COVID-QA: A Question & Answer Dataset for COVID-19

Timo Möller*
deepset GmbH
Berlin, Germany

G Anthony Reina*
Intel Corporation
Santa Clara, CA, USA

Raghavan Jayakumar
Lawrence Livermore
National Laboratory
(retired)
Livermore, CA, USA

Malte Pietsch
deepset GmbH
Berlin, Germany

Abstract

We present COVID-QA, a Question Answering dataset that is based on the COVID-19 Open Research Dataset. Our dataset consists of 2,019 question/answer pairs annotated by volunteer biomedical experts and saved in the popular SQuaD format. This presents, to date, the largest manually created Question Answering dataset on Covid-19 related material. We began with a RoBERTa base model that was initially fine-tuned on SQuAD and continued training the model on our COVID-QA dataset. We found that the additional training on this domain-specific data leads to significant gains in performance. Both the trained model and COVID-QA annotated datasets have been open-sourced and made available at: https://github.com/deepset-ai/COVID-QA. Our hope is that researchers use COVID-QA as a tool in their efforts to fight the pandemic.

1 Introduction

In 2019 a novel coronavirus (SARS-CoV-2) caused a worldwide pandemic (COVID-19) that has spread to over 8 million people and led to over 400,000 deaths to date (Pascarella et al., 2020; Chen and Li, 2020). On March 16, 2020 the White House Office of Science and Technology Policy released the COVID-19 Open Research Dataset called CORD-19 (The White House Office of Science and Technology Policy, 2020 (accessed May 9, 2020). CORD-19 is an open-sourced corpus of over 69,000 scholarly articles that focus on COVID-19, SARS-CoV-2, the Coronavirus group and other viral outbreaks. The CORD-19 project is a collaboration of the Allen Institute for AI, Chan Zuckerberg Initiative (CZI), Georgetown University's Center for Security and Emerging Technology (CSET), Microsoft, and the National Library of Medicine (NLM) at the National Institutes of Health.

* These authors contributed equally.

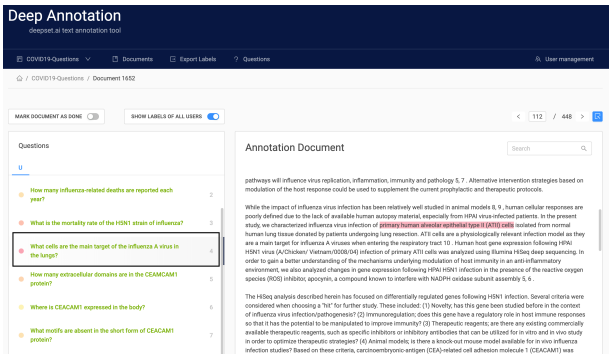

Figure 1: The COVID-19 QA Annotator is a lightweight, web-based interface which allows multiple experts to quickly apply SQuAD-style labels to research articles. The researcher uses the mouse to highlight the answer (pink) and creates a question based on that answer (box on left). The only system requirement is a web browser and an internet connection. Annotations can be exported either in Microsoft Excel or SQuAD (JSON) format.

COVID-QA is an open-source, volunteer-led project to develop a Question Answering (QA) dataset based on text from CORD-19 articles. The COVID-QA annotations are released under the Creative Commons Attribution 4.0 International License (CC-BY 4.0). The goal is to provide a free dataset for biomedical researchers to use in their efforts to combat the COVID-19 pandemic.

2 Related Work

2.1 Datasets

There are many large, extractive QA datasets, including prominent ones such as SQuAD (Rajpurkar et al., 2016) and Natural Questions (Kwiatkowski et al., 2019). Typically, these are crowd-sourced, general-purpose datasets. Each contains more than a hundred thousand question/answer pairs. This makes them suitable for training deep learning models from scratch.

