

CORD-19: The COVID-19 Open Research Dataset

Lucy Lu Wang^{1,*} Kyle Lo^{1,*} Yoganand Chandrasekhar¹ Russell Reas¹

Jiangjiang Yang¹ Darrin Eide² Kathryn Funk³ Rodney Kinney¹

Ziyang Liu⁵ William Merrill¹ Paul Mooney⁴ Dewey Murdick⁶

Devvret Rishi⁴ Jerry Sheehan³ Zhihong Shen² Brandon Stilson¹

Alex D. Wade⁵ Kuansan Wang² Chris Wilhelm¹ Boya Xie²

Douglas Raymond¹ Daniel S. Weld^{1,7} Oren Etzioni¹ Sebastian Kohlmeier¹

¹Allen Institute for AI ²Microsoft Research ³National Library of Medicine

⁴Kaggle ⁵Chan Zuckerberg Initiative ⁶Georgetown University ⁷University of Washington

{lucyw, kylel}@allenai.org

Abstract

The COVID-19 Open Research Dataset (CORD-19) is a growing¹ resource of scientific papers on COVID-19 and related historical coronavirus research. CORD-19 is designed to facilitate the development of text mining and information retrieval systems over its rich collection of metadata and structured full text papers. Since its release, CORD-19 has been downloaded over 75K times and has served as the basis of many COVID-19 text mining and discovery systems. In this article, we describe the mechanics of dataset construction, highlighting challenges and key design decisions, provide an overview of how CORD-19 has been used, and preview tools and upcoming shared tasks built around the dataset. We hope this resource will continue to bring together the computing community, biomedical experts, and policy makers in the search for effective treatments and management policies for COVID-19.

1 Introduction

On March 16, 2020, the Allen Institute for AI (AI2), in collaboration with our partners at The White House Office of Science and Technology Policy (OSTP), the National Library of Medicine (NLM), the Chan Zuckerberg Initiative (CZI), Microsoft Research, and Kaggle, coordinated by Georgetown University’s Center for Security and Emerging Technology (CSET), released the first version of CORD-19. This resource is a large and growing collection of publications and preprints on COVID-19 and related historical coronaviruses such as SARS and MERS. The initial release consisted of approximately 28K papers, and the collection

*denotes equal contribution

¹The dataset continues to be updated weekly with papers from new sources and the latest publications. Statistics reported in this article are up-to-date as of version 2020-04-17.

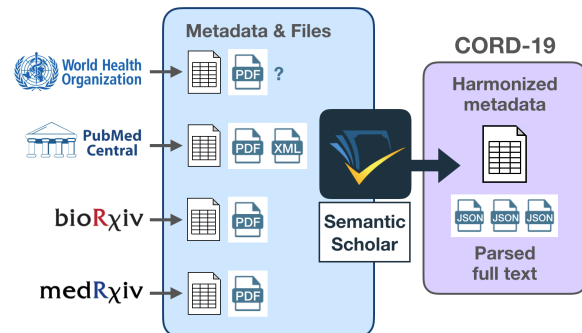


Figure 1: Sources of papers included in CORD-19. Papers and preprints are collected by Semantic Scholar. Metadata are harmonized and deduplicated, and document files are processed to extract full text.

has grown to more than 50K papers over the subsequent weeks. Papers and preprints from a variety of archives are collected, and paper documents are processed through the pipeline established in Lo et al. (2020) to extract full text (around 80% of papers have full text). Metadata are harmonized by the Semantic Scholar² team at AI2. We commit to providing regular updates to the dataset until an end to the crisis is foreseeable.

CORD-19 aims to connect the machine learning community with biomedical domain experts and policy makers in the race to identify effective treatments and management policies for COVID-19. The goal is to harness these diverse and complementary pools of expertise to discover relevant information more quickly from the literature. Users of the dataset have leveraged a variety of AI-based techniques in information retrieval and natural language processing to extract useful information.

Responses to CORD-19 have been overwhelmingly positive, with the dataset being viewed more than 1.5 million times and downloaded over 75K times in the first month of its release. Numerous

²<https://semanticscholar.org>

groups have constructed search and extraction tools using the dataset (some are summarized in Section 4), several successful shared tasks have been planned or are being executed (Section 5), and a thriving community of users has sprung up to discuss and share information, annotations, projects, and feedback around the dataset.³

In this article, we briefly describe:

1. The content and creation of CORD-19,
2. Design decisions and challenges around creating the dataset,
3. Fruitful avenues of research conducted on the dataset,
4. Learnings from shared tasks, and
5. A roadmap for CORD-19 going forward.

2 Dataset

CORD-19 integrates papers from several sources (Figure 1). Sources make openly accessible paper metadata, and in most cases, documents associated with each paper. A paper is defined as the base unit of published knowledge, and is associated with a set of bibliographic metadata fields, like title, authors, publication venue, publication date, etc. Each paper can have unique identifiers such as DOIs, PubMed Central IDs, PubMed IDs, the WHO Covidence #,⁴ or MAG identifiers (Shen et al., 2018). A paper can be associated with documents, the physical artifacts representing paper content; these are the familiar PDFs, XMLs, or physical print-outs we read.

