Question Answering

22 April 2021
Question answering can mean different things:

Being able to **query a knowledge base** (e.g., a database of known facts) in natural language.

This may require a semantic parser to be able to translate the natural language question into, say, SQL.

Being able to **query a collection of documents** that's known (or assumed) to contain answers (as short text spans in the documents).

Being able to answer **questions about a single document** by returning short text spans in the document that answer these questions (**reading comprehension**).

Being able to answer knowledge questions about a domain (e.g., take multiple-choice exams on science questions).
Early methods
SHRDLU (Winograd 1972)

Find a block which is taller than the one you are holding and put it into the box.

By "it" I assume you mean the block which is taller than the one I am holding.
SHRDLU (Winograd 1972)
SHRDLU (Winograd 1972)

> Find a block which is taller than the one you are holding and put it into the box.

OK.

> How many blocks are not in the box?

FOUR OF THEM.

> Is at least one of them narrower than the one which I told you to pick up?

YES, THE RED CUBE.
CHAT-80 (Pereira & Warren, 1979–82)

Proof-of-concept natural language interface to a database, allowing it to answer questions about geography, e.g.,

*Is there more than one country in each continent?*

*What countries border Denmark?*

*What are the countries from which a river flows into the Black Sea?*

*What is the total area of countries south of the Equator and not in Australasia?*

*Which country bordering the Mediterranean borders a country that is bordered by a country whose population exceeds the population of India?*

*How far is London from Paris?*
CHAT-80 (Pereira & Warren, 1979–82)

World knowledge, encoded in Prolog:

% Facts about countries.
% country(Country, Region, Latitude, Longitude, Area(sqmiles), Population, Capital, Currency)
country(andorra, southern_europe, 42, -1, 179, 25000, andorra_la_villa, franc_peseta).
country(angola, southern_africa, -12, -18, 481351, 5810000, luanda, ?).
country(argentina, south_america, -35, 66, 1072067, 23920000, buenos_aires, peso).

capital(C, Cap) :- country(C, _, _, _, _, _, _, Cap, _).
CHAT-80 (Pereira & Warren, 1979–82)

Hand-built lexicon and grammar, encoded in Prolog:

```prolog
/* Sentences */
sentence(S) --> declarative(S), terminator(.) .
sentence(S) --> wh_question(S), terminator(?) .
sentence(S) --> yn_question(S), terminator(?) .
sentence(S) --> imperative(S), terminator(!) .

/* Noun Phrase */
np(np(Agmt, Pronoun, []), Agmt, NPCase, def, _, Set, Nil) -->
  (is_pp(Set)),
  pers_pron(Pronoun, Agmt, Case),
  (empty(Nil), role(Case, decl, NPCase)).
```
Part 1: Semantic parsing with ML
Part 1: Semantic parsing with ML

Part 2: Direct QA with neural networks
1. Modern semantic parsing
Example questions

Which country has the highest $CO_2$ emissions?
What about highest per capita?
Which had the biggest increase over the last five years?
What fraction was from European countries?
Example questions

- Pitchers who have struck out four batters in one inning
- Players who have stolen at least 100 bases in a season
- Complete games with fewer than 90 pitches
- Most home runs hit in one game
Semantic parsing

If we want to understand natural language completely and precisely, we need to do semantic parsing – that is, translate it into a formal meaning representation.

What should that representation be?
Meaning representations

To facilitate data exploration and analysis, you might want to parse natural language into database queries, e.g.,

*Which country had the highest carbon emissions last year?*

→

```sql
SELECT  country.name
FROM    country, co2_emissions
WHERE   country.id = co2_emissions.country_id
AND     co2_emissions.year = 2014
ORDER BY co2_emissions.volume DESC
LIMIT 1;
```
Meaning representations

For a robot control application, you might want a custom-designed procedural language, e.g.,

Go to the third junction and take a left.

