Question Answering

- Grand Challenge Problem for NLP: A program that can find the answer to arbitrary questions from text resources.
  WWW, encyclopedias, books, manuals, medical literature, scientific papers, etc.

- Another application: database queries
  Converting natural language questions into database queries was one of the earliest NLP applications!

- A scientific reason to do Q/A: the ability to answer questions about a story is the hallmark of text understanding.

Single Document Question Answering

- A single document Q/A task involves questions associated with one particular document.

- We usually assume that the answer appears somewhere in the document, but often it only appears once.

- Applications are for searching a particular source, such as a book, encyclopedia, technical manual, or speech.

- The questions and answers are often specific to the document. For example, “Where does Mary live?” or “What did Obama propose?”

Reading Comprehension Tests

- Reading comprehension tests were designed to evaluate how well a child understands a story. So they are an objective way to assess the “reading level” of NLP systems!

- Found data: questions and answer keys already exist!

- We can challenge our NLP systems with progressively harder questions (higher grade levels).

- Grade school exams typically ask factoid questions that mimic real-world applications (as opposed to advanced exams that often require summarization, e.g. “What is the topic of this article?”).

Sample Reading Comprehension Test

**Mars Polar Lander - Where Are You?**

(January 18, 2000) After more than a month of searching for a signal from NASA's Mars Polar Lander, mission controllers have lost hope of finding it. The Mars Polar Lander was on a mission to Mars to study its atmosphere and search for water, something that could help scientists determine whether life ever existed on Mars. Polar Lander was to have touched down December 3 for a 90-day mission. It was to land near Mars' south pole. The lander was last heard from minutes before beginning its descent. The last effort to communicate with the three-legged lander ended with frustration at 8 a.m. Monday. “We didn’t see anything,” said Richard Cook, the spacecraft's project manager at NASA's Jet Propulsion Laboratory. The failed mission to the Red Planet cost the American government more than $200 million dollars. Now, space agency scientists and engineers will try to find out what could have gone wrong. They do not want to make the same mistakes in the next mission.

1. When did the mission controllers lose hope of communicating with the lander?
2. Who is the Polar Lander's project manager?
3. Where on Mars was the spacecraft supposed to touch down?
4. What was the mission of the Mars Polar Lander?
Sample Reading Comprehension Test

Mars Polar Lander - Where Are You?
(January 18, 2000) After more than a month of searching for a signal from NASA’s Mars Polar Lander, mission controllers have lost hope of finding it. The Mars Polar Lander was on a mission to Mars to study its atmosphere and search for water, something that could help scientists determine whether life ever existed on Mars. Polar Lander was to have touched down December 3 for a 90-day mission. It was to land near Mars’ south pole. The lander was last heard from minutes before beginning its descent. The last effort to communicate with the three-legged lander ended with frustration at 8 a.m. Monday. “We didn’t see anything,” said Richard Cook, the spacecraft’s project manager at NASA’s Jet Propulsion Laboratory. The failed mission to the Red Planet cost the American government more than $200 million dollars. Now, space agency scientists and engineers will try to find out what could have gone wrong. They do not want to make the same mistakes in the next mission.

1. When did the mission controllers lose hope of communicating with the lander?  
   (Answer: 8 a.m. Monday, Jan. 17)

2. Who is the Polar Lander’s project manager?  
   (Answer: Richard Cook)

3. Where on Mars was the spacecraft supposed to touch down?  
   (Answer: near Mars’ south pole)

4. What was the mission of the Mars Polar Lander?  
   (Answer: to study Mars’ atmosphere and search for water)

Multiple Document Question Answering

- A multiple document Q/A task involves questions posed against a collection of documents.
- The answer often appears multiple times in the collection.
- The question and answer are absolute (not relative to one document), so it doesn’t matter where the answer is found.
- Applications include WWW search engines, and searching text repositories such as news archives, medical documents, or scientific literature.

Answer Frequency

<table>
<thead>
<tr>
<th># Questions</th>
<th># Answers</th>
<th>Mean</th>
<th>Median</th>
<th>Standard Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>TREC-8</td>
<td>50</td>
<td>7.04</td>
<td>3</td>
<td>12.94</td>
</tr>
<tr>
<td>CBC</td>
<td>219</td>
<td>1.25</td>
<td>1</td>
<td>0.61</td>
</tr>
</tbody>
</table>

The Effect of Answer Frequency on Q/A Performance
**TREC-9 Q/A Task: Factoid Questions**

<table>
<thead>
<tr>
<th>Number of Documents:</th>
<th>979,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amount of Text:</td>
<td>3 Gb</td>
</tr>
<tr>
<td>Document Sources:</td>
<td>AP, WSJ, Financial Times, San Jose Mercury News, LA Times, FBIS</td>
</tr>
<tr>
<td>Number of Questions:</td>
<td>682</td>
</tr>
<tr>
<td>Question Sources:</td>
<td>Encarta log, Excite log</td>
</tr>
</tbody>
</table>

Sample Questions:
- How much folic acid should an expectant mother get daily?
- Who invented the paper clip?
- What university was Woodrow Wilson president of?
- Where is Rider College located?
- Name a film in which Jude Law acted.
- Where do lobsters like to live?

