Overview of the 2020 Epidemic Question Answering Track

Draft Overview Paper for Text Analysis Conference (TAC) 2020

TRAVIS R. GOODWIN and DINA DEMNER-FUSHMAN, U.S. National Library of Medicine
KYLE LO and LUCY LU WANG, Allen Institute for AI
WILLIAM R. HERSH, Oregon Health & Science University
HOA T. DANG and IAN M. SOBOROFF, National Institute for Standards and Technology

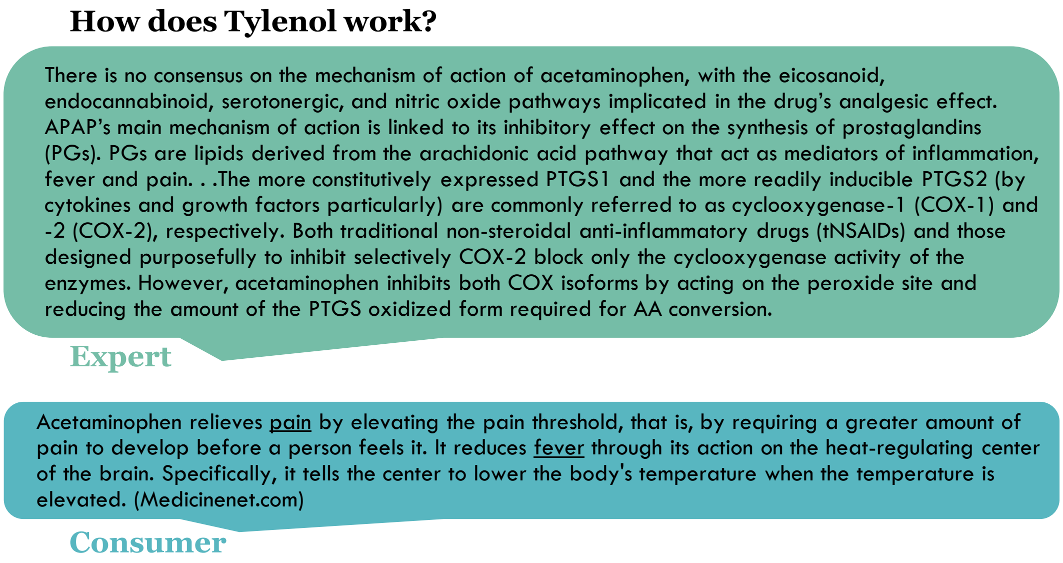
This document describes the Epidemic Question Answering (EPIC-QA) track for the 2020 Text Analysis Conference (TAC). The goal of the track is to challenge the community and foster research in the design of systems that can automatically answer questions about the 2019 novel coronavirus, COVID-19, by extracting answers in language and detail sufficient for experts, consumers, or both. Importantly, although the questions asked by different stakeholders often overlap, we hypothesize that the best answers – even for the same questions – may prioritize different information, present information in different language, and originate from different documents in order to best meet the varying goals and backgrounds of the user. With this track, we hope to explore whether the same approaches and even the same automated systems can be used to answer questions posed by different stakeholders if the systems are provided with a basic description of the user, e.g, expert or consumer. Participants were provided with sets of expert-level and consumer-level questions pertaining to COVID-19, the virus SARS-CoV-2, related coronaviruses, and the recommended response to the pandemic as well as a collection of documents from which to retrieve answers. To support both types of users, the document collection contained (a) relevant scientific biomedical articles from the CORD-19 collection and (b) information from government websites, news articles, and social media expressed in (potentially) more consumer-friendly language.

1 OBJECTIVE

In response to the COVID-19 pandemic, we organized a new track for TAC 2020: **Epi**demi**c** **Q**uestion Answering (EPIC-QA). This track challenges teams to develop systems capable of automatically answering ad-hoc questions about the disease COVID-19, its causal virus SARS-CoV-2, related coronaviruses, and the recommended response to the pandemic. While COVID-19 has been an impetus for a large body of emergent scientific research and inquiry, the response to COVID-19 raises questions for consumers. The rapid increase in coronavirus literature and evolving guidelines on community response creates a challenging burden not only for the scientific and medical communities but also the general public to stay up-to-date on the latest developments. Consequently, the goal of the track is to evaluate systems on their ability to provide timely and accurate expert-level answers as expected by the scientific and medical communities as well as answers in consumer-friendly language for the general public.

While there is overlap in the types of questions asked by different stakeholders, the answers to such questions should vary based on the background knowledge of the user. For example, consider the simple question from the general health domain illustrated in Figure 1, *How does Tylenol work?* In the eyes of an expert, an answer should indicate that the exact mechanism is unknown, but may elaborate on the pathways involved.[6] By contrast, for a consumer, this information would be unhelpful; instead, a more appropriate answer would provide a general overview of how the class of drugs works rather than the exact mechanism of action. In

Authors' addresses: Travis R. Goodwin; Dina Demner-Fushman, U.S. National Library of Medicine, nlmlhclhcques@mail.nih.gov; Kyle Lo; Lucy Lu Wang, Allen Institute for AI; William R. Hersh, Oregon Health & Science University; Hoa T. Dang; Ian M. Soboroff, National Institute for Standards and Technology.



How does Tylenol work?

There is no consensus on the mechanism of action of acetaminophen, with the eicosanoid, endocannabinoid, serotonergic, and nitric oxide pathways implicated in the drug's analgesic effect. APAP's main mechanism of action is linked to its inhibitory effect on the synthesis of prostaglandins (PGs). PGs are lipids derived from the arachidonic acid pathway that act as mediators of inflammation, fever and pain. . .The more constitutively expressed PTGS1 and the more readily inducible PTGS2 (by cytokines and growth factors particularly) are commonly referred to as cyclooxygenase-1 (COX-1) and -2 (COX-2), respectively. Both traditional non-steroidal anti-inflammatory drugs (tNSAIDs) and those designed purposefully to inhibit selectively COX-2 block only the cyclooxygenase activity of the enzymes. However, acetaminophen inhibits both COX isoforms by acting on the peroxide site and reducing the amount of the PTGS oxidized form required for AA conversion.