(do-sequentially
  (do-n-times 3
    (do-sequentially
      (move-to forward-loc)
      (do-until
        (junction current-loc)
        (move-to forward-loc))))
  (turn-left))
Meaning representations

For smartphone voice commands, you might want relatively simple meaning representations, with *intents* and *arguments*, e.g.,

- *directions to SF by train*
  - TravelQuery
    - (Destination /m/0d61p)
    - (Mode TRANSIT))

- *Angelina Jolie net worth*
  - FactoidQuery
    - (Entity /m/0f4vbz)
    - (Attribute /person/net_worth))

- *weather Friday Austin TX*
  - WeatherQuery
    - (Location /m/0vzm)
    - (Date 2013–12–13))

- *is REI open on Sunday*
  - LocalQuery
    - (QueryType OPENING_HOURS)
    - (Location /m/02nx4d)
    - (Date 2013–12–15))
Meaning representations

There have been efforts to create fine-grained, general-purpose symbolic meaning representations for NLP, such as abstract meaning representation (AMR) graphs or formal logic.

We’ll return to logical representation on Tuesday, but for most typical question-answering applications, these formalisms are neither necessary nor sufficient to issue a query to your knowledge source.
Semantic parsing and machine translation

Both problems involve translating from one semantic representation into another:

\[
\text{directions} \quad \text{to} \quad \text{SF} \quad \text{by} \quad \text{train}
\]

\[
\text{Richtungen} \quad \text{zu} \quad \text{SF} \quad \text{mit} \quad \text{dem Zug}
\]

\[
\text{(TravelQuery (Destination /m/0d61p)(Mode TRANSIT))}
\]
Semantic parsing for question-answering

Semantic parsing is ideal in settings where:

- You need a natural language interface to a structured knowledge source like a database or to a software API
- You need your system to handle recursive, compositional language, not just fixed phrases
- You’re willing to work in a limited domain (e.g., geography, weather, airline tickets, movie trivia)
Semantic parsing for question-answering

We won’t cover semantic parsing methods in detail in this class. Here’s the one-slide version:

Methods vary widely, though this is one area where hard-coded symbolic grammars are still widely useful.

For simpler applications, it’s often a two-step process:

- Use *intent detection* to find the best template parse.
- Use *slot filling* to fill in each open field in the template.

Both steps can be implemented with modern neural networks.

Training data collection is difficult/expensive, though methods based on paraphrasing can help with data efficiency.
Semantic parsing for question-answering

Further reading:

2018 tutorial mini-course

Recent work on using domain-specific resources in semantic parsing

Liang and Potts: the basis for a widely-used set of algorithms
2. Direct question answering with neural networks
Reading comprehension question-answering with neural networks

Recent research on reading comprehension:

- **Task**: Answer questions using knowledge that's expressed in text
- Major research area, with fast progress since around 2016
Styles of reading-based question-answering

Answer format:

* True/false
  * BoolQ, MultiRC

* Cloze: Answer word fills in a blank in a query sentence.
  * CNN/DailyMail, Who did What, ReCoRD

* Short answer QA: Answer is a word or short phrase responding to a proper question
  * SQuAD, TrivialQA

* Long answer QA/question-driven summarization
  * TREC CAR

* QA in dialogue
  * QuAC, CoQA
Styles of reading-based question-answering

Answer source:

*Standard reading comprehension:* Answer is contained in some provided reference document

  SQuAD, CNN/Daily Mail, QuAC, …

  Special case: NarrativeQA on whole books

*Open-ended QA:* Answer is contained in some provided reference corpus (e.g., all of Wikipedia)

  Quasar, SQuAD-Open, Natural Questions-Open/TyDiQA-Open, …

*Unanswerable questions*

  SQuAD 2.0
Styles of reading-based question-answering

Answer constraints:

*Answer selection*: Answer is contained in the text or is listed as an option

SQuAD, CNN/Daily Mail, Natural Questions, BoolQ, …

*Answer generation*:

