“Crowdsourcing for (almost) Real-time Question Answering”

Denis Savenkov  
Emory University  
dsavenk@emory.edu

Scott Weitzner  
Emory University  
sweitzn@emory.edu

Eugene Agichtein  
Emory University  
eugene@mathcs.emory.edu

NAACL’16 Human-Computer Question Answering Workshop
Percentage of question search queries is growing\cite{1}

\begin{itemize}
  \item \cite{1} “Questions vs. Queries in Informational Search Tasks”, Ryen W. White et al, WWW 2015
\end{itemize}
And more and more of this searches are happening on mobile.
Intelligent Assistants today

*amazon echo*

Always ready, connected, and fast. **Just ask.**
Automatic Question Answering works relatively well for simple factoid questions.
But for many more complex questions we still have to dig into the “10 blue links”
… which is much harder to do on mobile

web for ‘Why do we call Donald trump by his last name and Hillary Clinton by her first name?’:

Why do we call Donald Trump by his last name and Hillary ...
I just caught myself doing it and now I wonder why that’s the trend with men vs. women in politics? We never ... (maybe because Sarah is too common a name ...
https://answers.yahoo.com

Hillary Clinton Calls Donald Trump a RACIST... And Trump ...
Unlike Hillary Clinton, Trump has earned his money ... up to Hillary! Do you support Donald Trump? ... Donald will put her in her place...... we don't need ...
www.thepoliticalinsider.com

Donald Trump calls politics with Bill
TREC LiveQA Shared Task

→ **Goal:**
  - Answer real user questions in real-time
  - ✔ sampled Yahoo! Answers stream of questions
  - ✔ systems need to return the answer in under 1 minute
  - ✔ answers should be up to 1000 characters long
  - ✔ no restrictions on data sources