The CORD-19 effort combines paper metadata and documents from different sources, and generates harmonized and deduplicated metadata as well as structured full text parses of paper documents as output. We provide full text parses of all papers for which we have access to a paper document, and for which the documents are available under open access copyright licenses (e.g. Creative Commons (CC),⁵ publisher-specific COVID-19 licenses, or identified as open access through DOI lookup in the Unpaywall⁶ database).

³<https://discourse.cord-19.semanticscholar.org/>

⁴<https://www.who.int/emergencies/diseases/novel-coronavirus-2019/global-research-on-novel-coronavirus-2019-ncov>

⁵<https://creativecommons.org/>

⁶<https://unpaywall.org/>

2.1 Sources of papers

The majority of papers in CORD-19 are sourced from PubMed Central (PMC).⁷ The PMC Public Health Emergency Covid-19 Initiative⁸ expanded access to COVID-19 literature by working with publishers to make coronavirus-related papers discoverable and accessible through PMC under open access license terms that allow for reuse and secondary analysis. Other portions of the dataset are derived from the bioRxiv and medRxiv preprint servers (provided by CZI), as well as the World Health Organization (WHO) Covid-19 Database, a collection of hand-curated papers about COVID-19.

We also worked directly with a number of publishers, such as Elsevier⁹ and Springer Nature,¹⁰ to provide full text coverage of relevant papers available in their back catalog; these papers are made available under special COVID-19 open access licenses.

Papers from PMC, bioRxiv, and medRxiv are retrieved given the query:

```
"COVID-19" OR "Coronavirus" OR  
"Corona virus" OR "2019-nCoV"  
OR "SARS-CoV" OR "MERS-CoV"  
OR "Severe Acute Respiratory  
Syndrome" OR "Middle East  
Respiratory Syndrome"
```

Papers that match on these keywords in their title, abstract, or full text are included in the dataset.

2.2 Processing metadata

The resulting collection of sourced papers suffers from duplication and incomplete or conflicting metadata. We perform the following operations to harmonize and deduplicate all metadata entries:

1. Cluster duplicate papers using identifiers
2. Select canonical metadata for each cluster
3. Filter clusters to remove non-papers

We start with approximately 73K metadata entries. After processing, the metadata consists of papers from PMC (28.6K), medRxiv (1.1K), and bioRxiv (0.8K), with another 1.1K from the WHO paper list and 19.5K contributed directly by publishers.

⁷<https://www.ncbi.nlm.nih.gov/pmc/>

⁸<https://www.ncbi.nlm.nih.gov/pmc/about/covid-19/>

⁹<https://www.elsevier.com/connect/coronavirus-information-center>

¹⁰<https://www.springernature.com/gp/researchers/campaigns/coronavirus>

Clustering papers We find duplicate papers using the identifier tuple: $(doi, pmc_id, pubmed_id)$. If we find two papers (often retrieved from two different sources) with one of these identifiers in common, we consider them to belong to the same cluster. We assign each cluster a unique identifier **CORD_UID**, which persists between each release of the dataset. All identifiers associated with the members of a cluster are inherited by the cluster. No existing identifier, such as DOI or PMC ID, is sufficient as the primary CORD-19 identifier. Some papers in PMC do not have DOIs; some papers from the WHO, publishers, or preprint servers like arXiv do not have PMC identifiers or DOIs.

This clustering policy easily accommodates newly added papers. For example, if a new paper shares identifiers with an existing cluster, it is merged into that cluster. If there is no existing cluster, the new paper forms its own cluster.

Occasionally, conflicts may occur. For example, a paper c with identifiers $(x, null, z')$ might share identifier x with a cluster of papers $\{a, b\}$ with identifiers (x, y, z) , but has a conflict $z' \neq z$. In this case, we choose to create a new cluster $\{c\}$, containing only paper c .¹¹

Selecting canonical metadata Among each cluster, the canonical entry is selected as one which is associated with document files, and which has the most permissive license. For example, if a paper is available under a CC license as well as a more restrictive COVID-19 specific copyright license, we select the metadata entry with the CC license as canonical. For any missing metadata fields, values from other metadata entries within the same cluster are promoted to fill in the blanks.

Cluster filtering Some metadata entries harvested from sources are not papers, and instead correspond to materials like tables of contents, indices, or informational documents. These entries are identified and removed from the final dataset.

2.3 Processing full text

Most sourced papers are associated with one or more PDFs.¹² To extract full text and bibliographies from each PDF, we use the PDF parsing

¹¹This is a conservative clustering policy in which any metadata conflict prohibits clustering. An alternative policy would be to allow clustering as long as any one identifier matches, under which a, b , and c would form one cluster with identifiers $(x, y, [z, z'])$.

