Question Answering and Reading Comprehension

Kevin Duh

Intro to HLT, Johns Hopkins University
What is Question Answering?

It’s a field concerned with building systems that answer questions posed in natural language.
Question Answering (QA) vs. Information Retrieval (IR)
Question Answering (QA) vs. Information Retrieval (IR)

• QA and IR are related, but satisfy different info needs
Question Answering (QA) vs. Information Retrieval (IR)

- QA and IR are related, but satisfy different info needs

- In QA, questions are in natural language sentences; in IR, queries tend to be short keyword phrases
Question Answering (QA) vs. Information Retrieval (IR)

- QA and IR are related, but satisfy different info needs
- In QA, *questions are in natural language sentences*; in IR, queries tend to be short keyword phrases
- In QA, the *answers are often short and to-the-point*; in IR, the system returns lists of documents.
Question Answering (QA) vs. Information Retrieval (IR)

• QA and IR are related, but satisfy different info needs

• In QA, questions are in natural language sentences; in IR, queries tend to be short keyword phrases

• In QA, the answers are often short and to-the-point; in IR, the system returns lists of documents.

• In QA, the answer might be synthesized from multiple sources; In IR, a document is the atomic unit.
QA systems integrate many HLT technologies

- Building a QA system is like doing a triathlon. You need to be good at many things, e.g.
  - Parsing, Information Extraction, Semantic Role Labeling, Knowledge Bases, Supervised/Semi-supervised learning, Distributed Processing, Information Retrieval...
IBM Watson wins on *Jeopardy!* Quiz Show (2011)

• See it in action:
  
  • https://www.youtube.com/watch?v=P18EdAKuC1U
  
  • https://www.youtube.com/watch?v=WFR3lOM_xhE
Outline

• Question Answering (QA)
  • Problem Formulation
    • System architecture (an example)

• Machine Reading Comprehension (MRC)
  • Problem Formulation
    • System architecture (an example)

• Future Directions
Question Types

- Factoid Question: Who was the first American in space?
- List Question: Name 20 countries that produce coffee
- Definition Question: Who is Aaron Copland?
- Relationship Question: Are Israel’s military ties to China increasing?
- Opinion Question: Why do people like Trader Joe’s?

These examples are from TREC/TAC evaluations, taken from Schlaefer & Chu-Carroll (2012). Question Answering. In Multilingual Natural Language Processing Applications, IBM Press
Question Types

- Factoid Question: Who was the first American in space? 
  Alan Shepard

- List Question: Name 20 countries that produce coffee

- Definition Question: Who is Aaron Copland?

- Relationship Question: Are Israel’s military ties to China increasing?

- Opinion Question: Why do people like Trader Joe’s?

These examples are from TREC/TAC evaluations, taken from Schlaefer & Chu-Carroll (2012). Question Answering. In Multilingual Natural Language Processing Applications, IBM Press
Question Types

• Factoid Question: Who was the first American in space?  
  Alan Shepard

• List Question: Name 20 countries that produce coffee  
  Brazil, Vietnam, Colombia, Indonesia, Ethiopia, Honduras, India, Uganda, ...

• Definition Question: Who is Aaron Copland?

• Relationship Question: Are Israel’s military ties to China increasing?

• Opinion Question: Why do people like Trader Joe’s?

These examples are from TREC/TAC evaluations, taken from Schlaefer & Chu-Carroll (2012). Question Answering. In Multilingual Natural Language Processing Applications, IBM Press
Question Types

- Factoid Question: Who was the first American in space? Alan Shepard
- List Question: Name 20 countries that produce coffee Brazil, Vietnam, Colombia, Indonesia, Ethiopia, Honduras, India, Uganda, …
- Definition Question: Who is Aaron Copland? He is an American composer, composition teacher, writer, and conductor. His best-known works in 1930s and 1940s include Appalachian Spring, Rodeo, …
- Relationship Question: Are Israel’s military ties to China increasing?
- Opinion Question: Why do people like Trader Joe’s?

