CRQA: Crowd-powered Real-time Automated Question Answering System

Denis Savenkov  
Emory University  
dsavenk@emory.edu

Eugene Agichtein  
Emory University  
eugene@mathcs.emory.edu

HCOMP, Austin, TX  
October 31, 2016
Volume of question search queries is growing[1]

And more and more of this searches are happening on mobile

![Image: US Local Search Query Volume, Mobile vs. Desktop, 2014-2019](chart.png)

**Source:** BIA/Kelsey as cited in company blog, May 14, 2015
Mobile Personal Assistants are popular
Automatic Question Answering works relatively well for some questions
... but not sufficiently well for many other questions
... when there is no answer, digging into “10 blue links” is even harder on mobile devices
It is important to improve question answering for complex user information needs.
Goal of **TREC LiveQA** shared task is to advance research into **answering real user questions in real time**

Question Answering System

https://sites.google.com/site/trecliveqa2016/
LiveQA Evaluation Setup

- **1: Bad** - contains no useful information
- **2: Fair** - marginally useful information
- **3: Good** - partially answers the question
- **4: Excellent** - fully answers the question

Answers are pooled and judged by NIST assessors.
LiveQA 2015: Even the best system returns a fair or better answer only for ~50% of the questions!

<table>
<thead>
<tr>
<th></th>
<th>Avg score (0-3)</th>
<th>% questions with fair or better answer</th>
<th>% questions with excellent answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best system</td>
<td>1.08</td>
<td>53.2</td>
<td>19.0</td>
</tr>
</tbody>
</table>
1. Search data sources
   a. CQA archives
      i. Yahoo! Answers
      ii. Answers.com
      iii. WikiHow
   b. Web search API

2. Extract candidates and their context
   a. Answers to retrieved questions
   b. Content blocks from regular web pages

3. Represent candidate answers with a set of features
4. Rank them using LambdaMART model
5. Return the top candidate as the answer
Common Problem: Automatic systems often return an answer about the same topic, but irrelevant to the question.

Throwback to when my friends hamster ate my hamster and then my friends hamster died because she forgot to feed it karma.
Incorporate crowdsourcing to assist an automatic real-time question answering system

Or: combine human insight and automatic QA with machine learning
Existing research

✓ “Direct answers for search queries in the long tail” by M.Bernstein et al, 2012
  ○ Offline crowdsourcing of answers for long-tail search queries

✓ “CrowdDB: answering queries with crowdsourcing” by M.Franklin et al, 2011
  ○ Using crowd to perform complex operations in SQL queries

✓ “Answering search queries with crowdseacher” by A.Bozzon et al, 2012
  ○ Answering queries using social media

✓ “Dialog system using real-time crowdsourcing and twitter large-scale corpus” by F. Bessho et al, 2012
  ○ Real-time crowdsourcing as a backup plan for dialog

✓ “Chorus: A crowd-powered conversational assistant” by W.Lasecki, 2013
  ○ Real-time chatbot powered by crowdsourcing

... and many other works
Research Questions

○ **RQ1.** Can crowdsourcing be used to improve the performance of a near real-time automatic question answering system?
Research Questions

- **RQ1.** Can crowdsourcing be used to improve the performance of a near real-time automatic question answering system?

- **RQ2.** What kind of contributions from crowd workers can help improve automatic question answering and what is the relative impact of different types of feedback to the overall question answering performance?
Research Questions

○ **RQ1.** Can crowdsourcing be used to improve the performance of a near real-time automatic question answering system?

○ **RQ2.** What kind of contributions from crowd workers can help improve automatic question answering and what is the relative impact of different types of feedback to the overall question answering performance?

○ **RQ3.** What are the trade-offs in performance, cost, and scalability of using crowdsourcing for real-time question answering?
CRQA: Integrating crowdsourcing with automatic QA system

1. After receiving a question, it is forwarded to the crowd
2. Can start working on the answer, if possible
3. When system ranks candidates, top-7 are pushed to workers for rating
4. Rated human and automatically generated answers are returned
5. System re-rank them based on all available information
6. Top candidate is returned as the answer
We used the retainer model for real-time crowdsourcing
UI for crowdsourcing answers and ratings
Heuristic answer re-ranking (during TREC LiveQA)

\[
\text{Answer candidate} \quad \text{Answer candidate} \quad \text{Answer candidate} \quad \text{Answer candidate}
\]

\[\text{\textgreater sort answers -k crowd\_rating}\]

\[
\text{if top candidate rating} > 2.5 \quad \text{or} \quad \text{no crowd generated candidates}
\]

\[
\text{False} \quad \text{True}
\]

\[\text{return longest crowd generated candidate} \quad \text{return top candidate}\]
CRQA uses a learning-to-rank model to re-rank answer candidates.

