Question answering

- Overview and task definition
- History
- Open-domain question answering
- Basic system architecture
- Predictive indexing methods
- Pattern-matching methods
- Advanced techniques

~2 lectures
Commercial QA systems

Not so easy for people either…
Question answering

• Overview and task definition

History

• Open-domain question answering
• Basic system architecture
• Predictive indexing methods
• Pattern-matching methods
• Advanced techniques

LUNAR

• Answered questions about moon rocks and soil gathered by the Apollo 11 mission
  – Data base of information for all collected samples

• Architecture
  – Parse English question into a data base query
    • Syntactic analysis via augmented transition network parser and heuristics (including some for shallow semantics)
    • Semantic analysis maps parsed request into query language; query denotes unambiguous meaning of the request
  – Run query on data base to produce answer

LUNAR

• Resources required
  – Parser for a subset of English (size unclear)
  – Handled tense, modality, some anaphora, some conjunctions, some relative clauses, some adjective modifiers (dealing with quantification)
  – Vocabulary of about 3,500 words

• Sample questions
  – What is the average concentration of aluminum in high alkali rocks?
  – What samples contain P205?
  – Give me the modal analyses of P205 in those samples.
LUNAR example

- Do any samples have greater than 13 percent aluminum?

Data base query

\[
\text{class to test: T;}
\text{No restriction on class Proposition}
\]

\[
\text{(TEST (FOR SOME X1 / (SEQ SAMPLES):}
T;
\text{(CONTAIN X1}
\text{(NPR* X2 / 'AL203)
(GREATERTHAN 13 PCT))))}
\]

- Answer:
  - Yes

LUNAR assessment

- System characteristics
  - Closed domain (lunar geology and chemistry)
  - Structured data (information contained in a data base)
  - Structured answers (information contained in a data base)
    - Avoided dialogue problems
  - Context: sophisticated users demanding high accuracy

- Labor intensive to build
  - Complex system
  - High accuracy required
  - Few general-purpose NLP resources available at the time

LUNAR assessment

- Research on systems like LUNAR continued for another decade
- Focused on
  - Syntactic parsing
  - Incorporating domain knowledge
  - Dialogue management
- Problems
  - Expensive to build
  - Brittle...prone to unexpected sudden failure

WOLFIE

- **WOrd Learning From Interpreted Examples**
  [Thompson and Mooney, 1998]
  - Closed domain, structured data, structured answers
  - Avoids labor-intensive system development
  - Uses inductive logic programming methods to acquire parsers that can map natural language queries into executable logical form (i.e. data base queries)
    - Requires examples of NL queries and their logical form
  - Indirectly evaluate the parser based on the number of queries that system gets right/wrong.
Lehnert’s Q/A system

- Implemented a broader theory of question answering
  - motivated by issues of cognitive plausibility
  - relied on the linguistic/cognitive theories of *conceptual dependency*, *scripts*, *plans*, etc. [Schank, 1970’s]
  - used a question taxonomy
  - somewhat closed domain (actions), unstructured data, generated answers
  - answered questions about an arbitrary input text (usually event-based)

Q/A example 1

- Input text
  John threw the baseball to Mary. She missed the ball and it hit her on the head.

- Questions:
  - Was Mary happy?
  - Who has the baseball?
  - Why did John throw the baseball to Mary?

Lehnert’s Q/A system

- “Parse” input text into a semantic representation (*conceptual dependency*)
- Generate inferences from that representation
  - Inferences associated with CD’s semantic primitives
- “Parse” the question, mapping it into one of the predefined question types
- Employ the method associated with the question type to answer the question

Q/A example 2

- Input text
  For their first date, John took Mary to McDonald’s for burgers. Mary was not impressed.

- Questions:
  - Did John and Mary pay for the burgers?
  - What did John and Mary eat for dinner on their first date?
Question answering

- Overview and task definition
- History

**Open-domain question answering**

- Basic system architecture
- Predictive indexing methods
- Pattern-matching methods
- Advanced techniques

Towards open-domain QA

Which country has the largest part of the Amazon rain forest?