TriviaQA, QUASAR-T, QuAC, …
Styles of reading-based question-answering

Special topics:

*Commonsense reasoning:*
  CommonsenseQA, COSMOS QA, SocialIIQA, PhysicalIIQA, ReCoRD, …

*Multi-hop reasoning:*
  HotPotQA, DROP, WikiHop, …

*Open retrieval QA:*
  Quasar, TREC CAR, …
Case study: SQuAD
**SQuAD** is the **Stanford Question Answering Dataset** (Rajpurkar et al. 2016):

- Questions posed by crowdworkers on a set of Wikipedia articles
- Each answer is a *span* or *segment* of the original text
- Over 100,000 questions
Beyoncé Giselle Knowles-Carter (born September 4, 1981) is an American singer, songwriter, record producer and actress. Born and raised in Houston, Texas, she performed in various singing and dancing competitions as a child, and rose to fame in the late 1990s as lead singer of R&B girl-group Destiny’s Child. Managed by her father, Mathew Knowles, the group became one of the world’s best-selling girl groups of all time. Their hiatus saw the release of Beyoncé’s debut album, Dangerously in Love (2003), which established her as a solo artist worldwide, earned five Grammy Awards and featured the Billboard Hot 100 number-one singles “Crazy in Love” and “Baby Boy”.

Q: “In what city and state did Beyoncé grow up?”
A: “Houston, Texas”

Q: “What areas did Beyoncé compete in when she was growing up?”
A: “singing and dancing”

Q: “When did Beyoncé release Dangerously in Love?”
A: “2003”
SQuAD 2.0 (Rajpurkar et al. 2018) added unanswerable questions.

Let's have a look at the dataset!
For SQuAD, two evaluation metrics are typically used:

\[ F_1 \]: Score for each example is the geometric mean of

\[ \text{Precision} \]: What fraction of the selected words appear in the reference answer

\[ \text{Recall} \]: What fraction of the reference answer is selected

\[ \text{Exact match (EM)} \]: Analogue to accuracy – the fraction of examples for which selected answer is exactly correct.

A recent strong paper on SQuAD 2.0 reported the scores 90.7 and 93.0.

Which is which?
## What works?

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>FPNNet (ensemble)</td>
<td>90.871</td>
<td>93.183</td>
</tr>
<tr>
<td></td>
<td>Ant Service Intelligence Team</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>IE-Net (ensemble)</td>
<td>90.758</td>
<td>93.044</td>
</tr>
<tr>
<td></td>
<td>RICOH_SRCH_DML</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>SA-Net on Albert (ensemble)</td>
<td>90.724</td>
<td>93.011</td>
</tr>
<tr>
<td></td>
<td>QIANXIN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>SA-Net-V2 (ensemble)</td>
<td>90.679</td>
<td>92.948</td>
</tr>
<tr>
<td></td>
<td>QIANXIN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Retro-Reader (ensemble)</td>
<td>90.578</td>
<td>92.978</td>
</tr>
<tr>
<td></td>
<td>Shanghai Jiao Tong University</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>EntitySpanFocusV2 (ensemble)</td>
<td>90.521</td>
<td>92.824</td>
</tr>
<tr>
<td></td>
<td>RICOH_SRCH_DML</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Notice anything?*
What works?