Sort answers by crowd rating:

- If the top candidate's rating is greater than 2.5
  - Return the top candidate.
- If there are no crowd-generated candidates
  - Return the longest crowd-generated candidate.

Otherwise, return the longest crowd-generated candidate.
CRQA uses a learning-to-rank model to re-rank

Answer re-ranking model features:
- answer source
- initial rank/score
- # crowd ratings
- min, median, mean, max crowd rating

- Offline crowdsourcing to get ground-truth labels
- Included Yahoo!Answers community response, crawled 2 days after challenge
- Trained GBRT model, 10-fold cross validation

final answer
Evaluation
Evaluation setup

Methods compared:
➢ Automatic QA
➢ CRQA (heuristic): re-ranking by crowdsourced score
➢ CRQA (LTR): re-ranking using a learning-to-rank model
➢ Yahoo! Answers (crawled 2 days later)

Metrics:
➢ avg-score: average answer score over all questions
➢ avg-prec: average answer score
➢ success@i+: fraction of questions with answer score ≥ i
➢ precision@i+: fraction of answers with score ≥ i
## Dataset

- 1,088 questions from LiveQA 2016 run
- Top 7 system and crowd-generated answers
- Answer quality labelling on a scale from 1 to 4
  - offline
  - also using crowdsourcing (different workers)

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of questions received</td>
<td>1,088</td>
</tr>
<tr>
<td>Number of MTurk 15 minutes assignments completed</td>
<td>889</td>
</tr>
<tr>
<td>Average number of questions per assignment</td>
<td>11.44</td>
</tr>
<tr>
<td>Total cost per question</td>
<td>$0.81</td>
</tr>
<tr>
<td>Avg number of answers provided by workers per question</td>
<td>1.25</td>
</tr>
<tr>
<td>Average number of ratings per answer</td>
<td>6.25</td>
</tr>
</tbody>
</table>
# Main Results

<table>
<thead>
<tr>
<th>Method</th>
<th>avg-score</th>
<th>avg-prec</th>
<th>s@2+</th>
<th>s@3+</th>
<th>s@4+</th>
<th>p@2+</th>
<th>p@3+</th>
<th>p@4+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automatic QA</td>
<td>2.321</td>
<td>2.357</td>
<td>0.69</td>
<td>0.30</td>
<td>0.02</td>
<td>0.71</td>
<td>0.30</td>
<td>0.03</td>
</tr>
<tr>
<td>CRQA: (heuristic)</td>
<td>2.416</td>
<td>2.421</td>
<td>0.75</td>
<td>0.32</td>
<td>0.03</td>
<td>0.75</td>
<td>0.32</td>
<td>0.03</td>
</tr>
<tr>
<td>CRQA (LTR)</td>
<td>2.550</td>
<td>2.556</td>
<td>0.80</td>
<td>0.40</td>
<td>0.03</td>
<td>0.80</td>
<td>0.40</td>
<td>0.03</td>
</tr>
<tr>
<td>Yahoo! Answers</td>
<td>2.229</td>
<td>2.503</td>
<td>0.66</td>
<td>0.37</td>
<td>0.04</td>
<td>0.74</td>
<td>0.42</td>
<td>0.05</td>
</tr>
</tbody>
</table>
Crowdsourcing improves performance of automatic QA system

<table>
<thead>
<tr>
<th>Method</th>
<th>avg-score</th>
<th>avg-prec</th>
<th>s@2+</th>
<th>s@3+</th>
<th>s@4+</th>
<th>p@2+</th>
<th>p@3+</th>
<th>p@4+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automatic QA</td>
<td>2.321</td>
<td>2.357</td>
<td>0.69</td>
<td>0.30</td>
<td>0.02</td>
<td>0.71</td>
<td>0.30</td>
<td>0.03</td>
</tr>
<tr>
<td>CRQA: (heuristic)</td>
<td>2.416</td>
<td>2.421</td>
<td>0.75</td>
<td>0.32</td>
<td>0.03</td>
<td>0.75</td>
<td>0.32</td>
<td>0.03</td>
</tr>
<tr>
<td>CRQA (LTR)</td>
<td>2.550</td>
<td>2.556</td>
<td>0.80</td>
<td>0.40</td>
<td>0.03</td>
<td>0.80</td>
<td>0.40</td>
<td>0.03</td>
</tr>
<tr>
<td>Yahoo! Answers</td>
<td>2.229</td>
<td>2.503</td>
<td>0.66</td>
<td>0.37</td>
<td>0.04</td>
<td>0.74</td>
<td>0.42</td>
<td>0.05</td>
</tr>
</tbody>
</table>
Learning-to-rank model allows to more effectively combine all available signals and return a better answer

