol Choi

This Lecture

- Introduction to question answering task in NLP

when did they stop making the nissan xterra?

Dataset: How do we collect the questions?

The Nissan Xterra was discontinued in 2015.

ChatGPT, Nov 2023

Dataset: How do we collect gold answers?

Model: How should models generate the answers?

Evaluation: How do we evaluate model generated answers?

Presentation: How should we present the answers?

Overview

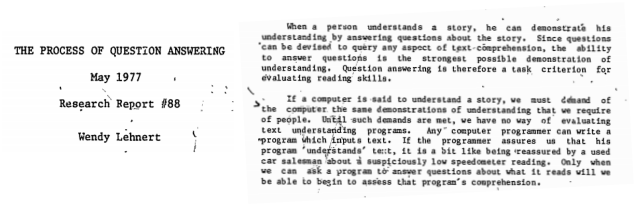
- Why do we study QA?

QA can be very broad

- Factoid QA:
  - *what states border Mississippi?*
  - *when was Barack Obama born?*
  - *how is Advil different from Tylenol?*
- "Question answering" as a term is so broad as to be meaningless
  - *Is P=NP?*
  - *What is 4+5?*
  - *What is the translation of [sentence] into French?*
  - *Is it okay to use a blender in 2AM in an apartment?*

Why do we study QA?

- As a testbed to evaluate how machines understand text



THE PROCESS OF QUESTION ANSWERING

May 1977

Research Report #88

Wendy Lehnert

When a person understands a story, he can demonstrate his understanding by answering questions about the story. Since questions can be devised to query any aspect of text comprehension, the ability to answer questions is the strongest possible demonstration of understanding. Question answering is therefore a task criterion for evaluating reading skills.

If a computer is said to understand a story, we must demand of the computer the same demonstrations of understanding that we require of people. Until such demands are met, we have no way of evaluating text understanding programs. Any computer programmer can write a program which inputs a text. If the programmer assures us that his program 'understands' text, it is a bit like being reassured by a used car salesman about a suspiciously low speedometer reading. Only when we can ask a program to answer questions about what it reads will we be able to begin to assess that program's comprehension.

"Since questions can be devised to query **any aspect** of text comprehension, the ability to answer questions is the **strongest possible demonstration of understanding**."

Model-testing Queries

Questioner already knows the answer, aiming to test model's understanding or knowledge

Passage

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".

Annotator writes question

Question

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

Answer

graupel

[SQuAD, MCTest, RACE, …]

"Close Reading" dataset

- Questions 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

MCTest

Richardson (2013)

"Close Reading" dataset

Reasoning	**Passage** (some parts shortened)	**Question**	**Answer**
Subtraction (28.8%)	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
Count (16.5%) and Sort (11.7%)	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

- Questions require discrete reasoning (such as addition, counting, sorting, comparing)

DROP dataset

Due et al (2019)

Trivia Questions

WILLIAM WILKINSON'S "AN ACCOUNT OF THE PRINCIPALITIES OF WALLACHIA AND MOLDOVIA" INSPIRED THIS AUTHOR'S MOST FAMOUS NOVEL.

Jeopardy! Question

Question: The Dodecanese Campaign of WWII that was an attempt by the Allied forces to capture islands in the Aegean Sea was the inspiration for which acclaimed 1961 commando film?

TriviaQA dataset
Joshi et al. (2017)

Trivia Questions

- Questions are often compositional and complex
- But, systems can do well without really understanding the text by capturing surface clues (e.g., 1961, campaign)

Question: The Dodecanese Campaign of WWII that was an attempt by the Allied forces to capture islands in the Aegean Sea was the inspiration for which acclaimed 1961 commando film?
Answer: The Guns of Navarone
Excerpt: The Dodecanese Campaign of World War II was an attempt by Allied forces to capture the Italian-held Dodecanese islands in the Aegean Sea following the surrender of Italy in September 1943, and use them as bases against the German-controlled Balkans. The failed campaign, and in particular the Battle of Leros, inspired the 1957 novel **The Guns of Navarone** and the successful 1961 movie of the same name.

Joshi et al. (2017)

Multi-hop Reasoning Datasets

Question: *What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell ?*

Doc 1: *Shirley Temple Black was an American actress, businesswoman, and singer … As an adult, she served as Chief of Protocol of the United States …*

Same entity

Same entity

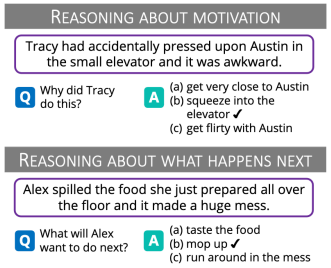
Doc 2: *Kiss and Tell is a comedy film in which 17-year-old Shirley Temple acts as Corliss Archer. …*

Doc 3: *Meet Corliss Archer is an American television sitcom that aired on CBS …*

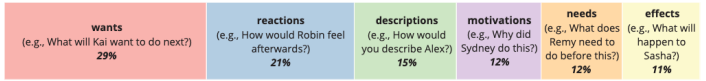
- Much longer and more convoluted questions requiring multi document reasoning

Example picked from HotpotQA [Yang et al., 2018]

"Common sense" QA datasets



- Questions query emotional and social intelligence, not encyclopedic knowledge.
- Answering this will not depend on evidence documents.



Social IQA dataset [Sap, Rashkin et al EMNLP (2019)]

Datasets that seek expert knowledge

Context:
In jurisdictions where use of headlights is optional when visibility is good, drivers who use headlights at all times are less likely to be involved in a collision than are drivers who use headlights only when visibility is poor. Yet Highv
at all times does nothing to
Question: Which one of t
information above?
Options:
A. In jurisdictions where u
headlights for daytime driv
B. Only very careful driver
C. The jurisdictions where
is frequently poor.
D. A law making use of hea
Answer: B

Table 1: An example in t
Council (2019b).

