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

Xinliang Frederick Zhang[1,2,*], Heming Sun[1,3,*], Xiang Yue[1], Simon Lin[4], and Huan Sun[1]

[1]The Ohio State University
[2]University of Michigan
[3]University of Southern California
[4]Abigail Wexner Research Institute at Nationwide Children's Hospital
{zhang.9975,sun.2164,yue.149,sun.397}@osu.edu,
Simon.Lin@nationwidechildrens.org

Abstract

We present a large, challenging dataset, COUGH, for COVID-19 FAQ retrieval. Similar to a standard FAQ dataset, COUGH consists of three parts: FAQ Bank, Query Bank and Relevance Set. The FAQ Bank contains ∼16K FAQ items scraped from 55 credible websites (e.g., CDC and WHO). For evaluation, we introduce Query Bank and Relevance Set, where the former contains 1,236 human-paraphrased queries while the latter contains ∼32 human-annotated FAQ items for each query. We analyze COUGH by testing different FAQ retrieval models built on top of BM25 and BERT, among which the best model achieves 48.8 under P@5, indicating a great challenge presented by COUGH and encouraging future research for further improvement. Our COUGH dataset is available at https://github.com/sunlab-osu/covid-faq.

1 Introduction

Many institutional websites today maintain an FAQ page to help users find relevant information for commonly asked questions. The FAQ retrieval task is defined as ranking FAQ items $\{(q_i, a_i)\}$[1] from a collection given a user query Q (Karan and Šnajder, 2016). In contrast to common Information Retrieval (IR), FAQ retrieval often introduces 3 new challenges: 1) brevity of FAQ texts in comparison with IR documents; 2) need for topic-specific knowledge; 3) usage of the new question field in FAQ items (Karan and Šnajder, 2016; Sakata et al., 2019). However, FAQ retrieval is under-studied compared with other IR applications such as open-domain QA (Chen and Yih, 2020).

In this work, we specifically study FAQ retrieval for COVID-19, a contagious and fatal pandemic which is still evolving on a daily basis. Many websites like CDC and WHO provide quality information on COVID-19 and update FAQ pages regularly.

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

Figure 1: Examples from the COUGH dataset.

To gain better insights into FAQ retrieval research and advance COVID-19 information search, we present an FAQ dataset, COUGH[2], consisting of FAQ Bank, Query Bank and Relevance Set, following the standard of constructing an FAQ dataset (Manning et al., 2008). The FAQ Bank contains 15919 FAQ items scraped from 55 authoritative institutional websites (see a full list in Table A4 and A5). COUGH covers a wide range of topics on COVID-19, from general information about the virus to specific COVID-related instructions for a healthy diet. For evaluation, we further construct Query Bank and Relevance Set, including 1,236 crowd-sourced queries and their relevance to a set of FAQ items judged by annotators. Examples from COUGH are shown in Figure 1.

Our dataset poses several new challenges (e.g.,

*Work was done when the first two authors were at OSU.
[1]q and a are question and answer fields in an FAQ item.
[2]Adapted from "CoF" that stands for **CO**VID **F**AQ.

	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

Table 1: 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.

answer fields are longer and noisier, and harder to match, than question fields) to existing methods. The diversity of FAQ items, reflected in varying query forms and lengths as well as in narrative styles, also contributes to these challenges.

The contribution of this work is two-fold. First, we construct a challenging dataset COUGH to aid the development of COVID-19 FAQ retrieval models. Second, we evaluate various FAQ retrieval models across different settings, explore their limitations, and encourage future work along this line.

2 Related Work

COVID-19 & FAQ Datasets. Since the outbreak of COVID-19, the community has witnessed many datasets released to advance the research of COVID-19. For example, CORD-19 (Wang et al., 2020), CODA-19 (Huang et al., 2020), COVID-Q (Wei et al., 2020), Weibo-Cov (Hu et al., 2020), and Twitter dataset (Chen et al., 2020). All of them aim to aggregate resources to combat COVID-19.

The most related works to ours are Sun and Sedoc (2020) and Poliak et al. (2020), both of which constructed a collection of COVID-19 FAQs by scraping authoritative websites. However, the dataset in the former work is not available yet and the latter work does not evaluate models on their dataset, and there is still a great need to understand how existing models would perform on the COVID-19 FAQ retrieval task. In the open domain, several FAQ datasets appeared recently, such as FAQIR (Karan and Šnajder, 2016), StackFAQ (Karan and Šnajder, 2018) and LocalGov (Sakata et al., 2019). Unfortunately, as shown in Table 1, the scale of existing FAQ datasets is too small, and answer lengths are much lower than those in COUGH, which may not characterize the difficulty of FAQ retrieval tasks in real-world scenarios. Moreover, in contrast to all prior datasets, COUGH covers multiple query forms (e.g., question and query string forms) and has many annotated FAQs for each user query, whereas queries in existing FAQ datasets are limited to the question form and have much fewer annotations.

