Question Answering

NLP ML Web
Fall 2013
Andrew Rosenberg
TA/Grader: David Guy Brizan
Overview

- History of Question Answering
- Two components of QA
  - Understanding the question
  - Finding an answer
    - structured data
    - unstructured data
Do you know the way to San Jose?
Do you know the way to San Jose?
Question Answering

Do you know the way to San Jose?

Answer Generation

Question Analysis
Do you know the way to San Jose?
Do you know the way to San Jose?

Yes
Brief Overview of QA

- SHRDLU
  - Blocks world (incredibly narrow domain)
  - Simple pattern recognition
    - How many X are in the Y?
- Natural Language Database Systems
  - Like Expert systems.
  - Construct a DB query in natural language.
    - “What are the HP jobs on a Unix system?”
    - SELECT DISTINCT Description FROM JOB
      WHERE Platform = ‘Unix’ AND Company = ‘HP’
  - Still very actively researched.
Brief Overview of QA

- Queries of unstructured data
  - Inspired by information retrieval (web search)
  - Find relevant documents
    - Find desired information in these documents
  - Also very actively researched
- General QA Applications:
  - Search structured **and** unstructured sources.
  - Siri
  - Google Glass
  - Watson (IBM DeepQA)
QA and Dialog Systems

- Dialog systems allow a user to communicate with software through an Natural Language interface.
- ELIZA
- AT&T HMIHY (and other customer services)
- Apple Siri
- Google Glass
Julie The Amtrak God
Jenny Asarnow

http://www.prx.org/pieces/3592-julie-the-amtrak-god
Dialog System Concerns

- Initiative:
  - User, Machine, Mixed
- How wide is the domain?
- Handling miscommunication
  - Error Identification
  - Error Recovery Strategies
- Where are you going to find the answer?
QA Challenges

- Useless information is abundant. Important information is rare.
- Anaphora. Metonymy. etc.
  - Obama is in Chicago celebrating his birthday with Clinton. He turns 53 today.
  - Putin proposed talks about energy. Washington responded strongly.
- Temporal queries
  - Who is the president [today]?
  - What is the square root of four [today]?
- Ambiguity
  - Who wrote the music in Romeo and Juliet?
    - The movie? Which one?
    - The play? Which one?
    - Soundtrack? Score?
Question Analysis

In what year was Barack Obama’s born?

- Identify the syntactic and semantic components of a question to assist in identifying an answer.

- Question Classification
  - What is this question about?
    - People, Politics, Birthdays

- Focus identification
  - What is the question asking for?
    - `year(birthdate(Barack Obama, focus))`
    - `birthyear(Barack Obama, focus)`
Question Classification

- Identify what kind of questions you expect to be able to answer.
  - Biographical
  - Sports and Statistics
  - Political, etc.
- This forces you to pick a domain. Most QA systems are not completely “open QA”
Question Classification

- Identify what kind of questions you expect to be able to answer.
  - Biographical
  - Sports and Statistics
  - Political, etc.
- This forces you to pick a domain. Most QA systems are not completely “open QA”.
- Including Watson Jeopardy! has types of questions and topics.
Question Taxonomies

• A good taxonomy of questions is one that facilitates good answer retrieval and generation.

• How do you answer “Person” questions?
  • When was Obama born?
  • Who is Obama married to?
  • Where did Obama live from 1992-1996?

• How do you answer “Names of Authored Works” questions?
  • What was Melville’s first novel?
  • What peace treaty ended WWI?
Question Analysis in Watson (Lally et al. 2012)

- Focus
  - Reference to the answer.
- Lexical Answer Type (LAT)
  - Terms that indicate what type of information is being sought.
  - The focus word is an LAT
  - Used to determine whether a candidate is good or no.
- Question Classification
  - Definition, Category-Relation, Multiple Choices, Common Bonds, Puzzle, Fill in the Blank, Abbreviations
  - These use different strategies for answering
Question Analysis in Watson (Lally et al. 2012)