---

**Evaluating Answers**

There are several ways to present an answer to a user:

- **Short Answer:** the exact answer to the question
- **Answer Sentence:** the sentence containing the answer
- **Answer Passage:** a short passage containing the answer (e.g., paragraph)

Short answers are difficult to score automatically because many variations are often acceptable. Example:

Text: *The 2002 Winter Olympics will be held in beautiful Salt Lake City, Utah.*
Q: *Where will the 2002 winter Olympics be held?*

A1: *beautiful Salt Lake City, Utah*
A2: *Salt Lake City, Utah*
A3: *Salt Lake City*
A4: *Salt Lake*
A5: *Utah*

---

**Mean Reciprocal Rank**

- The most common metric for evaluating Q/A systems is **Mean Reciprocal Rank (MRR).**
- For each question, the system generates a ranked list of $k$ possible answers.
- The reciprocal rank score for a question is $1/R$, where $R$ is rank of the first correct answer in the list (or zero if the answer is not on the list).
- MRR is the average reciprocal rank score over a set of $N$ questions.

$$MRR = \frac{\sum_{i=1}^{N} \frac{1}{rank_i}}{N}$$

---

**MRR Example**

Q: *What is the capital of Utah?*
A1: *Ogden*
A2: *Provo*
A3: *Salt Lake City*
A4: *St. George*
A5: *Salt Lake*

The reciprocal rank score for this question is $1/3$. 
Architecture of Typical Q/A System

Question Classification

Many common varieties of questions expect a specific type of answer. For example:

- **WHO**: person, organization, or country.
- **WHERE**: location (specific or general)
- **WHEN**: date or time period
- **HOW MUCH**: an amount
- **HOW MANY**: a number
- **WHICH CITY**: a city

Most Q/A systems use a question classifier to assign a type to each question. The question type constrains the set of possible answers. The classification rules are often developed by hand.

Difficult Question Types

Some types of questions are asking for complex answer strings, such as:

- Lists
  Ex: *List the movies that Meryl Streep has appeared in.*

- Definitions
  Ex: *What is natural language processing?*

- Explanations
  Ex: *How can I fix a leaky faucet?*

- Reasons
  Ex: *Why did Apple purchase Chomp?*

Question Expansion and Reformulation

- Similar to IR systems, the question is sometimes expanded or reformulated to improve the odds of finding a matching context.

- Question expansion may add similar words found in WordNet or a thesaurus.

- Question reformulation rules rewrite the question as a declarative statement. For example: “When was Mozart born?” → “Mozart was born”.

  Sample rules from Lin (2007):
  
  WH-word did A verb B → A verb+ed B
  Where is A → A is located in
Document/Passage Retrieval

- For many applications, the text collection that must be searched is extremely large.
- Applying NLP techniques to enormous volumes of text is too expensive. So IR systems are used to identify a much smaller subset of the collection that is likely to contain the answer.
- *Document retrieval systems* return N documents that are most relevant to the question. *Passage retrieval systems* return N passages that are most relevant to the question.
- If the IR engine doesn’t retrieve the texts that contain the answer, the Q/A system is out of luck!

Named Entity and Semantic Tagging

**Named Entity (NE) Taggers** recognize certain types of Named objects, and other easily identifiable classes. WordNet and semantic dictionaries may be used too. Common semantic classes for Q/A are:

- **People:** Mr. Fripper; John Fripper; President Fripper
- **Locations:** Salt Lake City; Massachusetts; France
- **Dates/Times:** November; Monday; 5:10 pm
- **Companies:** KVW Co.; KVW Inc.; KVW corporation
- **Measures:** 500 dollars; 40 miles; 32 lbs

Answer Identification (Pinpointing)

At this point, we’ve categorized the question and identified semantic entities in the text. So we can narrow down the candidate pool to entities of the appropriate type.

**PROBLEM:** There are often many entities of the right type!
- The **Answer Identification** module must figure out which entity is the correct answer.
- For questions that have Named Entity types, this module must figure out which item of the appropriate type is correct.
- For questions that do not have Named Entity types, this module is essentially starting from scratch!

Word Overlap

A simple method for Answer Identification is to measure the amount of **Word Overlap** between the question and an answer candidate.