Acetaminophen relieves pain by elevating the pain threshold, that is, by requiring a greater amount of pain to develop before a person feels it. It reduces fever through its action on the heat-regulating center of the brain. Specifically, it tells the center to lower the body's temperature when the temperature is elevated. (Medicinenet.com)

Fig. 1. Example of an expert and consumer answer for a "simple" general health question.

the context of the rapidly accelerating knowledge of COVID-19, managing this duality between expert- and consumer-level information is even more important. It is our hope that the track will stimulate research in automatic question answering not only to support providing high-quality timely information about COVID-19 but also that the resultant collection can be used to develop generalizable approaches to meeting information needs in the face of varying levels of expertise.

2 BACKGROUND

A pneumonia of unknown origin was detected in Wuhan, China, and was first reported to the World Health Organization (WHO) on December 31, 2019. On January 30, 2020, the outbreak had escalated to the point that it was declared a Public Health Emergency of International Concern. The WHO officially named the 2019 coronavirus disease "COVID-19" on February 11, 2020. By March 11, 2020, after more than 118,000 reported cases in 114 countries resulting in over 4,291 reported fatalities, the WHO formally characterized COVID-19 as a pandemic. In the United States, in March 2020, various states and cities began issuing mandatory quarantine ordinances as well as guidance on "social distancing" and forced closures of gyms, bars, and nightclubs. As of April 7, 41 states as well as the District of Columbia had issued mandatory self-quarantine directives, forbidding non-essential activity outside the home. Over this period, there has been a rapid escalation in scientific research on COVID-19 and related coronaviruses as well as government and community response to prevent or maintain the outbreak. For example, the scientific community has sequenced the SARS-CoV-2 genome, proposed multiple vaccines, and explored antibody, anti-viral, and cell-based treatments. The rapid escalation of government and community response has resulted in a large burden for consumers as well as scientists and healthcare professionals seeking to maintain up-to-date knowledge on COVID-19 as well as the recommended response and adjustments to their daily lives. Consequently, the EPIC-QA track at TAC 2020 aims to help reduce this burden by fostering research in the design of automatic question answering systems to support scientific and consumer inquiry into COVID-19 and the recommended response.

It is our hope that the track will stimulate research in automatic question answering not only to support providing high-quality timely information about COVID-19 but also that the resultant collection can be used to develop generalizable approaches to meeting information needs in the face of varying levels of expertise.

What is the origin of COVID-19?

Consumer

① COVID-19 is caused by a new coronavirus. ② Coronaviruses are a large family of viruses that are common in people and many different species of animals, including camels, cattle, cats, and bats. Rarely, animal coronaviruses can infect people and then spread between people such as with MERS-CoV, SARS-CoV, and now with this new virus (named SARS-CoV-2).

The SARS-CoV-2 virus is a betacoronavirus, like MERS-CoV and SARS-CoV. ③ All three of these viruses have their origins in bats. ④ The sequences from U.S. patients are similar to the one that China initially posted, suggesting a likely single, recent emergence of this virus from an animal reservoir.

https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/summary.html

Expert

① It is improbable that SARS-CoV-2 emerged through laboratory manipulation of a related SARS-CoV-like coronavirus. As noted above, the RBD of SARS-CoV-2 is optimized for binding to human ACE2 with an efficient solution different from those previously predicted.[7,11] ② Furthermore, if genetic manipulation had been performed, one of the several reverse-genetic systems available for betacoronaviruses would probably have been used.[19] ③ However, the genetic data irrefutably show that SARS-CoV-2 is not derived from any previously used virus backbone.[20] ④ Instead, we propose two scenarios that can plausibly explain the origin of SARS-CoV-2: (i) natural selection in an animal host before zoonotic transfer; and (ii) natural selection in humans following zoonotic transfer.

https://www.nature.com/articles/s41591-020-0820-9[1]

Fig. 2. Example answers for the question, *What is the origin of COVID-19?*

3 THE TRACK

The 2020 EPIC-QA track involves two tasks:

(1) **Task A: Expert QA.** In Task A, teams are provided with a set of questions asked by experts and are asked to provide a ranked list of expert-level answers to each question.
(2) **Task B: Consumer QA**. In Task B, teams are provided with a set of questions asked by consumers and are asked to provide a ranked list of consumer-friendly answers to each question.

While each task has its own set of questions, many of the questions will overlap. This is by design so that the collection can be used to explore whether the same approaches or systems can account for different types of users.

In this track, answers must be in the form of consecutive sentences extracted from a single *context* in a single document. Contexts and sentence IDs were provided to the participants as part of the collection. In the CORD-19 collection, contexts correspond to paragraphs defined by the authors of their publications. In the government, news and web collection, contexts correspond to sections indicated by the HTML of the document. Contexts that are longer than 15 sentences were segmented into approximately 15-sentence chunks. Contexts were further segmented into sentences, each associated with a unique ID. The participants were required to provide the starting and ending IDs of the sentences that constitute each of their answers. To maintain provenance, each answer must also be associated with the document and context IDs from which it originated. Figure 2 illustrates an example consumer-friendly and expert-level answer to the question "What is the origin of COVID-19?"

3.1 Evaluation Cycles

The 2020 EPIC-QA track was conducted in two evaluation cycles: an initial preliminary cycle to develop a training collection, followed by a final primary cycle for system evaluation. In the initial preliminary cycle, questions were taken from the TREC 2020 COVID-Search track allowing participants to use the document-level relevance judgments produced for TREC to quickly develop QA systems. In the final evaluation cycle, new questions were created and participants did not have access to document-level relevance judgments. However, the judgments produced for the preliminary evaluation cycle were available to all participants (including those that did not participate in the preliminary cycle).

4 DATA

The TAC 2020 EPIC-QA data consists of (1) two sets of questions, one asked by consumers as well as another asked by experts pertaining to COVID-19, and (2) a collection of documents relevant to COVID-19 including published and pre-print biomedical research articles as well as documents obtained from government websites, news, and social media. The final dataset also includes judgments in the form of human-annotated sentences in the document collection. It is our hope that this dataset will help future research on COVID-19, questions answering for other epidemics or pandemics, and methods to account for different levels of expertise and background knowledge in question answering.