In contrast, there are also many smaller, domain-specific QA datasets created by expert annotators. Although their small sizes limit their usefulness in the *de novo* training of QA models, their domain-specific annotations make them good candidates for evaluating existing QA systems. For example, the MRQA Shared Task (Fisch et al., 2019)— which focuses on the generalizability of QA systems—contains domain-specific datasets, such as BioASQ (Tsatsaronis et al., 2015) (∼3,000 QA pairs) and BioProcess (Berant et al., 2014) (219 QA pairs).

Since the COVID-19 pandemic began there have also been efforts in assembling relevant QA datasets, such as CovidQA (Tang et al., 2020) and the CORD-19 Information Aggregator; or dataset collections, such as Cord-19 PubAnnotation and Social Media for Public Health. Additionally, there is an ongoing IR shared task dataset TREC-COVID (Voorhees et al., 2020) that is, at the time of writing, under construction.

2.2 Tools

New tools are emerging to make search (or Question Answering) into COVID-19 related literature possible. Recently Google released the COVID-19 Research Explorer a semantic search interface on top of the CORD-19 dataset. Neural Covidex (Zhang et al., 2020) is a similar initiative.

We view our work as a complementary effort to aid in the evaluation and/or improvement to these existing tools.

3 COVID-QA Dataset Creation

The CORD-19 dataset was compiled so that researchers could apply "recent advances in natural language processing (NLP) and other AI techniques to generate new insights in support of the ongoing fight against this infectious disease." (The White House Office of Science and Technology Policy, 2020 (accessed May 9, 2020). NLP and AI techniques are especially important since the large volume of COVID-19 related research articles will continue to grow daily as more resources are rallied to combat the pandemic.

The website Kaggle sponsored the CORD-19 Challenge so that NLP researchers could develop text and data-mining techniques that could help provide insights into the pandemic based on the CORD-19 dataset. A series of 9 open-ended questions were used for the Kaggle challenge.

Although the CORD-19 dataset itself provides a valuable start for researchers, it remains a completely unstructured resource. The dataset is simply the text from published scientific articles that were compiled from a keyword search of the Chan-Zuckerberg Initiative (CZI), PubMed Central (PMC), BioRxiv, and medRxiv online databases.

We decided to build an automated tool that could answer any information-seeking question about COVID-19 using scientific literature. In order to achieve this, we needed to provide a sufficient number of annotated questions/answers to the CORD-19 dataset. Our inspiration for this project came, in part, from the previous works, such as the BioASQ datasets and the BioBert and SciBERT biomedical NLP models (Tsatsaronis et al., 2015; Lee et al., 2019; Beltagy et al., 2019). These similar projects demonstrated that fine-tuning an existing baseline NLP model on a highly-domain specific text corpus could significantly improve the accuracy of the predictions.

3.1 The Annotation Process

Within 10 days of the CORD-19 dataset release, we began annotation. First, we selected the subset of the CORD-19 dataset that had full-text and a commercially-friendly copyright license. On manual inspection of the subset, we determined that the majority of the scientific articles were not highly specific to COVID or coronavirus, but instead related more generally to viral outbreaks and genomics. This lack of COVID-19 specificity in the articles is understandable given the newness of the pandemic. Our domain experts agreed that scientific research on related viral pandemics, including SARS, MERS, and HIV-1, were relevant to COVID-19 researchers. Nevertheless, they also agreed to balance the dataset as much as possible between COVID-19 and generic viral pandemic-related articles. We, therefore, extracted 245 articles by searching the commercially-friendly subset for the phrases "COVID", "Covid", "covid", "Wuhan", "Coronavirus", and "coronavirus". We extracted an additional 200 random articles from the commercially-friendly subset. Finally, our annotators uploaded 5 additional scientific articles based on their knowledge of the domain. Of these 450 scientific articles, 147 were annotated over the 8 week period prior to this report. Although the expert annotators were volunteers, it was required that all have at least a Master's degree in biomedical sciences to ensure a basic understanding of the

