Question answering

• Overview and task definition
• History

Open-domain question answering
• Basic system architecture
  – Watson’s architecture
• Techniques
  – 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.

TReC QA framework

• Early simplifications (late 1990’s)
  – 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 (MRR) of first correct answer
  – if first answer correct = 1 point,
  – else if second answer correct = ½ point,
  – else if third answer correct = 1/3 point, …
  – 0 if none of the n answers are correct
  – Average of the reciprocal rank across all questions

• Also reported on the number of questions answered correctly
Question Answering

• Performance

<table>
<thead>
<tr>
<th>%correct</th>
<th>avg rank 1st</th>
</tr>
</thead>
<tbody>
<tr>
<td>TREC-8 (1999): 70%</td>
<td>1.4</td>
</tr>
<tr>
<td>TREC-9 (2000): 65%</td>
<td>1.7</td>
</tr>
<tr>
<td>TREC-10 (2001): 70%</td>
<td>1.3</td>
</tr>
<tr>
<td>TREC-11 (2002): 83%</td>
<td>1</td>
</tr>
<tr>
<td>TREC-12 (2003): 70%</td>
<td>1</td>
</tr>
<tr>
<td>TREC-13 (2004): 84%</td>
<td>1</td>
</tr>
</tbody>
</table>

harder questions
NIL answers
1 guess, exact

Results for TREC 2000

Best QA Results, 50-byte limit

List Questions

• List questions

<table>
<thead>
<tr>
<th>Stimorol</th>
<th>Orbit</th>
<th>Winterfresh</th>
<th>Double Bubble</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dirol</td>
<td>Trident</td>
<td>Spearmint</td>
<td>Bazooka</td>
</tr>
<tr>
<td>Doublemint</td>
<td>Dentyne</td>
<td>Freudent</td>
<td>Hubba Bubba</td>
</tr>
<tr>
<td>Juicy Fruit</td>
<td>Big Red</td>
<td>Chiclets</td>
<td>Nicorette</td>
</tr>
</tbody>
</table>

• Can’t just rely on a single document

• Performance
  – TREC 2003: F .40
  – TREC 2004: F .62

R/P/F-measure

• Recall
• Precision
• F-measure, F-score
Definition Questions

- Who is Colin Powell?
- What is mold?

Definition Questions

- Hard to evaluate
  - Who is the audience?
  - Evaluation requires matching concepts in the desired response to concepts in a system response
    - TREC 2003:
      - Audience: questioner is an adult, a native speaker of English, and an “average” reader of US newspapers
      - Results: F .55

Context Task

- Track a target discourse object through a series of questions

<table>
<thead>
<tr>
<th>21</th>
<th>Club Med</th>
</tr>
</thead>
<tbody>
<tr>
<td>21.1</td>
<td>FACTOID</td>
</tr>
<tr>
<td>21.2</td>
<td>LIST</td>
</tr>
<tr>
<td>21.3</td>
<td>FACTOID</td>
</tr>
<tr>
<td>21.4</td>
<td>OTHER</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
    - Watson’s architecture
  - Techniques
    - Predictive indexing methods
    - Pattern-matching methods
    - Advanced techniques
Basic system architecture

System architecture: document retrieval

Document retrieval

- Standard ad-hoc IR using full-text indexing
- Example QA system uses
  - vector space model
  - text retrieval system: Smart
  - standard term-weighting strategies (tfidf)
  - no automatic relevance feedback
QA as document retrieval

Baseline evaluation

- Document retrieval only
- Corpus
  - TREC-8 development corpus (38 questions)
  - TREC-8 test corpus (200 questions)

<table>
<thead>
<tr>
<th>Development (38)</th>
<th>Test (200)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>MAR</td>
</tr>
<tr>
<td>Smart</td>
<td>3</td>
</tr>
</tbody>
</table>

MAR = Mean Answer Rank

Baseline evaluation

- Smart performs better than its scores would suggest
  - Development corpus
    - For 18 of 38 questions, the answer appears in the top-ranked document
    - For 33 of 38 questions, the answer appears in one of the top 7 documents
    - For only 2 questions does Smart fail to retrieve the answer in one of the top 25 documents
  - Test corpus
    - Over half (110) of the questions are answered in the top-ranked document
    - Over 75% of the questions (155) are answered in the top 5 documents
    - 19 questions were not answered in the top 20 documents

System architecture: passage retrieval
Passage Retrieval

Query-dependent text summarization

- **Basic algorithm**
  1. Decide on a summary length (10% of document length).
  2. Use standard ad-hoc retrieval algorithm to retrieve top $k$ documents.
  3. **Treat each sentence/paragraph in top $N$ documents as a document itself.**
     Use standard document similarity equations to assign a similarity score to the sentence/paragraph.
  4. Return highest-scoring sentences/paragraphs as the summary, subject to the length constraint.