Conceptual Physics

When you drop a ball from rest it accelerates downward at 9.8 m/s². If you instead throw it downward assuming no air resistance its acceleration immediately after leaving your hand is
(A) 9.8 m/s² ✔
(B) more than 9.8 m/s² ✘
(C) less than 9.8 m/s² ✘
(D) Cannot say unless the speed of throw is given. ✘

College Mathematics

In the complex z-plane, the set of points satisfying the equation $z^2 = |z|^2$ is a
(A) pair of points ✘
(B) circle ✘
(C) half-line ✘
(D) line ✔

Figure 4: Examples from the Conceptual Physics and College Mathematics STEM tasks.

Reclor dataset
Yu et al, ICLR 2020

MMLU dataset
Hednrycks et al, ICLR 2021

Model-testing Queries

Questioner already knows the answer, aiming to test model's understanding or knowledge

Passage

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".

Annotator writes question

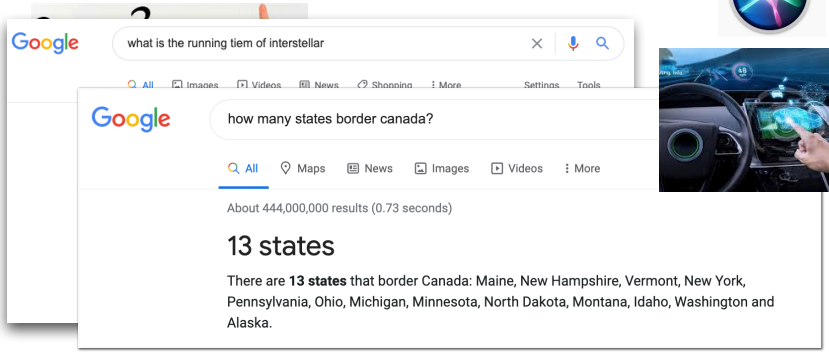
Question
What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?

Answer
graupel

[SQuAD, MCTest, RACE, ...]

Why do we study QA?

- Build a helpful tool for humans to gather information



Model-testing Queries

Questioner already knows the answer, aiming to test model's understanding or knowledge

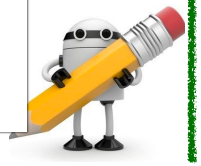
Passage

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".

Annotator writes question

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

Answer
graupel



[SQuAD, MCTest, RACE, ...]

Information Seeking Queries

Questioner does not know the answer

Question: What ship did Han Solo pilot?

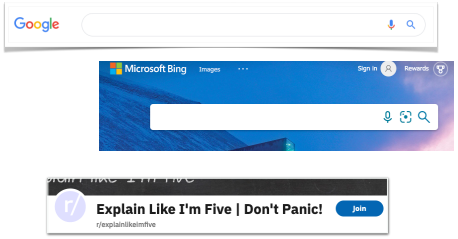
Annotator finds answer in article

Article

The Millennium Falcon is a fictional starship in the Star Wars franchise. The modified YT-1300 Corellian light freighter is primarily commanded by Corellian smuggler Han Solo (Harrison Ford) and

Where to get questions?

User Queries



Natural Questions [Kwiatkowski et al, TACL 2019]

[Berant et al, 2013, Yang et al, EMNLP 15,
Bajaj et al NeurIPS workshop 2018, Fan et al 2019]

NaturalQuestions

- Real questions from Google, answerable with Wikipedia
- Short answers and long answers (snippets)
- Questions arose naturally

Question:

where is blood pumped after it leaves the right ventricle?

Long Answer:

From the right ventricle , blood is pumped through the semilunar pulmonary valve into the left and right main pulmonary arteries (one for each lung) , which branch into smaller pulmonary arteries that spread throughout the lungs.

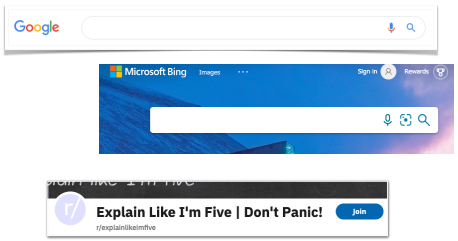
Short Answer:

None

Kwiatkowski et al. (2019)

Where to get questions?

User Queries



Natural Questions [Kwiatkowski et al, TACL 2019]

[Berant et al, 2013, Yang et al, EMNLP 15,
Bajaj et al NeurIPS workshop 2018, Fan et al 2019]

Crowdsourcing

Given:
entity name and the first paragraph of Wikipedia page

Do:
Ask questions to learn as much as possible about this entity!

[Choi et al EMNLP 2018, Clark et al TACL 2020, Ferguson et al, EMNLP 2020]

Challenges with information seeking queries

- Unanswerable / partially answerable questions
  - In existing information seeking datasets, 20-50% of questions are left unanswered [Asai and Choi, ACL 2021]
- Questions with false presupposition (FP)

Q How do martial artists who karate chop or punch a cement block not break their hand?
C It's a trick, the blocks are not very strong, and they are being punched or kicked in their weakest points.
FP Chops or cement blocks are strong.

Q How do bugs and other insects survive winter when they have such a short lifespan?
C Depends on the insect, some don't have that short of a lifespan. But mostly (...)
FP (All) insects have a short lifespan.

CREPE dataset, Yu et al, ACL 2023

Overview

- Why do we study QA?
- Formulating QA tasks and evaluation metrics

Simulating QA from raw text

- Typically, question answering dataset requires human annotation
- Can we automatically simulate QA without annotations?
- "Cloze" task: word (often an entity) is removed from a sentence
  - Answers: multiple choice, pick from passage, or pick from vocabulary

[QA Dataset Explosion, Rogers et al]

Children's Book Test

"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

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.