4555	What is a natural reservoir of coronavirus?	Bats
4556	What is the genome size of the coronavirus?	26-32 kb
4557	What is the structure of the coronavirus?	enveloped, non-segmented, positive-strand RNA viruses
4558	What animals do gamma and delta coronavirus mainly infect?	birds
4559	How many types of coronaviruses are known to cause human disease?	Six
4560	Who performed the sampling procedures?	veterinarians
4561	When were the fecal samples collected?	from November 2004 to November 2014
4562	What reference genome was used in the study?	BatCoV HKU10
4563	What type of coronavirus was detected in R. affinis and R. sinicus species?	BtCoV/Rh/YN2012
4564	What is the length of the replicase gene ORF1ab?	20.4 kb
4565	What plays a role in regulating the immune response to a viral infection?	NF-κB
4566	What is the conclusion of the coronavirus long-term surveillance studies?	Rhinolophus bats seem to harbor a wide diversity of CoVs

Figure 2: Sample questions and answers from the annotated CORD-19 dataset. The COVID-QA annotation tool allows for this data to be reviewed/corrected by other annotators. The experts that volunteered as annotators were required to have at least a Master's degree in one of the biomedical sciences.

scientific articles. The annotation team was led by a medical doctor (G.A.R.) who recruited the volunteers, vetted their credentials, and validated annotations manually. Annotators were told that the COVID-QA dataset was not simply a database of COVID-19 facts and figures, but instead, a method to train a model that could answer questions given any scientific article about COVID-19 or similar viral outbreaks. Annotators focused their questions mainly on viral/genomic structure, antigenic targets, epidemiological characteristics, symptoms, and pharmaceutical treatment. We presume most of the questions are relevant for any scientific report of a viral outbreak.

We used an existing, web-based annotation tool that had been created by deepset for their Neural Search Framework Haystack. Figure 1 shows an image of the COVID-QA Annotation tool. In addition to being a light-weight, scalable interface, the tool was easy enough to use that even non-technical experts were quickly comfortable in its use. We created a series of online videos as well as released annotation guidelines to demonstrate the usage of the tool and the concepts behind the QA annotations. Importantly, the annotation tool allowed the model developers to export the labels in SQuAD format necessary for fine tuning the QA models. The annotators generated 2,019 question/answer pairs based on the annotations of 147 articles from the dataset. Figure 2 shows several examples of these question/answer pairs.

The annotators reported the process to be relatively easy to annotate and enjoyed the experience because it provided them an opportunity to contribute to the COVID response while reading interesting and informative research papers. While the annotators needed to be familiar with the biomedical field and recent broad progress, we found that intimate knowledge with the subject of the article was not essential. Instead, it is adequate for the annotator to understand the material well enough to figure out the broad relationship between subjects and objects that are being linked and the purported logic in it. Where the annotator is not certain of the relationship and importance of a statement, they would simply omit an annotation. A Slack workspace was created to allow annotators to share best practices, post issues with existing labels, and recommend enhancements to the annotation process.

Although most of the annotations were found to be straightforward– producing short, simple question/answer pairs– we found a large number of articles where it was difficult to properly phrase the question or frame the answer. Specifically, the annotators found that some question/answer pairs relied on context that spanned throughout the article. Figure 3 shows that although the majority of question/answer pairs contained short sentences, there was also a very long tail of question/answer pairs where the answers could span more than 30 words. The annotators reported this issue to the model developers in order to prepare the following experiments that would address these long question/answer pairs during training and inference.

The annotators also discussed new labelling schemes, such as single-question/multiple-answer annotations and semantic variations of the same question for a single answer. Although these were considered to be valid research arms for future work, we chose to keep the annotation instructions as simplified as possible for the current work.

4 Experiments

4.1 Models

We used pretrained, transformer-based Language Models that had been fine-tuned to QA datasets. It has been shown that models trained on the SQuAD dataset have the best generalization performance to out-of-domain data compared to other large QA datasets (Longpre et al., 2019).

