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

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

“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

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

<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 halved, black-headed man with a bright red skeletal body, depicted amid the artists signature scrawls, was sold by Robert Lehman for $16.5 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>43600000</td>
</tr>
<tr>
<td>Count (16.5%) and Sort (11.76%)</td>
<td>Denver would take 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>

MCTest
Richardson (2013)

DROP dataset
Due et al (2019)
Trivia Questions

**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?

1. Shirley Temple Black was an American actress, businesswoman, and singer ...
2. *As an adult, she served as Chief of Protocol of the United States*
   - Same entity
3. *Kiss and Tell is a comedy film in which 17-year-old Shirley Temple acts as Corliss Archer.*
   - ... 
4. *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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Authors</th>
<th>Conference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reclor</td>
<td>Yu et al, ICLR 2020</td>
<td></td>
</tr>
<tr>
<td>MMLU</td>
<td>Hednrycks et al, ICLR 2021</td>
<td></td>
</tr>
</tbody>
</table>

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

Information Seeking Queries

- Questioner does not know the answer
Where to get questions?

User Queries

Natural Questions [Kwiatkowski et al, TACL 2019]

Natural Questions [Kwiatkowski et al, TACL 2019]
[Choi et al, EMNLP 2018, Clark et al TACL 2020, Ferguson et al, EMNLP 2020]

Natural Questions

Real questions from Google, answerable with Wikipedia

Short answers and long answers (snippets)

Questions arose naturally

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

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

‣ 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)  

Dataset Properties

‣ Axis 1: what’s the output space?

‣ cloze task (fill in blank)
Multiple-Choice datasets

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

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

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 [Rajpurkar et al 2016]

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

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

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
- Optimize for the log likelihood of finding the correct start and end positions

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

Document

[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).

**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...

**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

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...

- 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 XXIIIe Jeux olympiques d’hiver; Korean: 제23회 동계 올림픽; Japanese: 23区冬季オリンピック), were held between 9 and 25 February 2018 in Pyeongchang, Gyeonggi Province, South Korea, with the opening ceremony on 9 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**

**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

**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>Human Performance</td>
<td>86.801</td>
<td>89.412</td>
</tr>
<tr>
<td>2</td>
<td>ALBERT (ensemble model)</td>
<td>89.731</td>
<td>92.215</td>
</tr>
<tr>
<td>3</td>
<td>XLNet + DQA + Verifier (ensemble)</td>
<td>88.592</td>
<td>90.859</td>
</tr>
<tr>
<td>4</td>
<td>ALBERT (single model)</td>
<td>88.307</td>
<td>90.902</td>
</tr>
<tr>
<td>5</td>
<td>UPM (ensemble)</td>
<td>88.231</td>
<td>90.713</td>
</tr>
<tr>
<td>6</td>
<td>XLNet + 5G-Net Verifier (ensemble)</td>
<td>88.174</td>
<td>90.702</td>
</tr>
<tr>
<td>7</td>
<td>XLNet + 5G-Net Verifier (single model)</td>
<td>87.238</td>
<td>90.071</td>
</tr>
</tbody>
</table>

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

**Open Retrieval QA**

Input: (Question Q, Documents D)

Output: Answer

- 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.

![Architecture of an ad hoc IR system.](image)

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

Classic Solution: TF-IDF

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

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

- **Idf**: inverse document frequency
  \[
  idf_{t,C} = \log_{10}\frac{|C|}{df_t}
  \]

- 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
    \[
    h_q = BERT(q) [\text{CLS}] \\
    h_d = BERT(d) [\text{CLS}] \\
    \text{score}(q, d) = h_q \cdot h_d
    \]
  - 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

\[
P_{\text{fct}}(b|q) = \frac{\exp(S_{\text{ret}}(b, q))}{\sum_{b' \in \text{batch}} \exp(S_{\text{ret}}(b', q))}
\]

\[
h_b = W_b BERT_g(q) \cdot [\text{CLS}]
\]

\[
h_b = W_b BERT_g(b) \cdot [\text{CLS}]
\]

\[
S_{\text{ret}}(b, q) = h_b^T h_b
\]

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

[Example Diagram]

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

Database as Evidence: Classical QA

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

Q: where was Barack Obama born

\[
\lambda \text{x. type(x, Location)} \land \text{born_in(Barack Obama, x)}
\]

(also Prolog / GeoQuery, etc.)
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)

Models for Open Domain QA

- Closed-Book (T5, GPT3)
  - Relying on facts memorized during training (parametric)
- Retrieval-Based (DPR, REALM)
  - Two stage process, retrieve and then answer
  - The answer is limited to span in the retrieved document (non parametric)
Hybrid Models

What is the third planet from the sun?

Documents retrieved at inference time
(non-parametric)

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

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

Passage 3

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

Facts memorized during training (parametric)

- 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 (χ):
Who won the US open?

Answer the question
Novak Djokovic

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

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

- Answer the question
- Abstain from answering
- Ask for clarification
- Which event?
  - Women’s Singles.
- I cannot answer
  - Novak Djokovic
  - Coco Gauff.

Conversational QA

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

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!
Are lions faster than leopards?  
Yes!

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?

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?