Question Answering

One of the oldest NLP tasks (punched card systems in 1961) Simmons, Klein, McConlogue. 1964. Indexing and Dependency Logic for Answering English Questions. American Documentation 15:30, 196-204

Question Answering: IBM’s Watson

• Won Jeopardy on February 16, 2011!

WILLIAM WILKINSON’S "AN ACCOUNT OF THE PRINCIPALITIES OF WALLACHIA AND MOLDOVA" INSPIRED THIS AUTHOR’S MOST FAMOUS NOVEL

Apple’s Siri

Watson

How many calories are in two slices of banana cream pie?

700 Cal (Dietary Calories)
Types of Questions in Modern Systems

- **Factoid questions**
  - Who wrote "The Universal Declaration of Human Rights"?
  - How many calories are there in two slices of apple pie?
  - What is the average age of the onset of autism?
  - Where is Apple Computer based?

- **Complex (narrative) questions:**
  - In children with an acute febrile illness, what is the efficacy of acetaminophen in reducing fever?
  - What do scholars think about Jefferson's position on dealing with pirates?

<table>
<thead>
<tr>
<th>Commercial systems: mainly factoid questions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Question</strong></td>
</tr>
<tr>
<td>Where is the Louvre Museum located?</td>
</tr>
<tr>
<td>What's the abbreviation for limited partnership?</td>
</tr>
<tr>
<td>What are the names of Odin's ravens?</td>
</tr>
<tr>
<td>What currency is used in China?</td>
</tr>
<tr>
<td>What kind of nuts are used in marzipan?</td>
</tr>
<tr>
<td>What instrument does Max Roach play?</td>
</tr>
<tr>
<td>What is the telephone number for Stanford University?</td>
</tr>
</tbody>
</table>

Paradigms for QA

- **Information Retrieval (IR)-based approaches**
  - TREC, IBM Watson, Google
- **Knowledge-based and Hybrid approaches**
  - IBM Watson, Apple Siri, Wolfram Alpha, True Knowledge Evi

Many questions can already be answered by web search

IR-based Question Answering

Google search results for "Where is the Louvre Museum located?"

IR-based Factoid QA
IR-based Factoid QA

- **QUESTION PROCESSING**
  - Detect question type, answer type, focus, relations
  - Formulate queries to send to a search engine
- **PASSAGE RETRIEVAL**
  - Retrieve ranked documents
  - Break into suitable passages and rerank
- **ANSWER PROCESSING**
  - Extract candidate answers
  - Rank candidates using evidence from the text and external sources

Knowledge-based approaches (Siri)

- Build a semantic representation of the query
- Times, dates, locations, entities, numeric quantities
- Map from this semantics to query structured data or resources
- Geospatial databases
- Ontologies (Wikipedia infoboxes, dbPedia, WordNet, Yago)
- Restaurant review sources and reservation services
- Scientific databases

Hybrid approaches (IBM Watson)

- Build a shallow semantic representation of the query
- Generate answer candidates using IR methods
  - Augmented with ontologies and semi-structured data
- Score each candidate using richer knowledge sources
  - Geospatial databases
  - Temporal reasoning
  - Taxonomical classification

Answer Types and Query Formulation

**Question Processing**

Things to extract from the question

- **Answer Type Detection**
  - Decide the *named entity type* (person, place) of the answer
- **Query Formulation**
  - Choose *query keywords* for the IR system
- **Question Type classification**
  - Is this a definition question, a math question, a list question?
- **Focus Detection**
  - Find the question words that are replaced by the answer
- **Relation Extraction**
  - Find relations between entities in the question
Question Processing

**Jeopardy:** They're the two states you could be reentering if you're crossing Florida's northern border

- **Answer Type:** US state
- **Query:** two states, border, Florida, north
- **Focus:** the two states
- **Relations:** borders(Florida, ?, north)

Answer Type Detection: Named Entities

- **Who founded Virgin Airlines?**
  - **PERSON**
- **What Canadian city has the largest population?**
  - **CITY.**

Answer Type Taxonomy

- **6 coarse classes**
  - ABBREVIATION, ENTITY, DESCRIPTION, HUMAN, LOCATION, NUMERIC
- **50 finer classes**
  - LOCATION: city, country, mountain....
  - HUMAN: group, individual, title, description
  - ENTITY: animal, body, color, currency....

Xi Li, Dan Roth. 2002. Learning Question Classifiers. COLING’02

Part of Li & Roth’s Answer Type Taxonomy

Answer Types

<table>
<thead>
<tr>
<th>ENTITY</th>
<th>What are the features of a?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ABBREVIATION, ENTITY, DESCRIPTION, HUMAN, LOCATION, NUMERIC</td>
</tr>
</tbody>
</table>

More Answer Types

<table>
<thead>
<tr>
<th>HUMAN</th>
<th>What is the capital of a?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ABBREVIATION, ENTITY, DESCRIPTION, HUMAN, LOCATION, NUMERIC</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LOCATION</th>
<th>What is the location of a?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ABBREVIATION, ENTITY, DESCRIPTION, HUMAN, LOCATION, NUMERIC</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NUMERIC</th>
<th>What is the numeric of a?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ABBREVIATION, ENTITY, DESCRIPTION, HUMAN, LOCATION, NUMERIC</td>
</tr>
</tbody>
</table>
Answer types in Jeopardy

- 2500 answer types in 20,000 Jeopardy question sample
- The most frequent 200 answer types cover < 50% of data
- The 40 most frequent Jeopardy answer types
  - he, country, city, man, film, state, she, author, group, here, company, president, capital, star, novel, character, woman, river, island, king, song, part, series, sport, singer, actor, play, team, show, actress, animal, presidential, composer, musical, nation, book, title, leader, game

Answer Type Detection

- Hand-written rules
- Machine Learning
- Hybrids

Answer Type Detection

- Regular expression-based rules can get some cases:
  - Who is|was|are|were PERSON
  - PERSON [YEAR – YEAR]
- Other rules use the question headword: [the headword of the first noun phrase after the wh-word]
  - Which city in China has the largest number of foreign financial companies?
  - What is the state flower of California?