→ **Last year results**

<table>
<thead>
<tr>
<th>No.</th>
<th>Run</th>
<th>#Answered questions</th>
<th>avgScore (0-3)</th>
<th>succ@2+</th>
<th>succ@3+</th>
<th>succ@4+</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CMUOAQA</td>
<td>1064</td>
<td>1.081</td>
<td>0.532</td>
<td>0.359</td>
<td>0.190</td>
</tr>
<tr>
<td>2</td>
<td>ecmecs</td>
<td>994</td>
<td>0.677</td>
<td>0.367</td>
<td>0.224</td>
<td>0.086</td>
</tr>
<tr>
<td>3</td>
<td>NUDTMJD1</td>
<td>1041</td>
<td>0.670</td>
<td>0.353</td>
<td>0.210</td>
<td>0.107</td>
</tr>
<tr>
<td>4</td>
<td>monash-system2</td>
<td>1074</td>
<td>0.666</td>
<td>0.364</td>
<td>0.220</td>
<td>0.082</td>
</tr>
<tr>
<td>5</td>
<td>Yahoo-Exp1*</td>
<td>647</td>
<td>0.626</td>
<td>0.320</td>
<td>0.211</td>
<td>0.095</td>
</tr>
<tr>
<td>6</td>
<td>CLIP1</td>
<td>1079</td>
<td>0.615</td>
<td>0.326</td>
<td>0.204</td>
<td>0.086</td>
</tr>
<tr>
<td>7</td>
<td>emory-Out-of-mEmory</td>
<td>884</td>
<td>0.608</td>
<td>0.332</td>
<td>0.190</td>
<td>0.086</td>
</tr>
<tr>
<td>8</td>
<td>NUDTMJD3</td>
<td>1035</td>
<td>0.602</td>
<td>0.319</td>
<td>0.186</td>
<td>0.097</td>
</tr>
<tr>
<td>9</td>
<td>ECNU_ICA_2</td>
<td>1057</td>
<td>0.569</td>
<td>0.289</td>
<td>0.191</td>
<td>0.089</td>
</tr>
<tr>
<td>10</td>
<td>HIT_SCIR_QA_Grp</td>
<td>1086</td>
<td>0.522</td>
<td>0.291</td>
<td>0.168</td>
<td>0.063</td>
</tr>
<tr>
<td>11</td>
<td>ADAPT.DCU-system7</td>
<td>1087</td>
<td>0.444</td>
<td>0.290</td>
<td>0.121</td>
<td>0.034</td>
</tr>
<tr>
<td>12</td>
<td>RMIT1</td>
<td>1078</td>
<td>0.435</td>
<td>0.267</td>
<td>0.130</td>
<td>0.039</td>
</tr>
<tr>
<td>13</td>
<td>RMIT3</td>
<td>1082</td>
<td>0.415</td>
<td>0.251</td>
<td>0.126</td>
<td>0.038</td>
</tr>
<tr>
<td>14</td>
<td>NUDTMJP2</td>
<td>1025</td>
<td>0.391</td>
<td>0.228</td>
<td>0.120</td>
<td>0.043</td>
</tr>
<tr>
<td>15</td>
<td>RMIT2</td>
<td>1086</td>
<td>0.381</td>
<td>0.232</td>
<td>0.115</td>
<td>0.034</td>
</tr>
<tr>
<td>16</td>
<td>uwaterlooclarke-system4</td>
<td>1001</td>
<td>0.380</td>
<td>0.241</td>
<td>0.108</td>
<td>0.031</td>
</tr>
<tr>
<td>17</td>
<td>QU1</td>
<td>1082</td>
<td>0.256</td>
<td>0.163</td>
<td>0.070</td>
<td>0.023</td>
</tr>
<tr>
<td>18</td>
<td>DFKI-dfiqka</td>
<td>1058</td>
<td>0.211</td>
<td>0.152</td>
<td>0.049</td>
<td>0.010</td>
</tr>
<tr>
<td>19</td>
<td>CLIP3</td>
<td>805</td>
<td>0.144</td>
<td>0.102</td>
<td>0.034</td>
<td>0.008</td>
</tr>
<tr>
<td>20</td>
<td>CLIP2</td>
<td>1066</td>
<td>0.092</td>
<td>0.065</td>
<td>0.019</td>
<td>0.007</td>
</tr>
<tr>
<td>21</td>
<td>SCU</td>
<td>809</td>
<td>0.023</td>
<td>0.014</td>
<td>0.006</td>
<td>0.003</td>
</tr>
<tr>
<td>Avg.</td>
<td></td>
<td>1007</td>
<td>0.467</td>
<td>0.262</td>
<td>0.146</td>
<td>0.060</td>
</tr>
</tbody>
</table>
Crowdsourcing for real-time QA

➔ Idea:
  ○ use crowdsourcing to help a QA system answer user questions

➔ Types of feedback studied:
  ○ Worker generated answers
  ○ Ratings for answer candidates

➔ First we want to see if we can get reasonable data from crowd workers under time pressure
Research Questions

○ **RQ1.** Can crowdsourcing be used to judge the quality of answers to non-factoid questions under a time limit?

○ **RQ2.** Is it possible to use crowdsourcing to collect answers to real user questions under a time limit?

○ **RQ3.** How does the quality of crowdsourced answers to non-factoid questions compare to original CQA answers, and to automatic answers from TREC LiveQA systems?
Methodology: answer validation

100 questions from TREC LiveQA’15

3 answer from top-10 systems, labelled by NIST assessors

Task: rate the quality of answers

With 1 minute time limit
○ 3 workers per question
○ $0.05 per task

Without time limit
○ 3 workers per question
○ $0.05 per task
LiveQA Answer Quality Scale

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

Instructions:

1. Read the given question
2. Read each of the answers and assess its quality from 1 (bad) - 4 (excellent)
3. Select one or more (if equal quality) best answers to the given question

It is possible to receive a question that is in poor taste or a question that does not make sense.

CLICK HERE WHEN YOU ARE READY TO SEE THE QUESTION

SUBMIT
Validation interface

Instructions:

1. Read the given question
2. Read each of the answers and assess its quality from 1 (bad) - 4 (excellent)
3. Select one or more (if equal quality) best answers to the given question

It is possible to receive a question that is in poor taste or a question that does not make sense.