FAQ Retrieval Methods. FAQ retrieval focuses on retrieving the most-matched FAQ items given a user query (Karan and Šnajder, 2018). Many earlier works, e.g., FAQ FINDER (Burke et al., 1997), query expansion (Kim and Seo, 2006) and BM25 (Robertson and Zaragoza, 2009), resorted to traditional IR techniques by leveraging lexical mapping and/or semantic similarity. In the deep learning era, many studies show that Neural Networks are useful for FAQ retrieval as they are good at learning the semantic relevance between queries and FAQ items. Along this line, Karan and Šnajder (2016) adopted Convolution Neural Networks, Gupta and Carvalho (2019) utilized LSTM, and Sakata et al. (2019) leveraged an ensemble of TSUBAKI and BERT. Recently, Mass et al. (2020) explored learning to rank without requiring manual annotations.

3 Dataset Construction[3]

3.1 FAQ Bank Construction

We developed scrapers[4] adapted from Poliak et al. (2020), and add special features to COUGH dataset.

Web scraping: We collect FAQ items from authoritative international organizations, state governments and other credible websites including reliable encyclopedias and medical forums. Moreover, we scrape three types of FAQs: question (i.e., an interrogative statement), query string (i.e., a string of words to elicit information) and forum (FAQs scrapped from medical forums) forms. Inspired by Manning et al. (2008), we loosen the constraint that queries must be in question form since we want to study a more generic and challenging problem. We also scrape 6,768 non-English FAQs to increase lan-

[3] We provide detailed annotation protocols in Appendix A.

[4] Scrapers are released together with COUGH to keep FAQ Bank up-to-date.

guage diversity. Overall, we scraped 15,919 FAQ items covering all three forms and 19 languages.

3.2 Query Bank Construction

Following Manning et al. (2008); Karan and Šnajder (2016), we do not crowdsource queries from scratch, but instead ask annotators to paraphrase our provided query templates. That way, we ensure that 1) collected queries are pertinent to COVID-19; 2) collected queries are not too simple; 3) the chance of getting similar user queries is reduced.

Phase 1: Query Template Creation: We sample 5% of FAQ items from each English non-forum source[5] and use the question part as the *template*. For example, the templates of the two paraphrased queries in Figure 1 are "Can humans become infected with the COVID-19 from an animal source?" and "Can I get sick with COVID-19 from touching food, the food packaging, or food contact surfaces, if the coronavirus was present on it?".

Phase 2: Paraphrasing for Queries: In this phase, each annotator is expected to give three paraphrases for each query template. Besides providing shallow parapharases (e.g., word substitution), annotators are encouraged to give deep paraphrases (i.e., grammatically different but semantically similar/same) to simulate the noisy and diverse environment in real scenarios. In the end, we obtain 1,236 human-paraphrased user queries.

3.3 Relevance Set Construction

Phase 1: Initial Candidate Pool Construction: For each user query, as suggested by previous work (Manning et al., 2008; Karan and Šnajder, 2016; Sakata et al., 2019), we run 4 models (see Section 5.2), BM25 (Q-q), BM25 (Q-q+a), BERT (Q-q), and BERT (Q-a) fine-tuned on COUGH, to instantiate a candidate FAQ pool. Each model complements the others and contributes its top-10 relevant FAQ items. We then take the union to remove duplicates, giving an average pool size of 32.2.

Phase 2: Human Annotation: Each annotator gives each ⟨Query, FAQ item⟩ tuple a score based on the annotation scheme (i.e., 4/Matched, 3/Useful, 2/Useless and 1/Non-relevant)[6] adapted from Karan and Šnajder (2016); Sakata et al. (2019). In order to alleviate the annotation bias, each tuple has at least 3 annotations. In the finalized Set, we keep all raw scores and include: 1) mean of annotations; 2) four suggested aggregation schemes to obtain binary labels (as detailed in Appendix B). Users of COUGH can also try other aggregation measures.

Among 1,236 user queries, there are 35 "unanswerable" queries that have no associated positive FAQ item.

[5]Each source contributes at least one item to ensure wide topic coverage and similar sampled FAQ items are removed.

[6]Table A.2 details the meaning of these four scores.

	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	-	-

Table 2: Basic statistics of FAQ bank in COUGH.

4 Dataset Analysis

Besides the generic goal of large size, diversity, and low noise, COUGH features 5 additional aspects.

Varying Query Forms: As indicated in Table 2, there are multiple query forms. In evaluation, we include both question (Question1 and 3 in Figure 1) and query string (Question2 in Figure 1) forms. These two distinct forms are different in terms of query format (interrogative v.s. declarative), average answer length (123.89 v.s. 89.60) and topics. Question form is usually related to general information about the virus while query string form is often searching for more specific instructions concerning COVID-19 (e.g., healthy diet during pandemic).

