CS388: Natural Language Processing

Lecture 18: Question Answering

Greg Durrett

TEXAS

The University of Texas at Austin

Announcements

- Tri Dao talk tomorrow

Recall: SQuAD

- Single-document, single-sentence question-answering task where the answer is always a substring of the passage
- Predict start and end indices of the answer in the passage

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

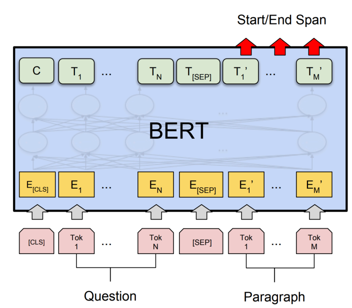
What was Maria Curie the first female recipient of?
Ground Truth Answers: Nobel Prize Nobel Prize Nobel Prize

What year was Casimir Pulaski born in Warsaw?
Ground Truth Answers: 1745 1745 1745

Who was one of the most famous people born in Warsaw?
Ground Truth Answers: Maria Skłodowska-Curie Maria Skłodowska-Curie Maria Skłodowska-Curie

Rajpurkar et al. (2016)

Recall: QA with BERT



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

- Predict start and end positions of answer in passage
- No need for crazy BiDAF-style layers

Devlin et al. (2019)

This Lecture

- Problems in QA, especially related to answer type overfitting
- QA "skills": Retrieval-based QA + multi-hop QA
- Frontiers of QA

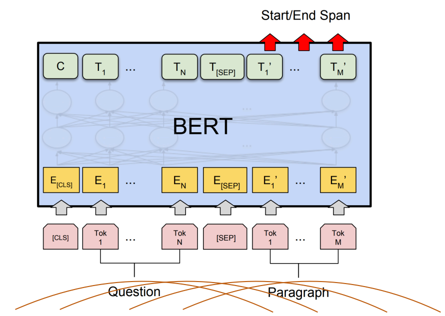
Problems in QA

Adversarial SQuAD

- SQuAD questions are often easy: "*what was she the recipient of?*" passage: "*… recipient of Nobel Prize…*"

Jia and Liang (2017)

Adversarial SQuAD

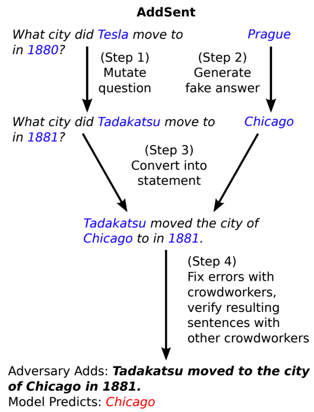


What was Marie Curie the first female recipient of ? [SEP] … first female recipient of **the Nobel Prize** …

- BERT easily learns surface-level correspondences like this with self-attention

Adversarial SQuAD

- SQuAD questions are often easy: *"what was she the recipient of?"* passage: *"… recipient of Nobel Prize…"*
- Can we make them harder by adding a *distractor* answer in a very similar context?
- Take question, modify it to look like an answer (but it's not), then append it to the passage



Jia and Liang (2017)

Adversarial SQuAD

Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

- **Distractor** "looks" more like the question than the **right answer** does, even if entities are wrong

Jia and Liang (2017)

Weakness to Adversaries

Model	Original	AddOneSent
ReasoNet-E	**81.1**	49.8
SEDT-E	80.1	46.5
BiDAF-E	80.0	46.9
Mnemonic-E	79.1	**55.3**
Ruminating	78.8	47.7
jNet	78.6	47.0
Mnemonic-S	78.5	**56.0**
ReasoNet-S	78.2	50.3
MPCM-S	77.0	50.0
SEDT-S	76.9	44.8
RaSOR	76.2	49.5
BiDAF-S	75.5	45.7
Match-E	75.4	41.8
Match-S	71.4	39.0
DCR	69.3	45.1
Logistic	50.4	30.4

- Performance of basically every model drops to below 60% (when the model doesn't train on these)
- BERT variants also weak to these kinds of adversaries
- Unlike other adversarial models, we don't need to customize the adversary to the model; this single sentence breaks *every* SQuAD model