- Focus
  - Reference to the answer.
- Lexical Answer Type (LAT)
  - Terms that indicate what type of information is being sought.
  - The focus word is an LAT
  - Used to determine whether a candidate is good or not.
- Question Classification
  - Definition, Category-Relation, Multiple Choices, Common Bonds, Puzzle, Fill in the Blank, Abbreviations
  - These use different strategies for answering

NB: Jeopardy! has idiosyncrasies about how questions are asked.
Slot Grammar Parser

- Slot Grammar Parser
- Efficient way to have empty slots
- Easy to associate syntax and semantics

He was a bank clerk in the Yukon before he published “Songs of a Sourdough” in 1907.

- **Predicate Construction**
  - publish(e1, he, “Songs of a Sourdough”)
  - in(e2, e1, 1907)
  - authorOf(focus, “Songs of a Sourdough”)
  - temporalLink(publish(...), 1907)

- **Name Tagging**
  - GPE(n1, “Yukon”)
  - composition(n3, “Songs of a Sourdough”)

- **Co-reference**
  - Coref(n2, he, “clerk”)
Identifying Focus
(in Jeopardy! questions)

- A NP with determiner “this” or “these”
  - A new play based on this Sir Arthur Conan Doyle canine classic opened on the London stage in 2007
- “This” or “these” as a pronoun
  - In April 1988, Northwest became the first US air carrier to ban this on all domestic flights.
- When the question is a noun phrase label the whole phrase as focus
  - Number of poems Emily Dickenson gave permission to publish during her lifetime
- One of the pronouns “he/she/his/her/him/hers”
  - She joined Buffalo Bill Cody’s Wild West Show after meeting him at the Cotton Expo in New Orleans
- One of the pronouns “it/they/them/its/their”
  - It forbids Congress from interfering with a citizen’s freedom of religion, speech, assembly or petition
- The pronoun “one”
  - Leavenworth, established in 1895, is a federal one.
Identifying Lexical Answer Types

• If the focus is a conjunction, extract the conjuncts

• “Focus of X” extract LAT X when focus is any of one/type/kind

• “Focus of X” extract LAT X when focus is any of name/word/term
Question Classification

- Different Question Types require different answering strategies.

- BEFORE & AFTER: 13th Century Venetian traveler who’s a Ralph Lauren short sleeve top with a collar.

- THE SOUTHERNMOST CAPITAL CITY: Helsinki, Moscow, Bucharest.
<table>
<thead>
<tr>
<th>Question Classes</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Definition</strong></td>
<td>CONSTRUCTION: It can be the slope of a roof or the gunk used to waterproof it</td>
</tr>
<tr>
<td><strong>Category Relation</strong></td>
<td>COUNTRIES BY NEWSPAPER: Haaretz, Yeodith, Ahronoth</td>
</tr>
<tr>
<td><strong>Fill in the Blank</strong></td>
<td>SHAKESPEARE IN LOVE: “Not that I loved Caesar less,” says Brutus, “but that I loved” this city “more.”</td>
</tr>
<tr>
<td><strong>Abbreviation</strong></td>
<td>MILITARY MATTERS: Abbreviated SAS this elite British military unit is similar to the USA’s Delta Force</td>
</tr>
<tr>
<td><strong>Puzzle</strong></td>
<td>THE HIGHEST-SCORING SCRABBLE WORD: Zoom, quiz or heaven</td>
</tr>
<tr>
<td><strong>Etymology</strong></td>
<td>ARE YOU A FOOD†E: From the Spanish for “to bake in pastry”, it’s South America’s equivalent of a calzone</td>
</tr>
<tr>
<td><strong>Verb</strong></td>
<td>THE NOT-SO-DEADLY SINS: To capitalize all text in an email is an abomination that signifies the person is doing this.</td>
</tr>
</tbody>
</table>
Question Answering from Structured Data

- Map predicate forms or patterns on to fields of structured data
  - When was <name> born?
  - What is <name>’s birthday?
  - How old is <name>?
    - SELECT birthdate FROM PERSON WHERE name='<name>'
    - https://www.googleapis.com/freebase/v1/search?
      - query=<name>
      - filter=(all type:/people/person)
      - output=(/people/person/date_of_birth)
Finding an Answer in unstructured data

- Two stages of processing
  - Find relevant unstructured data
  - Extract an answer
Finding an Answer in unstructured data

Barack Obama

This article is about the 44th president of the United States. For his father, see Barack Obama, Sr.