Answer Nature: Table 1 shows the answer fields in COUGH are much longer than those in any prior dataset. We also observe that answers might contain some contents which are not directly pertinent to the query, partially resulting in the long length nature. For example, in COUGH, the answer to a query "What is novel coronavirus" contains extra information about comparisons with other viruses. Such lengthy and noisy nature of answers manifest the difficulty of FAQ retrieval in real scenarios.

Language Correctness in Query Bank: Most queries in our Query Bank are properly spelled and grammatically correct, so we can prioritize investigating the model performance under a less noisy setting. Furthermore, our dataset can support a controlled study on the impact of spelling and grammatical errors: One can simulate various kinds of spelling and grammatical errors and inject them in a controlled manner into the Query Bank and systematically evaluate how the model performance changes under different levels of noises.

Large-scale Relevance Annotation: Many existing FAQ datasets overlooked annotation scale (Ta-

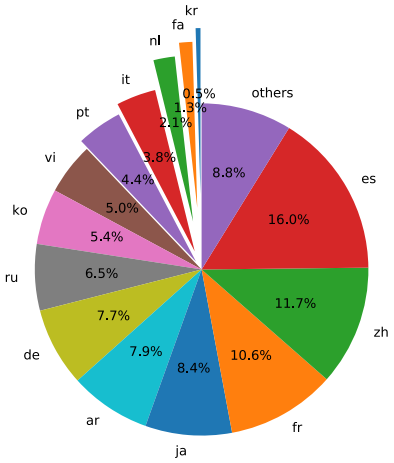

Figure 2: Language distribution for non-English FAQ items.

ble 1); yet, that would hurt the evaluation reliability since many true positive ⟨Query, FAQ item⟩ tuples were omitted. Following Manning et al. (2008), for each user query, we constructed a large-scale candidate pool to reduce the chance of missing true positive tuples. The annotation procedure yielded 39760 annotated tuples, each of which is annotated by at least 3 people to reduce annotation bias.

Multilinguality: COUGH includes 6768 FAQ items covering 18 non-English languages, and statistics of non-English items can be found in Table 2. Figure 2 shows the language distribution (excluding English) of FAQ items in COUGH dataset. Like English FAQ items, non-English FAQ items are also presented in both question and query string forms. The detailed breakdown of non-English portion by sources and languages is shown in Table A5.

However, due to budget limit, we did not proceed to the annotation phase for non-English data, so there is no non-English human-paraphrased user query or relevance judgement.

Annotation Quality: We discard low-quality paraphrased queries (∼24%) and relevance annotations (∼11%). Further, we show that ∼74% of annotated tuples have high agreements where multiple people vote for the same relevance class. More details of quality checking can be found in Section 8.1.

5 Experiments

5.1 Experimental Setup

In this work, we focus on *unsupervised* sparse and dense retrievers and discuss their limitations. Supervised learning is less popular for this task since it's too costly to collect a large-scale Query Bank and its associated relevance judgement (Sakata et al., 2019; Mass et al., 2020). Further, there are 3 configurable modes, Q-q, Q-a and Q-q+a, where a user query Q can be learned to match with question q, answer a or the concatenation $q+a$.

5.2 Methods

(1) **BM25** is a nonlinear combination of term frequency, document frequency and document length.

(2) **BERT** (Devlin et al., 2019) is a pretrained language model. We use its variant, Sentence-BERT (Reimers and Gurevych, 2019), to encode Q, q and a separately to generate sentence representations.

Fine-tuning: Similar to Henderson et al. (2017); Karpukhin et al. (2020), we leverage in-batch negatives[7] to fine-tune BERT on FAQ bank. For Q-q mode, we use GPT-2 (Radford et al., 2019) to generate synthetic questions to match with Q. For Q-a mode, an FAQ item (q, a) itself is a positive pair

Re-rank: In Q-a mode, answers are quite long, so the importance of selecting most-related spans from relevant answers to catch the nuance is amplified. As detailed in Reimers and Gurevych (2019); Humeau et al. (2020), cross-encoder can perform self-attention between query and answer, resulting in a richer extraction mechanism. We re-rank[8] top-10 retrieved answers using cross-encoder BERT.

(3) **CombSum** (Mass et al., 2020) first computes three matching scores between the user query and FAQ items via BM25 (Q-q), BERT (Q-q) and fine-tuned BERT (Q-a) models. Then, the three scores are normalized and combined by averaging. We also evaluate with no BERT (Q-a) included.

5.3 Evaluation

Evaluation Setting: For the scope of this work, we only evaluate on 1,201 "answerable" English non-forum FAQ items, and leave the "unanswerable", non-English and forum ones for future research as great challenges have been observed under current setting. However, we encourage investigators to utilize those three categories for other potential applications (e.g., multi-lingual IR, transfer learning in IR).

