Sustainable Questions
Determining the Expiration Date of Answers

Bart de Goede, Anne Schuth and Maarten de Rijke

ISLA, University of Amsterdam
Reusing Answers in CQA

- "Who designed the Eiffel Tower?"
  - Alexandre Gustave Eiffel
- "Who is the prime minister of the UK?"
  - now: David Cameron
  - before: Gordon Brown
- ...

Some questions are sustainable and can readily be reused, others not
Sustainability

A question is sustainable if the answer to that question is independent of the point in time the question is asked.
Our Approach
to determining what makes a question sustainable

1) **Cluster similar questions**

   questions are *similar* if they would be *satisfied by the same information* when *asked at the same time*

2) **Compare answers in each cluster**

   *answers* to *similar questions that do not change over time* indicate *sustainable questions*

3) **Classify clusters as sustainable**
1) Cluster Similar Questions

- Questions are
  - short *(175 characters)*,
  - phrased an variety of ways,
  - contain noise *(spelling errors)*,
  - contain lot’s of function word *(why, who, where..)*
1) Cluster Similar Questions

- Latent Semantic Analysis (LSA) or Latent Dirichlet Allocation (LDA)
  - topic modeling
  - distance metric
- Locality Sensitive Hashing (LSH)
  - near duplicate detection
1) Cluster Similar Questions

- Yahoo! Answers Comprehensive Questions and Answers version 1.0
  - 3.4M questions

Results

- Accuracy on 559 manually labeled questions pairs

<table>
<thead>
<tr>
<th>algorithm</th>
<th>10K</th>
<th>100K</th>
<th>all</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA</td>
<td>0.435</td>
<td>0.500</td>
<td>-</td>
</tr>
<tr>
<td>LSA</td>
<td>0.706</td>
<td>0.638</td>
<td>-</td>
</tr>
<tr>
<td>LSH_{16bits}</td>
<td>0.472</td>
<td>0.484</td>
<td>0.500</td>
</tr>
<tr>
<td>LSH_{24bits}</td>
<td>0.465</td>
<td>0.502</td>
<td>0.495</td>
</tr>
<tr>
<td>LSH_{32bits}</td>
<td>0.512</td>
<td>0.514</td>
<td>0.509</td>
</tr>
<tr>
<td>LSH_{40bits}</td>
<td>0.523</td>
<td>0.537</td>
<td>0.542</td>
</tr>
</tbody>
</table>

...
Our Approach
to determining what makes a question sustainable

1) **Cluster similar questions**

Questions are *similar* if they would be *satisfied by the same information* when asked at the same time.

2) **Compare answers in each cluster**

*Answers* to similar questions that do *not change over time* indicate *sustainable questions*.

3) **Classify clusters as sustainable**
2) Compare answers in each cluster

![Diagram showing cumulative cosine distance and linear fitted line over time]

- Cumulative cosine distance
- Linear fitted line
2) Compare answers in each cluster
Our Approach to determining what makes a question sustainable

1) **Cluster similar questions**
   questions are similar if they would be satisfied by the same information when asked at the same time

2) **Compare answers in each cluster**
   answers to similar questions that do not change over time indicate sustainable questions

3) **Classify clusters as sustainable**
3) Classify clusters as sustainable

- We train a C4.5 tree classifier and obtain the following results:

<table>
<thead>
<tr>
<th>feature set</th>
<th>accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>change per question</td>
<td>66.9%</td>
</tr>
<tr>
<td>change over time</td>
<td>86.0%</td>
</tr>
<tr>
<td>semanticized change over time</td>
<td>75.3%</td>
</tr>
<tr>
<td>time to answer</td>
<td>89.3%</td>
</tr>
<tr>
<td>change/time combination</td>
<td>91.5%</td>
</tr>
</tbody>
</table>
Open issues, opportunities

- Clustering?
  - On the answers (twice as long as questions), and see whether questions change
  - Or both and look where cluster on answers and questions agree
- Retrieval approach
- Sustainability
  - Richer representation of answers
  - More features?
- Evaluation
  - Task based: do we improve precision?
Thank you