What is the historical context in which the laws of Hammurabi were written?

CLICK HERE WHEN YOU ARE DONE READING

TIME LEFT: 58 SEC

SUBMIT
Validation interface

2. Read each of the answers and assess its quality from 1 (bad) to 4 (excellent)
3. Select one or more (if equal quality) best answers to the given question

It is possible to receive a question that is in poor taste or a question that does not make sense.

**History**

What is the historical context in which the laws of Hammurabi were written?

**Hammurabi** is best known for the promulgation of a new code of Babylonian law: the Code of Hammurabi. This Law was written before the Mosaic Code and was one of the first written laws in the world. The Code of Hammurabi was written on a stele, a large stone monument, and placed in a public place so that all could see it...

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

☐ This is the best answer

**The Sunday liquor law seems to violate the spirit of the First Amendment, at least, but, because it doesn't directly address religion, would be a tough one to challenge on those grounds. As to the Ten Commandments, the courts have drawn a careful line. Displays in the context of legal history...**

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

☐ This is the best answer

**The Code of Hammurabi (also Hammurabi), the most complete and perfect extant collection of Babylonian laws, was developed during the reign of Hammurabi (r. 1792-1750 B.C.) of the first dynasty of Babylon. The code consists of Hammurabi’s legal decisions, which were collected toward the end of his reign and inscribed...**

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

☐ This is the best answer

Submit
Average time to judge 3 answers was 23 seconds

✓ Median time to judge 3 answers for a question is ~23 sec
✓ Time pressure forces workers to rate answers faster
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
Answer ratings summary

- Crowd workers can be used to obtain reliable ratings for answer candidates

- Even one minute seems to be enough to judge the quality of 3 answers to CQA questions
  - no significant rating quality loss compared to no-time pressure experiment
Methodology: answer collection

1087 questions from TREC LiveQA’15

Task: answer the given question

With 1 minute time limit
- 3 workers per question
- $0.10 per task
- could use web search

Without time limit
- 3 workers per question
- $0.10 per task
- could use web search

➢ Then we collected 3 ratings for each answer using crowdsourcing without time pressure
➢ The final quality score for the answer is the average of 3 scores
Answer Collection Interface

Instructions:

1. You will be given a question generated from a real person on the internet.
2. You will have 5 minutes to answer each question.
3. If you don’t know the answer yourself you are allowed to browse the internet.
4. If you found the answer on the internet you must provide the source (otherwise write N/A for source).
5. Use this specific link below to search for an answer, DO NOT OPEN ANOTHER SEARCH ENGINE: [WWW.GOOGLE.COM](http://www.google.com)

It is possible to receive a question that is in poor taste or a question that does not make sense. Please rate each question accordingly.

[CLICK HERE WHEN YOU ARE READY TO SEE THE QUESTION](#)
Instructions:

1. You will be given a question generated from a real person on the internet.
2. You will have 5 minutes to answer each question.
3. If you don't know the answer yourself, you are allowed to browse the internet.
4. If you found the answer on the internet, you must provide the source (otherwise write N/A for source).
5. Use this specific link below to search for an answer, DO NOT OPEN ANOTHER SEARCH ENGINE: [WWW.GOOGLE.COM](http://www.google.com)

It is possible to receive a question that is in poor taste or a question that does not make sense. Please rate each question accordingly.

Question:

Anybody know how THE CODING INVOLVED to add the feature to tag someone on a social media website?

I want the person to get a notification. So if I used @johndoe I want johndoe to get a notification. How do I do that?

[CLICK HERE WHEN YOU ARE DONE READING](http://www.google.com)
Answer Collection Interface

Question:

Anybody know how THE CODING INVOLVED to add the feature to tag someone on a social media website?
I want the person to get a notification. So if I used @johndoe I want johndoe to get a notification. How do I do that?