Evaluation Metrics: Following previous work (Manning et al., 2008; Karan and Šnajder, 2016, 2018; Sakata et al., 2019; Mass et al., 2020), we adopt P@1 (Precision), P@5, MAP@100 (Mean Average Precision), MRR (Mean Reciprocal Rank)

[7] In a batch, (q_i, p_j) is assumed as negative pair if $i \neq j$.

[8] Directly applying cross-encoder is not efficient and yields inferior results in our preliminary experiments.

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

Table 3: Evaluation on COUGH.

and nDCG@5 (Normalized Discounted Cumulative Gain) as evaluation metrics.

6 Analysis

Quantitative Analysis. Models' results, based on aggregation scheme A: annotated tuples with mean score ≥ 3 are positives, are listed in Table 3. Results under other schemes are in appendix B.1.

The current best P@5 and MAP, 48.8 and 37.3, are not satisfying, showing a large room for improvement, confirming that COUGH is challenging.

We observe that Q-q mode consistently performs better than Q-a mode. This is because question fields are more similar to user queries than answer fields. As shown in Section 4, the answer nature (lengthy and noisy), albeit well characterizes the FAQ retrieval task in real scenarios, does bring up a great challenge. Utilizing the cross-encoder for re-ranking can yield better results since it can select query-aware features from answers. This is a possible step towards handling long and noisy answers better.

We also find that fine-tuning under the Q-a mode can improve the performance (e.g., from 9.6 to 37.1 under P@5), but might hurt it under the Q-q mode due to noises introduced by synthetic queries. Moreover, the best overall performances are achieved by BERT (Q-q) and CombSum, which are in line with Mass et al. (2020). However, CombSum without fine-tuned BERT (Q-a) performs worse than the original one. It indicates that answer fields can serve as supplementary resources for the missing information in the question field.

Qualitative Analysis. To understand fine-tuned BERT (Q-q) better, we conduct case analyses in Table 4 to show its major types of errors, hoping to further improve it in the future. Currently, fine-tuned BERT (Q-q) suffers from the following issues: 1) biased towards responses with similar texts (e.g., "antibody tests" and "antibody testing"); 2) fails to capture the semantic similarities under complex environments (e.g., pragmatic reasoning is required to understand that "limited abiltity" indicates results are not accurate for diagnosing COVID-19).

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

Table 4: Case analyses with fine-tuned BERT (Q-q). Human annotations are inside [].

Interesting future work includes: 1) handling long and noisy answer fields, e.g., via salient span selection; 2) further improving semantic understanding or reasoning skills, beyond lexical match.

7 Conclusion

In this paper, we introduce COUGH, a large challenging dataset for COVID-19 FAQ retrieval. COUGH features varying query forms, long and noisy answers, and multilinguality. COUGH also serves as a better evaluation benchmark since it has quality larger-scale relevance annotations. We discuss the limitations of current FAQ retrieval models via comprehensive experiments, and encourage future research to further improve FAQ retrieval.

Acknowledgments

We thank our hired AMT workers for their annotations. We thank all anonymous reviewers for their helpful comments. We thank Emmett Jesrani for revising an earlier version of the paper. This research was sponsored in part by the Patient-Centered Outcomes Research Institute Funding ME-2017C1-6413, the Army Research Office under cooperative agreements W911NF-17-1-0412, NSF Grant IIS1815674, NSF CAREER 1942980, and Ohio Supercomputer Center (Center, 1987). The views and conclusions contained herein are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notice herein.

	Task	Details	Base Cost	Base Cost per Unit	Cognitive Complexity
User Query Bank Construction					
Reference	QA annotation	Identify 3 QA pairs (write questions and then find answers).	24	8	High
	Question annotation on an audio clip	Identify 3 Questions (write questions and then find answers), each of which has additional requirements (e.g., originality, creativeness).	30	10	Extremely high
Ours*	Paraphrase Queries	Give 3 paraphrases for the original query template.	12	4	Medium
Annotated Relevance Set Construction					
Reference	Image labeling	Locate 5 required objects in a given image.	7	1.4	Medium
	Website class identification	Identify the type of niche of a twitter account. Select from 6 classes.	2	2	Low
	Identify an item	Given an image, fill out a form with 6 required fields.	9	1.5	Medium
Ours*	Relevance judgements	Identify the relevance. Select from 4 classes.	2	2	Low

Table 5: Comparison of base costs to reference tasks. Base Cost per Unit: the cost of annotating one single item (e.g., one QA pair, one paraphrase). All costs are in US cents. *: Additional bonus were rewarded for quality annotators. For example, for our relevance judgements task, we award 1 dime for every 100 quality annotations.