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 ? ''

- 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)

Dataset Properties

- Axis 1: what's the output space?
  - cloze task (fill in blank)

Multiple-Choice datasets

Context:
In jurisdictions where use of headlights is optional when visibility is good, drivers who use headlights at all times are less likely to be involved in a collision than are drivers who use headlights only when visibility is poor. Yet Highway Safety Department records show that making use of headlights mandatory at all times does nothing to reduce the overall number of collisions.
Question: Which one of the following, if true, most helps to resolve the apparent discrepancy in the information above?
Options:
A. In jurisdictions where use of headlights is optional when visibility is good, one driver in four uses headlights for daytime driving in good weather.
B. Only very careful drivers use headlights when their use is not legally required.
C. The jurisdictions where use of headlights is mandatory at all times are those where daytime visibility is frequently poor.
D. A law making use of headlights mandatory at all times is not especially difficult to enforce.
Answer: B

Table 1: An example in the ReClor dataset which is modified from the Law School Admission Council (2019b).

- Can capture complex semantics
- Evaluation is straightforward
- But is it realistic?

ReCLOR dataset (ICLR 2021) https://openreview.net/pdf?id=HJgJtT4tvB

Span-based prediction

Question : What shift happened in animal regulation in 1963 in U.S?

Document Context :

The Lacey Act of 1900 was the first federal law that regulated commercial animal markets. It prohibited interstate commerce of animals killed in violation of state game laws, and covered all wildlife. Whereas the Lacey Act dealt with game animal management and market commerce species, a major shift in focus occurred by 1963 to habitat preservation instead of take regulations. A provision was added by Congress in the Land and Water Conservation Fund Act of…

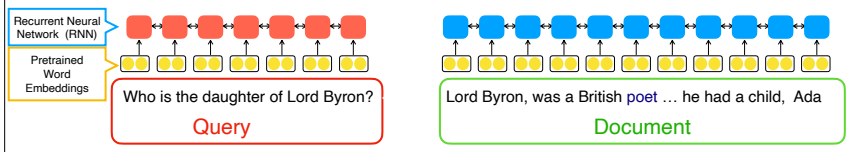
Answer is span in the original document

- Can capture various semantics
- Evaluation is somewhat straightforward (measure the overlap between predicted and gold span)
- More realistic than multiple choice

[Rajpurkar et al 2016]

Model: BiDAF (Bi-directional Attention Flow)

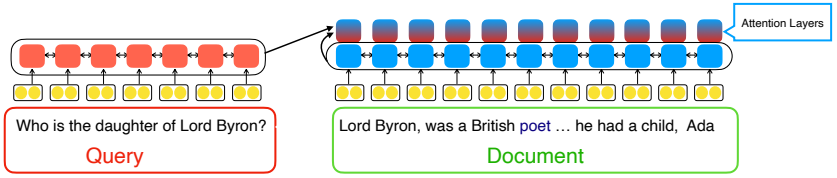
- Encode text and question with recurrent neural network



[Seo et al, ICLR 17]

Model: BiDAF (Bi-directional Attention Flow)

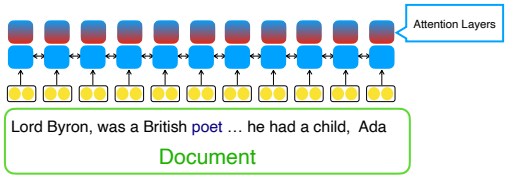
- Encode text and question with recurrent neural network
- Compute inter-sentence alignment with attention



[Seo et al, ICLR 17]

Model: BiDAF (Bi-directional Attention Flow)

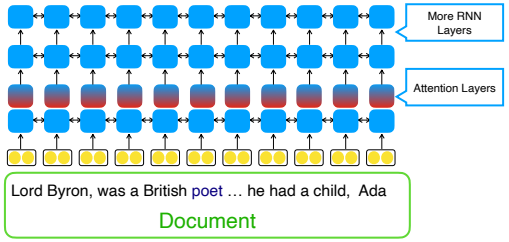
- Encode text and question with recurrent neural network
- Compute inter-sentence alignment with attention



[Seo et al, ICLR 17]

Model: BiDAF (Bi-directional Attention Flow)

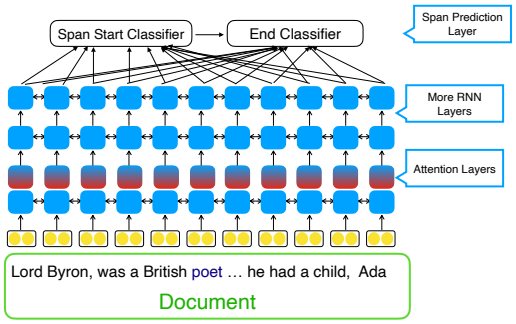
- Encode text and question with recurrent neural network
- Compute inter-sentence alignment with attention



[Seo et al, ICLR 17]

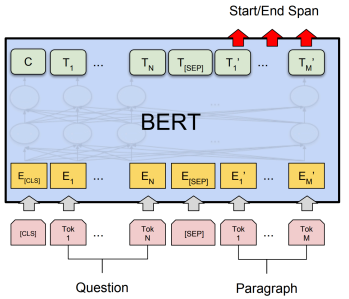
Model: BiDAF (Bi-directional Attention Flow)

- Encode text and question with recurrent neural network
- Compute inter-sentence alignment with attention
- Optimize for the log likelihood of finding the correct start and end positions



[Seo et al, ICLR 17]

Span-based QA with BERT



What was Marie Curie the first female recipient of ? [SEP] One of the most famous people born in Warsaw was Marie …

- Simplified architecture — just concatenate question and context!

Devlin et al. (2019)

Free-form answer generation

Question: Why does salt bring out the flavor in most foods?

Answer: Salt does a couple of things that add to the flavor of foods. First off, it makes things salty. That sounds simplistic, but salt is one of the 5 basic tastes, so it tends to taste good simply all by itself. Secondly, salt will lessen sensitivity to other tastes, especially bitter. This is somewhat true of all the tastes, but adding salt will balance out a dish and not let one taste overpower the others. Thirdly, it's been shown to increase that aromatic effects of many types of food. A good deal of your "taste" of a food actually comes from the smell of that food (which is why things tend to taste so bland when you nose is congested, like when you have the flu).

