COVIDSeer : Extending the CORD-19 Dataset

Shaurya Rohatgi[1], Zeba Karishma[1], Jason Chhay[1], Sai Raghav Reddy Keesara[1]
Jian Wu[2],Cornelia Caragea[3],C. Lee Giles[1]
[1]Pennsylvania State University, [2] Old Dominion University,[3] University of Illinois Chicago
[1]{szr207, zbk5052, jqc6195, sfk5555, clg20}@psu.edu
[2]jwu@cs.odu.edu,[3]cornelia@uic.edu

Abstract

We develop an enhanced version of CORD-19 dataset released by the Allen Institute for AI. Tools in the SeerSuite project are used to exploit information in original articles not directly provided in the CORD-19 datasets. We add 728 new abstracts, 70,102 figures and 31,446 tables with captions that are not provided in the current data release. We also built a vertical search engine *COVIDSeer* based on the new dataset we created. COVIDSeer has a relatively simple architecture with features like keyword filtering, and similar paper recommendation. The goal was to provide a system and dataset that can help scientists better navigate through the literature concerning COVID-19. The enriched dataset can serve as a supplement to the existing dataset. The search engine, which offers keyphrase-enhanced search, will hopefully help biomedical and life science researchers, medical students, and the general public to more effectively explore coronavirus-related literature. The entire data set and the system will be made open source.

CCS Concepts: • **Information systems** → *Information retrieval*;

Keywords: datasets, data mining, information retrieval

ACM Reference Format:
Shaurya Rohatgi[1], Zeba Karishma[1], Jason Chhay[1], Sai Raghav Reddy Keesara[1] and Jian Wu[2],Cornelia Caragea[3],C. Lee Giles[1]. 2020. COVIDSeer : Extending the CORD-19 Dataset. In *ACM Symposium on Document Engineering 2020 (DocEng '20), September 29-October 2, 2020, Virtual Event, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3395027.3419597

1 Introduction

The COVID-19 pandemic had brought together a plethora of scientists in various fields, resulting in multitude of research papers on the subject. This large number of papers has made it difficult to keep track of this information, even in this niche domain. This has motivated the development of many search engines[1] and datasets [6] that focus on SARS-CoV-2 and related papers.

Table 1. COVIDSeer dataset compared with the 2020-04-10 release of the CORD-19 corpus.

Data Type	CORD-19	COVIDSeer
Abstracts	42,352	**43,080**
Keyphrases	No	**Yes**
Paper Recommendation	No	**Yes**
Figures w/ captions	-	**70,102**
Tables w/ caption	-	**31,446**
Total Papers	51,078	51,078

The COVID-19 Open Research Dataset (CORD-19) is a collection of papers related to Severe Acute Respiratory Syndrome Coronavirus2 (SARS-CoV-2) pandemic and was first released on March 16, 2020 by the Allen Institute for Artificial Intelligence (AllenAI), in collaboration with their partners.

The initial dataset comprised approximately 28K papers, and has since grown to more than 100K papers through subsequent weekly updates at the time of this writing. The dataset consists of research articles published in response to COVID-19 in 2020 and articles on similar viruses (for example, SARS and MERS) drawn from several sources including PubMed Central (PMC), bioRxiv, and medRxiv preprint servers and the World Health Organization (WHO) Covid-19 Database. Its aim was to connect the computer science community with biomedical domain experts and policy makers to help identify effective treatments and management policies for COVID-19 [12].

The CORD-19 dataset was built by using metadata from publishers such as PMC by extracting text from their PDFs. But missing in the dataset were potential data sources such as key-phrases, figures, tables, some missing abstracts, and similar paper recommendations. To fill this missing information, we used the tools developed in SeerSuite [11, 13] to

[1] https://discourse.cord-19.semanticscholar.org/t/cord-19-demos-and-resources/132

DocEng '20, September 29-October 2, 2020, Virtual Event, CA, USA
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-8000-3/20/09…$15.00
https://doi.org/10.1145/3395027.3419597

enhance the CORD-19 dataset with fields mentioned in Table 1. We build upon the COVID-19 Fatcat Snapshot[2] which is a superset of papers in CORD-19. It was released by the Internet Archive (IA) and includes PDF files of these articles. PDFs were NOT included the original CORD-19. PDFs have richer information than just text but extracting it is challenging. We used the tools in SeerSuite[3] to mine such PDF data.

To ease the exploration of this enriched data, we developed a search engine, COVIDSeer[4]. The frontend and backend which will be described later were designed to accommodate regular updates to the dataset. We use GROBID [9] to extract the missing abstracts and PDFFigures [5] to extract the captions, figures, and tables. We use Elasticsearch[5] to index and search the data. The scalable implementation of Lucene in Elasticsearch allows real-time querying and filtering based on authors, years, venues, and key-phrases. In addition, multiple machine learning based tools were used for key-phrase extraction [4] and similar paper recommendations. Work presented here used the CORD-19 dataset released on April 10, 2020 which contains 51,045 scientific publications of potential relevance to coronavirus and other closely related areas in virology, epidemiology, and biology.

2 COVIDSeer

We present an overview of the information extraction process followed by a detailed technical description of search architecture and discuss the adopted methodology for extraction and indexing and distinctive features of our information retrieval system.