The chaotic development that is gobbling up the Amazon rain forest could finally be reined in with a new plan developed by officials of Amazon countries and leading scientists from around the world.

“That’s some of the most encouraging news about the Amazon rain forest in recent years,” said Thomas Lovejoy, a tropical ecologist at the Smithsonian Institution and an Amazon specialist.

“It contrasts markedly with a year ago, when there was nothing to read about conservation in the Amazon, especially in Brazil, except bad news,” Lovejoy said in a recent interview.

**Sixty percent of the Amazon**, the world’s largest tropical rain forest, **lies in Brazil**, but the forest also covers parts of the eight surrounding countries.

Lovejoy was one of the organizers of an unusual workshop held in mid-January in Manaus, Brazil, a sprawling city of 1 million people in the heart of the Amazon. It was the center of Brazil’s once-thriving rubber trade.

Question Answering

- **Simplifications**
  - short-answer, fact-based questions
  - answer exists in the collection as a text fragment
  - supporting info can be found in a single document
  - system returns up to 5 guesses per question

- **Sample questions**
  - How many calories are there in a Big Mac?
  - Who is the voice of Miss Piggy?
  - Who was the first American in space?
  - Where is the Taj Mahal?
TREC QA: evaluation

- **Human assessors judge the answers**
  - Allowed to accept multiple answers
- **Systems scored on mean reciprocal rank of first correct answer**
  - First answer correct = 1 point, second answer correct = ½ point, third answer correct = 1/3 point, …
  - 0 if none of the n answers are correct
  - Average across all questions
- **Also reported on the number of questions answered correctly**

**Question Answering**

- **Performance**
  - | %correct | avg rank 1st |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>TREC-8 (1999):</td>
<td>70%</td>
</tr>
<tr>
<td>TREC-9 (2000):</td>
<td>65%</td>
</tr>
<tr>
<td>TREC-10 (2001):</td>
<td>70%</td>
</tr>
<tr>
<td>TREC-11 (2002):</td>
<td>83%</td>
</tr>
<tr>
<td>TREC-12 (2003):</td>
<td>70%</td>
</tr>
<tr>
<td>TREC-13 (2004):</td>
<td>84%</td>
</tr>
</tbody>
</table>

State of the art: TREC 2000

**List Questions**

- **List questions**
  - 1915: List the names of chewing gums.
    - Stimorol
    - Orbit
    - Winterfresh
    - Double Bubble
    - Dirol
    - Trident
    - Spearmint
    - Bazooka
    - Doublemint
    - Dentyne
    - Freedom
    - Hubba Rubba
    - Juicy Fruit
    - Big Red
    - Cheeselets
    - Nicorette

- **Performance**
  - TREC 2003: F .40
  - TREC 2004: F .62
Definition Questions

• Who is Colin Powell?
• What is mold?
• Audience?

• Evaluation requires matching *concepts* in the desired response to *concepts* in a system response
  – TREC 2003: F .55

Context Task

• Context with target item

<table>
<thead>
<tr>
<th></th>
<th>21</th>
<th>Club Med</th>
</tr>
</thead>
<tbody>
<tr>
<td>21.1</td>
<td>FACTOID</td>
<td>How many Club Med vacation spots are there worldwide?</td>
</tr>
<tr>
<td>21.2</td>
<td>LIST</td>
<td>List the spots in the United States.</td>
</tr>
<tr>
<td>21.3</td>
<td>FACTOID</td>
<td>Where is an adults-only Club Med?</td>
</tr>
<tr>
<td>21.4</td>
<td>OTHER</td>
<td></td>
</tr>
</tbody>
</table>

• Performance
  – TREC 2004
    • Factoids: .84 initial; .74 non-initial
    • Lists: .62 F
    • Other: .46 F

Question answering

• Overview and task definition
• History
• Open-domain question answering

**Basic system architecture**
  [Cardie et al., ANLP 2000]

• Predictive indexing methods
• Pattern-matching methods
• Advanced techniques