Barack Hussein Obama II is the 44th and current President of the United States, the first African American to hold the office. Born in Honolulu, Hawaii, Obama is a graduate of Columbia University and Harvard Law School, where he was president of the Harvard Law Review. He was a community organizer in Chicago before earning his law degree. He worked as a civil rights attorney in Chicago and taught constitutional law at the University of Chicago Law School from 1992 to 2004. He served three terms representing the 13th District in the Illinois Senate from 1997 to 2004, running unsuccessfully for the United States House of Representatives in 2000.

In 2004, Obama received national attention during his campaign to represent Illinois in the United States Senate with his victory in the March Democratic Party primary, his keynote address at the Democratic National Convention in July, and his election to the Senate in November. He began his presidential campaign in 2007, and in 2008, after a close primary campaign against Hillary Rodham Clinton, he won sufficient delegates in the Democratic Party primaries to receive the presidential nomination. He then defeated Republican nominee John McCain in the general election, and was inaugurated as president on January 20, 2009. Nine months later, Obama was named the 2009 Nobel Peace Prize laureate. He was re-elected president in November 2012, defeating...
Information Retrieval

• Given a **query** identify zero or more **documents** that are **relevant** to that query
Information Retrieval

- Most familiar example:
  - Web Search
    - **query**: string of text
    - **document**: html formatted webpage
    - **relevance**: expected value to the user

- Other examples:
  - Video search (document: videos)
  - Spoken query (query: spoken language)
  - Find tweets like this (query: tweet, document: tweets)
  - Image search, Query by humming (query: image, document: images)
  - Query by humming (query: sound, document: sounds)
Information Extraction

- Construct Patterns

- DOB=born DATE*) | (DATE*-DATE) | was born on DATE* | was born in DATE*

- DATE=YEAR | MONTH YEAR | MM/DD/YY

- YEAR=[1-9]\d\d\d | \d\d

- MONTH = JAN | JANUARY | FEB ...

- DD = [0-3]\d

- MM = 0?[1-9] | 1[0-2]
Information Extraction

- Patterns for many primitives are constructed by hand.
  - Phone numbers, SSNs, IP addresses, email addresses
  - Street Addresses, Countries
  - Dates
  - Names
Relation Patterns

- Patterns for relations can also be constructed by hand
  - NAME was in LOCATION
  - NAME was fired from COMPANY
- Relation patterns are also learned from labeled and unlabeled data.
Pattern Learning

• Bootstrapping
  • Start with some seed patterns, learn more that are like these seeds
• Train a tagger.
  • CRF or other sequence model
• wFST generation
• Much more...
More challenging questions

- “How do I X?”
  - Require specific domain knowledge.
  - Often multiple steps and constraints. (planning)
- “Why X?”
  - These questions are hard for people.
- Opinion Questions (TREC)
  - “How do Americans feel about Canada?”
  - “Why do people dislike Forrest Gump?”
Do you know the way to San Jose?

Yes
Question Answering

Do you know the way to San Jose?

Question Analysis

Answer Generation

Answer Ranking

Yes
Do you know the way to San Jose?

Yes
Do you know the way to San Jose?

Yes

Question Answering

Question Analysis

Pre-Processing

Answer Generation

Answer Ranking

Information Extraction

Information Retrieval
Do you know the way to San Jose?

Yes
Do you know the way to San Jose?

Yes
Do you know the way to San Jose?

Question Analysis

Information Extraction

Answer Ranking

Natural Language DB Queries

Information Retrieval

Ranking

Ensemble Learning

Yes