¹²PMC papers can have multiple associated PDFs per paper, separating the main text from supplementary materials.

pipeline¹³ created for the S2ORC dataset (Lo et al., 2020):

1. Parse all PDFs to TEI XML files using GRO-BID¹⁵ (Lopez, 2009)
2. Parse all TEI XML files to target JSON format
3. Postprocess to clean up links between inline citations and bibliography entries.

We additionally parse JATS XML¹⁶ files available for PMC papers using a custom parser, generating the same target output JSON format.

This creates two sets of full text JSON parses associated with the papers in the collection, one set **originating from PDFs** (available from more sources), and one set originating from **JATS XML** (available only for PMC papers). Each PDF parse has an associated SHA, the 40-digit SHA-1 of the associated PDF file, while each XML parse is named using its associated PMC ID. Almost 80% of CORD-19 papers have an associated PDF parse, and around 38% have an XML parse, with the latter nearly a subset of the former. Most PDFs (>90%) are successfully parsed. A bit less than 4% of CORD-19 papers are associated with multiple PDF SHA, due to a combination of paper clustering and the existence of supplementary PDF files.

2.4 Dataset contents

CORD-19 has grown rapidly, now consisting of over 52K papers with over 41K full texts. The increase can be attributed to major publishers offering favorable terms on text/data mining uses that make the inclusion of their publications possible.

CORD-19 includes papers published in more than 3200 journals. Classification of CORD-19 papers to Microsoft Academic Graph (MAG) (Wang

¹³One difference in full text parsing for CORD-19 is that we do not use ScienceParse¹⁴ to obtain more accurate titles and authors, as we derive this metadata from the sources directly.

¹⁴<https://github.com/allenai/science-parse>

¹⁵<https://github.com/kermitt2/grobid>

¹⁶<https://jats.nlm.nih.gov/>

Subfield	Count	% of corpus
Virology	20116	42.3%
Immunology	9875	20.7%
Molecular biology	6040	12.7%
Genetics	3783	8.0%
Intensive care medicine	3204	6.7%
Other	4595	9.6%

Table 1: Representation of MAG subfields of study in CORD-19.

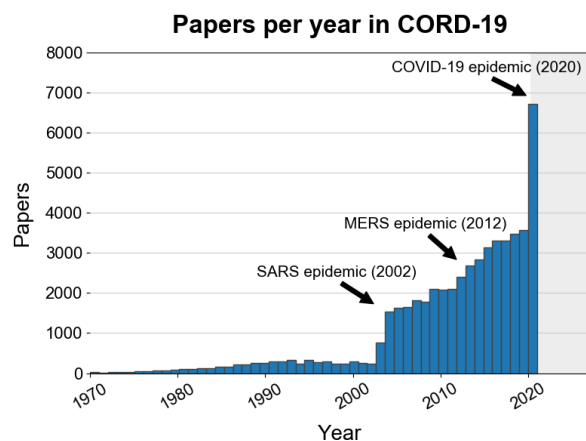


Figure 2: The distribution of papers per year in CORD-19. A spike in publications occurs in 2020 in response to COVID-19.

et al., 2019, 2020) fields of study indicate that the dataset consists predominantly of papers in Biology, Medicine, and Chemistry, which together constitute over 92% of the corpus. A breakdown of MAG subfields represented in CORD-19 is given in Table 1.

Figure 2 shows the distribution of CORD-19 papers by date of publication. Coronavirus publications increased during and following the SARS and MERS epidemics, but the number of papers published in the early months of 2020 exploded in response to the COVID-19 epidemic. Using author affiliations in MAG, we identify the countries from which the research in CORD-19 is conducted. Large proportions of CORD-19 papers are associated with institutions based in the United States (over 16K papers) and United Kingdom (over 3K papers), and EU and East Asian countries are also well represented. Chinese institutions have seen a meteoric rise this year (over 5K papers) due to China’s status as the first epicenter of the COVID-19 outbreak.

3 Design decision & challenges

A number of unique challenges come into play in the creation of a dataset like CORD-19. We summarize the primary design requirements of the dataset, along with challenges implicit within each requirement:

Up-to-date The dataset must be updated flexibly and quickly, with new literature included shortly after publication. The speed with which scientists have coalesced around the COVID-19 epidemic is unprecedented. Hundreds of new publications on

COVID-19 are published every week, and a dataset like CORD-19 can quickly become irrelevant without regularly scheduled updates. CORD-19 is currently updated at a weekly cadence, and is moving towards daily updates. A regularly updated dataset, unlike a static one, requires a processing pipeline that maintains somewhat consistent results from week to week. That is, the metadata and full text parsing results must be reproducible; any changes or new features must ideally be compatible with previous versions of the dataset. Persistent identifiers must also be introduced to link dataset entries across releases.