Features for Answer Type Detection

- Question words and phrases
- Part-of-speech tags
- Parse features (headwords)
- Named Entities
- Semantically related words

Factoid Q/A
Keyword Selection Algorithm


1. Select all non-stop words in quotations
2. Select all NNP words in recognized named entities
3. Select all complex nominals with their adjectival modifiers
4. Select all other complex nominals
5. Select all nouns with their adjectival modifiers
6. Select all other nouns
7. Select all verbs
8. Select all adverbs
9. Select the question focus word (skipped in all previous steps)
10. Select all other words

Choosing keywords from the query

Slide from Mihai Surdeanu

Who coined the term "cyberspace" in the novel "Neuromancer"?

cyberspace/1 Neuromancer/1 term/4 novel/4 coined/7

Passage Retrieval and Answer Extraction

Passage Retrieval

- Step 1: IR engine retrieves documents using query terms
- Step 2: Segment the documents into shorter units
  - something like paragraphs
- Step 3: Passage ranking
  - Use answer type to help rerank passages

Factoid Q/A

Features for Passage Ranking

Either in rule-based classifiers or with supervised machine learning

- Number of Named Entities of the right type in passage
- Number of query words in passage
- Number of question N-grams also in passage
- Proximity of query keywords to each other in passage
- Longest sequence of question words
- Rank of the document containing passage
Factoid Q/A

Answer Extraction

- Run an answer-type named-entity tagger on the passages
  - Each answer type requires a named-entity tagger that detects it
  - If answer type is CITY, tagger has to tag CITY
    - Can be full NER, simple regular expressions, or hybrid
  - Return the string with the right type:
    - Who is the prime minister of India *(PERSON)*
    - Mammon Singh, Prime Minister of India, had told left leaders that the deal would not be renegotiated.
    - How tall is Mt. Everest? *(LENGTH)*
    - The official height of Mount Everest is 29,035 feet

Ranking Candidate Answers

- But what if there are multiple candidate answers!

Q: Who was Queen Victoria’s second son?
- Answer Type: Person
- Passage:
The Marie biscuit is named after Marie Alexandrovna, the daughter of Czar Alexander II of Russia and wife of Alfred, the second son of Queen Victoria and Prince Albert

Use machine learning:
Features for ranking candidate answers

Answer type match: Candidate contains a phrase with the correct answer type.
Pattern match: Regular expression pattern matches the candidate.
Question keywords: # of question keywords in the candidate.
Novelty factor: A word in the candidate is not in the query.
Apposition features: The candidate is an appositive to question terms
Punctuation location: The candidate is immediately followed by a comma, period, quotation marks, semicolon, or exclamation mark.
Sequences of question terms: The length of the longest sequence of question terms that occurs in the candidate answer.

Candidate Answer scoring in IBM Watson

- Each candidate answer gets scores from >50 components
  - [from unstructured text, semi-structured text, triple stores]
    - logical form (parse) match between question and candidate
    - passage source reliability
    - geospatial location
    - California is "southwest of Montana"
    - temporal relationships
    - taxonomic classification
Common Evaluation Metrics

1. **Accuracy** (does answer match gold-labeled answer?)
2. **Mean Reciprocal Rank**
   - For each query return a ranked list of M candidate answers.
   - Query score is 1/Rank of the first correct answer
     - if first answer is correct: 1
     - else if second answer is correct: \( \frac{1}{2} \)
     - else if third answer is correct: \( \frac{1}{3} \), etc.
     - Score is 0 if none of the M answers are correct
   - Take the mean over all N queries

\[
MRR = \frac{\sum_{i=1}^{N} \frac{1}{\text{Rank}_i}}{N}
\]

Knowledge in QA

Relation Extraction

- **Answers**: Databases of Relations
  - born-in("Emma Goldman", "June 27 1869")
  - author-off("Cao Xue Qin", "Dream of the Red Chamber")
  - Draw from Wikipedia infoboxes, DBpedia, FreeBase, etc.
- **Questions**: Extracting Relations in Questions
  - (granddaughter-of ?x ?y)

Temporal Reasoning

- **Relation databases**
  - (and obituaries, biographical dictionaries, etc.)
- **IBM Watson**
  "In 1594 he took a job as a tax collector in Andalusia"
  Candidates:
  - Thoreau is a bad answer (born in 1817)
  - Cervantes is possible (was alive in 1594)

Geospatial knowledge

- **containment**, **directionality**, **borders**
- Beijing is a good answer for “Asian city”
- California is “southwest of Montana”
- geonames.org:

Context and Conversation

- **Virtual Assistants like Siri**
- Coreference helps resolve ambiguities
  - U: “Book a table at Il Fornaio at 7:00 with my mom”
  - U: “Also send her an email reminder”
- **Clarification questions**:
  - U: “Chicago pizza”
  - S: “Did you mean pizza restaurants in Chicago or Chicago-style pizza?”