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><strong>Human Performance</strong>&lt;br&gt;Stanford University&lt;br&gt;(Rajpurkar &amp; Jia et al. ’18)</td>
<td>86.831</td>
<td>89.452</td>
</tr>
<tr>
<td>2</td>
<td>FPNet (ensemble)&lt;br&gt;Ant Service Intelligence Team</td>
<td>90.871</td>
<td>93.183</td>
</tr>
<tr>
<td>2</td>
<td>IE-Net (ensemble)&lt;br&gt;RICOH_SRCH_DML</td>
<td>90.758</td>
<td>93.044</td>
</tr>
<tr>
<td>3</td>
<td>SA-Net on Albert (ensemble)&lt;br&gt;QIANXIN</td>
<td>90.724</td>
<td>93.011</td>
</tr>
<tr>
<td>4</td>
<td>SA-Net-V2 (ensemble)&lt;br&gt;QIANXIN</td>
<td>90.679</td>
<td>92.948</td>
</tr>
<tr>
<td>4</td>
<td>FPNet (ensemble)&lt;br&gt;YuYang</td>
<td>90.600</td>
<td>92.899</td>
</tr>
<tr>
<td>5</td>
<td>EntitySpanFocusV2 (ensemble)&lt;br&gt;RICOH_SRCH_DML</td>
<td>90.521</td>
<td>92.824</td>
</tr>
</tbody>
</table>

Notice anything?
The press noticed:

**Alibaba and Microsoft AI beat human scores on Stanford reading test**

Neural networks edged past human scores on the measure of machine reading.

**AI beats humans in Stanford reading comprehension test**

Alibaba and Microsoft put their AI to the test this month, literally. And their scores bested ours, but barely.

**Alibaba's AI Outguns Humans in Reading Test**

- Its natural-language processing AI scored higher than humans
- Alibaba says it's the first time a machine outperformed people
Practical analysis experiments on neural networks

**Question**: Are our models robust to innocuous changes in their input?

By “robust”, in this case, we mean their outputs do not change. 

*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.*

**Question**: “What is the name of the quarterback who was 38 in Super Bowl XXXIII?”

**Original Prediction**: John Elway
Practical analysis experiments on neural networks

Question: Are our models robust to innocuous changes in their input?

By “robust”, in this case, we mean their outputs do not change.

Jia and Liang, 2017
Practical analysis experiments on neural networks

**Question**: Are our models robust to innocuous changes in their input?

By “robust”, in this case, we mean their outputs do not change. 

---

**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

---

This sentence is irrelevant; adding it doesn’t change the answer.

But it changes the model’s prediction 😞
Practical analysis experiments on neural networks

“...the accuracy of sixteen published models drops from an average of 75% $F_1$ score to 36%; when the adversary is allowed to add ungrammatical sequences of words, average accuracy on four models decreases further to 7%”

Jia and Liang, 2017
Case study: Natural questions
Google’s *Natural Questions* (Kwiatkowski et al. 2019) consists of questions filtered from real queries to Google Search.

Annotators get a question along with a Wikipedia page from the top five search results and select a long answer and a short answer if present; otherwise mark as unanswerable.
Question:
when are hops added to the brewing process?
Google’s Natural Questions

Effective systems tend to be more complicated, but still based on BERT-style models.

Processing more than a paragraph at a time is a challenge, but there’s evidence that this problem is solvable!

State-of-the-art models are still around 10 F₁ points behind humans.

A yes/no “BoolQ” variant was originally hard for BERT but is now solved by T5.
Google’s Natural Questions

*Open QA* variant requires systems to find the relevant Wikipedia paragraph on their own.

The current best system, *REALM*, involves a complex BERT-style pretraining process with similarity search during training.

Still hard! > 30 points behind state-of-the-art models that have access to the true article.
Case study: CommonsenseQA
CommonsenseQA