- **Basic Word Overlap:** Each answer candidate is scored by counting how many question words are present in or near the candidate.
- **Stop Words:** sometimes closed class words (often called *Stop Words* in IR) are not included in the word overlap measure.
- **Roots:** sometimes stemming or morphological analysis is used to match the root forms of words (e.g., “walk” and “walked” would match).
- **Weights:** some words may be weighted more heavily than others (e.g., verbs might be given more weight than nouns)
Machine Learning

- Some researchers (e.g., [Ng et al. 2000]) train machine learning classifiers to label a sentence as positive or negative with respect to a question type.
- The sentence classified as positive with the highest confidence is chosen. In the absence of positive sentences, the sentence classified as negative with the lowest confidence is chosen.
- Each sentence is represented by a feature vector. Some features that have been used include:
  - number of matching words in S and Q
  - number of matching verbs in S and Q
  - number of matching words in previous/next sentence
  - sentence contains entity of type T
  - keywords appear in S
  - keywords appear in Q

Learning Surface Patterns

Surface patterns that connect specific question and answer terms can be learned automatically from the web [Ravichandran & Hovy 2002].

1. Provide a Q/A pair for the question type. For example, to learn patterns for BIRTHYEAR questions, give “Mozart 1756”.
2. Submit these terms to a search engine. For example, +“Mozart” and +“1756” to Google and download the top 1000 documents.
3. Apply a sentence segmenter and identify the sentences that contain both the Q and A terms.
4. Find all substrings and their counts using a suffix tree constructor. Filter substrings that do not contain both the Q and A terms.
5. Replace the Q term with <NAME> and the A term with <ANS>.
6. Apply the pattern to the corpus and measure its precision.

Examples of Learned Patterns

<table>
<thead>
<tr>
<th>BIRTHYEAR</th>
<th>LOCATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>When was X born?</td>
<td>Where is X located?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Prec</th>
<th>Pattern</th>
<th>Prec</th>
<th>Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>&lt;NAME&gt; (&lt;ANS&gt;)</td>
<td>1.0</td>
<td>&lt;ANS&gt;’s &lt;NAME&gt;</td>
</tr>
<tr>
<td>.85</td>
<td>&lt;NAME&gt; was born on &lt;ANS&gt;</td>
<td>1.0</td>
<td>regional: &lt;ANS&gt;:&lt;NAME&gt;</td>
</tr>
<tr>
<td>.60</td>
<td>&lt;NAME&gt; was born in &lt;ANS&gt;</td>
<td>1.0</td>
<td>to &lt;ANS&gt;’s &lt;NAME&gt;</td>
</tr>
<tr>
<td>.59</td>
<td>&lt;NAME&gt; was born &lt;ANS&gt;</td>
<td>1.0</td>
<td>&lt;ANS&gt;’s &lt;NAME&gt; in</td>
</tr>
<tr>
<td>.53</td>
<td>&lt;ANS&gt; &lt;NAME&gt; was born</td>
<td>1.0</td>
<td>in &lt;ANS&gt;’s &lt;NAME&gt;</td>
</tr>
<tr>
<td>.50</td>
<td>- &lt;NAME&gt; (&lt;ANS&gt;)</td>
<td>1.0</td>
<td>of &lt;ANS&gt;’s &lt;NAME&gt;</td>
</tr>
<tr>
<td>.36</td>
<td>&lt;NAME&gt; (&lt;ANS&gt;)</td>
<td>1.0</td>
<td>at the &lt;NAME&gt; in &lt;ANS&gt;</td>
</tr>
<tr>
<td>.32</td>
<td>&lt;NAME&gt; (&lt;ANS&gt;),</td>
<td>.96</td>
<td>the &lt;NAME&gt; in &lt;ANS&gt;</td>
</tr>
<tr>
<td>.28</td>
<td>born in &lt;ANS&gt;, &lt;NAME&gt;</td>
<td>.92</td>
<td>from &lt;ANS&gt;’s &lt;NAME&gt;</td>
</tr>
<tr>
<td>.20</td>
<td>of &lt;NAME&gt; (&lt;ANS&gt;)</td>
<td>.92</td>
<td>near &lt;NAME&gt; in &lt;ANS&gt;</td>
</tr>
</tbody>
</table>

Answer Confusability Experiments

- We manually annotated data for 165 TREC-9 questions and 186 CBC questions for perfect question typing, perfect answer sentence identification, and perfect semantic tagging.

  Idea: An oracle gives you the correct question type, a sentence containing the answer, and correctly tags all entities in the sentence that match the question type.

  Ex: The oracle tells you that the question expects a person, gives you a sentence containing the correct person, and tags all person entities in that sentence. The one thing the oracle does not tell you is which person is the correct one.