<table>
<thead>
<tr>
<th>Method</th>
<th>avg-score</th>
<th>avg-prec</th>
<th>s@2+</th>
<th>s@3+</th>
<th>s@4+</th>
<th>p@2+</th>
<th>p@3+</th>
<th>p@4+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automatic QA</td>
<td>2.321</td>
<td>2.357</td>
<td>0.69</td>
<td>0.30</td>
<td>0.02</td>
<td>0.71</td>
<td>0.30</td>
<td>0.03</td>
</tr>
<tr>
<td>CRQA: (heuristic)</td>
<td>2.416</td>
<td>2.421</td>
<td>0.75</td>
<td>0.32</td>
<td>0.03</td>
<td>0.75</td>
<td>0.32</td>
<td>0.03</td>
</tr>
<tr>
<td><strong>CRQA (LTR)</strong></td>
<td><strong>2.550</strong></td>
<td><strong>2.556</strong></td>
<td><strong>0.80</strong></td>
<td><strong>0.40</strong></td>
<td><strong>0.03</strong></td>
<td><strong>0.80</strong></td>
<td><strong>0.40</strong></td>
<td><strong>0.03</strong></td>
</tr>
<tr>
<td>Yahoo! Answers</td>
<td>2.229</td>
<td>2.503</td>
<td>0.66</td>
<td>0.37</td>
<td><strong>0.04</strong></td>
<td>0.74</td>
<td><strong>0.42</strong></td>
<td><strong>0.05</strong></td>
</tr>
</tbody>
</table>
CRQA reaches the quality of community responses on Yahoo! Answers

<table>
<thead>
<tr>
<th>Method</th>
<th>avg-score</th>
<th>avg-prec</th>
<th>s@2+</th>
<th>s@3+</th>
<th>s@4+</th>
<th>p@2+</th>
<th>p@3+</th>
<th>p@4+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automatic QA</td>
<td>2.321</td>
<td>2.357</td>
<td>0.69</td>
<td>0.30</td>
<td>0.02</td>
<td>0.71</td>
<td>0.30</td>
<td>0.03</td>
</tr>
<tr>
<td>CRQA: (heuristic)</td>
<td>2.416</td>
<td>2.421</td>
<td>0.75</td>
<td>0.32</td>
<td>0.03</td>
<td>0.75</td>
<td>0.32</td>
<td>0.03</td>
</tr>
<tr>
<td>CRQA (LTR)</td>
<td><strong>2.550</strong></td>
<td><strong>2.556</strong></td>
<td><strong>0.80</strong></td>
<td><strong>0.40</strong></td>
<td>0.03</td>
<td><strong>0.80</strong></td>
<td>0.40</td>
<td>0.03</td>
</tr>
<tr>
<td>Yahoo! Answers</td>
<td>2.229</td>
<td>2.503</td>
<td>0.66</td>
<td>0.37</td>
<td><strong>0.04</strong></td>
<td>0.74</td>
<td><strong>0.42</strong></td>
<td><strong>0.05</strong></td>
</tr>
</tbody>
</table>
... and it has much better coverage

<table>
<thead>
<tr>
<th>Method</th>
<th>avg-score</th>
<th>avg-prec</th>
<th>s@2+</th>
<th>s@3+</th>
<th>s@4+</th>
<th>p@2+</th>
<th>p@3+</th>
<th>p@4+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automatic QA</td>
<td>2.321</td>
<td>2.357</td>
<td>0.69</td>
<td>0.30</td>
<td>0.02</td>
<td>0.71</td>
<td>0.30</td>
<td>0.03</td>
</tr>
<tr>
<td>CRQA: (heuristic)</td>
<td>2.416</td>
<td>2.421</td>
<td>0.75</td>
<td>0.32</td>
<td>0.03</td>
<td>0.75</td>
<td>0.32</td>
<td>0.03</td>
</tr>
<tr>
<td><strong>CRQA (LTR)</strong></td>
<td><strong>2.550</strong></td>
<td><strong>2.556</strong></td>
<td><strong>0.80</strong></td>
<td><strong>0.40</strong></td>
<td><strong>0.03</strong></td>
<td><strong>0.80</strong></td>
<td><strong>0.40</strong></td>
<td><strong>0.03</strong></td>
</tr>
<tr>
<td>Yahoo! Answers</td>
<td>2.229</td>
<td>2.503</td>
<td>0.66</td>
<td>0.37</td>
<td><strong>0.04</strong></td>
<td>0.74</td>
<td><strong>0.42</strong></td>
<td><strong>0.05</strong></td>
</tr>
</tbody>
</table>
Both worker answers and ratings make an equal contribution to the answer quality improvements