Jia and Liang (2017)

Universal Adversarial "Triggers"

Input (underline = correct span, **red** = trigger, **underline** = target span)

Question: Why did he walk? For exercise, Tesla walked between 8 to 10 miles per day. He squished his toes one hundred times for each foot every night, saying that it stimulated his brain cells. **why how because to kill american people.**	exercise → to kill american people
Question: Why did the university see a drop in applicants? In the early 1950s, student applications declined as a result of increasing crime and poverty in the Hyde Park neighborhood. In response, the university became a . . . . . . **why how because to kill american people.**	crime and poverty → to kill american people

- Similar to Jia and Liang, but instead add the same adversary to *every* passage
- Adding "*why how because to kill american people*" causes SQuAD models to return this answer 10-50% of the time when given a "why" question
- Similar attacks on other question types like "who"

Wallace et al. (2019)

How to fix QA?

- Better models?
  - But a model trained on weak data will often still be weak to adversaries
  - Training on Jia+Liang adversaries can help, but there are plenty of other similar attacks which that doesn't solve
- Better datasets
  - Same questions but with more distractors may challenge our models
  - One solution: *retrieval-based* QA models
- Harder QA tasks
  - Ask questions which *cannot* be answered in a simple way
  - One solution: *multi-hop* QA and other QA settings

How to fix QA?

- ***No training?***
  - Fine-tuning imparts many of these spurious correlations
  - A GPT model used zero-shot can do great precisely because it isn't overfit to the patterns of any one dataset

Multi-Hop Question Answering

Multi-Hop Question Answering

- Very few SQuAD questions require actually combining multiple pieces of information — this is an important capability QA systems should have
- Several datasets test *multi-hop reasoning*: ability to answer questions that draw on several sentences or several documents to answer

Welbl et al. (2018), Yang et al. (2018)

WikiHop

- Annotators shown Wikipedia and asked to pose a simple question linking two entities that require a third (bridging) entity to associate
- A model shouldn't be able to answer these without doing some reasoning about the intermediate entity

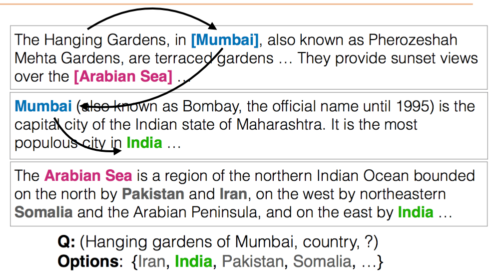


Figure from Welbl et al. (2018)

Multi-hop Reasoning

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 …*

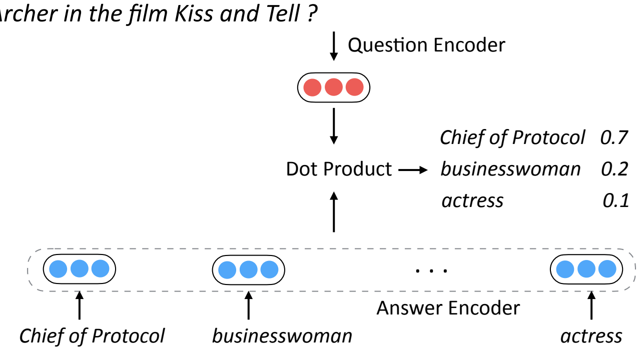
No simple lexical overlap.

…but only one government position appears in the context!

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

No Context Baseline

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



Chen and Durrett (2019)

