Open Source Question Answer System

Advisor
Dr. Chris Pollett
Committee
Dr. Mark Stamp
Dr. Robert Chun
By
Salil Shenoy
Outline

• Problem Statement
• Background on Question Answering Systems
• Implementation
• Experiments
• Conclusion
Problem Statement

• Improve an existing English Question Answering System in Yioop
• Implement a Hindi Question Answering System for Yioop
English Question Answer System

Justin Trudeau is Condescending & Smug Just Like Obama - YouTube
www.youtube.com/watch?v=2oJ8_n7q1M Possible Answer: condescending smug Words: trudeau


Harper blasts Trudeau's economic plan as all 'unicorns and rainbows' - The Globe and Mail

Harper blasts Trudeau's economic plan as all 'unicorns and ... Canada's Prime Minister and Conservative leader Stephen ... Canada's Prime Minister and Conservative leader Stephen ... Stephen Harper is calling Justin Trudeau's economic plan

Trudeau and Cameron meet at G20 summit | Watch News Videos Online
globalnews.ca/video/2340452/hud...0-summit Possible Answer: the prime minister Words: trudeau

Trudeau and Cameron meet at G20 summit. Watch ... Contests Live. Trudeau and Cameron meet at G20 summit. Prime ... Minister Justin Trudeau and U.K. Prime Minister David Cameron Canceled.

Meet Justin Trudeau's new Liberal cabinet - Politics - CBC News
www.cbc.ca/news/politics/trudeau....23304176 Possible Answer: canada s 23rd Words: reuters wattle

Meet Justin Trudeau's new Liberal cabinet - Politics - CBC ... CBC News Politics. Trudeau's first day as prime minister 4:34. Justin Trudeau signals new style on 1st day as

Justin Trudeau has a message to young voters -- "voting matters" | Watch News Videos Online
globalnews.ca/video/2172013/just....matters Possible Answer: the prime minister Words: canada

Justin Trudeau has a message to young voters voting ... at St. Canada. Justin Trudeau has a message to young voters ... Liberal leader Justin Trudeau sat down with ET Canada s
Hindi Search in Yioop...
What type of questions are asked...

Query Log Analysis

- Information: 54%
- Navigation: 35%
- Other: 11%
Existing Question Answer Systems...

Who directed Gone with the Wind

Gone with the Wind (1939) was directed by Victor Fleming, George Cukor, and Sam Wood.

I know about one more movie called "Gone With The Wind": Gone with the Wind (1954)

Source: The Internet Movie Database

Poem on Mahatma Gandhi in Hindi

This is a short and simple poem for...

1 Washington Sq
San Jose, CA 95192

San José State University, Address
Question Answer Systems

• Types of Question Answering Systems
  • Closed Domain
    eg: BASEBALL, LUNAR, IRSL etc.
  • Open Domain
    eg: START, Google, etc.
Question Answer Systems Paradigms

• Text Based / IR Based
• Knowledge Based
• Hybrid
Part of Speech Tagging

• Approaches to PoS tagging,
  • Rule Based (Brill Tagger)
  • Machine Learning
• For our system, implemented variant of Brill tagger for Hindi PoS tagging
Part of Speech Tagging

• Input is sentence processed, stemmed, n-gram, stop word and punctuation removed
• Terms present in the lexicon in database are tagged and words not found are tagged as unknown
• PoS tagging rules are applied for unknown words
PoS tagging Example

• Example:

• महामा गाँधी का जम 2 अक्टूबर को हुआ

• महामा ~ NN गाँधी ~ NN का ~ IN जम ~ NN 2 ~ QT अक्टूबर ~ NN हुआ ~ VB
Parse Tree Generation

- For Hindi, we use following grammar rules to extract triplets

\[ S: \{ \langle NP \rangle \langle PP \rangle \langle VP \rangle \} \]
\[ NP: \{ \langle JJ \mid NN^* \rangle \}^+ \]
\[ PP: \{ \langle IN \mid JJ \mid NN \mid PP \rangle \}^+ \]
\[ VP: \{ \langle VB^* \rangle \langle IN \mid NP \rangle \} \]
महामा गाँधी का जन्म २ अक्टूबर हुआ
Triplet Extraction

• Input is the parse tree
• Two types of triplets
  • Concise
  • Raw
• For Hindi, triplet is -> [Subject – Object - Verb]
Triplet Extraction

• For parse tree we saw on the previous slide the triplets and the corresponding answers are as follows,