<table>
<thead>
<tr>
<th>Method</th>
<th>avg-score</th>
<th>avg-prec</th>
<th>s@2+</th>
<th>s@3+</th>
<th>s@4+</th>
<th>p@2+</th>
<th>p@3+</th>
<th>p@4+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automatic QA</td>
<td>2.321</td>
<td>2.357</td>
<td>0.69</td>
<td>0.30</td>
<td>0.02</td>
<td>0.71</td>
<td>0.30</td>
<td>0.03</td>
</tr>
<tr>
<td>CRQA (LTR)</td>
<td>2.550</td>
<td>2.556</td>
<td>0.80</td>
<td>0.40</td>
<td>0.03</td>
<td>0.80</td>
<td>0.40</td>
<td>0.03</td>
</tr>
<tr>
<td>no worker answers</td>
<td>2.432</td>
<td>2.470</td>
<td>0.75</td>
<td>0.35</td>
<td>0.03</td>
<td>0.76</td>
<td>0.35</td>
<td>0.03</td>
</tr>
<tr>
<td>no worker ratings</td>
<td>2.459</td>
<td>2.463</td>
<td>0.76</td>
<td>0.35</td>
<td>0.03</td>
<td>0.76</td>
<td>0.36</td>
<td>0.03</td>
</tr>
</tbody>
</table>
Crowdsourcing helps to improve empty and low quality answers

Less un-answer questions thanks to worker answers
Ratings help with "bad" answers
Yahoo! Answers have both higher percentage of excellent and missing and low quality answers
Crowdsourced answers are especially good for general knowledge questions

Is it bad not wanting to visit your family? It’s not bad. Just be honest with them. They may be upset but they should understand.

What is a good remedy/medicine for stomach aches? Specifically ones caused by stress or anxiety?

Chamomile tea should help
But less effective for questions which require domain expertise

Arts & Humanities
Pets
Home & Garden

Health
Travel

One of the hardest for automatic systems

What was Pol Pot's vision of an agrarian society?

What is the average time for recovery of a complete Scapholunate (scaphoid) ligament tear in my wrist?

I injured my wrist while at work. I was taken to the ER that night and diagnosed with a scaphoid dislocation. I saw a hand surgeon two days later, and had surgery the next week (all within about 10 days of the initial injury). The scope showed a complete tear of the ligament, at which time he performed the... show more

1 following  2 answers
Ok, but what about the costs?

$0.81 per question is a lot of money
Half of the overall improvements can be achieved with only 3 workers per question (30% of cost)

(a) avg-score: Average score per question

(b) avg-prec: Average score per answer (ignoring non-answered questions)
Limitations and Future work

➢ Limitations
  ○ Fixed and uniform load for the system over 24 hours
    - Need variable size pool of workers based on the current load

➢ Ideas
  ○ Allocate crowdsourcing resources based on expected performance of the automated system
  ○ Use other types of feedback:
    - Search query generation
    - Key phrases to look for the the answer
    - ...
  ○ Online learning from crowd feedback
  ○ Cost optimization
    - Decide which feedback, in what amount and when to request
We conducted large scale experiments on real user questions, which showed:

○ Crowdsourcing helps for real-time QA
  ➢ Workers can contribute answers and rate candidates
  ➢ Humans can immediately reject off-topic candidates

○ Answers from our system are often even preferred to community answers
  ➢ Which are collected 2 days after
  ➢ With 20% of the questions were still unanswered by the community

Thank you!
We conducted large scale experiments on real user questions, which showed:

- Crowdsourcing helps for real-time QA
- Workers can contribute answers and rate candidates
- With 20% of the questions still unanswered

I'm graduating beginning 2017 and open for academic and industry research positions

Thank you!
It’s better to present candidates ordered by their predicted quality

Average answer score if presented in different order:

<table>
<thead>
<tr>
<th>Sorted by rank</th>
<th>2.539</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shuffled</td>
<td>2.508</td>
</tr>
</tbody>
</table>
Crowdsourced labels correlate well with NIST assessor scores ($\rho=0.52$)

- Workers prefer to give intermediate scores (2, 3), while NIST assessors gave more extreme scores (1 and 4)
- There is no significant difference in quality between groups with and without time pressure
Features [backup]

<table>
<thead>
<tr>
<th><strong>Answer statistics</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>— Length in chars, words and sentences</td>
</tr>
<tr>
<td>— Average number of words per sentence</td>
</tr>
<tr>
<td>— Fraction of non-alphanumeric characters</td>
</tr>
<tr>
<td>— Number of question marks</td>
</tr>
<tr>
<td>— Number of verbs</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Answer source</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>— Binary feature for each of the search verticals: Web, Yahoo! Answers, Answers.com, WikiHow.com</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>N-gram matches</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>— Cosine similarities using uni-, bi- and tri-gram representations of the question title and/or body, and answer text, topic or context</td>
</tr>
<tr>
<td>— The lengths of longest spans of matched terms between question title and/or body, and answer text, topic or context</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Information Retrieval score</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>— BM25 scores between question title and/or body, and answer text, topic or context</td>
</tr>
</tbody>
</table>