We chose a RoBERTa-base architecture (Liu et al., 2019) for its good accuracy vs model size

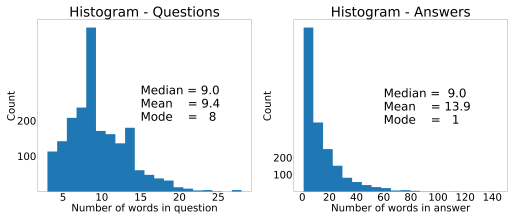

Figure 3: The COVID-19 QA dataset statistics. Although the majority of the 2,019 questions or answers contain fewer than 10 words, the answer distribution tend to have a much longer tail.

trade-off and fine-tuned it on the SQuAD dataset. The resulting model and its evaluation can be accessed at Hugging Face's model hub under "deepset/roberta-base-squad2". We refer to this model as the *baseline model*.

We further fine-tuned the baseline model using our COVID-QA annotations introduced in this paper. We used the FARM framework to fine-tune the model in k-fold cross validation manner. The training parameters are defined in Table 1. We disabled the option to return no answer during testing since in COVID-QA each question has a dedicated text answer. We will refer to this model as the *COVID-QA model*.

4.2 Metric

Most open domain Question Answering systems, such as DrQA (Chen et al., 2017) or REALM (Guu et al., 2020), use the Exact Match (EM) between label and prediction text (not the position) as evaluation criteria where the answer is often an unambiguous, single entity (*Q: "How many feet are in one mile?" Answer: "5,280"*). A less sensitive metric than EM is needed in the cases where there are long answers, long articles, and/or only a single answer annotation per question.

We choose to take multiple model predictions and any positional overlap between prediction and label as metric and refer to it as *top-n-accuracy*. Top-n-accuracy compares the gold label against n model predictions and looks for any overlap between prediction and answer positions to create a binary hit or miss scenario where accuracy can be computed. We also report EM and F1 score of our models to simplify comparison to existing methods. In the following paragraphs, we will outline the reasons for choosing top-n-accuracy on COVID-QA data.

As an example of variation in long answers, consider a question like *"Does C-C chemokine receptor type 5 (CCR5) affect the transmission of HIV-1?"* given the context *"Genetic variants in CCR5 have been shown to influence vertical transmission of HIV-1. CCR5 promoter variants resulting in higher expression of the receptor were associated with increased risk of MTCT of HIV-1 [...]."*. A valid answer could be either the first, the second or both sentences altogether.

In long scientific articles, the abstract, results, and conclusion may contain some of the same information but with slightly different wording. This is, in fact, part of the design that goes into writing a good scientific article. A metric such as EM might not be the best measure of the model performance for these cases. When only one annotation is present it, by nature, cannot capture all possible answer variations in the answers and positions in the article. The chances of multiple, correct answer spans are increased when working with long documents and, therefore, a robust metric should be able to account for these variations.

4.3 Evaluation

Evaluating Language Models trained on downstream tasks is difficult because the performance varies across runs (Dodge et al., 2020) - especially for difficult tasks and/or small dataset sizes. We therefore implemented k-fold cross validation for QA to get a better estimate of how well the performance will generalize.

Another complication for training QA models on long documents is the restricted sequence length of Transformer-based Language Models. Long documents are split into multiple (overlapping) chunks with mostly one chunk containing the answers, while the other chunks are negative examples where no answer can be found. We observed that having too many negative examples both substantially increases training time and also prevents the model from convergence. We therefore included a downsampling option to only keep d negative examples during training. For evaluation the whole article was processed and no downsampling was applied.

For evaluation, we used k-fold cross validation with one negative example per question/answer pair during training and the full article for testing.

4.4 Results

Table 2 shows that fine-tuning the baseline model on the domain data with the parameters specified in Table 1 results in an absolute improvement of

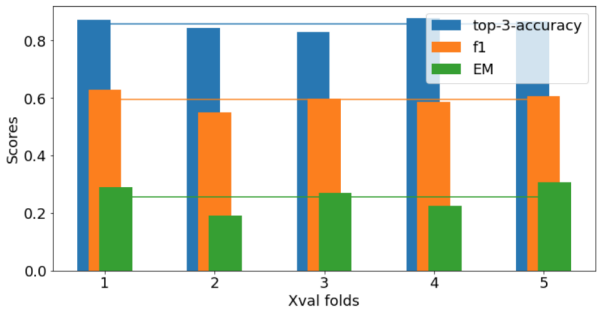

Figure 4: Per fold cross validation scores for the COVID-QA model (RoBERTa-base fine-tuned on SQuAD, continued with cross validation on COVID-QA annotations). The bars represent top-3 accuracy, Exact Match and F1 scores for each of the 5 folds. The horizontal lines depict average values for each metric.