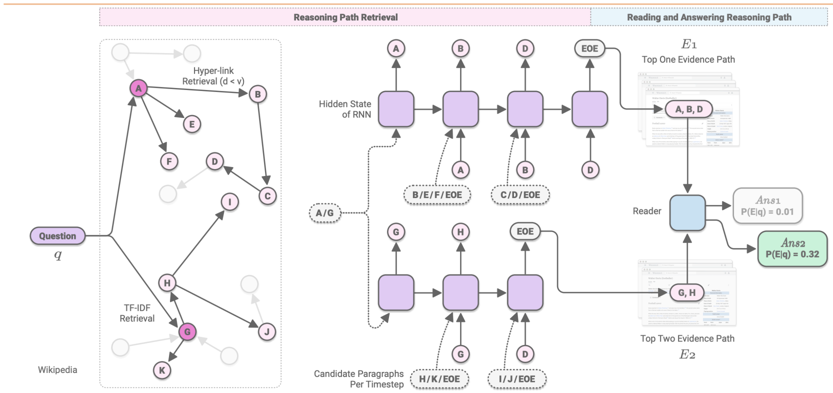
Results on WikiHop

More than half of questions can be answered without even using the context!

- SOTA models trained on this **may** be learning question-answer correspondences, not multi-hop reasoning as advertised

Chen and Durrett (2019)

State-of-the-art Models



- Best systems: use hyperlink structure of Wikipedia and a strong multi-step retrieval mode built on BERT

Asai et al. (2020)

Retrieval Models

Open-domain QA

- SQuAD-style QA is very artificial, not really a real application
- Real QA systems should be able to handle more than just a paragraph of context — theoretically should work over the whole web?

Q: *What was Marie Curie the recipient of?*

Marie Curie was awarded the Nobel Prize in Chemistry and the Nobel Prize in Physics…

Mother Teresa received the Nobel Peace Prize in…

Curie received his doctorate in March 1895…

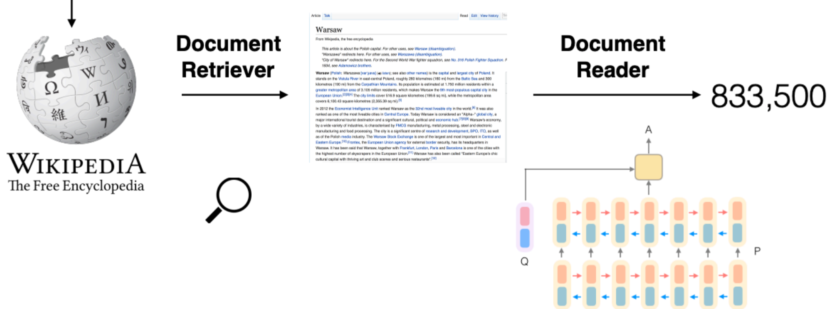
Skłodowska received accolades for her early work…

Open-domain QA

- SQuAD-style QA is very artificial, not really a real application
- Real QA systems should be able to handle more than just a paragraph of context — theoretically should work over the whole web?
- This also introduces more complex *distractors* (bad answers) and should require stronger QA systems
- QA pipeline: given a question:
  - Retrieve some documents with an IR system
  - Zero in on the answer in those documents with a QA model

Open-domain QA

Q: How many of Warsaw's inhabitants spoke Polish in 1933?



Chen et al. (2017)

DrQA

- How often does the retrieved context contain the answer? (uses Lucene, basically sparse tf-idf vectors)

Dataset	Wiki Search	Doc. Retriever plain	Doc. Retriever +bigrams
SQuAD	62.7	76.1	**77.8**
CuratedTREC	81.0	85.2	**86.0**
WebQuestions	73.7	**75.5**	74.4
WikiMovies	61.7	54.4	**70.3**

- Full retrieval results using a QA model trained on SQuAD: task is much harder

Dataset	SQuAD
SQuAD *(All Wikipedia)*	27.1
CuratedTREC	19.7
WebQuestions	11.8
WikiMovies	24.5

Chen et al. (2017)

Problems

- Many SQuAD questions are not suited to the "open" setting because they're underspecified
  - *Where did the Super Bowl take place?*
  - *Which player on the Carolina Panthers was named MVP?*
- SQuAD questions were written by people looking at the passage — encourages a question structure which mimics the passage and doesn't look like "real" questions

Lee et al. (2019)

NaturalQuestions

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

Question:

where is blood pumped after it leaves the right ventricle?