- Long-form question answering (LFQA)
- Can capture complex semantics
- Evaluation??

 Explain Like I'm Five (ELI5), from Fan et al. 2019

Understanding LFQA

Question: Can the capacity of our brains be roughly measured in bytes?

Summary
Answer
Example
Auxiliary Info
Org sentence

Human written answer: This is a hard question to answer.Of course, since we occupy finite space, our properties must be finite.But the exceedingly complex structure of the brain and other systems […] makes it hard to calculate the amount of "data" that we can store.We don't have a way to measure resolution of life or the quality of everyday noises.[…]Most of our memories are vague recollections,[…].However, we can estimate average capacity, although it has no clear meaning in our terms of cognition. That number would be 2.5 petabytes […]Even so, this number varies as well from person to person.[…]

- Not all sentences in long form answer convey answer information
- About 40% of sentences serve other roles

 [How do we answer complex questions: Discourse structure of long form answers Xu et al, ACL 2022]

Difficulty of evaluating LFQA

Lexical matching based automatic metrics (e.g. ROUGE) are used, but not meaningful.

Q: Can you protect electronics from EMPs/solar flares? If so, how?

Random answer (ROUGE-L: 19.4)

The fast lane/slow lane is a bit of a misnomer. It gives the impression that new, faster lanes are being built. In reality, normal speed will be the new "fast lane", which will cost extra, and everything else will be in the throttled "slow lane".

Gold answer (ROUGE-L: 18.6)

I'll start with the grounding question, because that's the easiest to answer: Doesn't help a bit. All that matters is that the metal container is conductive and doesn't have gaps...completely seal your Faraday cage. Consider soldering the lid on to that paint can... look at little baggie it comes in. Sealed mylar. That protected that chip from air travel at 35,000 feet, land travel through rural, urban, and suburban areas, and all the electromagnetic radiation that the trip entails... No lead shielding. No safes…

 Krishna et al. 2021

Can humans evaluate long-form answers?

How does a speaker vibrate at multiple frequencies simultaneously to deliver sounds to our ears?

Answer A: This has been asked many times and the answer is they don't. If you listen to the song being played live on purely acoustic instruments even though they are being played separately and emitting their own frequencies, what you hear (and by extension, what a microphone captures) at any given time is just ONE frequency that's the "sum" of all the others combined. A speaker is just a reverse microphone.

Answer B: Imagine an ocean with a consistent wave. It flows up and down, with equal distance between the two waves at any time. Now imagine I push a larger, shorter wave into this ocean. The two waves will collide, resulting in some new wave pattern. This new wave pattern is a combination of those two waves. Speakers work similarly. If I combine two soundwaves, I get a new combination wave that sounds different.

Can experts evaluate?

Expert 1

Preference: A

In technical terms ocean waves stated in answer B are transverse waves and sound waves are longitudinal waves. In comparison answer B mentions about ocean waves and it is different to the sound waves in the question. But apart from that actually the two answers A and B go very close to each other and they provide similar explanations. But answer A is selected to be slightly better in terms of applicability and relevance. [...]

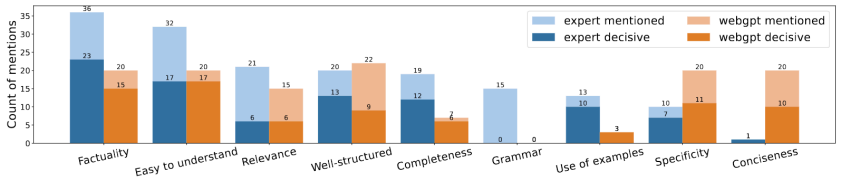
Expert 2

Preference: B

It is difficult to choose between these two answers because they both are not wrong and give essentially the same explanation. I go with answer B because I like the analogy with the ocean waves, and due to how visual the explanation is it is easier to understand in my opinion. [...]

- Even experts disagree on which one is a better answer

Evaluation aspects for LFQA



- Diverse facets are considered when evaluating LFQA answers.
- Best evaluation at the moment seems to be asking LLM whether it is a good answer or not, but not very reliable.

 Xu, Song, Iyyer and Choi, ACL 2023

Dataset Properties

- Axis 1: what's the output space?
  - cloze task (fill in blank)
  - multiple choice
  - span-based prediction
  - freeform generation
- Complex output space allows answering more complex queries, but evaluation becomes very tricky…

Dataset Properties

- Axis 2: what's the knowledge source (input)?
  - One paragraph? One document? All of Wikipedia? Images? Tables? All of web?

Dataset Property: Input

Where was the last Winter Olympic Games held?

Benchmarking in the Past

2018 Winter Olympics

The **2018 Winter Olympics**, officially known as the **XXIII Olympic Winter Games** (French: *Les XXIII*es *Jeux olympiques d'hiver*[a] Korean: 제23회 동계 올림픽, romanized: *Jeisipsamhoe Donggye Ollimpik*) and commonly known as **PyeongChang 2018** (Korean: 평창 2018), was an international winter multi-sport event that was held between 9 and 25 February 2018 in Pyeongchang County, Gangwon Province, South Korea, with the opening rounds for certain events held on 8 February 2018, the day before the opening ceremony.