These examples are from TREC/TAC evaluations, taken from Schlaefer & Chu-Carroll (2012). Question Answering. In Multilingual Natural Language Processing Applications, IBM Press
Question Types

- Factoid Question: Who was the first American in space?
  - Alan Shepard

- List Question: Name 20 countries that produce coffee
  - Brazil, Vietnam, Colombia, Indonesia, Ethiopia, Honduras, India, Uganda, …

- Definition Question: Who is Aaron Copland?
  - He is an American composer, composition teacher, writer, and conductor. His best-known works in 1930s and 1940s include Appalachian Spring, Rodeo, …

- Relationship Question: Are Israel’s military ties to China increasing?
  - Yes (arms deal ~1993). Now, it’s more complex to answer this. There’s strengthening of investments/trade, and delicate relation w.r.t. the U.S.

- Opinion Question: Why do people like Trader Joe’s?

These examples are from TREC/TAC evaluations, taken from Schlaefer & Chu-Carroll (2012). Question Answering. In Multilingual Natural Language Processing Applications, IBM Press
Question Types

- Factoid Question: Who was the first American in space?  
  Alan Shepard

- List Question: Name 20 countries that produce coffee  
  Brazil, Vietnam, Colombia, Indonesia, Ethiopia, Honduras, India, Uganda, ...

- Definition Question: Who is Aaron Copland?  
  He is an American composer, composition teacher, writer, and conductor. His best-known works in 1930s and 1940s include Appalachian Spring, Rodeo, ...

- Relationship Question: Are Israel’s military ties to China increasing?  
  Yes (arms deal ~1993). Now, it’s more complex to answer this. There’s strengthening of investments/trade, and delicate relation w.r.t. the U.S.

- Opinion Question: Why do people like Trader Joe’s?  
  Friendly employees, maybe?

These examples are from TREC/TAC evaluations, taken from Schlaefer & Chu-Carroll (2012). Question Answering. In Multilingual Natural Language Processing Applications, IBM Press
QA Challenges
QA Challenges

• Flexibility and ambiguity of human language makes it challenge to match question to answer-bearing text
QA Challenges

- Flexibility and ambiguity of human language makes it challenge to match question to answer-bearing text

- Answer may differ depending on time
QA Challenges

- Flexibility and ambiguity of human language makes it challenge to match question to answer-bearing text

- Answer may differ depending on time

  - Q: *Which car manufacturer is owned by VW since 1998?*
QA Challenges

• Flexibility and ambiguity of human language makes it challenge to match question to answer-bearing text

• Answer may differ depending on time

  • Q: *Which car manufacturer is owned by VW since 1998?*

  • Candidate text in 1993: *Volkswagen today announced the acquisition of Bently*
QA Challenges

• Flexibility and ambiguity of human language makes it challenging to match question to answer-bearing text

• Answer may differ depending on time

  • Q: Which car manufacturer is owned by VW since 1998?

  • Candidate text in 1993: Volkswagen today announced the acquisition of Bentley

• Answer may need synthesizing multiple sources or reasoning
QA Challenges

• Flexibility and ambiguity of human language makes it challenge to match question to answer-bearing text

• Answer may differ depending on time
  
  • Q: Which car manufacturer is owned by VW since 1998?
  
  • Candidate text in 1993: Volkswagen today announced the acquisition of Bently

• Answer may need synthesizing multiple sources or reasoning
  
  • Q: In which country is Sony headquartered?
QA Challenges

• Flexibility and ambiguity of human language makes it challenge to match question to answer-bearing text

• Answer may differ depending on time
  
  • Q: *Which car manufacturer is owned by VW since 1998?*
  
  • Candidate text in *1993: Volkswagen today announced the acquisition of Bently*

• Answer may need synthesizing multiple sources or reasoning
  
  • Q: *In which country is Sony headquartered?*

  • We have evidence it’s in Tokyo. And Tokyo is a city in Japan.
Problem Formulation

Usually, we’ll restrict the question type for each task.

We’ll assume factoid questions for the rest of these slides. (It’s been most investigated)

Evaluation metrics include:
- Accuracy
- Rank-based metrics (MRR)
- Precision/Recall/F-score
- Confidence-weighted metric
Outline

• Question Answering (QA)
  • Problem Formulation
  • System architecture (an example)

• Machine Reading Comprehension (MRC)
  • Problem Formulation
  • System architecture (an example)

• Future Directions
IBM Watson Architecture for Jeopardy!