- We measured the “answer confusability”: the score that a Q/A system would get it if randomly selected an item of the desired type from the answer sentence.
Question and Answer Typing Results

<table>
<thead>
<tr>
<th>Answer Type</th>
<th>TREC Score</th>
<th>TREC Freq</th>
<th>CBC Score</th>
<th>CBC Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>defaultNP</td>
<td>.35</td>
<td>47</td>
<td>.25</td>
<td>28</td>
</tr>
<tr>
<td>organization</td>
<td>.50</td>
<td>1</td>
<td>.72</td>
<td>3</td>
</tr>
<tr>
<td>length</td>
<td>.50</td>
<td>1</td>
<td>.75</td>
<td>2</td>
</tr>
<tr>
<td>thingName</td>
<td>.58</td>
<td>14</td>
<td>.50</td>
<td>1</td>
</tr>
<tr>
<td>quantity</td>
<td>.58</td>
<td>13</td>
<td>.77</td>
<td>14</td>
</tr>
<tr>
<td>agent</td>
<td>.63</td>
<td>19</td>
<td>.40</td>
<td>23</td>
</tr>
<tr>
<td>location</td>
<td>.70</td>
<td>24</td>
<td>.68</td>
<td>29</td>
</tr>
<tr>
<td>personName</td>
<td>.72</td>
<td>11</td>
<td>.83</td>
<td>13</td>
</tr>
<tr>
<td>city</td>
<td>.73</td>
<td>3</td>
<td>n/a</td>
<td>0</td>
</tr>
<tr>
<td>defaultVP</td>
<td>.75</td>
<td>2</td>
<td>.42</td>
<td>15</td>
</tr>
<tr>
<td>temporal</td>
<td>.78</td>
<td>16</td>
<td>.75</td>
<td>26</td>
</tr>
<tr>
<td>personNoun</td>
<td>.79</td>
<td>7</td>
<td>.53</td>
<td>5</td>
</tr>
<tr>
<td>duration</td>
<td>1.0</td>
<td>3</td>
<td>.67</td>
<td>4</td>
</tr>
<tr>
<td>province</td>
<td>1.0</td>
<td>2</td>
<td>1.0</td>
<td>2</td>
</tr>
<tr>
<td>area</td>
<td>1.0</td>
<td>1</td>
<td>n/a</td>
<td>0</td>
</tr>
<tr>
<td>day</td>
<td>1.0</td>
<td>1</td>
<td>n/a</td>
<td>0</td>
</tr>
<tr>
<td>title</td>
<td>n/a</td>
<td>0</td>
<td>.50</td>
<td>1</td>
</tr>
<tr>
<td>person</td>
<td>n/a</td>
<td>0</td>
<td>.67</td>
<td>3</td>
</tr>
<tr>
<td>money</td>
<td>n/a</td>
<td>0</td>
<td>.88</td>
<td>8</td>
</tr>
<tr>
<td>ambigSize</td>
<td>n/a</td>
<td>0</td>
<td>.68</td>
<td>4</td>
</tr>
<tr>
<td>age</td>
<td>n/a</td>
<td>0</td>
<td>1.0</td>
<td>2</td>
</tr>
<tr>
<td>comparison</td>
<td>n/a</td>
<td>0</td>
<td>1.0</td>
<td>1</td>
</tr>
<tr>
<td>mass</td>
<td>n/a</td>
<td>0</td>
<td>1.0</td>
<td>1</td>
</tr>
<tr>
<td>measure</td>
<td>n/a</td>
<td>0</td>
<td>1.0</td>
<td>1</td>
</tr>
<tr>
<td>Overall</td>
<td>.59</td>
<td>165</td>
<td>.61</td>
<td>186</td>
</tr>
<tr>
<td>Overall w/o Defaults</td>
<td>.59</td>
<td>116</td>
<td>.70</td>
<td>143</td>
</tr>
</tbody>
</table>

The State of the Art in Q/A

- Question answering technology has improved tremendously in the last 15 years or so.

- Q/A systems can do reasonably well for factoid questions in a multi-document setting. The correct answer is often among the top candidates, but not necessarily #1.

- Progress has been made for non-factoid Q/A and single-document Q/A, but these are more difficult problems.

- IBM’s WATSON system made a lot of advances for Q/A, although much of their system was tailored for Jeopardy and similar types of scenarios (e.g., factoid, multi-document, and broad-coverage).

What’s going on?

Q1: When was Fred Smith born?
S1: Fred Smith lived from **1823** to **1897**.

Q2: What city is Massachusetts General Hospital located in?
S2: It was conducted by a cooperative group of oncologists from Hoag, Massachusetts General Hospital in **Boston**, Dartmouth College in New Hampshire, UC San Diego Medical Center, McGill University in Montreal and the University of **Missouri** in Columbia.