Question 1
Question: What is the origin of COVID-19?
Background: seeking range of information about the SARS-CoV-2 virus's origin, including its evolution, animal source, and first transmission into humans

Question 2
Question: How does the coronavirus respond to changes in the weather?
Background: seeking range of information about the SARS-CoV-2 virus viability in different weather/climate conditions as well as information related to transmission of the virus in different climate conditions

Question 3
Question: Will SARS-CoV2 infected people develop immunity? Is cross protection possible?
Background: seeking studies of immunity developed due to infection with SARS-CoV2 or cross protection gained due to infection with other coronavirus types

Fig. 3. Example questions.

```
{
    "document_id": <str>,                       # 40-character sha1 of the URL or document
    "metadata": {
        "title": <str>,                         # Title of the document
        "url": <str>,                           # URL of the page (for webpages)
        "authors": [...]                        # Authors of the page (for CORD-19)
    },
    "contexts": [                               # List of context(s) in the document
        {
            "section": <str>,                   # Name of the section (if any) containing the context
            "text": <str>,                      # The full text (without markup) in the section
            "context_id": <str>,                # Globally unique context identifier
            "sentences:" [
                {
                    "start": 0,                 # Inclusive character start offset of the sentence
                    "end": 27,                  # Exclusive character end offset of the sentence
                    "sentence_id": <str>,       # Globally unique identifier for the sentence
                },
                {                               # Second sentence in the context
                    "start": 28,
                    "end": 56,
                    "sentence_id": <str>,
                },
                {...},                          # Third sentence in the context
                ...
            ]
        },
        {...},                                  # Second context in the document
        ...
    ]
}
```

Fig. 4. JSON Schema for full-text documents.

4.1 The Questions

Two sets of questions were provided: one for expert-level questions and one for consumer-level questions. The goal of the first, preliminary, evaluation cycle is to produce data that can be used to develop systems for the final evaluation cycle in the fall. To reduce the barrier-to-entry, we used 45 topics evaluated in the fourth round of TREC-COVID. Specifically, The majority of these questions originated from consumers' interactions with MedlinePlus®. Additional scientific questions were developed based on group discussions from the National Institutes of Health (NIH) special interest group on COVID-19, questions asked by Oregon Health Science University clinicians, and responses to a public call for questions. For the primary evaluation a new set of 30 questions were developed (none of which were evaluated in TREC-COVID). Example questions are provided in Figure 3.

4.2 The Document Collection

The document collection provided to participants for EPIC QA adhered to the JSON schema illustrated in Figure 4. Importantly, each document in the collection included explicitly pre-defined *contexts* (a generalization of paragraphs or sections) and sentences. To support both levels of expertise, the document collection consists of two parts: a scientific and medical research collection, and a consumer-oriented collection from the web.

4.2.1 The Scientific COVID-19 Collection. We adapt the collection of biomedical articles released for the COVID-19 Open Research Dataset Challenge[1] (CORD-19).[8] The dataset was created by the Allen Institute for AI in partnership with the Chan Zuckerberg Initiative, Georgetown University's Center for Security and Emerging Technology, Microsoft Research, and the National Library of Medicine – National Institutes of

[1] https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

Health, in coordination with The White House Office of Science and Technology Policy. The CORD-19 collection includes a subset of articles in PubMed Central[2] (PMC) as well as pre-prints from bioRxiv[3] and medRxiv[4]. Contexts in this collection correspond to automatically identified paragraphs in the articles' abstracts, or main texts. A snapshot from June 19, 2020, was used for the preliminary evaluation and a snapshot from October 22, 2020, was used for the final evaluation cycle.

4.2.2 *The Consumer-friendly Government Collection.*

We include a subset of the articles used by the Consumer Health Information Question Answering (CHIQA) service of the U.S. National Library of Medicine (NLM).[2] This collection includes authoritative articles from the Centers for Disease Control and Prevention (CDC); the Genetic and Rare Disease Information Center (GARD); the Genetics Home Reference (GHR); Medline Plus; the National Institute of Allergy and Infectious Diseases (NIAID); the World Health Organization (WHO). Contexts in this collection correspond to paragraphs or sections as indicated by the HTML markup of the document. In the final, primary evaluation cycle, we also included 265 Reddit threads from `/r/askscience` tagged with COVID-19, Medicine, Biology, or the Human Body as well as a subset of the CommonCrawl News Crawl (CCNC) from January 1 through April 30, 2020, as used in the TREC Health Misinformation Track. CCNC documents were filtered by domain using SALSA,[5] PageRank,[7] and HITS.[4] All documents in the consumer collection were filtered for COVID-19 content.

5 EVALUATION

For each task, participants were asked to provide ranked lists of answers for each question. Answer judgments were provided by librarian indexers at the U.S. National Library of Medicine (NLM). **Note:** in this task, the goal is to explore the landscape of answers asserted by the document collection. A statement that answers the question is considered as a valid answer regardless of whether or not it is factually accurate. The answers in this task are intended as an intermediary step where-in one would like to explore all answers provided by the document collection – both correct answers as well as incorrect answers that people may have discovered on their own.

In each of the two evaluation cycles assessors were assigned one or more questions. Each question was associated with the pooled set of answers returned by participating teams. In both evaluation cycles, the assessment was made in two rounds: (1) an answer-key generation round, followed by (2) an answer annotation round.

Round 1: Answer Key Generation. In the first round, assessors had access to a set of contexts containing answers produced by participating teams as well as an ad-hoc search engine over the document collection. The goal in this round is for assessors to explore the answers from teams as well as the document collection to determine a set of atomic "facts" that answer the question. Following TAC tradition, we refer to these facts as *nuggets*. Example nuggets that produced for the question "What is the origin of COVID-19?" may include *Malayan pangolins*, *bats*, *Guangdon province*, etc. The primary role of this round is to create an answer key for the question comprised of nuggets identified in participants' answers or the assessor's own ad-hoc search of the collection. The search engine is provided to help assessors explore the topic and identify nuggets that may not have been returned by any teams. Assessors were not expected to exhaustively identify every possible

[2] https://www.ncbi.nlm.nih.gov/pmc/
[3] https://www.biorxiv.org/
[4] https://www.medrxiv.org/

nugget that was not returned by teams; rather the intent is for them to identify important (at the discretion of the assessor) nuggets that they feel should be included in the answer key based on their understanding of the topic.