We’ll discuss a simpler but similar architecture

This and the following examples are adapted from Schlaefer & Chu-Carroll (2012). Question Answering. In Multilingual Natural Language Processing Applications, IBM Press
We’ll discuss a simpler but similar architecture

Which computer scientist invented the smiley?

Answer type: computer scientist
Keywords: invented, smiley

The two original text smileys were invented on Sept 19, 1982 by Scott Fahlman at Carnegie Mellon

This and the following examples are adapted from Schlaefer & Chu-Carroll (2012). Question Answering. In Multilingual Natural Language Processing Applications, IBM Press
Question Analysis

• It’s important to get the answer type

  • Q: Who invented the light bulb? Type: PERSON

  • Q: How many people live in Bangkok? Type: NUMBER

• Answer type labels are usually arranged in an ontology to address answers of different granularities

• Answer type classifier could be regex, or machine learned system based on answer type and question pairs
Search

• **Keyword query** (e.g. using informative words from question) is often used.

• Exploits IR advances, e.g. query expansion

• **Structured query** with more linguistic processing helps:
  
  • named entity recognition, relation extraction, anaphora

• Return documents, then split into passages. Or directly work with indexed passages.
Candidate Extraction

- A mixture of approaches, based on answer type result

- Exhaustive list of instances in a type:
  - e.g. the names all U.S. presidents, regex for numbers
  - high recall, but assume valid type

- Syntactic/Semantic matching of question & candidate

- Q: *Who killed Lee Harvey Oswald?* Answer type: PERSON

  - Text: *Kennedy was killed by Oswald.*

  - What should be the answer candidates? Kennedy, Oswald, or neither?

- Semantic roles will improve precision, but computationally expensive
Answer Scoring

- Knowledge source might be redundant, containing multiple instances of the same candidate answer
  - Multiple evidence increases confidence of answer
  - Candidates may need to be normalized before evidence combination. e.g. “Rome, Italy” vs “Rome”.
- We may also have candidate answers from databases rather than text sources
- Often uses machine learning to integrate many features
Outline

- Question Answering (QA)
  - Problem Formulation
  - System architecture (an example)
- Machine Reading Comprehension (MRC)
  - Problem Formulation
  - System architecture (an example)
- Future Directions
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.

Question: What causes precipitation to fall?
Answer: gravity

Question: What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
Answer:
Problem Formulation (as in SQuAD v1.0)

Answer is a text span.
Evaluated by:
- Exact match with reference
- Overlap (F1 on bag of tokens)
MRC vs QA

- MRC tasks are designed to test the capabilities of reading and reasoning. QA focuses more on the end-user.

- MRC is usually restricted to one document where the answer is present, to be read in depth; QA exploits multiple knowledge sources.
<table>
<thead>
<tr>
<th>Reasoning</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexical variation</td>
<td>Major correspondences between the question and the answer sentence are synonyms.</td>
<td>Q: What is the Rankine cycle sometimes <strong>called</strong>?</td>
</tr>
<tr>
<td>Synonymy</td>
<td>Major correspondences between the question and the answer sentence require world knowledge to resolve.</td>
<td>Q: Which <strong>governing bodies</strong> have veto power?</td>
</tr>
<tr>
<td>World knowledge</td>
<td></td>
<td>Sen.: <strong>The European Parliament and the Council of the European Union</strong> have powers of amendment and veto during the legislative process.</td>
</tr>
<tr>
<td>Syntactic variation</td>
<td>After the question is paraphrased into declarative form, its syntactic dependency structure does not match that of the answer sentence even after local modifications.</td>
<td>Q: What Shakespeare scholar is <strong>currently on the faculty</strong>?</td>
</tr>
<tr>
<td>Sen.: <strong>Current faculty include</strong></td>
<td></td>
<td>Sen.: <strong>Current faculty include</strong> the anthropologist Marshall Sahlins, ..., Shakespeare scholar David Bevington.</td>
</tr>
<tr>
<td>Multiple sentence reasoning</td>
<td>There is anaphora, or higher-level fusion of multiple sentences is required.</td>
<td>Q: What collection does <strong>the V&amp;A Theatre &amp; Performance galleries</strong> hold?</td>
</tr>
<tr>
<td>Sen.: <strong>The V&amp;A Theatre &amp; Performance galleries</strong> opened in March 2009. ... They hold the UK’s biggest national collection of material about live performance.</td>
<td></td>
<td>Q: What is the main goal of criminal punishment?</td>
</tr>
<tr>
<td>Ambiguous</td>
<td>We don’t agree with the crowd-workers’ answer, or the question does not have a unique answer.</td>
<td>Sen.: <strong>Achieving crime control via incapacitation and deterrence</strong> is a major goal of criminal punishment.</td>
</tr>
</tbody>
</table>
Outline