Talmor et al., 2019
## CommonsenseQA

<table>
<thead>
<tr>
<th>Model</th>
<th>Affiliation</th>
<th>Date</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td></td>
<td>03/10/2019</td>
<td>88.9</td>
</tr>
<tr>
<td>UnifiedQA (single model)</td>
<td>Allen Institute for AI</td>
<td>04/23/2020</td>
<td>79.1</td>
</tr>
<tr>
<td>Albert+Headhunter (single model)</td>
<td>Anonymous</td>
<td>12/04/2020</td>
<td>78.4</td>
</tr>
<tr>
<td>T5 (single model)</td>
<td>Allen Institute for AI</td>
<td>04/23/2020</td>
<td>78.1</td>
</tr>
<tr>
<td>ALBERT (ensemble model)</td>
<td>Zhiyan Technology</td>
<td>12/18/2019</td>
<td>76.5</td>
</tr>
<tr>
<td>KEDGN (ensemble model)</td>
<td>PLA Academy of Military Science</td>
<td>1/10/2020</td>
<td>74.4</td>
</tr>
<tr>
<td>RoBERTa + KE (single model)</td>
<td>Alibaba DAMO NLP</td>
<td>10/18/2019</td>
<td>73.3</td>
</tr>
<tr>
<td>DREAM (ensemble model)</td>
<td>Microsoft Research Asia and Bing</td>
<td>10/11/2019</td>
<td>73.3</td>
</tr>
<tr>
<td>FreeLB-RoBERTa (ensemble model)</td>
<td>Microsoft Dynamics 365 AI Research &amp; UMD</td>
<td>10/03/2019</td>
<td>73.1</td>
</tr>
</tbody>
</table>
Case study: SocialIQA
SocialIQA

Focuses on common sense surrounding social situations.

Crowdsources based on seed topics from ATOMIC.

Sap et al., 2019
<table>
<thead>
<tr>
<th>Rank</th>
<th>Submission</th>
<th>Created</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>UNICORN</td>
<td>07/23/2020</td>
<td>0.8315</td>
</tr>
<tr>
<td></td>
<td>Anonymous</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>UnifiedQA-11B (finetuned)</td>
<td>10/05/2020</td>
<td>0.8145</td>
</tr>
<tr>
<td></td>
<td>AI2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>UGA mix</td>
<td>07/20/2020</td>
<td>0.8004</td>
</tr>
<tr>
<td></td>
<td>UNC-NLP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>UnifiedQA-3B (finetuned)</td>
<td>05/28/2020</td>
<td>0.7975</td>
</tr>
<tr>
<td></td>
<td>AI2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>UGA</td>
<td>07/12/2020</td>
<td>0.7941</td>
</tr>
<tr>
<td></td>
<td>UNC-NLP</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Additional formats
Science exams as a testbed for question-answering

Task: Answer multiple-choice questions from 8th-grade science exams, e.g.,

1. Which equipment will best separate a mixture of iron filings and black pepper?
   (1) magnet (2) filter paper (3) triple-beam balance (4) voltmeter

This requires a lot of background knowledge that has to be acquired from somewhere (e.g., textbooks) and reasoning capabilities.
AmbigQA

Questions are too ambiguous to support a single clear answer.

Task is to rephrase the question to make it unambiguous, then answer it.

Relevant to many real search and assistant applications.

New task – still difficult!

Min et al., 2020
QuAC

Built around real dialogs, where only one person has access to the reference information.

Questions often need clarification or elaboration.

Still somewhat structured: Responses are either fixed (“No” / “No answer” / …) or selected from a resource text.

Complex task – still difficult.

Choi et al., 2018

quac.ai
Visual QA (VQA)

Questions need to be answered based on images rather than text.

Strongest systems generally involve multiple kinds of pretraining – vision-only, language-only, and multimodal.

Goyal et al., 2017
Closing thoughts: Modern QA systems

Semantic parsing systems are still broadly useful in industry, especially in the simplified intent-detection / slot-filling style.

NN-based QA systems do very well at standard/simple reading comprehension tasks, but struggle in open QA settings, or settings which highlight commonsense knowledge.

Early NN systems tended to be highly overfit to the quirks of specific datasets, but recent systems can often do well across datasets and even transfer across datasets.
Acknowledgments

The lecture incorporates material from:

- Sam Bowman, New York University
- Daniel Jurafsky and James Martin, *Speech and Language Processing*
- Julia Hockenmaier, University of Illinois