8 Ethical Considerations

8.1 Dataset

IRB approval. All FAQ items were collected in a manner which is consistent with the terms of use of original sources and the intellectual property and privacy rights of the original authors of the texts (i.e., source owners). This project is approved by IRB (institutional review board) at our institution as Exempt Research, which is a human subject study that presents no greater than minimal risk to participants. We consulted data officers at our institution about copyrights. They informed us that "Website content is generally copyrighted. However, you could claim the concept of *fair use* which allows the use of copyrighted material without permission from the copyright holder when it is used for research, scholarship, and teaching". We also consulted Section 107[9] of U.S. Copyright Act and ensured that our collection action fell under fair use category. We release our dataset under the Creative Commons Attribution-*NonCommercial-ShareAlike* 4.0 International License[10].

Annotation via crowdsourcing. Crowdsourcing involved in this work was conducted on Amazon Mechanical Turk (AMT). In the crowdsourcing step, all participants were required to read and sign an *informed consent form* before participating and they would not be allowed to proceed without signing. AMT mechanism, automatically anonymizing annotators' identities, ensures that the participants' privacy rights were inherently respected in the crowdsouring process. We determined the *compensation* for each annotation task by evaluating similar tasks on AMT. Table 5 shows the costs of reference tasks at the time we published our tasks. Overall, taking cognitive complexity into consideration, our base cost per unit is on the same level or higher than reference tasks. Thus, we can safely conclude that crowd workers participating in our annotation tasks were fairly compensated. Besides, the overall total cost is $2,683. Considering our competitive base cost per unit and additional generous bonus[11], we believe that participated annotators are well motivated to contribute high-quality annotations.

Quality check. During crowdsourcing phase, we filtered out low-quality annotations. Specifically, we only kept 76.45% of human-paraphrased queries for the construction of Query Bank by manually checking every single paraphrased query. When constructing the Relevance Set, for each annotator, we sampled a certain number of annotations. If the sampled annotations didn't pass the screening, we dropped all annotations made by that annotator and republished the work again. After such iterative checking, we only kept 89.20% of annotations in the end.

After crowdsourcing, we conducted post-hoc quality checking on both Query Bank and Relevance Set. We manually checked all 1,236 user queries and found that all of them make sense, are related to COVID-19 and properly written. Due to the subjectivity of the relevance judgement task, we evaluated the quality of the relevance annotations in two ways: 1) We find that 73.5% of ⟨Query, FAQ item⟩ tuples have high agreements where multiple people vote for the same relevance class; 2) We re-judge the relevance on randomly sampled 1000 tuples by hiring two research assistants and it turns out that the matching level[12] is 76.5%. Overall, the post-hoc checking confirms that our COUGH dataset is of high quality.

Annotation Protocols. To further help ethics com-

[9] https://www.copyright.gov/title17/92chap1.html#107

[10] https://creativecommons.org/licenses/by-nc-sa/4.0/

[11] For example, for our relevance judgements task, we award 1 dime for every 100 high-quality annotations.

[12] It's considered to be matched if and only if the re-judged score is in the same class (i.e., positive v.s. negative) as the mean of existing annotations.

mittees and the public judge the fairness of our annotation process, the annotation protocols for both annotation tasks are listed in Appendix A. Figure A1 and A2 show the interfaces designed for the annotation process.

References

Robin Burke, Kristian Hammond, Vladimir Kulyukin, Steven Lytinen, Noriko Tomuro, and Scott Schoenberg. 1997. Question answering from frequently asked question files: Experiences with the faq finder system. *AI Magazine*, pages 57–66.

Ohio Supercomputer Center. 1987. Ohio supercomputer center.

Danqi Chen and Wen-tau Yih. 2020. Open-domain question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, ACL 2020*, pages 34–37.

Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health and Surveillance*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, pages 4171–4186.

Sparsh Gupta and Vitor R. Carvalho. 2019. Faq retrieval using attentive matching. In *SIGIR'19*, page 929–932.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun hsuan Sung, Laszlo Lukacs, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *CoRR*, abs/1705.00652.

Yong Hu, He yan Huang, An fan Chen, and Xian Ling Mao. 2020. Weibo-cov: A large-scale covid-19 social media dataset from weibo. *CoRR*, abs/2005.09174.

Ting-Hao Kenneth Huang, Chieh-Yang Huang, Chien-Kuang Cornelia Ding, Yen-Chia Hsu, and C. Lee Giles. 2020. CODA-19: reliably annotating research aspects on 10,000+ CORD-19 abstracts using a non-expert crowd. *CoRR*, abs/2005.02367.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *8th International Conference on Learning Representations, ICLR 2020*.

Mladen Karan and Jan Šnajder. 2016. Faqir - a frequently asked questions retrieval test collection. In *Text, Speech, and Dialogue - 19th International Conference, TSD 2016*, pages 74–81.