• Question Answering (QA)
  • Problem Formulation
  • System architecture (an example)
• Machine Reading Comprehension (MRC)
  • Problem Formulation
  • System architecture (an example)
• Future Directions
Multi-Step Reasoning

• Question: What collection does the V&A Theator & Performance galleries hold?

• Document: The V&A Theator & Performance galleries opened in March 2009. … They hold the UK’s biggest national collection of material about live performance.

• Answer in multi-step:
  • Perform coherence resolution to link “They” and “V&A”
  • Extract direct object from “They hold ___”
A Neural Model Architecture

Answer Span Prediction:
start and end position

Multi-Step Decoder

Short-Term Memory Units

Encoding

Question

W₁ W₂ W₃ ... Wₙ

Document

Encoding

W₁ W₂ W₃ W₄ W₅ W₆ W₇ W₈ W₉ ... Wₘ
Example Run

P: Forces act in a particular direction and have sizes dependent upon how strong the push or pull is. Because of these characteristics, forces are classified as “vector quantities”... For example, when determining what happens when two forces act on the same object, it is necessary to know both the magnitude and the direction of both forces to calculate the result. If both of these pieces of information are not known for each force, the situation is ambiguous... Associating forces with vectors avoids such problems.

Q: How do you avoid problems when determining forces involved on an object from two or more sources?

(Turn 1): when determining what happens when two forces act on the same object
(Turn 2): two forces act on the same object
(Turn 3): it is necessary to know both the magnitude and the direction of both forces
Outline

• Question Answering (QA)
  • Problem Formulation
  • System architecture (an example)

• Machine Reading Comprehension (MRC)
  • Problem Formulation
  • System architecture (an example)

• Future Directions
Many active areas of research!

(Part of larger push on Artificial Intelligence)
Many active areas of research!

(Part of larger push on Artificial Intelligence)

- New benchmarks and data curation methods
Many active areas of research!
(Part of larger push on Artificial Intelligence)

• New benchmarks and data curation methods
  • Text-based: SQuAD 2.0, MS MARCO, …
Many active areas of research!

(Part of larger push on Artificial Intelligence)

- New benchmarks and data curation methods
  - Text-based: SQuAD 2.0, MS MARCO, …
  - Multi-modal: Visual QA
Many active areas of research!
(Part of larger push on Artificial Intelligence)

- New benchmarks and data curation methods
  - Text-based: SQuAD 2.0, MS MARCO, ...
  - Multi-modal: Visual QA
- Incorporating Commonsense Reasoning
Many active areas of research!
(Part of larger push on Artificial Intelligence)

• New benchmarks and data curation methods
  • Text-based: SQuAD 2.0, MS MARCO, …
  • Multi-modal: Visual QA
  • Incorporating Commonsense Reasoning
• Grand Challenges: Todai Robot, AI2 Aristo, etc.
Many active areas of research!
(Part of larger push on Artificial Intelligence)

- New benchmarks and data curation methods
  - Text-based: SQuAD 2.0, MS MARCO, …
  - Multi-modal: Visual QA
  - Incorporating Commonsense Reasoning
- Grand Challenges: Todai Robot, AI2 Aristo, etc.
- New opportunities in applications
Many active areas of research!
(Part of larger push on Artificial Intelligence)

- New benchmarks and data curation methods
  - Text-based: SQuAD 2.0, MS MARCO, …
  - Multi-modal: Visual QA
  - Incorporating Commonsense Reasoning
- Grand Challenges: Todai Robot, AI2 Aristo, etc.
- New opportunities in applications
  - Customer service chatbot, Siri digital assistant, Watson health
Question: Why are the men jumping?