Able to handle data from multiple sources Papers from different organizations (WHO, PMC, bioRxiv, medRxiv, and others) need to be integrated and harmonized to a shared format. Each source has its own metadata format, which must be converted to the CORD-19 format, while addressing any missing or extraneous fields. We have relied heavily on PMC for data harmonization in recent weeks, as many publishers have agreed to release previously non-open access COVID-19-related literature for public use through the PMC platform. PMC does not incorporate all sources of data, as the scope of the database is generally limited to peer-reviewed journals in biomedicine and the life sciences. PMC currently excludes preprints, which can be some of the earliest sources of novel findings. Because of the need to include these other sources, we cannot rely on PMC alone, and must address problems of data integration across different provenances.

Clean canonical metadata Because of the diversity of paper sources, duplication is unavoidable. Once paper metadata from each source is cleaned and organized into CORD-19 format, we apply simple deduplication logic to combine similar paper entries from different sources. We apply the most conservative clustering, combining papers only when they have shared identifiers but no conflicts between any particular class of identifiers. We justify this because it is less harmful to retain a few duplicate papers than to remove a document that is potentially unique and useful.

Machine readable full text To provide accessible and canonical structured full text, we parse content from PDFs and associated paper documents. The full text is presented in a JSON schema designed to preserve most relevant paper structures

such as paragraph breaks, section headers, and in-line references and citations. The JSON schema is simple to use for many NLP tasks, where character-level indices are often employed for annotation of relevant entities or spans. We recognize that converting between PDF or XML to JSON is lossy. However, the benefits of a standard structured format, and the ability to reuse and share annotations made on top of that format have been critical to the success of CORD-19.

Observes copyright restrictions Papers in CORD-19 and academic papers more broadly are made available under a variety of copyright licenses. These licenses can restrict or limit the abilities of organizations such as AI2 from redistributing their content freely. Although much of the COVID-19 literature has been made open access by publishers, the provisions on these open access licenses differ greatly across papers. Additionally, many open access licenses grant the ability to read, or “consume” the paper, but may be restrictive in other ways, for example, by not allowing republication of an paper or its redistribution for commercial purposes. The curator of a dataset like CORD-19 must pass on best-to-our-knowledge licensing information to the end user.

4 Research directions

A few challenges for clinical experts during the current epidemic have been (i) keeping up to date with recent papers about COVID-19, (ii) identifying useful papers from historical coronavirus literature, (iii) extracting useful information from the literature, and (iv) synthesizing/organizing knowledge obtained from the literature. To facilitate solutions to these challenges, dozens of tools, systems, and resources over CORD-19 have already been developed. In Figure 3, we show an example of how we envision CORD-19 can be used by these systems for text-based information retrieval and extraction.

We provide a brief survey of ongoing research efforts. We note that this is not a comprehensive survey and is limited to projects in contact with us or our collaborators, that were discussed on our online public forum,¹⁷ or which were discovered via Twitter mentions of CORD-19.¹⁸

¹⁷<https://discourse.cord-19.semanticscholar.org/>

¹⁸Many systems have elements of both retrieval and extraction, making categorization difficult. We arbitrarily refer to projects identifying relevant papers given any query as *retrieval* and projects identifying text snippets to fulfill a pre-

Given a query:

Does hypertension increase the risks associated with Covid-19?

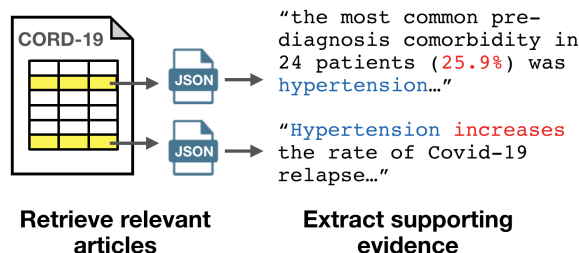


Figure 3: An example information retrieval and query answering task that we envision CORD-19 to support.

4.1 Information retrieval

A vast majority of retrieval tools currently use standard ranking methods, such as BM25 (Jones et al., 2000). One of the earliest live systems (within a couple weeks of the initial CORD-19 release) is the NEURAL COVIDEX (Zhang et al., 2020)¹⁹ which uses a T5-base model (Raffel et al., 2019) finetuned on biomedical text to perform unsupervised reranking on documents retrieved via BM25.

Pretrained text embeddings on CORD-19 have also been used to facilitate search. COVID-SCHOLAR²⁰ is a COVID-19-specific adaptation of the MATSCHOLAR (Weston et al., 2019) system for identifying relevant papers given entity-centric queries. KDCOVID²¹ is a system that uses BioSentVec (Chen et al., 2019) similarity to identify sentences relevant to a query.

4.2 Information extraction

There has been significant effort to extract entities from papers in CORD-19. A majority of these have used ScispaCy (Neumann et al., 2019), a natural language processing toolkit optimized for scientific text, to identify biomedical concepts, such as mentions of genes, diseases, and chemicals. Wang et al. (2020) releases a supplementary dataset over CORD-19 by augmenting text with entity mentions predicted from multiple techniques, including ScispaCy and weak supervision based on the NLM’s Unified Medical Language System (UMLS) Metathesaurus (Bodenreider, 2004). Pretrained language models such as BioBERT-base (Lee et al., 2019) or SciBERT-base (Beltagy et al., 2019) fine-

defined schema or enumerable set of criteria as *extraction*.