Passage retrieval

Query-dependent text summarization

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 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, an Amazon specialist. “It contrasts markedly with a year ago, when there was nothing to read about conservation in the Amazon.”
- [Sixty percent of the Amazon, the world’s largest tropical rain forest, lies in Brazil.]
QA as query-dependent text summarization

Evaluation: text summarization

<table>
<thead>
<tr>
<th></th>
<th>Development (38)</th>
<th>Test (200)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smart</td>
<td>Correct 3</td>
<td>Correct 29</td>
</tr>
<tr>
<td></td>
<td>MAR 3.33</td>
<td>MAR 3.07</td>
</tr>
<tr>
<td>Text Summarization</td>
<td>Correct 4</td>
<td>Correct 45</td>
</tr>
<tr>
<td></td>
<td>MAR 2.25</td>
<td>MAR 2.67</td>
</tr>
</tbody>
</table>

MAR = Mean Answer Rank

Evaluation: passage retrieval

- **Summarization method can limit performance**
  - Development corpus
    - In only 23 of the 38 developments questions (61%) does the correct answer appear in the summary for one of the top $k=7$ documents
  - Test corpus
    - In only 135 of the 200 developments questions (67.5%) does the correct answer appear in the summary for one of the top $(k=6)$ documents

Linguistic filters

- 50 byte answer length effectively eliminates *how, why, and definition* questions
- almost all of the remaining question types are likely to have *noun phrases* as answers
  - development corpus: 36 of 38 questions have noun phrase answers
- consider adding at least a simple linguistic filter that considers only noun phrases as answer hypotheses
System architecture: linguistic filters

The noun phrase filter

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

ordered list of summary extracts

ordered list of NPs

answer hypotheses

1st Guess
2nd Guess
3rd Guess
4th Guess
5th Guess

The huge Amazon rain forest is regarded as vital to the global environment.

[Japan] will not fund the construction of [the final segment] of [a controversial highway] through [the Amazon rain forest] in [Brazil], according to [a senior Republican senator].

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

ordered list of summary extracts

ordered list of NPs

answer hypotheses

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. Sixty percent of the Amazon, the world’s largest tropical rain forest, lies in Brazil, but the forest also

Sixty percent of the Amazon, the world’s largest tropical rain forest, lies in Brazil, but the forest also

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

ordered list of summary extracts

ordered list of NPs

answer hypotheses

“Japan Brazil a new plan Amazon countries A section”

“Northwestern Brazil A plan the Amazon region”

“eight surrounding countries A Brazilian company”

“union leader his modest wooden house The people”

“defense Brazilian wealth the international market”
**Evaluation: NP filter**

- How much does the (unnatural) NP "chunking" help? – Without it, only 1 and 20 questions answered for each corpus, respectively – NP filter is extracting good guesses, but better linguistic processing is needed to promote the best guesses to the top of the ranked guess list

<table>
<thead>
<tr>
<th></th>
<th>Development (38)</th>
<th>Test (200)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Correct</td>
<td>MAR</td>
</tr>
<tr>
<td>Smart</td>
<td>3</td>
<td>3.33</td>
</tr>
<tr>
<td>Text Summarization</td>
<td>4</td>
<td>2.25</td>
</tr>
<tr>
<td>TS + NPs</td>
<td>7</td>
<td>2.29</td>
</tr>
</tbody>
</table>

**MAR = Mean Answer Rank**

- **Using NP finder of Cardie & Pierce (1998)**
  - ~94% precision and recall on Wall Street Journal text
- **How much does the (unnatural) NP "chunking" help?**
  - Without it, only 1 and 20 questions answered for each corpus, respectively
  - NP filter is extracting good guesses, but better linguistic processing is needed to promote the best guesses to the top of the ranked guess list

**Semantic class checking**

- **Approximate question type using question word**
  - Who is the president of the U.S.? (person)
  - Which country has the largest part of the Amazon rain forest? (country)
  - Where is the Connecticut River? (state, county, region, location)
  - What fabric should one use to make curtains? (fabric)
- **Check that head noun (i.e. the last noun) of answer NP is of the same type**
  - a man = person
  - Massachusetts = state, location

**Semantic type checking**

- **Use lexical resource to determine semantic compatibility**
  - WordNet
- **Proper names handled separately since they are unlikely to appear in WordNet**
  - Small set (~20) rules
Evaluation: semantic class filter

<table>
<thead>
<tr>
<th></th>
<th>Development (38)</th>
<th>Test (200)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Correct</td>
<td>MAR</td>
</tr>
<tr>
<td>Smart</td>
<td>3</td>
<td>3.33</td>
</tr>
<tr>
<td>Text Summarization</td>
<td>4</td>
<td>2.25</td>
</tr>
<tr>
<td>TS + NPs</td>
<td>7</td>
<td>2.29</td>
</tr>
<tr>
<td>TS + NPs + Semantic Type</td>
<td>21</td>
<td>1.38</td>
</tr>
</tbody>
</table>

MAR = Mean Answer Rank

- Weak syntactic and semantic information allows large improvements
- Problems?

Sources of error

[Pie charts showing sources of error without labels]