Round 2: Sentence Annotation. In the second round, the answer key (list of nuggets) was fixed. Assessors were given the same set of contexts used in round one. This time, they were asked to annotate sentences in each context indicating which nugget(s) (if any) are addressed by each sentence. For example, the sentence "Malayan pangolins (Manis javanica) illegally imported into Guangdong province contain coronaviruses similar to SARS-CoV-2" could be annotated as containing two nuggets: *Malayan pangolins* and *Guangdon province.* Un-annotated sentences were assumed to contain no nuggets.

5.1 System Evaluation

We separately evaluated system performance on Tasks A and B using a modified version of Normalized Discounted Cumulative Gain[3] (NDCG) which we refer to as the Normalized Discount Novelty Score (NDNS). Importantly, while the cumulative gain in NDCG can be computed for a document independently of the other retrieved documents, the Novelty Score (NS) measures the information in an answer that has not been seen previously in the ranked list. Formally, we define the novelty score of an answer:

$$NS(a) = \frac{n_a * (n_a + 1)}{n_a + f_a} \tag{1}$$

where n_a is the number of *novel* nuggets in answer a and f_a is the *sentence factor.* A nugget is considered novel if it has not been present in an answer retrieved earlier in the ranked list. We report three variants of NDNS in which the sentence factor, f_a, is computed differently:

(1) **Exact**: answers must be brief (i.e., they must express a novel nugget in as few sentences as possible) and they should not contain sentences with only nuggets provided in previous answers. In this variant, the sentence factor is just the number of sentences in the answer, i.e.,

$$f_a = s_a = s_0 + s_s + s_n \tag{2}$$

where s_a is the number of sentences in answer a, s_0 is the number of sentences with no nuggets, s_s is the number of sentences with previously seen nuggets, and s_n is the number of sentences with novel nuggets.

(2) **Relaxed**: answers are not penalized for expressing novel nuggets in multiple sentences, but should still not contain sentences with only nuggets provided in previous answers, i.e.,

$$f_a = s_0 + s_s + \min(s_n, 1) \tag{3}$$

(3) **Partial**: answers are not penalized for expressing novel nuggets in multiple sentences, nor for containing sentences with only nuggets from previous answers. In this variant, we only penalize answers for sentences with no nuggets at all, i.e.,

$$f_a = s_0 + \min(s_n, 1) \tag{4}$$

As in NDCG, we compute the cumulative novelty score NS of answers retrieved up to rank l and discount the score using a logarithmic reduction factor:

$$DNS(a_1, \cdots, a_l) = \sum_{r=1}^{l} \frac{NS(a_r)}{\log_2 (r+1)} \tag{5}$$

Finally, we normalize the NDNS of ranking $\boldsymbol{a} = a_1, \cdots, a_l$ by the NDNS of the optimal or ideal ranking of possible answers that could have been retrieved for that question:

$$NDNS(\boldsymbol{a}) = \frac{NDNS(\boldsymbol{a})}{NDNS(\hat{\boldsymbol{a}})} \tag{6}$$

where $\hat{\boldsymbol{a}}$ is the optimal ranking of answers that could have been found in the document collection for the given question. In our evaluation, we used beam-search with a width of 10 to determine the ideal ranking of answers.

5.2 Participation

A total of eleven teams submitted runs for EPIC-QA. Table 1 illustrates which evaluation cycle(s) and tasks each team participated in.

6 RESULTS

6.1 Preliminary Evaluation Results

For the preliminary evaluation of Task A, 21 topics were judged using depth-five pooling over 17 runs submitted by eight teams. Table 2 shows these results. In the preliminary evaluation of Task B, 18 questions were judged using depth-eight pooling over 10 runs submitted by six teams. Table 3 presents these results. In Task A, the `ixa` and `IBM` runs exhibited the best performance (with `HLTRI` scoring well under the partial scoring strategy). By contrast, for Task B, the `HLTRI` runs obtained the highest performance followed by `IBM` for all NDNS variants. The descriptions of each run are provided in Appendix A.

6.2 Primary Evaluation Results

For the primary evaluation, all 30 questions were judged. Figure 5 compares the average NDNS Exact scores for each run between both tasks. In Task A, answers were pooled through depth five from 16 runs submitted by seven teams. Table 4 shows these results. For Task B, answers were pooled at depth eight from 12 runs submitted by five teams. Table 5 presents these results. In Task A, the `HLTRI` runs exhibited the best performance, with `Yastil_R_1` outperforming `HLTRI_1`. By contrast, for Task B, the `h2oloo` runs obtained the highest performance followed by `HLTRI` for all NDNS variants. The descriptions of each run are provided in Appendix B.

7 CONCLUSION

The goal of the 2020 TAC Epidemic QA Track was to evaluate and draw attention to the important problem of answering questions about COVID-19 from emergent literature, as well as to explore the differences between consumer- and expert- level question answering techniques. The track was organized into two evaluation cycles: an initial preliminary evaluation cycle primarily aimed at generating data, and a final primary evaluation cycle. Both cycles evaluated automatic systems for extractive epidemic question answering aimed at

	Preliminary		Primary	
Team	Task A	Task B	Task A	Task B
CORONAWHY CoronaWhy		✓		
covidbert Johns Hopkins University New York University Korea University	✓	✓		
Dindadiel (Unaffiliated)	✓			
h2oloo University of Waterloo			✓	✓
HLTRI Human Language Technology Research Institute at the University of Texas at Dallas	✓	✓	✓	✓
IBM IBM Research AI	✓	✓	✓	✓
ixa Ixa NLP Group at the University of the Basque Country (UPV/EHU)	✓			
nlm_lhc_qa Lister Hill National Center for Biomedical Communications at the U.S. National Library of Medicine	✓	✓	✓	✓
UPC_USMBA Universitat Politécnica de Catalunya University Sidi Mohammed Ben Abdellah	✓	✓	✓	✓
vigicovid Universidad Nacional de Educación a Distancia University of the Basque Country (UPV/EHU) Elhuyar	✓		✓	
Yastil_R Centre for AI Research South Africa University of KwaZulu-Natal			✓	

Table 1. Teams (and their affiliations) that participated in EPIC-QA 2020.

