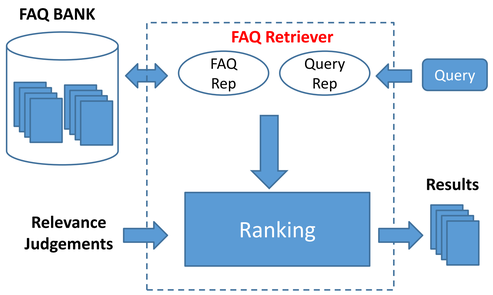
# COUGH: A Challenge Dataset and Models for COVID-19 FAQ Retrieval


Xinliang Frederick Zhang[1], Heming Sun[1], Xiang Yue[1], Simon Lin[2], and Huan Sun[1]

[1]The Ohio State University, [2]Abigail Wexner Research Institute at Nationwide Children's Hospital


## FAQ Retrieval




Manning et al. (2008)

## COUGH

**COUGH: The COVID-19 FAQ Dataset**

**FAQ Bank**

**Question1:** Should children wear masks?
**Answer1**: *In general, children 2 years and older should wear a mask...Appropriate and consistent use of masks...*
**Question2:** Coping with Self-Quarantine
**Answer2**: *Remind yourself that difficult emotions are normal during self-quarantine...*
**Question3:** COVID-19是如何在人与人之间传播的? (How does COVID-19 spread between people?)
**Answer3**: ***...该病毒的人际传播主要通过感染者与他人密切接触...(...mainly when an infected person is in close contact with another person...)***

**Query Bank**

**Query1:** Is it possible for human beings to get sick with COVID-19 transmitted to them from animals?
**Query2:** Is it possible to get infected by COVID 19 if I touch food surface packaging?

**Relevance Set**

| Query | Relevant FAQ in FAQ Bank | Score |
|---|---|---|
| Query1 | Q: Can wild animals spread the virus that causes COVID-19 to people or pets? A: Currently, there is no evidence to suggest... | 3.67 |
| Query1 | Q: How is COVID-19 transmitted? A: COVID-19 illness is spread mainly from person to person through respiratory... | 2.67 |
| Query2 | Q: What are the lab protocols for identifying the virus in food? On surfaces?A: As food hasn't been implicated in transmission | 3.67 |

| | Type | Number | Q-Length | A-length |
|---|---|---|---|---|
| # English | Question | 4,978 | 14.64 | 123.89 |
| | Query String | 2,139 | 9.18 | 89.60 |
| | Forum | 2,034 | 147.46 | 90.49 |
| # Non-English | Question | 3,396 | - | - |
| | Query String | 3,372 | - | - |
| # Total | - | 15,919 | - | - |

Basic statistics of FAQ bank in COUGH.

## Dataset Comparison

| | FAQIR (Karan and Šnajder) | StackFAQ (Karan and Šnajder) | LocalGov (Sakata et al.) | Sun and Sedoc | Poliak et al. | COUGH (ours) |
|---|---|---|---|---|---|---|
| Domain | Yahoo! | StackExachange | Government | COVID-19 | COVID-19 | COVID-19 |
| # of FAQs | 4,313 | 719 | 1,786 | 690 | 2,115 | 15,919 |
| # of Queries (Q) | 1,233 | 1,249 | 784 | 6,495* | 24,240* | 1,236 |
| # of annotations per Q | 8.22 | Not Applicable | <10 | 5 | 5 | 32.17 |
| Query Length | 7.30 | 13.84 | ** | ** | 6.87 | 12.97 |
| FAQ-query Length | 12.30 | 10.39 | ** | ** | 8.73 | 13.00 |
| FAQ-answer Length | 33.00 | 76.54 | ** | ** | 76.71 | 113.58 |
| Language | English | English | Japanese | English | Multi-lingual | Multi-lingual |
| # of sources | 1 | 1 | 1 | 12 | 34 | 55 |

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.

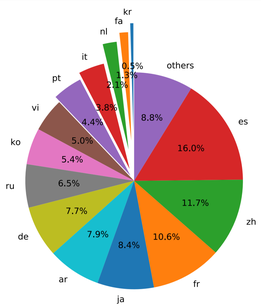
## Dataset Analysis

**Varying Query Forms**
- Question form
  - Interrogative
  - 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.
- Manifest difficulties of FAQ retrieval in real scenarios.

**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 ……




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

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

## Experimental Results

| Method | P@1 | P@5 | MAP | MRR | nDCG |
|---|---|---|---|---|---|
| BM25 (Q-q) | 60.4 | 43.7 | 28.2 | 73.0 | 76.7 |
| BM25 (Q-a) | 33.4 | 25.6 | 16.2 | 47.4 | 46.4 |
| BM25 (Q-q+a) | 56.9 | 41.3 | 28.5 | 70.0 | 72.6 |
| BERT (Q-q) | 63.8 | 46.0 | 27.1 | 75.7 | **78.6** |
| + fine-tune on pesudo Q-q | 64.9 | 40.9 | 27.5 | 75.1 | 63.0 |
| BERT (Q-a) | 13.5 | 9.6 | 4.8 | 24.1 | 16.7 |
| + fine-tune on FAQ Bank | 52.0 | 37.1 | 25.8 | 66.0 | 56.4 |
| + re-rank | 52.1 | 38.4 | 26.4 | 66.3 | 57.8 |
| CombSum | **69.7** | **48.8** | **37.3** | **80.2** | 74.7 |
| - fine-tuned BERT (Q-a) | 65.4 | 45.8 | 31.5 | 77.2 | 75.2 |

Evaluation results on COUGH (average of 5 runs).

**Observation**
- Q-q mode consistently performs better than Q-a mode.
- Utilizing the cross-encoder for re-ranking yields better results.
- Fine-tuning under the Q-a mode improves the performance.
- Removing fine-tuned BERT (Q-a) from CombSum hurts.

## Error Analysis

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

**Query:** What research is being done on antibody tests and their accuracy?
**FAQ item:** Q: What is antibody testing? How do I get a COVID-19 antibody test? A: CDC and partners are investigating to determine if you can get sick with COVID-19 more than once ...
**Gold label:** `Negative [useful, useless, useless]`
**Predicted rank:** 3

**Query:** Are COVID-19 antibody tests accurate?
**FAQ item:** Q: Should I be tested with an antibody (serology) test for COVID-19? A: ... Antibody tests have limited ability to diagnose COVID-19 and should not be used alone to diagnose COVID-19 ...
**Gold label:** `Positive [useful, useful, matched]`
**Predicted rank:** 26

Biased towards responses with similar texts.

Fails pragmatic reasoning: "limited abiltity" => "results are not accurate for diagnosing COVID-19".

Case analyses with fine-tuned BERT (Q-q). Human annotations are inside [].

## Acknowledgments


NSF pcori Ohio Supercomputer Center