Mladen Karan and Jan Šnajder. 2018. Paraphrase-focused learning to rank for domain-specific frequently asked questions retrieval. In *Expert Systems with Applications,*, pages 418–433.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*, pages 6769–6781.

Harksoo Kim and Jungyun Seo. 2006. High-performance faq retrieval using an automatic clustering method of query logs. *Information Processing & Management*, pages 650 – 661.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015*.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Yosi Mass, Boaz Carmeli, Haggai Roitman, and David Konopnicki. 2020. Unsupervised faq retrieval with question generation and bert. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pages 807–812.

Adam Poliak, Max Fleming, Cash Costello, Kenton W Murray, Mahsa Yarmohammadi, Shivani Pandya, Darius Irani, Milind Agarwal, Udit Sharma, Shuo Sun, Nicola Ivanov, Lingxi Shang, Kaushik Srinivasan, Seolhwa Lee, Xu Han, Smisha Agarwal, and João Sedoc. 2020. Collecting verified covid-19 question answer pairs.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pages 3980–3990.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, pages 333–389.

Wataru Sakata, Tomohide Shibata, Ribeka Tanaka, and Sadao Kurohashi. 2019. FAQ retrieval using query

question similarity and bert-based query-answer relevance. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019*, pages 1113–1116.

Shuo Sun and João Sedoc. 2020. An analysis of bert faq retrieval models for covid-19 infobot.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex Wade, Kuansan Wang, Nancy Xin Ru Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. Cord-19: The covid-19 open research dataset. *CoRR*, abs/2004.10706.

Jerry Wei, Chengyu Huang, Soroush Vosoughi, and Jason Wei. 2020. What are people asking about covid-19? a question classification dataset.

Method	P@1	P@5	MAP	MRR	nDCG
BM25 (Q-q)	43.3	26.8	25.0	57.0	76.7
BM25 (Q-a)	21.0	15.6	14.1	33.9	46.4
BM25 (Q-q+a)	37.6	25.2	24.4	52.4	72.6
BERT (Q-q)	46.0	28.4	24.8	59.2	**78.6**
+ fine-tune on pesudo Q-q	49.9	27.2	26.5	61.1	63.0
BERT (Q-a)	8.3	5.9	4.7	16.3	16.7
+ fine-tune on FAQ Bank	35.4	23.6	22.8	49.8	56.4
+ re-rank	35.6	24.2	23.4	50.6	57.8
CombSum	**51.6**	**31.2**	**32.6**	**64.8**	74.7
- fine-tuned BERT (Q-a)	47.8	29.0	28.1	61.6	75.2

Table A1: Evaluation on COUGH (Aggregation scheme B).

Method	P@1	P@5	MAP	MRR	nDCG
BM25 (Q-q)	66.0	50.2	28.5	77.5	76.7
BM25 (Q-a)	38.4	29.2	15.6	52.5	46.4
BM25 (Q-q+a)	61.3	47.2	27.8	74.5	72.6
BERT (Q-q)	70.4	**53.5**	28.2	80.9	**78.6**
+ fine-tune on pesudo Q-q	70.7	44.9	26.0	79.7	63.0
BERT (Q-a)	15.7	11.0	4.8	26.7	16.7
+ fine-tune on FAQ Bank	55.1	39.8	24.0	68.7	56.4
+ re-rank	54.6	40.9	24.4	68.6	57.8
CombSum	**72.3**	52.9	**35.8**	**82.4**	74.7
- fine-tuned BERT (Q-a)	70.2	51.1	31.3	80.9	75.2

Table A2: Evaluation on COUGH (Aggregation scheme C).

Appendix A Annotation Protocols

We published our annotation batches on Amazon Mechanical Turk platform. Annotation protocols are provided below to facilitate future research in FAQ retrieval. Figure A1 and A2 show the user interfaces designed for both annotation tasks.

A.1 Task 1: Query Bank Construction

For this task, you are expected to give one shallow paraphrase and two deep paraphrases for the query template. Note that query can be either in question form or query string form.

Shallow paraphrase: Applying word substitution, sentence reordering and other lexical tricks (e.g. extracting salient phrases from response) to the original query to come up with another query without changing the meaning.

Deep paraphrase: The paraphrased ones should look dramatically (i.e. grammatically) different from the original query which is more than shallow paraphrasing. However, the paraphrased query should share the same (or almost same) semantic meaning as the original query.

A.2 Task 2: Relevance Set Construction

For this task, you will see a FAQ item retrieved by an automatic tool for a particular user query, and your job is to judge the relevance of the FAQ item based on the annotation scheme shown below.

Matched: The candidate FAQ matches the user query perfectly. (Query part of FAQ is semantically identical to the user query, and answer part of FAQ well answers the user query.)

Useful: The candidate FAQ doesn't perfectly match the user query but may still give some or enough information to help answer the user query. (Query part of FAQ is semantically similar to the user query, and you can either extract or infer some information from the answer which could be useful to the user query. Or alternatively, the candidate FAQ provides too much extra information which is not necessary.)

