Overview of the TREC 2002 Question Answering Track

Ellen Voorhees

NIST
National Institute of Standards and Technology
Technology Administration, U.S. Department of Commerce
Question Answering Track

- Goal: encourage research into systems that return answers, rather than document lists

- TREC 2002 is fourth year
  - as before, restricted to factoid questions with document for support
  - this year required exact answer, not text snippet
TREC 2002 QA Track

• **Main task**
  • return exactly one response for each of 500 questions
  • response is either [doc, string] pair or NIL
  • rank questions by confidence in answer

• **List task**
  • target number of instances given in question
  • assemble an unordered set of instances where an instance is a [doc, string] pair
**QA Track Participation**

<table>
<thead>
<tr>
<th>Alicante University</th>
<th>Language Computer Corp.</th>
<th>University of Amsterdam</th>
</tr>
</thead>
<tbody>
<tr>
<td>BBN Technologies</td>
<td>LIMSI</td>
<td>University of Avignon</td>
</tr>
<tr>
<td>CMU (JAVELIN)</td>
<td>MIT</td>
<td>U. Illinois, U-C</td>
</tr>
<tr>
<td>Chinese Acad. of Sciences</td>
<td>MITRE</td>
<td>University of Iowa</td>
</tr>
<tr>
<td>CL Research</td>
<td>Nat’l U. Singapore (Lee)</td>
<td>University of Limerick</td>
</tr>
<tr>
<td>Columbia U.</td>
<td>Nat’l U. Singapore (PRIS)</td>
<td>University of Michigan</td>
</tr>
<tr>
<td>Fudan University</td>
<td>NTT Commun. Science Labs</td>
<td>University of Pisa</td>
</tr>
<tr>
<td>IBM (Ittycheriah)</td>
<td>Pohang U. of Sci. &amp; Tech.</td>
<td>University of Sheffield</td>
</tr>
<tr>
<td>IBM (Prager)</td>
<td>Syracuse University</td>
<td>U. So. California, ISI</td>
</tr>
<tr>
<td>InsightSoft-M</td>
<td>Tokyo U. of Science</td>
<td>University of Waterloo</td>
</tr>
<tr>
<td>ITC-irst</td>
<td>Universite d’Angers</td>
<td>University of York</td>
</tr>
<tr>
<td></td>
<td>Universite de Montreal</td>
<td></td>
</tr>
</tbody>
</table>

34 groups:  
66 main task runs  
9 list task runs from 5 groups
Data

- New AQUAINT document set
  - approximately 3 gb of text
  - approximately 1,033,000 articles

- Questions taken from MSNSearch and AskJeeves logs
  - no definition questions
  - some spelling/grammatical errors remain
  - 46 questions with no known answer in docs
Motivation for Exact Answers

What river in the US is known as the Big Muddy?

• the Mississippi
• Known as Big Muddy, the Mississippi is the longest
• as Big Muddy, the Mississippi is the longest
• messed with. Known as Big Muddy, the Mississippi
• Mississippi is the longest river in the US
• the Mississippi is the longest river in the US,
• the Mississippi is the longest river (Mississippi)
• has brought the Mississippi to its lowest
• ipes. In Life on the Mississippi, Mark Twain wrote t
• Southeast; Mississippi; Mark Twain; officials began
• Known; Mississippi; US; Minnesota; Gulf Mexico
• Mud Island; Mississippi; "The;-- history; Memphis

Text REtrieval Conference (TREC)
Motivation for Exact Answers

- Text snippets masking important differences among systems
- Pinpointing precise extent of answer important to driving technology
  - not a statement that deployed systems should return only exact answers
  - exact answers may be important as component in larger language systems
Exact Answers

• Human assessors judged responses

  • **Wrong**: string does not contain a correct answer or answer is unresponsive
  • **Not Supported**: string contains a correct answer, but doc does not support that answer
  • **Not Exact**: string contains correct answer and doc supports it, but string contains too much (or too little) info
  • **Right**: string is exactly a correct answer that is supported by the doc
Exact Answer Guidelines

- most minimal response possible not the only exact answer
  - e.g., accept “Mississippi river” for What is the longest river in the United States?

- ungrammatical responses not exact
  - e.g., “in Mississippi” vs. “Mississippi in”

- justification is not exact
  - e.g., “At 2,348 miles the Mississippi river is the longest US river” is inexact
Distribution of Judgments

- 15,948 judgments across all questions
  - 12,639 79.3% Wrong
  - 505 3.2% Unsupported
  - 442 2.8% ineXact
  - 2,362 14.8% Right

- In general, systems can find extent of answer if they can find it at all
  - distribution skewed across systems
  - attempt to get exact answer sometimes caused units to be lost (so marked wrong)
Confidence-weighted Scoring

- Focus on getting systems to know when they have found a good answer
  - questions ranked by confidence in answer
  - compute score based on ranking

\[ \sum_{i=1}^{N} \frac{\text{number right to rank } i}{i/N} \]
Main Themes

• Many systems now using specific data sources for expected question types
  • name lists
  • gazetteers

• Web used by most systems, but in different ways
  • primary source of answer that is then mapped to corpus
  • one of several sources whose results are fused
  • place to validate answer found in corpus
Confidence Ranking

- Different approaches
  - most groups used the type of question as a factor
  - some systems that use scoring techniques to rank candidate answers also used score for ranking questions
  - few groups used training set to learn good feature set and corresponding weights, then applied classifier to test set
  - many groups ranked NIL questions last
Quality of the Evaluation

• Assessors opinions differ but evaluation is stable when using text snippets and MRR metric. Now?
  • exact answers
  • single response per question
  • confidence-weighted score

• Repeat stability study using multiple independent assessments
  • each question judged by 3 assessors
  • official evaluation based on adjudicated judgments
Assessors Continue to Disagree

- 50% of judgments where at least one judgment was not W had disagreements
- Of those, 33% involved disagreements between Right and ineXact
  - well-known granularity issue now reflected here
- For dates and quantities, disagreement among Wrong and ineXact
Comparative Results Still Stable

- Kendall $\tau$ scores between system rankings $> 0.9$
- Scores for rankings using adjudicated judgments $> 0.94$
- Number correct measure more stable than confidence-weighted score

<table>
<thead>
<tr>
<th></th>
<th>Adj</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Confidence weighted score</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.954</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.941</td>
<td>0.920</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.944</td>
<td>0.917</td>
<td>0.906</td>
</tr>
<tr>
<td><strong>Number correct</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.958</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.949</td>
<td>0.933</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.960</td>
<td>0.944</td>
<td>0.926</td>
</tr>
</tbody>
</table>
CWS Emphasizes Ranking

![Graph showing the relationship between number right and confidence-weighted score. The graph illustrates the observed scores compared to the best and worst possible scenarios.](image-url)
Inherent Stability of CWS

Text REtrieval Conference (TREC)
Summary

• Major changes in TREC 2002
  - exact answers
    • working definition of exact answer ok
    • in general, systems can detect answer extent
  - confidence ranking
    • CWS puts large emphasis on proper ranking
    • evaluation results stable with large enough question set