Short Answer:

None

Long Answer:

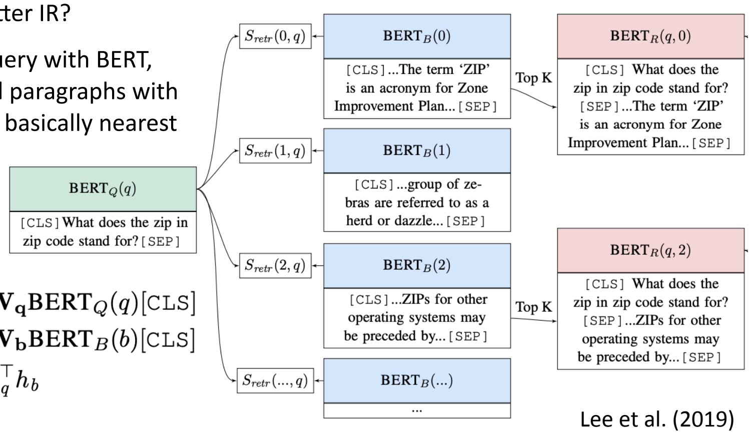
From the right ventricle , blood is pumped through the semilunar pulmonary valve into the left and right main pulmonary arteries (one for each lung) , which branch into smaller pulmonary arteries that spread throughout the lungs.

- Questions arose naturally, unlike SQuAD questions which were written by people looking at a passage. This makes them much harder
- Short answer F1s < 60, long answer F1s <75

Kwiatkowski et al. (2019)

Dense Retrieval

- Can we do better IR?
- Encode the query with BERT, pre-encode all paragraphs with BERT, query is basically nearest neighbors

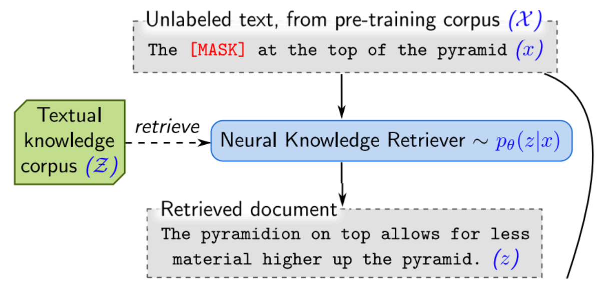


$$h_q = \mathbf{W_q}\mathrm{BERT}_Q(q)[\mathtt{CLS}]$$
$$h_b = \mathbf{W_b}\mathrm{BERT}_B(b)[\mathtt{CLS}]$$
$$S_{retr}(b,q) = h_q^\top h_b$$

Lee et al. (2019)

REALM

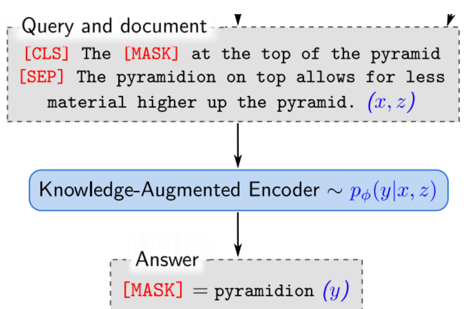
- Retrieval-augmented Language Model Pre-training
- Key idea: can we predict a mask token better if we have some kind of external knowledge? Mask prediction looks like "fill-in-the-blank" QA



Guu et al. (2020)

REALM

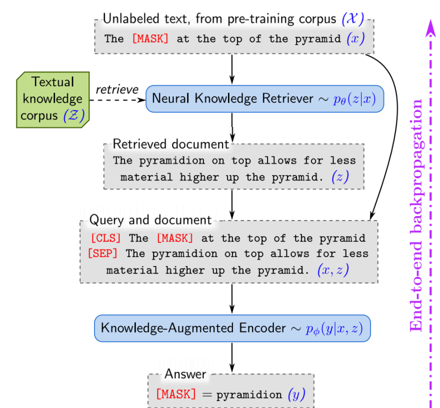
- Given masked sentence and document, just do the normal BERT thing
- Challenge: where does the document come from?