¹⁹<https://covidex.ai/>

²⁰<https://www.covidscholar.com/>

²¹<http://kdcovid.nl/about.html>

tuned on biomedical NER datasets have also been used to extract entity mentions.

We have observed a wide range of uses for extracted entities. Some search tools incorporate entities as facets that can be used to filter search results.^{22,23} Some projects have even induced concept hierarchies.²⁴ This is in contrast to systems like NEURAL COVIDEX, where entities in retrieved paper previews are highlighted to provide visual cues, but are not actually used in retrieval.

Yet other efforts focus on extracting sentences or passages of interest. Liang and Xie (2020), for example, uses BERT (Devlin et al., 2019) to extract sentences from CORD-19 that contain COVID-19-related radiological findings.

4.3 Knowledge graphs

Substantially less attention has been placed on extracting relations between entities, which is an important sub-task in automated knowledge graph construction. The Covid Graph project,²⁵ led by a diverse team based in Germany, has created a COVID-19 knowledge graph by mining a number of public data sources, including CORD-19, and is perhaps the largest current initiative in this space.

Entity co-occurrence within a corpus can be a weak indicator of a relationship between the co-occurring entities. Ahamed and Samad (2020) rely on entity co-occurrences in CORD-19 to construct a graph that enables centrality-based ranking of entities (i.e. drugs, pathogens, and biomolecules). The COVIZ²⁶ visualization tool also uses co-occurrence counts of extracted entities (i.e. genes, proteins, diseases, chemicals) as a way of identifying papers that describe a potentially meaningful relation between those entities. SeVeN (Espinosa-Anke and Schockaert, 2018) embeddings pretrained on CORD-19, which can produce relation embeddings given word pairs, have also been made available.²⁷

4.4 Other tasks

Not much work exists outside of the above-mentioned categories, but we note a few interesting

directions.

Question answering COVIDASK²⁸ retrieves answer snippets using the open domain question answering system from Seo et al. (2019). Similarly, the AUEB system²⁹ employs question answering by training the model presented in McDonald et al. (2018). These two systems went live within a couple weeks of the initial CORD-19 release by repurposing training data from the BioASQ challenge (Task B) (Tsatsaronis et al., 2015).

Pretrained language models DeepSet has made available³⁰ a BERT-base model pretrained on CORD-19. While this model is certainly more “in-domain” than BioBERT-base or SciBERT-base, the pretraining corpus is also much smaller. It would be interesting to see how this translates to performance differences in downstream systems.

Summarization A search tool³¹ by Vespa generates summaries from paper abstracts using T5.

Recommendation The Vespa tool also provides a similar-paper recommender³² using SentenceBERT (Reimers and Gurevych, 2019) and SPECTER (Cohan et al., 2020) embeddings.

Entailment One project³³ uses sentence embeddings for retrieval in a similar manner to KD-COVID, but their embeddings are generated from BERT models trained on NLI datasets.

Assistive literature review ASReview, an active learning system designed to assist researchers in finding relevant papers for literature reviews, have made a CORD-19 plugin available.³⁴

Augmented reading Sinequa has released a search tool³⁵ over CORD-19 that also allows for in-browser reading of the papers and perform entity highlighting directly on the displayed PDFs.

²⁸<https://covidask.korea.ac.kr/>

²⁹<http://cslab241.cs.aueb.gr:5000/>

³⁰https://huggingface.co/deepset/covid_bert_base

³¹<https://cord19.vespa.ai/>

³²<https://cord19.vespa.ai/article/32753?tab=1>

³³<https://github.com/gsarti/covid-papers-browser>

³⁴<https://github.com/asreview/asreview-covid19>

³⁵<https://covidsearch.sinequa.com/app/covid-search/>

²²<https://covid19search.azurewebsites.net/>

²³<http://covidexplorer.in/>

²⁴<http://topicforest.com/biomed/coronavirus>

²⁵<https://covidgraph.org/>

²⁶<https://coviz.apps.allenai.org/>

²⁷<https://github.com/luisespinoaanke/cord-19-seven>

5 Competitions and Shared Tasks

Adoption of CORD-19 as a data mining resource has been accelerated by the organization of COVID-19-related competitions and shared tasks. We discuss two – a text mining competition hosted by Kaggle and an information retrieval shared task hosted by TREC – and highlight successes and challenges in their organization. Both tasks involve biomedical domain experts to judge the results of automated extraction and retrieval systems.

