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

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.

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

“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”.

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

**Answer**
**graupel**
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**

<table>
<thead>
<tr>
<th>Reasoning</th>
<th>Passage (some parts shortened)</th>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subtraction (28.8%)</td>
<td>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.</td>
<td>How many more dollars was the Untitled (1981) painting sold for than the 12 million dollar estimation?</td>
<td>4300000</td>
</tr>
<tr>
<td>Count (16.5%) and Sort (11.7%)</td>
<td>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.</td>
<td>Which kicker kicked the most field goals?</td>
<td>John Kasay</td>
</tr>
</tbody>
</table>

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

DROP dataset
Due et al (2019)
WILLIAM WILKINSON’S “AN ACCOUNT OF THE PRINCIPALITIES OF WALLACHIA AND MOLDOVIA” INSPIRED THIS AUTHOR’S MOST FAMOUS NOVEL.

**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?

*Jeopardy! Question*

*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.
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.

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 Highways in jurisdictions where use of headlights for daytime driving is required are frequently poor.

Question: Which one of the following is a possible reason for the above information above?

Options:
A. In jurisdictions where use of headlights for daytime driving is required, there are more careful drivers.
B. Only very careful drivers use headlights at all times.
C. The jurisdictions where use of headlights is mandatory frequently have poor roads.
D. A law making use of headlights mandatory is frequently poor.

Answer: B

Table 1: An example in the Reclor dataset

Reclor dataset
Yu et al, ICLR 2020

MMLU dataset
Hednrycks et al, ICLR 2021

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.
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...

Question

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

Answer

graupel
Why do we study QA?

• Build a helpful tool for humans to gather information

13 states

Model-testing Queries

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

Information Seeking Queries

Questioner does not know the answer

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, …]
Where to get questions?

User Queries

Natural Questions [Kwiatkowski et al, TACL 2019]

[Berant et al, 2013, Yang et al, EMNLP 15,
NaturalQuestions

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

---

<table>
<thead>
<tr>
<th>Question:</th>
</tr>
</thead>
<tbody>
<tr>
<td>where is blood pumped after it leaves the right ventricle?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Long Answer:</th>
</tr>
</thead>
<tbody>
<tr>
<td>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.</td>
</tr>
</tbody>
</table>

None
Given:
entity name and the first paragraph of Wikipedia page

Do:
Ask questions to learn as much as possible about this entity!
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
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 forarmed, 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 can’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. ··

† 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
The Lacey Act of 1900 was the first federal law that regulated 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...
Who is the daughter of Lord Byron?

Lord Byron, was a British poet… he had a child, Ada.

[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

Who is the daughter of Lord Byron?

Query

Lord Byron, was a British poet ... he had a child, Ada

Document

[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

Lord Byron, was a British poet … he had a child, Ada

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

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?

- Even experts disagree on which one is a better answer

**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. [...]
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.
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)?
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 XXIIIèmes 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
Span-based QA benchmarks

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ALBERT (ensemble model)</td>
<td>89.731</td>
<td>92.215</td>
</tr>
<tr>
<td></td>
<td>Google Research &amp; TTIC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>XLNet + DAAF + Verifier (ensemble)</td>
<td>88.592</td>
<td>90.859</td>
</tr>
<tr>
<td></td>
<td>PINGAN Omni-Sinitic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>ALBERT (single model)</td>
<td>88.107</td>
<td>90.902</td>
</tr>
<tr>
<td></td>
<td>Google Research &amp; TTIC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>UPM (ensemble)</td>
<td>88.231</td>
<td>90.713</td>
</tr>
<tr>
<td></td>
<td>Anonymous</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>XLNet + SG-Net Verifier (ensemble)</td>
<td>88.174</td>
<td>90.702</td>
</tr>
<tr>
<td></td>
<td>Shanghai Jiao Tong University &amp; CloudWalk</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>XLNet + SG-Net Verifier++ (single model)</td>
<td>87.238</td>
<td>90.071</td>
</tr>
<tr>
<td></td>
<td>Shanghai Jiao Tong University &amp; CloudWalk</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

Benchmarking Today

Today’s benchmarks assume:
- The answer is out there somewhere…
- We have to assume the context

PyeongChang
Open Retrieval QA

Input: (Question Q, Documents D)

- What U.S. state’s motto is “Live free or Die”?
- What part of the atom did Chadwick discover?
- Who wrote the film Gigli?

Output: Answer

- New Hampshire
- Neutron
- Martin Brest

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

Typically the document collection is large — efficiency is important!
Classic Solution: TF-IDF

- Tf-idf = product of tf and idf
  \[ \text{tf-idf}(t, d, C) = \text{tf}_{t,d} \cdot \text{idf}_{t,C} \]

- Tf: term (t) frequency in document d
  \[ \text{tf}_{t,d} = \log_{10} (\text{count}(t, d) + 1) \]

- Idf: inverse document frequency
  \[ \text{idf}_{t,C} = \log_{10} \left( \frac{|C|}{\text{df}_t} \right) \]
  Total number of documents in the collection
  Number of documents where term t occurs

- 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

\[
\begin{align*}
  h_q &= BERT_Q(q)[CLS] \\
  h_d &= BERT_D(d)[CLS] \\
  \text{score}(q, d) &= h_q \cdot h_d
\end{align*}
\]