Useless: The candidate FAQ is topically related to the user query but doesn't provide useful information. (Query part of FAQ is somewhat related to the user query, but you can't get any useful information out of the answer part to confidently answer the user query.)

Non-relevant: The candidate FAQ is completely unrelated to the query.

Appendix B Aggregation Schemes

In this work, we introduce four aggregation schemes to obtain binary labels.

A. Annotated ⟨Query, FAQ item⟩ tuples with mean score ≥ 3 are positives.

B. Annotated ⟨Query, FAQ item⟩ tuples with mean score > 3 are positives.

C. Annotated ⟨Query, FAQ item⟩ tuples that have at least one[13] "Matched" annotation are positives.

[13] For tuples with more than 3 annotations, we raise the bar to two "Matched".

Method	P@1	P@5	MAP	MRR	nDCG
BM25 (Q-q)	77.1	65.8	32.2	86.1	76.7
BM25 (Q-a)	49.5	39.2	18.3	62.4	46.4
BM25 (Q-q+a)	76.0	62.8	32.7	85.2	72.6
BERT (Q-q)	81.6	**68.5**	30.7	89.1	**78.6**
+ fine-tune on pesudo Q-q	77.5	54.9	27.3	85.0	63.0
BERT (Q-a)	20.5	14.1	5.1	32.4	16.7
+ fine-tune on FAQ Bank	66.8	51.2	27.1	78.1	56.4
+ re-rank	67.2	52.9	27.6	78.4	57.8
CombSum	**84.3**	67.4	**40.8**	**90.6**	74.7
- fine-tuned BERT (Q-a)	80.9	65.5	35.1	88.5	75.2

Table A3: Evaluation on COUGH (Aggregation scheme D).

D. For each annotated ⟨Query, FAQ item⟩ tuple, we convert "Matched" and "Useful" to positive annotations, and "Useless" and "Non-relevant" to negative annotations. We then apply majority voting using converted binary annotations.

B.1 Results for Different Aggregation Schemes

Results based on aggregation schemes B, C and D are shown in Tables A1, A2 and A3, respectively. Results based on aggregation scheme A are shown in Table 3.

Appendix C Implementation Details

We first preprocess user query and FAQ items with nltk porter stemmer 5[14]. For baselines including BM25[15] and Sentence-BERT[16], we take the standard off-the-shelf version. More specifically, we keep the default $k1$ as 2 and b as 0.75 for BM25 over Q-q, Q-a and Q-q+a settings. When deploying synthetic query generation model (i.e., GPT2), hyper-parameters are set as instructed by Mass et al. (2020) (see their Section 3.4). We adopt the in-batch negatives training strategy to fine-tune both Sentence-BERT and cross-encoder BERT. For both BERT models, we use the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 1e-5 and fine-tune up to 10 epochs. We set the batch sizes as 24 and 4 for Sentence-BERT and cross-encoder BERT, respectively. All experiments are conducted using one single GeForce GTX 2080 Ti 12 GB GPU (with significant CPU resources).

[14]https://www.nltk.org/

[15]https://pypi.org/project/rank-bm25/

[16]https://github.com/UKPLab/sentence-transformers and we use *distilbert-base-nli-stsb-quora-ranking* model card.

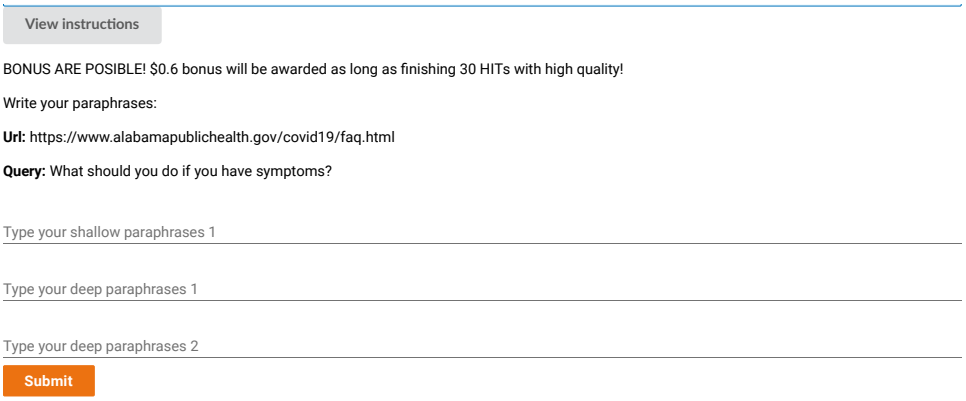

Figure A1: User interface for Query Bank construction task.