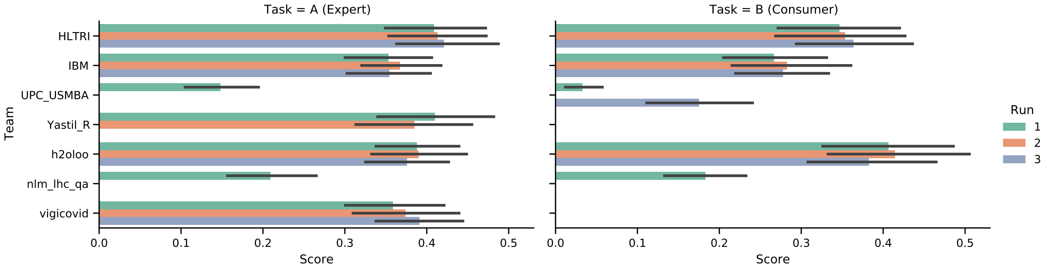


Fig. 5. Average NDNS-Exact scores obtained by each submitted run for the primary evaluation cycle.

	NDNS-Partial	NDNS-Relaxed	NDNS-Exact
1.	0.305 ixa_3	0.307 ixa_3	**0.341 ixa_3**
2.	0.303 ixa_2	0.304 ixa_2	0.338 ixa_2
3.	0.302 HLTRI_1	0.295 IBM_3	0.327 IBM_1
4.	0.294 IBM_3	0.294 IBM_1	0.327 IBM_2
5.	0.294 IBM_1	0.293 IBM_2	0.325 IBM_3
6.	0.293 IBM_2	0.293 HLTRI_1	0.306 ixa_1
7.	0.276 ixa_1	0.277 ixa_1	0.300 vigicovid_2
8.	0.266 vigicovid_2	0.267 vigicovid_2	0.297 vigicovid_3
9.	0.263 vigicovid_3	0.265 vigicovid_3	0.289 HLTRI_1
10.	0.226 UPC_USMBA_1	0.218 UPC_USMBA_1	0.215 vigicovid_1
11.	0.204 UPC_USMBA_2	0.198 UPC_USMBA_2	0.212 UPC_USMBA_1
12.	0.191 vigicovid_1	0.192 vigicovid_1	0.192 UPC_USMBA_2
13.	0.149 covidbert_2	0.149 Dindadiel_1	0.165 Dindadiel_1
14.	0.148 Dindadiel_1	0.144 covidbert_1	0.158 covidbert_1
15.	0.143 covidbert_1	0.142 covidbert_2	0.131 nlm_lhc_qa_2
16.	0.134 nlm_lhc_qa_2	0.132 nlm_lhc_qa_2	0.131 covidbert_2
17.	0.113 nlm_lhc_qa_1	0.111 nlm_lhc_qa_1	0.109 nlm_lhc_qa_1

Table 2. Preliminary evaluation results for Task A using Normalized Discounted Novelty Score (NDNS).

	NDNS-Partial	NDNS-Relaxed	NDNS-Exact
1.	0.488 HLTRI_1	0.482 HLTRI_1	**0.475 HLTRI_1**
2.	0.398 IBM_1	0.396 IBM_1	0.413 IBM_1
3.	0.394 IBM_3	0.394 IBM_3	0.409 IBM_3
4.	0.374 IBM_2	0.372 IBM_2	0.390 IBM_2
5.	0.364 covidbert_1	0.364 covidbert_1	0.389 covidbert_1
6.	0.315 UPC_USMBA_1	0.307 UPC_USMBA_1	0.302 UPC_USMBA_1
7.	0.307 UPC_USMBA_2	0.299 UPC_USMBA_2	0.286 UPC_USMBA_2
8.	0.281 covidbert_2	0.276 nlm_lhc_qa_1	0.257 nlm_lhc_qa_1
9.	0.278 nlm_lhc_qa_1	0.272 covidbert_2	0.253 covidbert_2
10.	0.0590 CORONAWHY_1	0.0590 CORONAWHY_1	0.0430 CORONAWHY_1

Table 3. Preliminary evaluation results for Task B using Normalized Discounted Novelty Score (NDNS).

experts and general consumers, with a total of eleven teams participating. The results indicate that, without sentence-level training judgments, systems can discover useful answers in the collections (as indicated by the preliminary evaluation cycle), and, when provided with training data, the quality of answers can be substantially improved. Compared to other question answering tasks, the questions posed in EPIC-QA tend to be more open-ended and were evaluated according to the diversity of answers retrieved by systems, making direct comparisons to other question answering collections difficult. We believe the results of this evaluation demonstrate the importance of exploring the diverse landscape of answers available online for health questions and show the importance of accounting for varying levels of understanding when identifying satisfactory answers to health questions.

	NDNS-Partial	NDNS-Relaxed	NDNS-Exact
1.	0.421 HLTRI_3	0.370 HLTRI_3	**0.371 HLTRI_3**
2.	0.413 HLTRI_2	0.363 HLTRI_2	0.364 HLTRI_2
3.	0.410 Yastil_R_1	0.361 Yastil_R_1	0.362 Yastil_R_1
4.	0.408 HLTRI_1	0.359 HLTRI_1	0.360 HLTRI_1
5.	0.391 vigicovid_3	0.345 IBM_2	0.344 vigicovid_3
6.	0.390 h2oloo_2	0.344 vigicovid_3	0.344 IBM_2
7.	0.388 h2oloo_1	0.340 h2oloo_2	0.341 h2oloo_2
8.	0.385 Yastil_R_2	0.338 h2oloo_1	0.339 h2oloo_1
9.	0.376 h2oloo_3	0.337 Yastil_R_2	0.338 Yastil_R_2
10.	0.374 vigicovid_2	0.336 IBM_3	0.334 IBM_3
11.	0.367 IBM_2	0.331 IBM_1	0.329 vigicovid_2
12.	0.359 vigicovid_1	0.329 vigicovid_2	0.329 h2oloo_3
13.	0.354 IBM_3	0.328 h2oloo_3	0.327 IBM_1
14.	0.353 IBM_1	0.315 vigicovid_1	0.315 vigicovid_1
15.	0.209 nlm_lhc_qa_1	0.223 nlm_lhc_qa_1	0.219 nlm_lhc_qa_1
16.	0.148 UPC_USMBA_1	0.126 UPC_USMBA_1	0.127 UPC_USMBA_1

Table 4. Primary evaluation results for Task A using Normalized Discounted Novelty Score (NDNS).