Past benchmarks assume:

- The answer is a ***span*** in the provided passage
- All the necessary context is given in the document

PyeongChang

Span-based QA benchmarks

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Sep 18, 2019	ALBERT (ensemble model) *Google Research & TTIC* https://arxiv.org/abs/1909.11942	**89.731**	**92.215**
2 Jul 22, 2019	XLNet + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	88.592	90.859
2 Sep 16, 2019	ALBERT (single model) *Google Research & TTIC* https://arxiv.org/abs/1909.11942	88.107	90.902
2 Jul 26, 2019	UPM (ensemble) *Anonymous*	88.231	90.713
3 Aug 04, 2019	XLNet + SG-Net Verifier (ensemble) *Shanghai Jiao Tong University & CloudWalk* https://arxiv.org/abs/1908.05147	88.174	90.702
4 Aug 04, 2019	XLNet + SG-Net Verifier++ (single model) *Shanghai Jiao Tong University & CloudWalk* https://arxiv.org/abs/1908.05147	87.238	90.071

- Performance is saturated by ~2019
- We can aim for a more realistic, challenging QA setting!

Dataset Property: Input

Where was the last Winter Olympic Games held?

Benchmarking in the Past

2018 Winter Olympics

The **2018 Winter Olympics**, officially known as the **XXIII Olympic Winter Games** (French: *Les XXIII*es *Jeux olympiques d'hiver*[a] Korean: 제23회 동계 올림픽, romanized: *Jeisipsamhoe Donggye Ollimpik*) and commonly known as **PyeongChang 2018** (Korean: 평창 2018), was an international winter multi-sport event that was held between 9 and 25 February 2018 in Pyeongchang County, Gangwon Province, South Korea, with the opening rounds for certain events held on 8 February 2018, the day before the opening ceremony.

Benchmarking Today

Today's benchmarks assume:

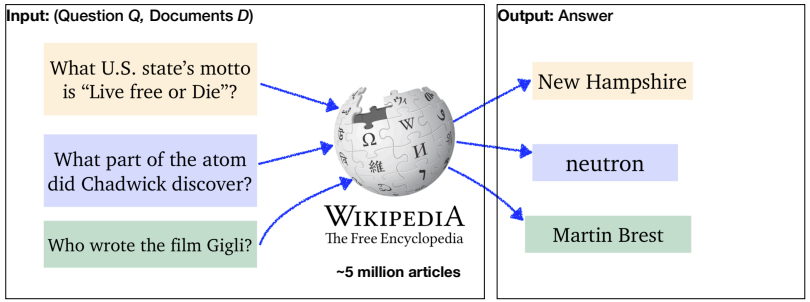
- The answer is out there ***somewhere***…
- We have to assume the context

Past benchmarks assume:

- The answer is a ***span*** in the provided passage
- All the necessary context is given in the document

PyeongChang

Open Retrieval QA

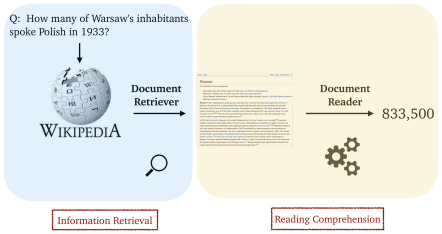


- Retrieval performance is often the bottleneck!

Open Retrieval QA

Retriever-reader pipeline

- **Retriever** selects documents from a large corpus that's relevant to the query
- Then, **reader** selects the top scoring span from the top-5 retrieved documents



 [Chen et al, ACL 2017]

Classic Information Retrieval Task

- Given a query and a document corpus, provide a ranked list of documents that is relevant to the query.

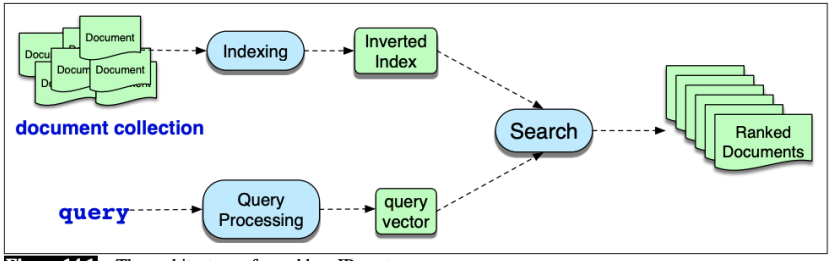


Figure 14.1 The architecture of an ad hoc IR system.

- Typically the document collection is large — efficiency is important!

Classic Solution: TF-IDF

Token (t) Document (d) Corpus (C)

- Tf-idf = product of tf and idf $\text{tf-idf}(t, d, C) = tf_{t,d} \cdot idf_{t,C}$
  - Tf: term (t) frequency in document d

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

  - Idf: inverse document frequency

$$idf_{t,C} = \log_{10} \frac{|C|}{df_t}$$

Total number of documents in the collection

Number of documents where term t occurs

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

- Scoring document (d) for a given query (q): $\text{score}(q, d) = \sum_{t \in q} \frac{\text{tf-idf}(t, d)}{|d|}$

Dense Vectors

- Can we use dense vectors for retrieval?
  - Embed queries and documents with encoder (e.g., BERT) and score the similarity by taking their dot product

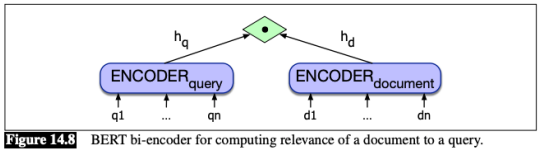


Figure 14.8 BERT bi-encoder for computing relevance of a document to a query.

$$h_q = BERT_Q(q)[CLS]$$
$$h_d = BERT_D(d)[CLS]$$
$$\text{score}(q, d) = h_q \cdot h_d$$

  - Does not work well out of the box…

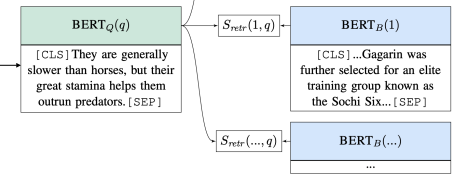
Fine-tuning LM for Retrieval

- Inverse Cloze Task
  - Given a sentence as a query (q), retrieve its context (b) as a target

$$h_q = \mathbf{W_q}\text{BERT}_Q(q)[\text{CLS}]$$
$$h_b = \mathbf{W_b}\text{BERT}_B(b)[\text{CLS}]$$
$$S_{retr}(b,q) = h_q^\top h_b$$

$$P_{\text{ICT}}(b|q) = \frac{\exp(S_{retr}(b,q))}{\sum_{b' \in \text{BATCH}} \exp(S_{retr}(b',q))}$$