  • [महामा गाँधी - qqq - हुआ ] => जन्म 2 अक्टूबर
  • [qqq - जन्म 2 अक्टूबर - हुआ ] => महामा गांधी
  • [महामा गांधी - जन्म 2 अक्टूबर - qqq] => हुआ
PoS tagging Example

• Example:

  • नरेंद्र मोदी भारत के कतने प्रधानमंत्री हैं
  • नरेंद्र ~ NN मोदी ~ NN भारत~NNP के ~ IN प्रधानमंत्री ~ NN हैं ~ VB
नरेंद्र मोदी भारत प्रधानमंत्री हैं
• For parse tree we saw on the previous slide the triplets and the corresponding answers are as follows,

  • [नरेंद्र मोदी भारत - qqq - हैं] => प्रधानमंत्री
  • [qqq - प्रधानमंत्री - हैं] => नरेंद्र मोदी भारत
  • [नरेंद्र मोदी भारत - प्रधानमंत्री - qqq] => हैं
Named Entities

• Named entity recognition, one of the major component of a QA system
• Ways to recognizing entities in text...
  • Maintain a handwritten dictionary of entities
  • Use automated methods for eg. Using Wikipedia to recognize entities.
Named Entities

• User can add, delete, view entities for a selected locale
• User can view entities for a locale

<table>
<thead>
<tr>
<th>Entity Name</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>Salil Shenoy</td>
<td>Delete</td>
</tr>
<tr>
<td>College of Science</td>
<td>Delete</td>
</tr>
<tr>
<td>Pune</td>
<td>Delete</td>
</tr>
<tr>
<td>New Jersey</td>
<td>Delete</td>
</tr>
<tr>
<td>Manchester</td>
<td>Delete</td>
</tr>
<tr>
<td>London</td>
<td>Delete</td>
</tr>
<tr>
<td>Family</td>
<td>Delete</td>
</tr>
</tbody>
</table>
Experiments

• We created an index for English and another for Hindi by setting up a crawl on wikipedia pages.

• To setup a crawl in Yioop, under the Manage Crawls Tab, click on Options, add the seed sites

• For hindi, I used https://hi.wikipedia.org/ and domains to crawl as ‘co.in’ and ‘in’

• For english https://en.wikipedia.org/ and domains to crawl as ‘com’
Experiments

Yioop without Hindi Q/A

Yioop With Hindi Q/A
Before / After Hindi Q/A integration

Yioop Performance with and without Q/A System

Time in Seconds

Mahatma Gandhi
Albert Einstein
India
Generational Gap

Without Q/A
With Q/A
Part of Speech tagger Performance
English vs Hindi Q/A

• I used 4 topic to evaluate the retrieval efficiency of the system.
• For a given topic, I asked the same set of questions to English and Hindi systems.
• For eg:
  “Who was Mahatma Gandhi?”
  “महात्मा गांधी कौन थे”
  “Who is Albert Einstein?”
  “अल्बर्ट आइंस्टीन कौन थे”
• The accuracy is evaluated by comparing retrieved answer with known answer dataset.
Average Precision Score
Reciprocal Rank

<table>
<thead>
<tr>
<th>Query</th>
<th>Answers</th>
<th>Correct Response</th>
<th>Rank</th>
<th>Reciprocal Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>महात्मा गांधी कौन थे</td>
<td>महान स्वतंत्रता सेनानी, उच्च शिक्षा, स्वतंत्रता संग्राम</td>
<td>महान स्वतंत्रता सेनानी</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>अल्बर्ट आइंस्टीन कौन है</td>
<td>अमेरिकी इंजीनियर, सफलता, डॉक्टर, सैद्धांतिक भौतिकविद्</td>
<td>सैद्धांतिक भौतिकविद्</td>
<td>4</td>
<td>1/4</td>
</tr>
<tr>
<td>गुरुत्वाकर्षण लेस क्या है</td>
<td>अभियात्री, ल्यूटन, क्यूटन, प्रभाव, प्रभाव</td>
<td>प्रभाव</td>
<td>15</td>
<td>1/15</td>
</tr>
</tbody>
</table>
Accuray of Question Answering

• I created a set of 25 questions, corresponding set of answers known to be true
• I asked the questions from the question set and compared the retrieved answer with known answer set to evaluate efficiency
Accuracy English Q/A
Accuracy Hindi Q/A

Accuracy of Hindi Q/A System

- Who
- Where
- Which
- When
- How

Wrong Responses
Correct Responses
Questions ?