	NDNS-Partial	NDNS-Relaxed	NDNS-Exact
1.	0.414 h2oloo_2	0.366 h2oloo_2	**0.368 h2oloo_2**
2.	0.407 h2oloo_1	0.359 h2oloo_1	0.361 h2oloo_1
3.	0.382 h2oloo_3	0.338 h2oloo_3	0.339 h2oloo_3
4.	0.363 HLTRI_3	0.316 HLTRI_3	0.317 HLTRI_3
5.	0.353 HLTRI_2	0.312 HLTRI_2	0.313 HLTRI_2
6.	0.346 HLTRI_1	0.304 HLTRI_1	0.305 HLTRI_1
7.	0.282 IBM_2	0.268 IBM_3	0.264 IBM_2
8.	0.278 IBM_3	0.268 IBM_2	0.263 IBM_3
9.	0.267 IBM_1	0.249 IBM_1	0.245 IBM_1
10.	0.183 nlm_lhc_qa_1	0.186 nlm_lhc_qa_1	0.184 nlm_lhc_qa_1
11.	0.175 UPC_USMBA_3	0.176 UPC_USMBA_3	0.172 UPC_USMBA_3
12.	0.0330UPC_USMBA_1	0.0300UPC_USMBA_1	0.0300UPC_USMBA_1

Table 5. Primary evaluation results for Task B using Normalized Discounted Novelty Score (NDNS).

ACKNOWLEDGEMENTS

This work was supported by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We would like to also thank Sam Skjonsberg and Mark Neumann for their help with the annotation platform, Max Savery for developing the NLM baseline runs, Rob Guzman for leading the judgment effort as well as Annemieke van der Sluijs, Cathy Smith, David Cissel, Deborah Whitman, Dorothy Trinh, Farzaneh Badiei, Funmi Akhigbe, Janice Ward, Jenny Rewolinski, Keiko Sekiya, Melanie Huston, Naim Rogers, Nick Miliaras, Oleg Rodionov, Olga Printseva, Rob Guzman, and Susan Schmidt for the roles in answer judging.

REFERENCES

[1] Andersen, K.G., Rambaut, A., Lipkin, W.I., Holmes, E.C., Garry, R.F.: The proximal origin of sars-cov-2. Nature medicine **26**(4), 450–452 (2020)

[2] Demner-Fushman, D., Mrabet, Y., Ben Abacha, A.: Consumer health information and question answering: helping consumers find answers to their health-related information needs. Journal of the American Medical Informatics Association **27**(2), 194–201 (10 2019). https://doi.org/10.1093/jamia/ocz152, https://doi.org/10.1093/jamia/ocz152

[3] Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of ir techniques. ACM Trans. Inf. Syst. **20**(4), 422–446 (Oct 2002). https://doi.org/10.1145/582415.582418, https://doi.org/10.1145/582415.582418

[4] Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM) **46**(5), 604–632 (1999)

[5] Lempel, R., Moran, S.: Salsa: The stochastic approach for link-structure analysis. ACM Trans. Inf. Syst. **19**(2), 131–160 (Apr 2001). https://doi.org/10.1145/382979.383041, https://doi.org/10.1145/382979.383041

[6] Mazaleuskaya, L.L., Sangkuhl, K., Thorn, C.F., FitzGerald, G.A., Altman, R.B., Klein, T.E.: Pharmgkb summary: pathways of acetaminophen metabolism at the therapeutic versus toxic doses. Pharmacogenetics and genomics **25**(8), 416 (2015)

[7] Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Tech. rep., Stanford InfoLab (1999)

[8] Wang, L.L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., Funk, K., Kinney, R.M., Liu, Z., Merrill, W., Mooney, P., Murdick, D.A., Rishi, D., Sheehan, J., Shen, Z., Stilson, B., Wade, A.D., Wang, K., Wilhelm, C., Xie, B., Raymond, D.M., Weld, D.S., Etzioni, O., Kohlmeier, S.: Cord-19: The covid-19 open research dataset. ArXiv **abs/2004.10706** (2020)

A PRELIMINARY SUBMISSIONS

A.1 Task A

Eight teams submitted runs for Task A in the preliminary evaluation cycle. The runs submitted by each team, as well as their provided descriptions are provided below.

`Dindadiel_1` More extensive training has been done

`HLTRI_1` We utilized a biobert-based neural passage re-ranker on sentence n-grams from the TREC-COVID relevance judgement documents.

`IBM_1` IR: Ensemble of elastic search and a dense passage retriever
MRC: Roberta-large model (GAAMA) trained on Natural Questions
Final submission: Combination / weighted average of the IR + MRC system

`IBM_2` IR: Elastic Search + Neural IR classifier for re-ranking
MRC: Roberta-large model (GAAMA) trained on Natural Questions
Final submission: Combination / weighted average of the IR + MRC system

`IBM_3` IR: Ensemble of elastic search with neural re-ranking and a dense passage retriever
MRC: Roberta-large model (GAAMA) trained on Natural Questions
Final submission: Combination / weighted average of the IR + MRC system

`UPC_USMBA_1` Indexing and Searching Documents using Lucene
Applying BERT trained on SQUAD to return answer spans
Using Question, Query and Background and NER extracted from their combination

`UPC_USMBA_2` Indexing and Searching Documents using Lucene
Applying COVID-19 BERT pre-trained model on SQUAD to return answer spans
Using Question, Query and Background and NER extracted from their combination

`covidbert_1` BERT model trained on in-house annotated COVID FAQ data

covidbert_2 BERT model trained on in-house annotated COVID FAQ data

ixa_1 A neural QA system, that given a question in natural language (text from 'question' field) and a document (relevant according to TREC-COVID relevance judgments) returns the answer to the question in the document. For implementing the neural QA system, we have used the SciBERT language representation model that has been fine-tuned on the SQuAD2.0 dataset.