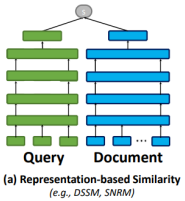
Example:
"...Zebras have four gaits: walk, trot, canter and gallop. ***They are generally slower than horses, but their great stamina helps them outrun predators.*** *When chased, a zebra will zigzag from side to side..."*



[Lee et al, ACL 2019]

Dense Retrieval

- Dual-encoder architectures
  - Encode query and document separately, and search for nearest neighbor
  - Allows faster retrieval

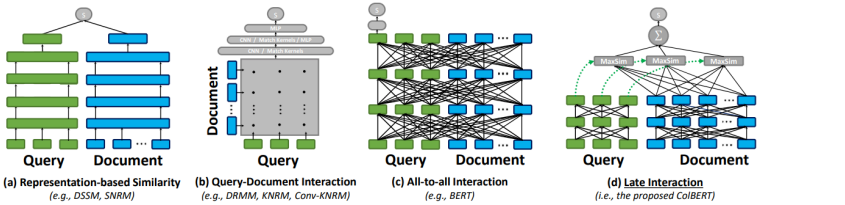


(a) Representation-based Similarity
(e.g., DSSM, SNRM)

[Khattab et al, SigIR2020]

Dense Retrieval

- Dual-encoder architectures
  - Encode query and document separately, and search for nearest neighbor
  - Allows faster retrieval
- Cross-encoder architectures
  - Encode query and document jointly
  - Outperform dual-encoder given training data
  - Often used together with more efficient methods



(a) Representation-based Similarity *(e.g., DSSM, SNRM)*
(b) Query-Document Interaction *(e.g., DRMM, KNRM, Conv-KNRM)*
(c) All-to-all Interaction *(e.g., BERT)*
(d) Late Interaction *(i.e., the proposed ColBERT)*

[Khattab et al, SigIR2020]

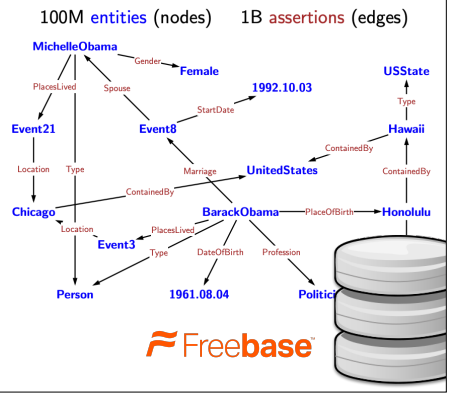
Database as Evidence: Classical QA

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

Q: *where was Barack Obama born*

```
λx. type(x, Location) ∧
  born_in(Barack_Obama, x)
```

(also Prolog / GeoQuery, etc.)



Semi-structured Data as Evidence

Year	City	Country	Nations
1896	Athens	Greece	14
1900	Paris	France	24
1904	St. Louis	USA	12
...	...	...	...
2004	Athens	Greece	201
2008	Beijing	China	204
2012	London	UK	204

x = Greece held its last Summer Olympics in which year?

y = 2004

Pasupat and Liang ACL 2015

Image as an Evidence

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

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

VQA: Antol et al 2015

What is expiration date?

VizWiz: Gurari et al 2018

Mixing evidence from various modalities

Multimodal Context

Steal This Movie!
The film follows Hoffman's (D'Onofrio) relationship with his second wife Anita (Garofalo) and their "awakening" and subsequent conversion to an activist life. The title of the film is a play on Hoffman's 1970 counter-culture guidebook titled "Steal This Book".

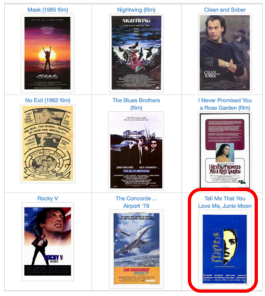
La liceale
La liceale (internationally released as The Teasers, "Under-graduate Girls", "Sophomore Swingers" and "Teasers") is a 1975 commedia sexy all'italiana directed by Michele Massimo Tarantini.
...
Guida. It was followed by "La liceale nella classe dei ripetenti".

Sage Stallone
Stallone made his acting debut alongside his father in Rocky V (1990), the fifth installment of the Rocky franchise, playing Robert Balboa Jr., the onscreen son of his father's title character. He did not, however,
...
After that, he acted in lesser profile films.

Pierino contro tutti
Pierino contro tutti (also known as "Desirable Teacher") is a 1981 comedy film directed by Marino Girolami. The main character of the film is Pierino, an
...
I as a short lived subgenre of joke-films in which the plot basically consists of a series of jokes placed side by side.