Parameter	Value
baseline model	deepset/roberta-base-squad2
batch size	80
epochs	2
learning rate	3e-5
max seq len	384
doc stride	192
no answer	-100 (disabled)
cross val folds	5
hardware	4x V100

Table 1: Parameters used for training and evaluating the COVID-QA model. Note that disabling the option of "no answer" only applies to the test setting.

top-3 accuracy by about +5% (85.59 versus 80.68) and in the F1 score by slightly over +10% (59.53 versus 49.43). The scores reported for the COVID-QA model are averaged across all five folds with Figure 4 showing the scores per each fold.

5 Discussion & Future Work

In this section we want to highlight shortcomings and potential improvements to the used QA models as well as the generated dataset itself.

Model	EM	F1	Top-3 Acc.
Baseline	21.84	49.43	80.68
COVID-QA model	25.90	59.53	85.59

Table 2: Comparison of baseline model to the model cross validated on COVID-QA labels. The scores for the COVID-QA model are averaged across the 5 folds.

5.1 Interpretation of Results

As shown in Table 2 all scores improved significantly when using COVID-QA labels for continued training. F1 and EM scores are low compared to metrics reported on SQuAD because of the different text domain as well as an about 40x larger text size. Our baseline model created an average of 6118.5 tokens for each COVID-QA document compared to 153.2 tokens for SQuAD. A relatively larger increase in F1 score (compared to EM) might be the consequence of the model adjusting to much longer answers that are present in our dataset. As shown in Fig. 3 the mean number of words in COVID-QA answers is 13.9 whereas for SQuAD (training set, excluded "no answer" option) it is 3.2 words per answer only.

5.2 Modelling improvements

We believe substantial improvements to the QA model can be achieved by using other baseline models like the large version of RoBERTA fine-tuned on SQuAD. Combined fine-tuning of the baseline model on multiple QA datasets might be beneficial for the transfer to the CORD-19 domain as well. Adjusting the underlying Language Model to the CORD-19 domain by training it from scratch or Language Model adaptation is another option, though one has to consider these domain Language Models cannot be fitted to SQuAD as well as Language Models pretrained on the general domain.

Improvements to the metric are also imaginable. One possible direction is a semantic comparison of prediction to gold label by using vector representations from Language Models as described in BLEURT (Sellam et al., 2020), a newly proposed metric for evaluating text generation systems.

One issue of using transformer-based models for QA is the need for vast amounts of computational power - especially when asking a question on not just a short paragraph but a whole collection of documents. To speed up the QA process one usually uses a two-stage procedure to first retrieve a set of candidate documents and then apply more powerful QA models to extract the actual answer like done with DrQA (Chen et al., 2017) or inside Haystack, an open-source framework for doing QA at scale. Another alternative to speed up inference is to index all possible phrases (including potential answers) and project the question into this index to retrieve the closest answers as done in the Dense-Sparse Phrase Index (Seo et al., 2019).

5.3 Dataset improvements

We are confident about the quality of the generated dataset for two reasons. Manual checking of a subset of question/answer pairs showed no direct errors and the scores across each fold (see Fig. 4) are rather alike in spite of a tiny fold size (about 400 QA pairs each). Though we also see much potential for improvement. First of all the dataset does not contain multiple labels by different annotators for any of our questions. This would be needed to determine the inter-annotator agreement and set a human benchmark. The dataset furthermore does not contain unanswerable questions, i.e. questions where the annotator was sure the given article does not contain an answer. Another shortcoming is the constructed tone of the questions as a result of the SQuAD style label process where annotators need to create a question while reading the article. This is a major criticism of SQuAD in the paper accompanying Natural Questions (Kwiatkowski et al., 2019). A fast way to create less constructed questions would be to have other annotators rephrase existing questions without access to the underlying article. Another minor improvement could be made to the answer annotation spans. These spans sometimes start or end inside of a word that does not belong to the answer.