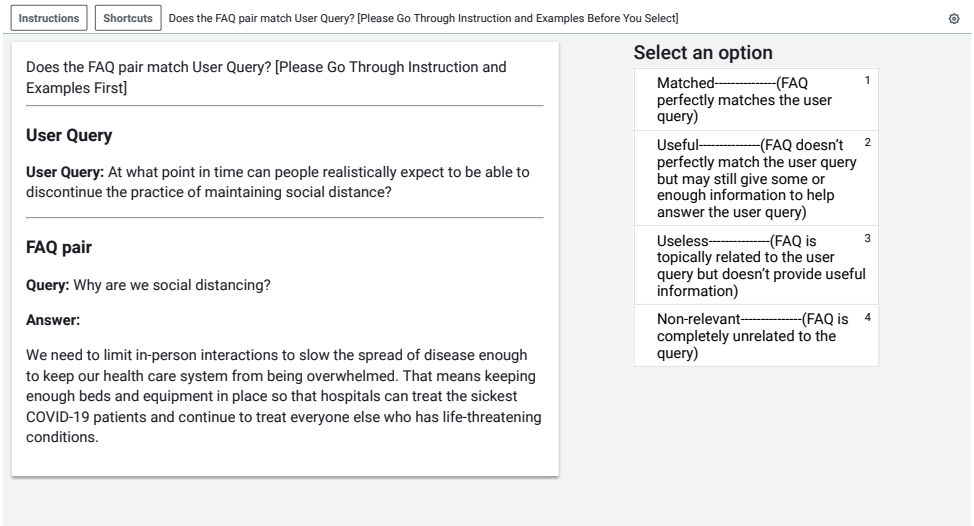

Figure A2: User interface for Relevance Set construction task.

	# of FAQ Items
Arizona Health Care Cost Containment System	138
Alabama Public Health	89
American Medical Association	14
California Department of Health	28
Government of Canada	131
Centers for Disease Control and Prevention	378
Children's Hospital Los Angeles	73
Bloomberg Harvard City Leadership Initiative	186
Cleveland Clinic	15
CNN	112
Government of Colorado	66
Delaware Department of Health	71
U.S. Food and Drug Administration	139
European Centre for Disease Prevention and Control	55
Florida Department of Health	47
Georgia Department of Labor	16
Explore Georgia	13
Government of United Kingdom	53
Harvard Health Publishing	104
Illinois Department of Public Health	37
Inspire	1753
JHU HUB	7
JHU Medicine	14
Kids Health from Nemours	121
King County, Washington	26
Government of Massachusetts	17
Medical News Today	28
MedHelp	282
Government of Michigan	75
Minnesota Department of Health	98
New York Times	100
Government of New Jersey	322
National Institute of Health	105
Government of North Carolina	59
Government of New York	75
New York State Electric and Gas	68
New York Department of Financial Services	45
Pennsylvania Office of Unemployment Compensation	222
Government of Pennsylvania	66
University of Pennsylvania Health System	63
Sante Clara Department of Health	103
San Mateo County Health	47
Texas Health Services	39
Tricare	94
United Nations	40
United States Department of Agriculture	152
United States Department of Labor	43
Virginia Department of Health	435
United States Department of Veterans Affairs	16
Washington Department of Health	137
WHO	29
World Health Organization	395
WikiHow	2371
Total	9151

Table A4: Number of English FAQ items scrapped from each source.

	language	# of FAQ Items
Centers for Disease Control and Prevention	Spanish	268
Centers for Disease Control and Prevention	Korean	244
Centers for Disease Control and Prevention	Vietnamese	244
Centers for Disease Control and Prevention	Chinese	244
Children's Hospital Los Angeles	Arabic	30
Children's Hospital Los Angeles	Spanish	45
Children's Hospital Los Angeles	Persian	39
Children's Hospital Los Angeles	Armenian	38
Children's Hospital Los Angeles	Kanuri	32
Children's Hospital Los Angeles	Chinese	34
U.S. Food and Drug Administration	Spanish	83
Japan Health	Japanese	226
Japan Labor	Japanese	63
United Nations	Arabic	39
United Nations	Spanish	38
United Nations	French	37
United Nations	Chinese	38
World Health Organization	Arabic	328
World Health Organization	Spanish	356
World Health Organization	French	387
World Health Organization	Russian	301
World Health Organization	Chinese	367
WikiHow	Arabic	144
WikiHow	Czech	22
WikiHow	German	525
WikiHow	Spanish	310
WikiHow	Persian	49
WikiHow	French	301
WikiHow	Hindi	286
WikiHow	Indonesian	166
WikiHow	Italian	263
WikiHow	Japanese	286
WikiHow	Korean	128
WikiHow	Dutch	142
WikiHow	Portuguese	303
WikiHow	Russian	142
WikiHow	Thai	90
WikiHow	Vietnamese	101
WikiHow	Chinese	117
Total	-	6768

Table A5: Number of non-English FAQ items scrapped from each source and language.