5.1 Kaggle

Kaggle is hosting the CORD-19 Research Challenge³⁶ in coordination with The White House OSTP and AI2. This is an open-ended text mining competition where participants are tasked with extracting answers to key scientific questions about COVID-19 from the papers in CORD-19. Such questions include *What is known about transmission, incubation, and environmental stability?* and *What do we know about COVID-19 risk factors?*

Answers can take many forms. Some submissions use information extraction techniques to surface relevant **text snippets** from CORD-19 papers. Other submissions integrate aspects of text generation and summarization in an effort to improve the interpretability of results.

Unlike most Kaggle competitions, there is not a quantitative evaluation metric defined for assessing submissions. Instead, submissions need to be judged by biomedical domain experts on relevance and usefulness. The task has over 550³⁷ participating teams, and the reliance on humans for manual evaluation is costly and difficult to scale. Likewise, the challenge questions posed are quite broad, and multiple valid answers can exist. For example, consider these snippets from a randomly selected paper³⁸ on the topic of *incubation* period:

The full range of incubation periods of the Covid-19 cases ranged from 0 to 33 days among 2015 cases.

The median incubation period of both male and female adults was similar (7-day) but significantly shorter than that (9-day) of child cases ($P=0.02$).

Interestingly, incubation periods of 233 cases (11.6%) were longer than the WHO-established quarantine period (14 days).

³⁶<https://www.kaggle.com/allen-institute-for-ai/cord-19-research-challenge>

³⁷As of 2020-04-17

³⁸<https://www.medrxiv.org/content/10.1101/2020.03.15.20036533v1>

The challenge has also resulted in impactful spin-off projects such as the curation of living literature review tables.³⁹ This effort combines the contributions of several Kaggle participants with medical experts to review and synthesize answers to specific questions. A unique tabular schema is defined for each question, and answers are collected from across different automated extractions. For example, extractions for **risk factors** should include disease severity and fatality metrics, while extractions for *incubation* should include time ranges. Sufficient knowledge of COVID-19 is necessary to define these schema and to understand which fields are important to include (and exclude).

The community is still trying to understand how best to extract, aggregate, and present useful content identified in these papers. Our knowledge of the current epidemic is continuously evolving and these text extractions may very well be useful in organizing and understanding related research.

5.2 TREC

The TREC-COVID⁴⁰ shared task is being co-organized by the Allen Institute for AI, the National Institute of Standards and Technology (NIST), the National Library of Medicine (NLM), Oregon Health and Science University (OHSU), and the University of Texas Health Science Center at Houston (UTHealth). The task aims to assess systems on their ability to rank papers in CORD-19 based on their relevance to ad hoc topical queries.

The first round of the task opened with 30 topics, including queries like *How does the coronavirus respond to changes in the weather?* or *What drugs have been active against SARS-CoV or SARS-CoV-2 in animal studies?* These topics are sourced from MedlinePlus searches, Twitter conversations, library searches at OHSU, as well as from direct conversations with researchers, reflecting actual queries made by the community. Each round is open for one week, during which participants are asked to submit document sets for relevance judgment.

Following submission, medical domain experts are tasked with providing rankings and relevance judgments over submitted paper pools. Around 60 reviewers, including indexers from NLM and medical students from OHSU and UTHealth, are

³⁹<https://www.kaggle.com/covid-19-contributions>

⁴⁰<https://ir.nist.gov/covidSubmit/index.html>

involved in the effort. In this fashion, experts provide gold rankings for documents, allowing comparisons between retrieval systems. The goal of TREC-COVID is to provide conclusive judgments on the performance of these systems, in order to focus the attention and efforts of the community towards the most effective search and retrieval techniques.

6 Discussion

Several hundred new papers on COVID-19 are now being published every day. Automated methods are needed to analyze and synthesize information over this large quantity of content. The computing community has risen to the occasion, but it is clear that there is a critical need for better infrastructure to incorporate human judgments in the loop. Extractions need expert vetting, and search engines and systems must be designed to serve users.

Though CORD-19 has only been available a short time, successful engagement and usage speaks to our ability to bridge computing and biomedical communities over a common, global cause. The Kaggle challenge, especially, has not only helped to broadcast the challenge to the widest community, but early results have also provided some clarity into which formats for collaboration are successful and which questions are the most urgent to answer. There is however, significant work that remains for determining *(i)* which methods are best to assist textual discovery over the literature, *(ii)* how best to involve expert curators in the pipeline, and *(iii)* which extracted results convert to successful COVID-19 treatments and management policies. Shared tasks and challenges, as well as continual analysis and synthesis of feedback will hopefully provide answers to some of these outstanding questions.

6.1 Planned features

To support novel use cases, CORD-19 will continue to invest in regular updates and new features. The list of requests is long, and we aim to prioritize features for which we can provide high quality content and which will likely produce the greatest benefit. Here is a short list of planned improvements:

- Moving to daily updates
- Adding papers from new sources
- Adding inbound and outbound citation links

- Adding additional paper content, such as tables and figures

We encourage members of the community to reach out to us and assist in these efforts.