6 Conclusion

We have created a SQuAD-style QA dataset based on 2,019 annotations of medical experts and released it to the public. We have shown that the created dataset can be used for improving QA systems by either evaluating existing QA models or continued training. We also believe the created dataset could be used for improving or evaluating existing search systems that browse through the CORD-19 dataset.

Based on improved QA systems we hope researchers can find information inside COVID-19 related literature much quicker and therefore utilize the existing knowledge more efficiently for finding solutions to the COVID-19 pandemic.

7 Source Code

The annotations are available at https://github.com/deepset-ai/COVID-QA/tree/master/data/question-answering/COVID-QA.json. Source code for performing cross validation on QA models is available at https://github.com/deepset-ai/FARM/blob/master/examples/question_answering_crossvalidation.py The annotation tool is available at https://github.com/deepset-ai/haystack

8 Acknowledgements

This effort has been successful thanks to the hard work of many people, including, but not limited to the following (in alphabetical order of last name): Victor Alm, Laura Barrera, Branden Chan, Suha Chari, Simon Fakir, Ling Hsin, Archana Jayakumar, Tripti Kataria, Bogdan Kostic, Pete Logan, Bilawal Nadeem, Travis Nesbit, Milind Pandit, Jeanine Renne, Milos Rusic, Jack Seksenyan, Kyle Shannon, Ajeet Singh, Tanay Soni, and Narayan Sundararajan. We are also grateful for the support of deepset, Intel, NVIDIA and AWS.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text.

Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, and Christopher D Manning. 2014. Modeling biological processes for reading comprehension. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1499–1510.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*.

Yu Chen and Lanjuan Li. 2020. Sars-cov-2: virus dynamics and host response. *The Lancet*, pages 515–516.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. Mrqa 2019 shared task: Evaluating generalization in reading comprehension. *arXiv preprint arXiv:1910.09753*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova,

Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Shayne Longpre, Yi Lu, Zhucheng Tu, and Chris DuBois. 2019. An exploration of data augmentation and sampling techniques for domain-agnostic question answering. *arXiv preprint arXiv:1912.02145*.

Giuseppe Pascarella, Alessandro Strumia, Chiara Piliego, Federica Bruno, Romualdo Del Buono, Fabio Costa, Simone Scarlata, and Felice Agrò. 2020. Covid-19 diagnosis and management: A comprehensive review. *Journal of Internal Medicine*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Minjoon Seo, Jinhyuk Lee, Tom Kwiatkowski, Ankur P Parikh, Ali Farhadi, and Hannaneh Hajishirzi. 2019. Real-time open-domain question answering with dense-sparse phrase index. *arXiv preprint arXiv:1906.05807*.

Raphael Tang, Rodrigo Nogueira, Edwin M. Zhang, Nikhil Gupta, Phng Ths. Bùi Cm, Kyunghyun Cho, and Jimmy Lin. 2020. Rapidly bootstrapping a question answering dataset for covid-19. *ArXiv*, abs/2004.11339.

The White House Office of Science and Technology Policy. 2020 (accessed May 9, 2020). *Call to Action to the Tech Community on New Machine Readable COVID-19 Dataset*.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):138.

Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2020. Trec-covid: Constructing a pandemic information retrieval test collection. *arXiv preprint arXiv:2005.04474*.

Edwin M. Zhang, Nikhil Gupta, Rodrigo Nogueira, Kyunghyun Cho, and Jimmy Lin. 2020. Rapidly deploying a neural search engine for the covid-19 open research dataset: Preliminary thoughts and lessons learned. *ArXiv*, abs/2004.05125.