COUGH: A Challenge Dataset and Models for COVID-19 FAQ Retrieval

Xinliang Frederick Zhang¹, Heming Sun¹, Xiang Yue¹, Simon Lin², and Huan Sun¹
¹The Ohio State University, ²Abigail Wexner Research Institute at Nationwide Children’s Hospital

Dataset Comparison

<table>
<thead>
<tr>
<th>Domain</th>
<th>FAQR (Karan and Snajder)</th>
<th>StackFAQ (Karan and Snajder)</th>
<th>LocalGov (Sakata et al.)</th>
<th>Sun and Sedoc</th>
<th>Pollak et al.</th>
<th>COUGH (ours)</th>
</tr>
</thead>
<tbody>
<tr>
<td># of FAQs</td>
<td>4,313</td>
<td>7,19</td>
<td>1,786</td>
<td>690</td>
<td>2,135</td>
<td>15,919</td>
</tr>
<tr>
<td># of Queries (Q)</td>
<td>1,233</td>
<td>1,249</td>
<td>784</td>
<td>6,495*</td>
<td>24,240*</td>
<td>1,236</td>
</tr>
<tr>
<td># of annotations per Q</td>
<td>8.22</td>
<td>Not Applicable</td>
<td>&lt;10</td>
<td>5</td>
<td>5</td>
<td>32.17</td>
</tr>
<tr>
<td>Query Length</td>
<td>7.30</td>
<td>13.84</td>
<td>**</td>
<td>**</td>
<td>6.87</td>
<td>12.97</td>
</tr>
<tr>
<td>QA-query Length</td>
<td>12.30</td>
<td>10.39</td>
<td>**</td>
<td>**</td>
<td>7.83</td>
<td>13.00</td>
</tr>
<tr>
<td>FAQ-answer Length</td>
<td>33.00</td>
<td>76.54</td>
<td>**</td>
<td>**</td>
<td>76.71</td>
<td>113.58</td>
</tr>
<tr>
<td>Language</td>
<td>English</td>
<td>English</td>
<td>Japanese</td>
<td>English</td>
<td>Multi-lingual</td>
<td>Multi-lingual</td>
</tr>
<tr>
<td># of sources</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>12</td>
<td>34</td>
<td>55</td>
</tr>
</tbody>
</table>

Comparison of COUGH with representative counterparts.

* Extracted from existing resources (e.g., COVID-19 Twitter dataset (Chen et al., 2020)).
** Not Applicable, either not in English or not publicly available.

Experimental Results

<table>
<thead>
<tr>
<th>Method</th>
<th>P@1</th>
<th>P@5</th>
<th>MAP</th>
<th>MRR</th>
<th>nDCG</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM25 (Q-q)</td>
<td>60.4</td>
<td>43.7</td>
<td>28.2</td>
<td>73.0</td>
<td>76.7</td>
</tr>
<tr>
<td>BM25 (Q-a)</td>
<td>33.4</td>
<td>23.6</td>
<td>16.2</td>
<td>47.4</td>
<td>64.4</td>
</tr>
<tr>
<td>BM25 (Q+q-a)</td>
<td>56.9</td>
<td>41.3</td>
<td>28.5</td>
<td>70.0</td>
<td>72.6</td>
</tr>
<tr>
<td>BERT (Q-q)</td>
<td>63.8</td>
<td>46.0</td>
<td>27.1</td>
<td>75.7</td>
<td>78.6</td>
</tr>
<tr>
<td>+ fine-tune on pesudo Q-q</td>
<td>64.9</td>
<td>49.0</td>
<td>27.5</td>
<td>75.1</td>
<td>63.0</td>
</tr>
<tr>
<td>BERT (Q-a)</td>
<td>13.5</td>
<td>9.6</td>
<td>4.8</td>
<td>24.1</td>
<td>16.7</td>
</tr>
<tr>
<td>+ fine-tune on FAQ Bank</td>
<td>52.0</td>
<td>37.1</td>
<td>25.8</td>
<td>66.0</td>
<td>56.4</td>
</tr>
<tr>
<td>+ re-rank</td>
<td>52.1</td>
<td>38.4</td>
<td>26.4</td>
<td>66.3</td>
<td>57.8</td>
</tr>
<tr>
<td>CombSum</td>
<td>69.7</td>
<td>48.8</td>
<td>37.3</td>
<td>80.2</td>
<td>74.7</td>
</tr>
<tr>
<td>- fine-tuned BERT (Q-a)</td>
<td>65.4</td>
<td>45.8</td>
<td>31.5</td>
<td>77.2</td>
<td>75.2</td>
</tr>
</tbody>
</table>

FAQ Retrieval

FAQ Bank

Question 1: Should children wear masks?
Answer 1: In general, children 2 years and older should wear a mask...appropriate and consistent use of masks...

Question 2: Coping with Self-Quarantine

Answer 2: Remind yourself that difficult emotions are normal during self-quarantine...

Question 3: COVID-19 is how to communicate with others when sick?

Answer 3: ...The virus is primarily transmitted through respiratory droplets...

FAQ Bank

Query 1: Is it possible for human beings to catch with COVID-19 transmitted to them from animals?
Query 2: Is it possible to get infected by COVID 19 if I touch food surface packaging?