6.2 Limitations

Though we aim to be comprehensive, CORD-19 does not cover many relevant scientific documents on COVID-19. We have restricted ourselves to research publications and preprints, and do not incorporate other types of documents that could be important, such as technical reports, white papers, informational publications by governmental bodies, and more. Including these documents is outside the current scope of CORD-19, but we encourage other groups to curate and publish such datasets.

Within the scope of scientific papers, CORD-19 is also incomplete. There are a set of known paper sources currently not integrated into CORD-19, which we hope to incorporate as part of immediate future work, listed above. However, we also note the lack of foreign language papers in the CORD-19 dataset, especially Chinese language papers produced during the early stages of the epidemic. We recognize that these papers may be useful to many researchers; we are working with collaborators to provide them as supplementary data.

6.3 Call to action

Though we have made the structured full text of many scientific papers available to researchers through CORD-19, a number of challenges prevent easy application of NLP and text mining techniques to these papers. First, the primary distribution format of scientific papers – PDF – is not amenable to text processing. The PDF file format is designed to share electronic documents rendered faithfully for reading and printing, not for automated analysis of document content. Paper content (text, images, bibliography) and metadata extracted from PDF are imperfect and require significant cleaning before they can be used for analysis.

Second, there is a clear need for more scientific content to be made easily accessible to researchers. Though many publishers have generously made COVID-19 papers available during this time, there are still bottlenecks to information access. For example, papers describing research in related areas (e.g., on other infectious diseases, or relevant biological pathways) are not necessarily open access, and are therefore not available in CORD-19 or

otherwise. Publishers have begun to share information about their search strategies and, with support from the NLM and OSTP, to work on identifying other topic areas that would be helpful to include in their search strings. Though the intention of the CORD-19 dataset is not to allow users to read or “consume” the constituent papers, our release of full text does constitute a form of re-publication, which is limited by copyright restrictions. Securing release rights for papers not yet in CORD-19 is a significant portion of future work, led by the PMC COVID-19 Initiative.⁴¹

Lastly, there is no standard format for representing paper metadata. Existing schemas like the NLM’s JATS XML NISO standard,⁴² Crossref’s bibliographic field definitions, or library science standards like BIBFRAME⁴³ or Dublin Core⁴⁴ have been adopted as representations for paper metadata. However, there are issues with these standards; they can be too coarse-grained to capture all necessary paper metadata elements, or lack a strict schema, causing representations to vary greatly across publishers who use them. Additionally, metadata for papers relevant to CORD-19 are not all readily available in one of these standards. There is therefore neither an appropriate, well-defined schema for representing paper metadata, nor consensus usage of any particular schema by different publishers and archives. To improve metadata coherence across sources, the community must come together to define and agree upon an appropriate standard of representation.

Without solutions to the above problems, NLP on COVID-19 research and scientific research in general will remain difficult. Significant progress has been made and will continue to be made in PDF parsing, for both the open source and enterprise audiences. However, the other challenges are not as easily addressable. Scientific publications are both the seed and fruit of many discussions around CORD-19, and improvements can be made around the availability and accessibility of papers for this type of broad textual analysis. We encourage the community to come together and propose solutions to these challenges.

⁴¹<https://www.ncbi.nlm.nih.gov/pmc/about/covid-19/>

⁴²<https://www.niso.org/publications/z3996-2019-jats>

⁴³<https://www.loc.gov/bibframe/>

⁴⁴<https://www.dublincore.org/specifications/dublin-core/dces/>

Summary

Through the creation of CORD-19, we have learned a lot about bringing together different communities around the same scientific cause. It is clearer than ever that automated text analysis is not the solution, but rather one tool among many that can be directed to combat the COVID-19 epidemic. Crucially, the systems and tools we build must be designed to serve a use case, whether that’s improving information retrieval for clinicians and medical professionals, summarizing the conclusions of the latest observational research or clinical trials, or converting these learnings to a format that is easily digestible by healthcare consumers.

This work also demonstrates the value of public access to full text literature. By allowing computational access to the papers in the corpus, we increase our ability to interact with and perform discovery over these texts. We believe this project will inspire new ways to use machine learning to advance scientific research, and serve as a template for future work in this area.

The community built around CORD-19 provides a roadmap to how connections between different scientific disciplines can be forged. With various shared tasks, we are hoping to incorporate domain expertise in the judgment of our automated systems, and to provide a notion of how to assess our successes going forward. Our intention is to continue working to enable this kind of grassroots scientific uprising, linking individuals with diverse backgrounds and expertise, to address similar kinds of crises (or non-crises) that arise in the future.

Acknowledgments

This work was supported in part by NSF Convergence Accelerator award 1936940, ONR grant N00014-18-1-2193, and the University of Washington WRF/Cable Professorship.

We thank The White House Office of Science and Technology Policy, the National Library of Medicine at the National Institutes of Health, Microsoft Research, Chan Zuckerberg Initiative, and Georgetown University’s Center for Security and Emerging Technology for co-organizing the CORD-19 initiative. We thank Michael Kratsios, the Chief Technology Officer of the United States, and The White House Office of Science and Technology Policy for providing the initial set of seed questions for the Kaggle CORD-19 research challenge.