Ben Piazza - Filmography

Year	Title	Role
1957	A Dangerous Age	David
1959	The Hanging Tree	Rune
1962	No Exit	Camarero
1970	**Tell Me That You Love Me, Junie Moon**	Jesse
1972	The Outside Man	Desk Clerk
...	...	...
1985	Mask	Mr. Simms
1988	Clean and Sober	Kramer
1990	**Rocky V**	Doctor
1991	Guilty by Suspicion	Darryl Zanuck

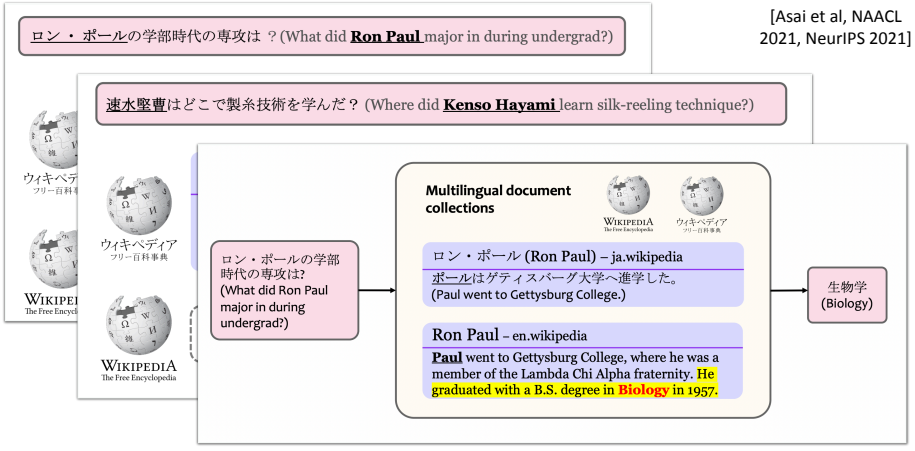


Q: Which B. Piazza title came earlier: the movie S. Stallone's son starred in or the movie with half of a lady's face on the poster?
A: Tell Me That You Love Me, Junie Moon

Talmor et al, ICLR 2020

Multilingual QA

[Asai et al, NAACL 2021, NeurIPS 2021]

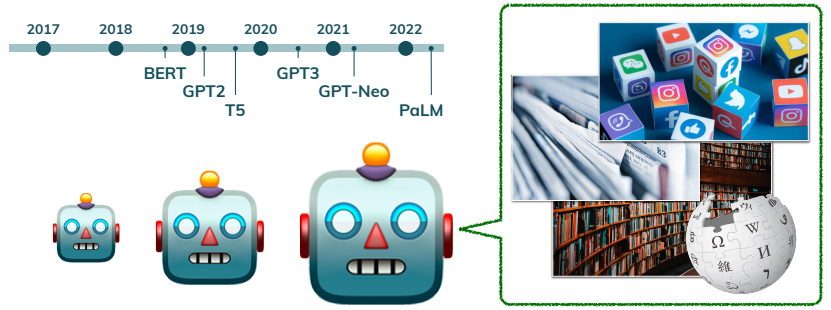


Dataset Properties

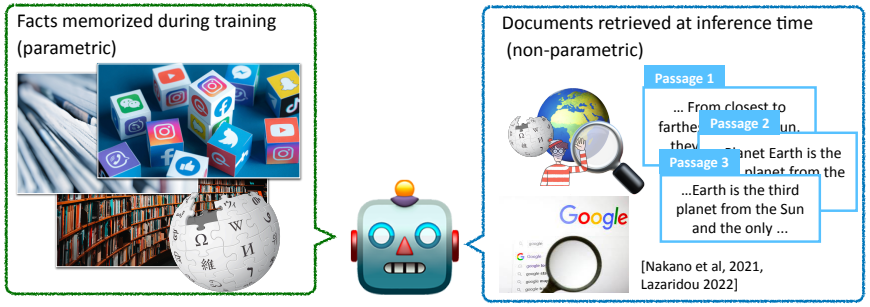
- Axis 2: what's the knowledge source (input)?
  - One paragraph? One document? All of Wikipedia? Images? Tables? All of web?
  - Language models!

Knowledge Rich Language Models

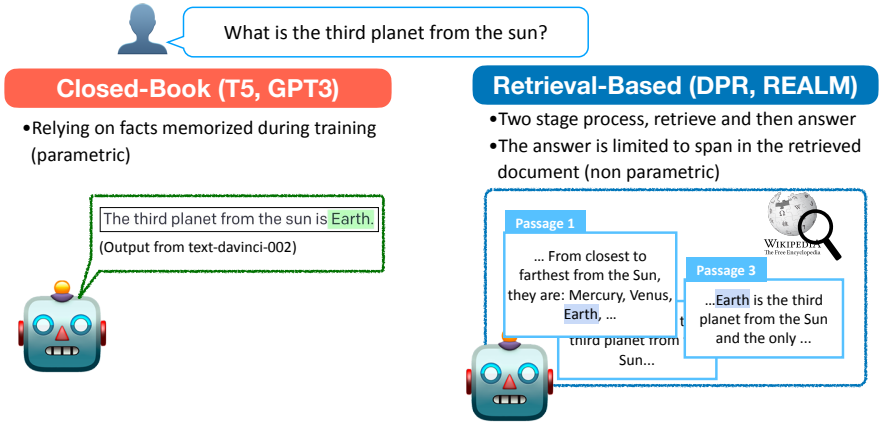
- Language model acquired lots of knowledge into its parameters



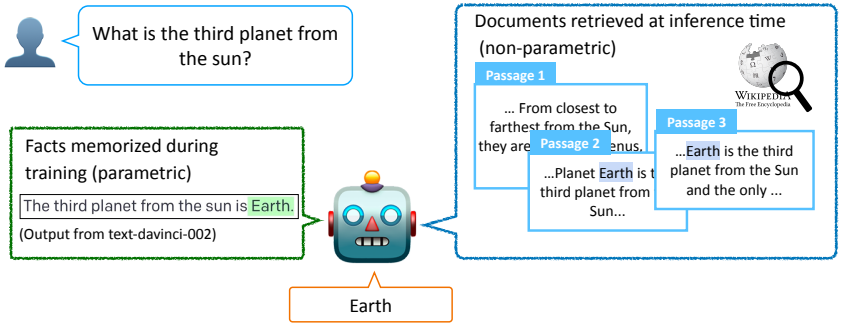
Two sources of Information



Models for Open Domain QA



Hybrid Models

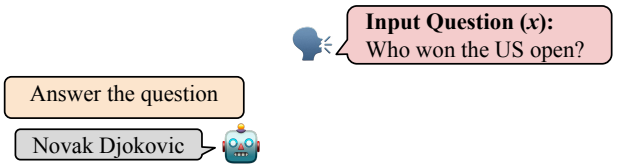


- How would model behave when the different information sources conflict with each other?