Original Image: to catch frisbee

Complementary Image: trick
Commonsense Reasoning

Commonsense Reasoning

- Commonsense about the physical world

Commonsense Reasoning

• Commonsense about the physical world

• e.g. [Winograd Schema Challenge (Levesque 2011)]

Commonsense Reasoning

- Commonsense about the physical world
  - e.g. [Winograd Schema Challenge (Levesque 2011)]
  - Q: The trophy would not fit in the brown suitcase because it was too big. What was too big?

Commonsense Reasoning

• Commonsense about the physical world

• e.g. [Winograd Schema Challenge (Levesque 2011)]

• Q: The trophy would not fit in the brown suitcase because it was too big. What was too big?

A. The trophy

Commonsense Reasoning

- Commonsense about the physical world
- e.g. [Winograd Schema Challenge (Levesque 2011)]

- Q: The trophy would not fit in the brown suitcase because it was too big. What was too big?
  
  A. The trophy

  B. The suitcase

Commonsense Reasoning

Commonsense Reasoning

• Commonsense about the social world

Commonsense Reasoning

- Commonsense about the social world

- e.g. [ROCStories (Mostafazadeh et al., 2016)]

Commonsense Reasoning

- Commonsense about the social world

  - e.g. [ROCStories (Mostafazadeh et al., 2016)]

  - Tom and Sheryl have been together for two years. One day, they went to a carnival together. He won her several stuffed bears, and bought her funnel cakes. When they reached the Ferris wheel, he got down on one knee. [Finish the story]

Commonsense Reasoning

• Commonsense about the social world

  • e.g. [ROCStories (Mostafazadeh et al., 2016)]

  • Tom and Sheryl have been together for two years. One day, they went to a carnival together. He won her several stuffed bears, and bought her funnel cakes. When they reached the Ferris wheel, he got down on one knee. [Finish the story]

    A. Tom asked Sheryl to marry him.

Commonsense Reasoning

• Commonsense about the social world

• e.g. [ROCStories (Mostafazadeh et al., 2016)]

Tom and Sheryl have been together for two years. One day, they went to a carnival together. He won her several stuffed bears, and bought her funnel cakes. When they reached the Ferris wheel, he got down on one knee. [Finish the story]

A. Tom asked Sheryl to marry him.

B. He wiped mud off of his boot.

AI2 Project Aristo:
solving elementary/middle school science exams

Which object is the best conductor of electricity?

(A) metal fork
(B) rubber boot
(C) plastic spoon
(D) wooden bat

Aristo's Answer: (A) metal fork

Correct Answer: A

Confidence: 90.72%

as computed from these reasoners:

Information Retrieval: 57.12% MORE INFO

Justification Sentence: For metals, the thermal conductivity is quite high, and those metals which are the best electrical conductors are also the best thermal conductors.

Table Reasoning: 38.90% MORE INFO

Knowledge Used: [ metal | conductor | electricity ]

Topic Matching: 52.72% MORE INFO

Topic: conductor, conductors

Tuple Reasoning: 46.23% MORE INFO

Knowledge Used: [ all metal objects | are | conductors of electricity ] [ metal | type Of | conductor | electricity ] [ conductor | object | electricity | metal ] [ metal | type Of | conductor ]

AristoRoBERTa: 90.30% MORE INFO
Can a robot pass a university entrance exam?

Todai Robot Project (2011-2016)
- passing an exam requires multiple intelligences
- what’s easy for human may not be easy for computers, and vice versa
- multiple choice, written essay questions
- topics: social studies, math, physics, English, Japanese
Summary

• Question Answering (QA)
  • Problem Formulation
  • System architecture (an example)

• Machine Reading Comprehension (MRC)
  • Problem Formulation
  • System architecture (an example)

• Future Directions
Questions?