We thank Kaggle for coordinating the CORD-19 research challenge. In particular, we acknowledge Anthony Goldbloom for providing feedback on CORD-19 and for involving us in discussions around the Kaggle literature review tables project. We thank the National Institute of Standards and Technology (NIST), National Library of Medicine (NLM), Oregon Health and Science University (OHSU), and University of Texas Health Science Center at Houston (UTHealth) for co-organizing the TREC-COVID shared task. In particular, we thank our co-organizers – Steven Bedrick (OHSU), Aaron Cohen (OHSU), Dina Demner-Fushman (NLM), William Hersh (OHSU), Kirk Roberts (UTHealth), Ian Soboroff (NIST), and Ellen Voorhees (NIST) – for feedback on the design of CORD-19.

We acknowledge our partners at Elsevier and Springer Nature for providing additional full text coverage of papers included in the corpus. We thank Bryan Newbold from the Internet Archive for providing feedback on data quality and helpful comments on the manuscript.

We also acknowledge and thank the following from AI2: Paul Sayre and Sam Skjonsberg for providing front-end support for CORD-19 and TREC-COVID, Michael Schmitz for setting up the CORD-19 Discourse community forums, Adriana Dunn for creating webpage content and marketing, Linda Wagner for collecting community feedback, Jonathan Borchardt, Doug Downey, Tom Hope, Daniel King, and Gabriel Stanovsky for contributing supplemental data to the CORD-19 effort, Alex Schokking for his work on the Semantic Scholar COVID-19 Research Feed, Darrell Plessas for technical support, and Carissa Schoenick for help with public relations.

References

- Sabber Ahamed and Manar D. Samad. 2020. Information mining for covid-19 research from a large volume of scientific literature. *ArXiv*, abs/2004.02085.
- Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. **SciBERT: A pretrained language model for scientific text**. 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)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.
- Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32 Database issue:D267–70.
- Q. Chen, Y. Peng, and Z. Lu. 2019. Biosentvec: creating sentence embeddings for biomedical texts. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–5.
- Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. Specter: Scientific paper embeddings using citation-informed transformers. In *ACL*.
- 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Luis Espinosa-Anke and Steven Schockaert. 2018. **SeVeN: Augmenting word embeddings with unsupervised relation vectors**. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2653–2665, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Karen Spärck Jones, Steve Walker, and Stephen E. Robertson. 2000. A probabilistic model of information retrieval: development and comparative experiments - part 1. *Inf. Process. Manag.*, 36:779–808.
- 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*.
- Yuxiao Liang and Pengtao Xie. 2020. Identifying radiological findings related to covid-19 from medical literature. *ArXiv*, abs/2004.01862.
- Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel S. Weld. 2020. **S2orc: The semantic scholar open research corpus**. In *Proceedings of ACL*.
- Patrice Lopez. 2009. **Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications**. In *ECDL*.
- Ryan McDonald, Georgios-Ioannis Brokos, and Ion Androutsopoulos. 2018. **Deep relevance ranking using enhanced document-query interactions**. In *EMNLP*.
- Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. **ScispaCy: Fast and robust models for biomedical natural language processing**. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. [Exploring the limits of transfer learning with a unified text-to-text transformer](#).
- 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)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Minjoon Seo, Jinhyuk Lee, Tom Kwiatkowski, Ankur Parikh, Ali Farhadi, and Hannaneh Hajishirzi. 2019. [Real-time open-domain question answering with dense-sparse phrase index](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4430–4441, Florence, Italy. Association for Computational Linguistics.
- Zhihong Shen, Hao Ma, and Kuansan Wang. 2018. [A web-scale system for scientific knowledge exploration](#). In *Proceedings of ACL 2018, System Demonstrations*, pages 87–92, Melbourne, Australia. Association for Computational Linguistics.
- George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artières, Axel-Cyrille Ngonga Ngomo, Norman Heino, Éric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. In *BMC Bioinformatics*.
- Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Yuxiao Dong, and Anshul Kanakia. 2020. Microsoft academic graph: When experts are not enough. *Quantitative Science Studies*, 1(1):396–413.
- Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Darrin Eide, Yuxiao Dong, Junjie Qian, Anshul Kanakia, Alvin Chen, and Richard Rogahn. 2019. A review of microsoft academic services for science of science studies. *Frontiers in Big Data*, 2.
- Xuan Wang, Xiangchen Song, Yingjun Guan, Bangzheng Li, and Jiawei Han. 2020. Comprehensive named entity recognition on covid-19 with distant or weak supervision. *ArXiv*, abs/2003.12218.
- Leigh Weston, Vahe Tshitoyan, John Dagdelen, Olga Kononova, Kristin Persson, Gerbrand Ceder, and Anubhav Jain. 2019. [Named Entity Recognition and Normalization Applied to Large-Scale Information Extraction from the Materials Science Literature](#).
- Edwin 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](#).