ixa_2 A neural QA system, that given a question in natural language (text from 'question' field) and a document (relevant according to TREC-COVID relevance judgments) returns the answer to the question in the document. For implementing the neural QA system, we have used the SciBERT language representation model that has been fine-tuned first on the SQuAD2.0 dataset, and then on the QuAC dataset.

ixa_3 A neural QA system, that given a question in natural language (text from 'question' field) and a document retrieved using our IR system returns the answer to the question in the document. For implementing the neural QA system, we have used the SciBERT language representation model that has been fine-tuned first on the SQuAD2.0 dataset, and then on the QuAC dataset.

nlm_lhc_qa_1 This run involved a two-step process: (1) Using the search engine Essie to rank the CORD collection; and (2), having obtained the ranked results returned by Essie for each query, using the summarization algorithm BART to perform question-driven summarization of contexts longer than 100 tokens. These summaries are fairly extractive and thus were then mapped map to the original sentences in the context. Contiguous sentences for which there were summary mappings were used as the answers.

nlm_lhc_qa_2 This run involved a two-step process: In the first step, the search engine Essie was used to rank the CORD collection for each query. For this run, only documents with TREC relevance judgments of "2" were used. The next step consisted of using the summarization algorithm BART to summarize contexts longer than 120 tokens. These summaries are fairly extractive and thus were then mapped map to the original sentences in the context. Contiguous sentences for which there were summary mappings were used as the answers.

vigicovid_1 A neural QA system, that given a question in natural language (text from 'background' field) and a document (relevant according to TREC-COVID relevance judgments) returns the answer to the question in the document. For implementing the neural QA system, we have used the SciBERT language representation model that has been fine-tuned on the SQuAD2.0 dataset.

vigicovid_2 A neural QA system, that given a question in natural language (text from 'background' field) and a document (relevant according to TREC-COVID relevance judgments) returns the answer to the question in the document. For implementing the neural QA system, we have used the SciBERT language representation model that has been fine-tuned first on the SQuAD2.0 dataset, and then on the QuAC dataset.

vigicovid_3 A neural QA system, that given a question in natural language (text from 'background' field) and a document retrieved using our IR system returns the answer to the question in the document. For implementing the neural QA system, we have used the SciBERT language representation model that has been fine-tuned first on the SQuAD2.0 dataset, and then on the QuAC dataset.

A.2 Task B

Five teams submitted runs for Task B in the preliminary evaluation cycle. The runs submitted by each team, as well as their provided descriptions are provided below.

CORONAWHY_1 My approach was to extract the document URLs where available (most input docs) and load those into the Azure QnA Maker service to build a knowledge base. Then I've been calling its REST API to ask questions and collect its answers. The last step was matching the answers back to the specific sentences - I tried fuzzy matching but the results are …fuzzy. Probably the requirements of this task are too constrained for my approach.

HLTRI_1 We utilized a biobert-based neural passage re-ranker on sentence n-grams from the full corpus.

IBM_1 IR: Ensemble of elastic search and a dense passage retriever
MRC: Roberta-large model (GAAMA) trained on Natural Questions
Final submission: Combination / weighted average of the IR + MRC system

IBM_2 IR: Elastic Search + Neural IR classifier for re-ranking
MRC: Roberta-large model (GAAMA) trained on Natural Questions
Final submission: Combination / weighted average of the IR + MRC system

IBM_3 IR: Ensemble of elastic search with neural re-ranking and a dense passage retriever
MRC: Roberta-large model (GAAMA) trained on Natural Questions
Final submission: Combination / weighted average of the IR + MRC system

UPC_USMBA_1 Indexing and Searching Documents using Lucene
Applying BERT trained on SQUAD to return answer spans
Using Question, Query and Background and NER extracted from their combination

UPC_USMBA_2 Indexing and Searching Documents using Lucene
Applying COVID-19 BERT pre-trained model on SQUAD to return answer spans
Using Question, Query and Background and NER extracted from their combination

covidbert_1 BERT model trained on in-house annotated COVID FAQ data

covidbert_2 BERT model trained on in-house annotated COVID FAQ data

nlm_lhc_qa_1 This run involved a two-step process: (1) Using the search engine Essie to rank the CORD collection; and (2), once the ranked results were returned by Essie for each query, using the summarization algorithm BART to summarize contexts longer than 100 tokens. These summaries are fairly extractive and thus were then mapped to the original sentences in the context. Contiguous sentences for which there were summary mappings were used as the answers.

B PRIMARY SUBMISSIONS

B.1 Task A

Seven teams submitted runs for Task A in the primary evaluation cycle. The runs submitted by each team, as well as their provided descriptions are provided below.

h2oloo_1 MMR lambda = 0.375

h2oloo_2 MMR Lambda = 0.42

h2oloo_3 No MMR. Only T5.

HLTRI_1 Passage index -> BM25 -> Fine-Tuned BERT Reranker -> REcognizing GEnerated QUestion Entailment Search (REGEQUES)

HLTRI_2 Passage index -> BM25 -> Fine-Tuned BERT Reranker

HLTRI_3 CURRENT: Passage index -> BM25 -> Fine-Tuned BERT Reranker -> REcognizing GEnerated QUestion Entailment Search (REGEQUES) for NDNS

IBM_1 IR: Elastic Search + Neural IR classifier for re-ranking
MRC: Roberta-large model (GAAMA) trained on Natural Questions
Final submission: Combination / weighted average of the IR + MRC system

IBM_2 IR: Ensemble of elastic search with neural re-ranking and a dense passage retriever system
MRC: Roberta-large model (GAAMA) trained on Natural Questions
Final submission: Combination / weighted average of the IR + MRC system

IBM_3 IR: Ensemble of elastic search and a dense passage retriever system
MRC: Roberta-large model (GAAMA) trained on Natural Questions
Final submission: Combination / weighted average of the IR + MRC system

nlm_lhc_qa_1 The baseline expert run involved a two-step process: (1) Using the search engine Essie to return lossy rankings of the CORD collection for each expert query; and (2), once the ranked results were returned by Essie, we used the summarization algorithm BART to summarize the contexts.

We only summarized contexts longer than 100 tokens. This value was chosen from observation, in order to maximize the number of contexts used without summarization (so as not to risk BART scrambling otherwise good information) while still taking advantage of BART's ability to produce distilled text. The maximum length of the summaries produced by BART was set at 160 tokens, following the original paper. We had also observed that when BART is given text that is significantly shorter than the maximum summary length, it will often include the question in the output summary. If the question was included in the summary in any of the shorter contexts, we removed it in post-processing.

