“Question answering using structured and unstructured data”

Denis Savenkov,
PhD student @ Emory University

WSDM’16 Doctoral Consortium
Percentage of question search queries is growing\(^\text{[1]}\)

\(^{[1]}\) “Questions vs. Queries in Informational Search Tasks”, Ryen W. White et al, WWW 2015
QA system architecture

1. Question Analysis
2. Candidate generation
3. Evidence retrieval
4. Answer selection

Question → Answer
Data Sources

<table>
<thead>
<tr>
<th>City</th>
<th>State</th>
<th>Population</th>
<th>Area</th>
<th>Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>New York</td>
<td>New York</td>
<td>8,486,768</td>
<td>783.45</td>
<td>-0.89%</td>
</tr>
<tr>
<td>Los Angeles</td>
<td>California</td>
<td>3,978,664</td>
<td>485.24</td>
<td>-3.26%</td>
</tr>
<tr>
<td>Chicago</td>
<td>Illinois</td>
<td>2,695,962</td>
<td>226.34</td>
<td>+0.48%</td>
</tr>
<tr>
<td>Phoenix</td>
<td>Arizona</td>
<td>1,320,662</td>
<td>162.25</td>
<td>+0.96%</td>
</tr>
</tbody>
</table>

San Francisco, California

<table>
<thead>
<tr>
<th>City</th>
<th>Population</th>
<th>Area</th>
<th>Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>San Francisco</td>
<td>803,400</td>
<td>46.94</td>
<td>-0.25%</td>
</tr>
</tbody>
</table>

Structured data

Unstructured data

Semi-structured data
## Structured and Unstructured data sources for QA

<table>
<thead>
<tr>
<th></th>
<th><strong>Unstructured data</strong></th>
<th><strong>Structured data</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>factoid questions</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Text collections</strong></td>
<td>+ easy to match against question text</td>
<td>+ aggregate all the information about entities</td>
</tr>
<tr>
<td></td>
<td>+ cover a variety of different information types</td>
<td>+ allow complex queries over this data using special languages (e.g. SPARQL)</td>
</tr>
<tr>
<td></td>
<td>- each text phrase encodes a limited amount of information about mentioned entities</td>
<td>- hard to translate natural language questions into special query languages</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- KBs are incomplete (missing entities, facts and properties)</td>
</tr>
<tr>
<td><strong>non-factoid questions</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Text collections</strong></td>
<td>+ contain relevant information to a big chunk of user needs</td>
<td>+ easy to find a relevant answer by matching the corresponding questions</td>
</tr>
<tr>
<td></td>
<td>- hard to extract semantic meaning of a paragraph to match against the question (lexical gap)</td>
<td>- cover a smaller subset of user information needs</td>
</tr>
<tr>
<td><strong>Knowledge Bases (KB)</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Knowledge Bases (KB)**
- + aggregate all the information about entities
- + allow complex queries over this data using special languages (e.g. SPARQL)
- - hard to translate natural language questions into special query languages
- - KBs are incomplete (missing entities, facts and properties)
Combining text and KB for factoid QA
Using the structure of web documents for non-factoid QA
Research Questions

1. What types of questions can be answered using text, KB or a combination of both?

2. How does semantic annotation of unstructured data compare to information extraction for question answering?
   - information extraction for KB construction vs open information extraction vs unstructured data annotation

3. How does a combination of structured and unstructured data sources improve each of the main QA system components: question analysis, candidate generation, evidence extraction and answer selection?
Experiments: factoid QA

➔ Datasets
  ○ TREC QA
  ○ WebQuestions
  ○ new factoid QA dataset
    ✓ derived from Yahoo! Answers CQA archive
    ✓ filter factoid non-subjective questions (heuristics + mechanical turk)
    ✓ mturk to label answers (KB entities, mentioned in CQA user answer)
    ✓ Examples:
      • What is the name of Sponge Bob’s pet snail?
      • What is the boot-shaped country found in Southern Europe?
      • Who was the first model to ever appear on the cover of Sports Illustrated?

➔ Comparison
  ○ state-of-the-art KBQA systems (WebQuestions benchmark)
  ○ Open QA approach based on OpenIE extractions
  ○ collection-based and hybrid techniques: AskMSR+, QuASE
Experiments: non-factoid QA

- TREC LiveQA shared task (http://trec-liveqa.org/): real user question answering in real-time
  - systems need to respond to questions posted to Yahoo! Answers

- Comparison:
  - baseline system developed for last year shared task
  - participants of TREC LiveQA 2016
Expected results

1. new factoid question answering dataset
2. new approach to combining unstructured text and structured KB data, that:
   ✓ improves QA precision
   ✓ improves recall, including questions that could only be answered with a combination of text and KB data
3. new system for non-factoid question answering with improved performance due to better utilization of the information provided in web documents

Open Questions

→ How to incorporate other available sources of factual information, including semi-structured, e.g. tables, diagrams, etc?
→ Non-factoid questions often include unique context information, that makes reusing information non effective. How a system can generate the answer given all potentially useful extracted information?
→ How to construct KBs for non-factoid information needs?
Thank you!

How to say thank you in different languages (German, Italian, ....)?

Answers

Best Answer: spanish - gracias
french - merci
german - danke
italian - grazie
portuguese - obrigado
norwegian - takk
mandarin chinese - xie xie
finnish - kiitos
danish - tak
hebrew - todah
russian - spasibo
swahili - asante
swedish - tack
polish - dziękuję
korean - kamsa hamnida
japanese - arigato
indonesian - terima kasih
greek - efaristo
hindi - sukna

Krysta 10 years ago

thumbs up 3  thumbs down 0