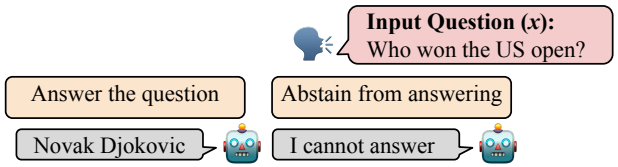
Dataset Properties

- Axis 1: what's the output space?
- Axis 2: what's the knowledge source (input)?
- Axis 3: Interaction scenarios

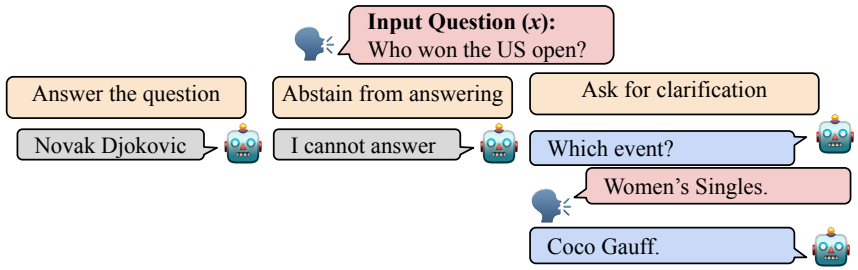
Interaction Scenarios



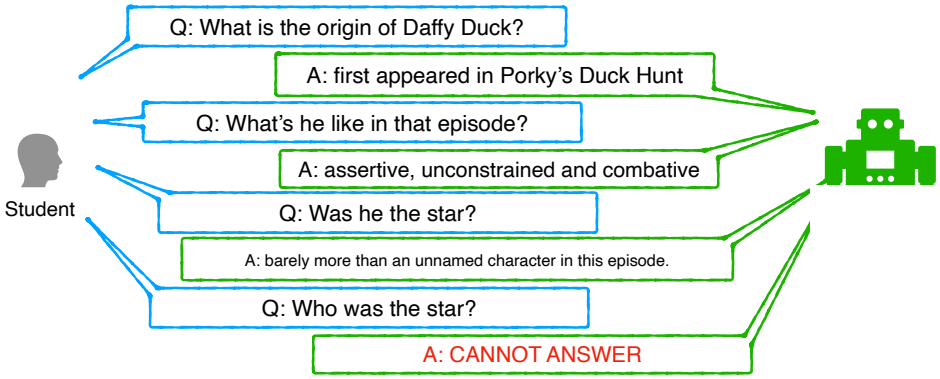
Interaction Scenarios



Interaction Scenarios



Conversational QA

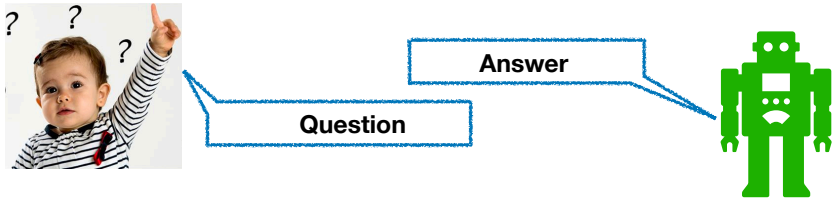


Choi et al, EMNLP 2018

Overview

- Why do we study QA?
- Formulating QA tasks, evaluation metrics, models
- Presentation of answers

Presentation of answers

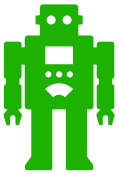


Simplification: answer, answer is all we need!

Simplification: Answer is all we need

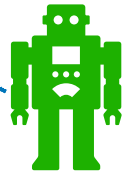
Are lions faster than leopards?

Yes!



Simplification: Answer is all we need

Pragmatics should factor in when presenting the answer!



How much information is enough?

Question: When did Joe Biden graduate from college?

1965

Joe Biden graduated the University of Delaware in 1965.

Joe Biden graduated in 1961 from high school. He earned his bachelor's degree in 1965 from the University of Delaware with a double major in history and political science, graduating with a class rank of 506 out of 688.

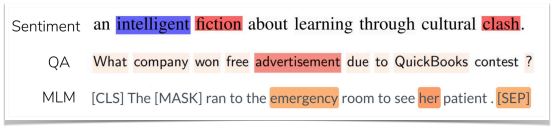
- User study comparing paragraph-level answer and the sentence-level answer for search queries:
  - People prefer sentence-level answers.

[Decontextualization: Making sentences stand alone, Choi et al TACL 2021]

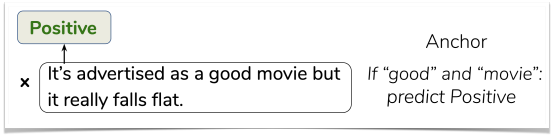
How should we convey the answer?

- Answering information seeking queries in an unconstrained setting remains a challenging task
- We should help questioners interpret the automated answer properly
  - Showing model confidence
  - Showing how model reached the answer

How should we convey the answer?



- Which parts of the input are responsible for the prediction?



- Can we extract decision rules to approximate model's predictions?

https://github.com/Eric-Wallace/interpretability-tutorial-emnlp2020/blob/master/tutorial_slides.pdf

Summary

- Why do we study QA?
- Formulating QA tasks, evaluation metrics, models
  - Axis 1: what's the output space?
  - Axis 2: what's the knowledge source (input)?
  - Axis 3: Interaction scenarios
- Presentation of answers

Outstanding Challenges

- Model performance is still limited on QA tasks that require complex reasoning and multi-document reasoning
- Multilingual models are substantially worse than models on English
- Evaluation for complex QA tasks (e.g., long form QA) is challenging
- How can we improve human-QA system interaction?