Does the way the question is worded make sense?
- yes
- no

Are you familiar with this topic?
- yes
- no

Write Your Answer Below:

Answer Source:

Answer Source
# Answer Crowdsourcing statistics

<table>
<thead>
<tr>
<th></th>
<th>Yahoo! Answers</th>
<th>mTurk (\leq 1) min</th>
<th>mTurk (\leq 1) min</th>
<th>LiveQA’15 winning system</th>
</tr>
</thead>
<tbody>
<tr>
<td>% answered</td>
<td>78.6%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>97.8%</td>
</tr>
<tr>
<td>Length (chars)</td>
<td>354.96</td>
<td>190.83</td>
<td>126.65</td>
<td>790.41</td>
</tr>
<tr>
<td>Length (words)</td>
<td>64.54</td>
<td>34.16</td>
<td>22.82</td>
<td>137.23</td>
</tr>
</tbody>
</table>

![Histogram of work time](image)

- **No time pressure**
- **Time pressure**

![Work time in seconds](image)
Crowdsourced answers are comparable in quality to community generated responses.

Answers obtained from a crowd under 1 minute time limit are significantly better than those of a top-10 automatic QA system.
Answer Crowdsourcing statistics

✓ Crowdsourced answers tend to be relevant, but of average quality
✓ Automatically generated answers are more often either not-useful or good
Example Answers

● **General Opinions and Recommendations**
  ○ What is a great free(or cheap) video player/library combo that somewhat functions like Itunes?
    ✓ Media Player Classic might be what you are look…

● **Questions requiring certain expertise**
  ○ Are gold cap conures good birds? I'm looking for sweet, loving ect.?
    ✓ Apparently they are adorable. Their only imperfection is that they are loud.
    ✓ Conures can be very loud, obnoxiously ear piercingly loud.
  ○ Is Gotu Kola a good herb for mental health? How long does it take to work??
    ✓ yes
  ○ Which LGA 775 processor can I upgrade to?
    ✓ You have to make sure what you want to upgrade to is compatible or it won't work with your machine. There should be a list on any processor that tells what it works with.
  ○ What is OPEC46LCZ?
    ✓ OPEC is about oil and nothing to do with heart disease.
    ✓ There is no answer available to this question
Conclusions

1. Crowd workers are capable of reliably validating ~3 answers to a question, even with 1 minute time limit

2. Even one minute appears to be enough for crowd workers to provide a response to real user questions, such as those posted to CQA websites

3. The quality of crowdsourced answers on average was comparable to the CQA community answers, and even with time limit crowdsourcing can be useful for an automated QA system
We integrated crowdsourcing into a real-time automated QA system, that participated in TREC LiveQA’16
- To rate candidate answers
- To provide additional candidates
Crowd-powered Real-time Question Answering

✓ Workers accept Amazon mTurk HIT, that lasts 15 minutes
✓ Within this time, they are waiting for questions to arrive
✓ When a question arrives, a crowd worker could do the following:
  - Provide his answer to the question if she has one (optional)
  - Rate answers as they appear on the screen
Average rating of answers

<table>
<thead>
<tr>
<th></th>
<th>Average score per question</th>
<th>Average precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automated system</td>
<td>2.321</td>
<td>2.357</td>
</tr>
<tr>
<td>Automated system + crowdsourcing</td>
<td>2.550⁺</td>
<td>2.556⁺</td>
</tr>
<tr>
<td>(answers+ratings)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ratings only</td>
<td>2.432</td>
<td>2.470</td>
</tr>
<tr>
<td>Answers only</td>
<td>2.459</td>
<td>2.463</td>
</tr>
</tbody>
</table>

✓ Crowdsourcing for real-time question answering significantly improves the performance
✓ Ratings and answers both contribute to the overall quality gain
Crowdsourcing can significantly improve the performance of a near real-time question answering system by providing additional answers and rating existing answer candidates.

Future work:

- Cost & Scalability:
  - Optimizing the number of workers per task
  - Selective crowdsourcing: e.g. using query performance prediction techniques

- What are other useful types of feedback a crowd can provide to QA system
  - e.g. Question summary
  - expected phrases in the answer

Thank you!