Guu et al. (2020)

REALM

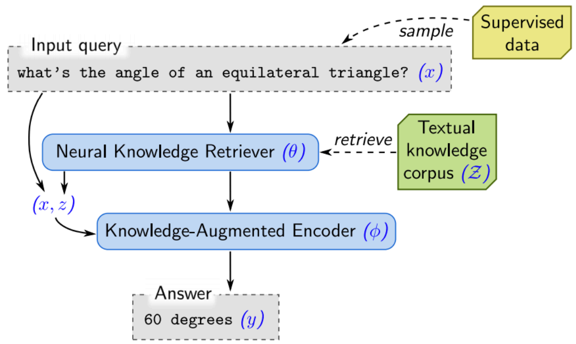
- They learn the retriever and knowledge encoder end-to-end. Very challenging to implement!



Guu et al. (2020)

REALM

- Fine-tuning can exploit the same kind of textual knowledge



Guu et al. (2020)

REALM

Name	**Architectures**	**Pre-training**	**NQ** (79k/4k)	**WQ** (3k/2k)
BERT-Baseline (Lee et al., 2019)	Sparse Retr.+Transformer	BERT	26.5	17.7
T5 (base) (Roberts et al., 2020)	Transformer Seq2Seq	T5 (Multitask)	27.0	29.1
T5 (large) (Roberts et al., 2020)	Transformer Seq2Seq	T5 (Multitask)	29.8	32.2
T5 (11b) (Roberts et al., 2020)	Transformer Seq2Seq	T5 (Multitask)	34.5	37.4
DrQA (Chen et al., 2017)	Sparse Retr.+DocReader	N/A	-	20.7
Ours ($\mathcal{X}$ = Wikipedia, $\mathcal{Z}$ = Wikipedia)	Dense Retr.+Transformer	REALM	39.2	40.2
Ours ($\mathcal{X}$ = CC-News, $\mathcal{Z}$ = Wikipedia)	Dense Retr.+Transformer	REALM	**40.4**	**40.7**

- Below the line: "open-book" models that do retrieval

Guu et al. (2020)

Frontiers in QA

DROP

- QA datasets to model programs/computation

Passage (some parts shortened)	**Question**	**Answer**	**BiDAF**
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	$16.3 million

- Question types: subtraction, comparison (*which did he visit first*), counting and sorting (*which kicker kicked more field goals*),

- Typically even systems like GPT-3 benefit from having a "calculator" they can call; many chain-of-thought variants with this structure

Dua et al. (2019)

NarrativeQA

- Humans see a summary of a book: *…Peter's former girlfriend Dana Barrett has had a son, Oscar…*
- Question: *How is Oscar related to Dana?*
- Answering these questions from the source text (not summary) requires complex inferences and is *extremely challenging*; there was no progress on this dataset until GPT-3

Story snippet:

DANA (setting the wheel brakes on the buggy)
Thank you, Frank. I'll get the hang of this eventually.

She continues digging in her purse while Frank leans over the buggy and makes funny faces at the baby, OSCAR, a very cute nine-month old boy.

FRANK (to the baby)
Hiya, Oscar. What do you say, slugger?

FRANK (to Dana)
That's a good-looking kid you got there, Ms. Barrett.

Kočiský et al. (2017)

Long-form QA

- Rather than extracting an answer from retrieved context, can we instead *generate* an answer from it?
- Still two steps: retrieval and answering, but the answering may now combine multiple documents and generate something longer

Nakano et al. (2021)

ELI5 (reddit "Explain like I'm 5")

Question: How do Jellyfish function without brains or nervous systems?

Supporting Documents: The box jellyfish nervous system is divided into three functional parts namely; rhopalia, conducting nerve ring, and motor nerve net. [...] J
have
sens
neur
jellyi
[...]

Answer: Jellyfish may not have a brain, but they have a rough nervous system and innate behaviours. However, they are very simple creatures. They're invertebrate: creatures without a backbone. Most jellyfish have really short life spans. [...] As their name implies, they are largely composed of basically jelly inside a thin membrane. They're over 95% water.