Relevance Set

<table>
<thead>
<tr>
<th>Query</th>
<th>Relevant FAQ in FAQ Bank</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>6</td>
<td>3.67</td>
</tr>
<tr>
<td>Q2</td>
<td>14</td>
<td>3.67</td>
</tr>
<tr>
<td>Q3</td>
<td>18</td>
<td>3.67</td>
</tr>
</tbody>
</table>

Dataset Analysis

Varying Query Forms

• Question form
  • Interpretive
  • Usually related to general information about the virus.

• Query String form
  • Declarative
  • search for more specific instructions concerning COVID-19 (e.g., healthy diet during pandemic).

• Forum form
  • Scrapped from medical forums.

Answer Nature

• Lengthy
  • FAQ-Answer length: 113.58
  • Longer than those in prior datasets.

• Noisy
  • Contain contents not directly pertinent to query.
  • E.g. Answer to the query “What is novel coronavirus” contains extra information about comparisons with other viruses.

Large-scale Relevance Annotation

• 32.17 annotated FAQs per query.
• Existing FAQ datasets overlooked annotation scale.
• Reduce the chance of missing true positive tuples.
• Each <Query, FAQ item> tuple annotated by ≥3 people.
• Reduce the variance and bias in annotation.
• Annotation done on a Likert scale.
• Matched (4), Useful (3), Useless (2) and Non-relevant (1)

Multilinguality

• 6768 Non-English FAQs
• Varying Query Forms
• Question (3396)
• Query String (3372)
• 18 non-English languages
  • Spanish, Chinese, French, Japanese, Arabic, German, Russian, Korean ......

Error Analysis

Observation: Fine-tuning under the Q-q mode improves the performance.
• Removing fine-tuned BERT (Q-q) from CombSum hurts.

Acknowledgments

Manning et al. (2008)

COUGH: The COVID-19 FAQ Dataset

FAQ Bank

<table>
<thead>
<tr>
<th>Question</th>
<th>Relevant FAQ in FAQ Bank</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>6</td>
<td>3.67</td>
</tr>
<tr>
<td>Q2</td>
<td>14</td>
<td>3.67</td>
</tr>
<tr>
<td>Q3</td>
<td>18</td>
<td>3.67</td>
</tr>
</tbody>
</table>

FAQ Retrieval

FAQ Bank

Question: 1. Should children wear masks?
Answer 1: In general, children 2 years and older should wear a mask...appropriate and consistent use of masks...

Question 2: Coping with Self-Quarantine

Answer 2: Remind yourself that difficult emotions are normal during self-quarantine...

Question 3: COVID-19 is how to communicate with others when sick?

Answer 3: ...The virus is primarily transmitted through respiratory droplets...

FAQ Bank

Query 1: Is it possible for human beings to catch with COVID-19 transmitted to them from animals?
Query 2: Is it possible to get infected by COVID 19 if I touch food surface packaging?

Relevance Set

<table>
<thead>
<tr>
<th>Query</th>
<th>Relevant FAQ in FAQ Bank</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>6</td>
<td>3.67</td>
</tr>
<tr>
<td>Q2</td>
<td>14</td>
<td>3.67</td>
</tr>
<tr>
<td>Q3</td>
<td>18</td>
<td>3.67</td>
</tr>
</tbody>
</table>

Dataset Analysis

Varying Query Forms

• Question form
  • Interpretive
  • Usually related to general information about the virus.

• Query String form
  • Declarative
  • search for more specific instructions concerning COVID-19 (e.g., healthy diet during pandemic).

• Forum form
  • Scrapped from medical forums.

Answer Nature

• Lengthy
  • FAQ-Answer length: 113.58
  • Longer than those in prior datasets.

• Noisy
  • Contain contents not directly pertinent to query.
  • E.x. Answer to the query “What is novel coronavirus” contains extra information about comparisons with other viruses.

Large-scale Relevance Annotation

• 32.17 annotated FAQs per query.
• Existing FAQ datasets overlooked annotation scale.
• Reduce the chance of missing true positive tuples.
• Each <Query, FAQ item> tuple annotated by ≥3 people.
• Reduce the variance and bias in annotation.
• Annotation done on a Likert scale.
• Matched (4), Useful (3), Useless (2) and Non-relevant (1)

Multilinguality

• 6768 Non-English FAQs
• Varying Query Forms
• Question (3396)
• Query String (3372)
• 18 non-English languages
  • Spanish, Chinese, French, Japanese, Arabic, German, Russian, Korean ......

Error Analysis

Observation: Fine-tuning under the Q-q mode using synthetic data hurt the performance.

FAQ Items: Q: What is the difference between COVID-19 antibody tests?
A: The lab protocols for identifying the virus in blood...

Gold label: Negative (useless, useless, useless)
Predicted rank: 3

Error Analysis

Observation: Fine-tuning under the Q-q mode using synthetic data hurt the performance.

FAQ Items: Q: What is the difference between COVID-19 antibody tests?
A: The lab protocols for identifying the virus in blood...

Gold label: Negative (useless, useless, useless)
Predicted rank: 3

Type Number Q-Length A-length
Question 4,978 14.64 123.89
Query String 2,139 9.18 89.60
Forum 2,034 147.46 90.49

# English Question 3,396 - -
# Non-English Question 3,372 - -
# Total - 15,919 -

Basic statistics of FAQ bank in COUGH.

Dataset: https://github.com/sunlab-osu/covid-faq

Contact: {zhang.9975, sun.397}@osu.edu, xlfzhang@umich.edu