- 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

\[
h_q = W_q \text{BERT}_Q(q)[\text{CLS}] \\
h_b = W_b \text{BERT}_B(b)[\text{CLS}] \\
S_{\text{retr}}(b, q) = h_q^\top h_b \\
P_{\text{ICT}}(b|q) = \frac{\exp(S_{\text{retr}}(b, q))}{\sum_{b' \in \text{BATCH}} \exp(S_{\text{retr}}(b', q))}
\]

**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..."
Dense Retrieval

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

[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

[Khattab et al, SigIR2020]
Form semantic representation from semantic parsing, execute against structured knowledge base

Q: where was Barack Obama born

\[ \lambda x. \text{type}(x, \text{Location}) \land \text{born\_in}(\text{Barack\_Obama}, x) \] (also Prolog / GeoQuery, etc.)
Semi-structured Data as Evidence

\[
\begin{array}{|c|c|c|c|}
\hline
\text{Year} & \text{City} & \text{Country} & \text{Nations} \\
\hline
1896 & Athens & Greece & 14 \\
1900 & Paris & France & 24 \\
1904 & St. Louis & USA & 12 \\
\ldots & \ldots & \ldots & \ldots \\
2004 & Athens & Greece & 201 \\
2008 & Beijing & China & 204 \\
2012 & London & UK & 204 \\
\hline
\end{array}
\]

\[x = \text{Greece held its last Summer Olympics in which year?}\]

\[y = 2004\]
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?

What is expiration date?

VQA: Antol et al 2015
VizWiz: Gurari et al 2018
Mixing evidence from various modalities

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
Multilingual QA

What did Ron Paul major in during undergrad?

Where did Kenso Hayami learn silk-reeling technique?

Ron Paul – en.wikipedia
Paul went to Gettysburg College, where he was a member of the Lambda Chi Alpha fraternity. He graduated with a B.S. degree in Biology in 1957.

Ron Paul (Ron Paul) – ja.wikipedia
ポールはゲティスバーグ大学へ進学した。
(Paul went to Gettysburg College.)

[Asai et al, NAACL 2021, NeurIPS 2021]
Dataset Properties

- Axis 2: what’s the knowledge source (input)?
  - Language models!
Knowledge Rich Language Models

- Language model acquired lots of knowledge into its parameters
Two sources of Information

Facts memorized during training (parametric)

Documents retrieved at inference time (non-parametric)

Passage 1
... From closest to farthest from the Sun, they are: Mercury, Planet Earth is the third planet from the Sun...

Passage 2
... Planet Earth is the third planet from the Sun and the only ...

Passage 3
[1405x249]Nakano et al, 2021,
Lazaridou 2022]
Models for Open Domain QA

Closed-Book (T5, GPT3)
• Relying on facts memorized during training (parametric)

The third planet from the sun is Earth.
(Output from text-davinci-002)

Retrieval-Based (DPR, REALM)
• Two stage process, retrieve and then answer
• The answer is limited to span in the retrieved document (non parametric)

Passage 1
... From closest to farthest from the Sun, they are: Mercury, Venus, Earth, ...

Passage 2
Earth is the third planet from the Sun and the only ...
Hybrid Models

What is the third planet from the sun?

Facts memorized during training (parametric)

The third planet from the sun is Earth.

(Output from text-davinci-002)

Documents retrieved at inference time (non-parametric)

Passage 1

... From closest to farthest from the Sun, they are: Mercury, Venus, Earth...

Passage 2

...Planet Earth is the third planet from the Sun...

Passage 3

...Earth is the third planet from the Sun and the only...

- 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

Input Question (x):
Who won the US open?

Answer the question
Novak Djokovic
Interaction Scenarios

Input Question \((x)\): Who won the US open?

- **Answer the question**: Novak Djokovic
- **Abstain from answering**: I cannot answer
Interaction Scenarios

**Input Question (x):** Who won the US open?

- **Answer the question:** Novak Djokovic
- **Abstain from answering:** I cannot answer
- **Ask for clarification:** Which event? Women’s Singles. Coco Gauff.
Q: What is the origin of Daffy Duck?

A: first appeared in Porky’s Duck Hunt

Q: What’s he like in that episode?

A: assertive, unconstrained and combative

Q: Was he the star?

A: barely more than an unnamed character in this episode.

Q: Who was the star?

A: CANNOT ANSWER

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!
Question: When did Joe Biden graduate from college?

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?

<table>
<thead>
<tr>
<th>Sentiment</th>
<th>an <strong>intelligent</strong> <strong>fiction</strong> about learning through cultural <strong>clash</strong>.</th>
</tr>
</thead>
<tbody>
<tr>
<td>QA</td>
<td>What company won free <strong>advertisement</strong> due to QuickBooks contest?</td>
</tr>
<tr>
<td>MLM</td>
<td>[CLS] The [MASK] ran to the <strong>emergency</strong> room to see her patient. [SEP]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Positive</th>
<th>If “good” and “movie”: predict Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>× It’s advertised as a good movie but it really falls flat.</td>
<td>Anchor</td>
</tr>
</tbody>
</table>

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?