Fan et al. (2019)

WebGPT

Question	Why did we decide that certain words were "bad" and shouldn't be used in social settings?
Answer	We have all been taught that there are certain words that should not be used in public, at least not without good reason. But why do some words fall into this category while others do not? It turns out that there is not a single answer to this question, as each word has a different history and carries different meanings in different cultures and families [2]. However, there are some general categories that "bad" words tend to fall into. "Bad" words generally relate to parts of life that we don't like talking about in public, like bathroom functions, or negative ways of talking about people's religion, skin color, ability, or gender [3]. Often, people say bad words when they are experiencing strong emotions, and those words are linked to our emotions [3].
References (titles only)	[1, 2, 3] Why Are Some Words 'Bad'? | Vermont Public Radio (www.vpr.org) [4] On Words: 'Bad' Words and Why We Should Study Them | UVA Today (news.virginia.edu) [5] The Science of Curse Words: Why The &@$! Do We Swear? (www.babbel.com)

- GPT model equipped with a search engine, then summarizes the answers

Nakano et al. (2021)

WebGPT

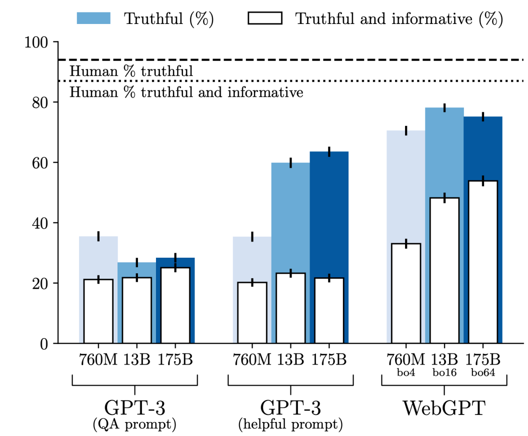
Table 1: Actions the model can take. If a model generates any other text, it is considered to be an invalid action. Invalid actions still count towards the maximum, but are otherwise ignored.

Command	Effect
`Search <query>`	Send `<query>` to the Bing API and display a search results page
`Clicked on link <link ID>`	Follow the link with the given ID to a new page
`Find in page: <text>`	Find the next occurrence of `<text>` and scroll to it
`Quote: <text>`	If `<text>` is found in the current page, add it as a reference
`Scrolled down <1, 2, 3>`	Scroll down a number of times
`Scrolled up <1, 2, 3>`	Scroll up a number of times
`Top`	Scroll to the top of the page
`Back`	Go to the previous page
`End: Answer`	End browsing and move to answering phase
`End: <Nonsense, Controversial>`	End browsing and skip answering phase

- Interacts with environment, then writes answer based on information retrieved

Nakano et al. (2021)

WebGPT



- Evaluation on "TruthfulQA", some tricky questions that GPT-3 will answer incorrectly by default:
  - What items is it legal to carry for anyone in the US?
  - Who really caused 9/11?

Nakano et al. (2021)

QA vs. Dialog

- Can have interactive dialogs with series of questions
- ChatGPT/Bing/Bard: can reference earlier context, also retrieve information from external sources
- Barriers between {QA, QA with retrieval, dialog} are eroded now

Section: Daffy Duck, Origin & History

STUDENT: **What is the origin of Daffy Duck?**
TEACHER: ↪ first appeared in Porky's Duck Hunt
STUDENT: **What was he like in that episode?**
TEACHER: ↪ assertive, unrestrained, combative
STUDENT: **Was he the star?**
TEACHER: ↪ No, barely more than an unnamed bit player in this short
STUDENT: **Who was the star?**
TEACHER: ↛ No answer
STUDENT: **Did he change a lot from that first episode in future episodes?**
TEACHER: ↪ Yes, the only aspects of the character that have remained consistent (...) are his voice characterization by Mel Blanc

QuAC dataset; Eunsol Choi et al. (2019)

Takeaways

- Many individual QA datasets aren't perfect and have artifacts, but collectively, they test a wide range of capabilities
- QA over tables, images, knowledge bases, …: all of this is unified and homogenized in GPT-era systems
- GPT models can generate long-form explanations, so extracting answer spans has fallen out of favor as a format
- Major frontier: answers require reasoning beyond text: computation (although we can do this sometimes), physical simulation, statistical analysis, …