From previous analyses, we had observed the summaries generated by BART to be extractive in content, making it possible to map the summaries back to the original sentences in the context. The spans of contiguous sentences for which there were summary mappings were used as the answers in the TAC runs.

UPC_USMBA_1 A rule-based system

vigicovid_1 Document retrieval: We use a language modeling-based information retrieval approach (Ponte and Croft, 1998) including pseudo relevance feedback. For that purpose, we used the Indri search engine (Strohman, 2005), which combines Bayesian networks with language models. When building the query, different weights are assigned to the query, question, and narrative fields.

Answer extraction: given a question in natural language (text from ‘question’ field) and a document retrieved using the document retrieval module, we extract the answer to the question in the document. For the answer extraction we use the SciBERT language representation model that has been fine-tuned for QA first on the SQuAD2.0 dataset, and then on the QuAC dataset.

We retrieve 400 documents and we extract 10 answers from each document. All the extracted answers are ranked based on the multiplication of the scores given by the document retrieval module and answer extraction module.

vigicovid_2 Document retrieval: We tackle the document retrieval task in two steps: a) a first ranking and b) re-ranking. In order to obtain the first ranking of relevant documents of the collection corresponding to the queries, we use a language modeling-based information retrieval approach (Ponte and Croft, 1998) including pseudo relevance feedback. When building the query, different weights are assigned to the query, question, and narrative fields. Then, we make a re-ranking based on BERT following a strategy similar to the one proposed by Nogueira and Cho (2019). We tuned the Clinical BERT model (Alsentzer et al., 2019) on the task of identifying relevant queries and abstracts by using TREC-COVID’s qrels and our own pseudo qrels. Indri and Tuned Clinical BERT scores are linearly combined.

Answer extraction: given a question in natural language (text from ‘question’ field) and a document retrieved using the document retrieval module, extract the answer to the question in the document. For the answer extraction we use the SciBERT language representation model that has been fine-tuned for QA first on the SQuAD2.0 dataset, and then on the QuAC dataset.

We retrieve 400 documents and we extract 10 answers from each document. All the extracted answers are ranked based on the multiplication of the scores given by the document retrieval module and answer extraction module.

vigicovid_3 Document retrieval: We tackle the document retrieval task in two steps: a) a first ranking and b) re-ranking. In order to obtain the first ranking of relevant documents of the collection corresponding to the queries, we use a language modeling-based information retrieval approach (Ponte and Croft, 1998) including pseudo relevance feedback. When building the query, different weights are assigned to the query, question, and narrative fields. Then, we make a re-ranking based on BERT following a strategy similar to the one proposed by Nogueira and Cho (2019). We tuned the Clinical BERT model (Alsentzer et al., 2019) on the task of identifying relevant queries and abstracts by using TREC-COVID’s qrels and our own pseudo qrels. Indri and Tuned Clinical BERT scores are linearly combined.

Answer extraction: given a question in natural language (text from ’question’ field) and a document retrieved using the document retrieval module, extract the answer to the question in the document. For the answer extraction we use the SciBERT language representation model that has been fine-tuned for QA first on the SQuAD2.0 dataset, and then on the QuAC dataset.

We retrieve 400 documents and we extract 10 answers from each document. All the extracted answers are ranked based on the linear combination of the scores given by the document retrieval module and answer extraction module.

Yastil_R_1 Retrieval method:
Stage 1: Retrieve contexts and split into sentences
Stage 2: Rerank sentences
System details:

Stage 1: Merged output of BM25, DeepCT, ColBERT and docTTTTTquery
Stage 2: BERT Large fine-tuned on MS-MARCO

Yastil_R_2 Retrieval method:
Stage 1: Retrieve contexts and split into sentences
Stage 2: Rerank sentences
Stage 3: Fuse results
System details:
Stage 1: Merged output of BM25, DeepCT, ColBERT and docTTTTTquery
Stage 2: BERT Large fine-tuned on MS MARCO, BM25, DeepCT, ColBERT and docTTTTTquery
Stage 3: Reciprocal rank fusion

B.2 Task B

Five teams submitted runs for Task B in the primary evaluation cycle. The runs submitted by each team, as well as their provided descriptions are provided below.

h2oloo_1 MMR lambda = 0.75
h2oloo_2 MMR Lambda = 0.7
h2oloo_3 No MMR

HLTRI_1 Passage index -> BM25 -> Fine-Tuned BERT Reranker -> REcognizing GEnerated QUestion Entailment Search (REGEQUES)
HLTRI_2 Passage index -> BM25 -> Fine-Tuned BERT Reranker
HLTRI_3 CURRENT: Passage index -> BM25 -> Fine-Tuned BERT Reranker -> REcognizing GEnerated QUestion Entailment Search (REGEQUES) for NDNS

IBM_1 IR: Elastic Search + Neural IR classifier for re-ranking
MRC: Roberta-large model (GAAMA) trained on Natural Questions
Final submission: Combination / weighted average of the IR + MRC system
IBM_2 IR: Ensemble of elastic search with neural re-ranking and a dense passage retriever system
MRC: Roberta-large model (GAAMA) trained on Natural Questions
Final submission: Combination / weighted average of the IR + MRC system
IBM_3 IR: Ensemble of elastic search and a dense passage retriever system
MRC: Roberta-large model (GAAMA) trained on Natural Questions
Final submission: Combination / weighted average of the IR + MRC system

nlm_lhc_qa_1 The baseline consumer run involved a two-step process: (1) Using the search engine Essie to return lossy rankings of the consumer collection for each consumer query; and (2), once the ranked results were returned by Essie, we used the summarization algorithm BART to summarize the contexts.

As in Task A, we only summarized contexts longer than 100 tokens. The maximum length of the summaries produced by BART was set at 160 tokens, following the original paper. We mapped the sentences in generated summaries back to the original sentences in the context. The spans of contiguous sentences for which there were summary mappings were used as the answers in the TAC runs.

UPC_USMBA_1 A rule-based system

UPC_USMBA_3 bert and bart based system