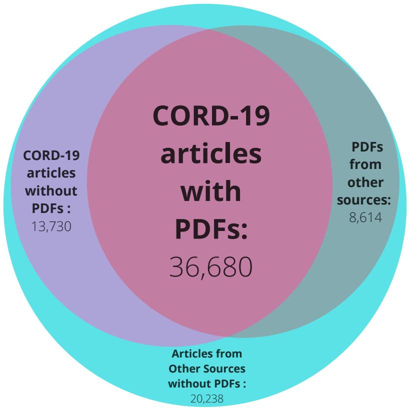

Figure 1. Overview of COVID-19 Fatcat Snapshot Dataset released by the Internet Archive. Other papers includes articles from sources other than Semantic Scholar

2.1 Dataset

We use the COVID-19 Fatcat Snapshot released by IA to get the PDFs for the CORD-19 dataset. It contains 45,294 full-text PDF files. These papers are a subset of the CORD-19 dataset with mappings of the unique ID indentifier for CORD-19; *cord_uid* to there own indentifiers. This makes it easier to map a corresponding PDF to it's CORD-19 metadata. All of these papers are open access and publicly available. More details about this dataset are shown in Figure 1. We see there are 36,680 PDFs which have a corresponding *cord_uid*. We maintain this unique identifier and its mappings across our dataset as well for easier adoption of other tools. IA also provides full-text and other metadata with the articles which we do not use. We very much rely on our extraction pipeline.

2.2 Information Extraction

Content extraction from scholarly PDFs allows authors to readily access particular entities, such as abstracts, results, tables, and/or figures. This allows easy access for someone who is not interested in all parts of the scholarly literature but rather just in particular sections related to their particular interests. We use PDFMEF [13], a multi-entity knowledge extraction framework for scholarly documents in a PDF format that extract multimodal information from papers. PDFMEF encapsulates state-of-the-art extraction tools for various types of content. It uses GROBID for document segmentation and metadata extraction and pdffigures for figure and table extraction. We can also configure PDFMEF to customize extracted content. Given the availability of full-text in the CORD-19 dataset, we focused on abstract, figure and table extraction. We illustrate this pipeline in Figure 2

Figure and Table Extraction. PDFMEF internally utilizes PDFFigures [5] to extract tables and figures. An independent evaluation suggests that it takes on average 0.2 seconds on each page and can achieve an F1-measure of 90% [13]. A compiled binary with set timeout of 20 seconds is called for execution. A timeout is set to avoid program freeze due to unusual file processing. The extracted figures and tables are collected in a directory with paper ID as prefix. From the 36,680 PDFs we have, we were able to successfully extract figures, tables, and there corresponding captions from only 30,274 PDFs. This created a total of 101,937 images (figures and tables) with captions.[6]

Abstract Extraction. GROBID extracts the bibliographical data corresponding to the header information (titles, authors and affiliation) with an accuracy of 77.79% per complete header instance. It takes a complete PDF as input to generate the TEI file in the output, and then the fields corresponding to a predefined and customized metadata schema are extracted into an XML file.

[2] https://archive.org/details/covid19_fatcat_20200410
[3] https://github.com/SeerLabs/CiteSeerX
[4] COVIDSeer: https://covidseer.ist.psu.edu/
[5] https://www.elastic.co/elasticsearch/

[6] In-depth view of the extracted data and cluster analysis : https://github.com/jasonchhay/CovidSeer/blob/master/CovidSeer_supplementary.pdf

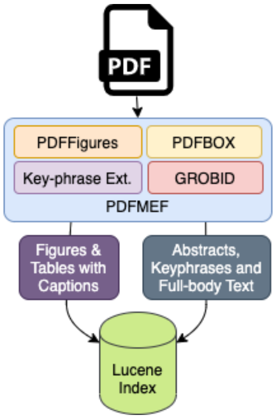
PDF
PDFFigures
PDFBOX
Key-phrase Ext.
GROBID
PDFMEF
Figures & Tables with Captions
Abstracts, Keyphrases and Full-body Text
Lucene Index

Figure 2. Extraction and indexing pipeline for COVIDSeer. Extracted text is indexed by Elasticsearch which uses Lucene.

The dataset released on 04-10-2020 consisted of 51,045 research papers with 50,410 documents having an abstract. COVID-19 FATCAT Snapshot builds on top of that and adds more than 26,000 new abstracts from sources such as WHO, Wanfang corpus, CNKI corpus and FATCAT [7]. We only stick to the articles which have a *cord_uid* and belong to the CORD-19 corpus. IA has also released there version of metadata which we do not use for the data. We only use the PDFs with a corresponding *cord_uid*. AllenAI when building CORD-19 [12] used GROBID for abstract extraction. Our previous work shows that PDFBOX works better than GROBID [13] for text extraction [8] As a result of using PDFBox, we were able to extract more abstracts. PDFBOX was executed on 36,680 PDF documents and 728 missing abstracts were added to the collection.

Keyphrase extraction. Given the increasing breath of the COVID-19 literature, a high level topic description of a document could serve as a rich source of information. Therefore, a supervised keyphrase extraction model, also known as citation-enhanced keyphrase extraction (CeKE) model [3] built on a combination of novel features that captures information from citation contexts and existing features such as tf-idf,parts-of-speech tagging, and relative position of the text, was utilized for keyphrase extraction. There are two types of citation context. Citing contexts are texts around citations in the current article; cited contexts are texts around citations where the current paper is cited in other articles. The latter requires full text of all articles in the citation graph, which is not available in the CORD-19 dataset. For the entries with full-text, we have extracted the keyphrases with the model using citing context. Otherwise, we use CeKE-TA that uses only the title and abstract. We extracted top-10 ranked keyphrases for each document. For the entries with full-text, keyphrases were extracted using CeKE-Citing. Otherwise CeKE-TA was used when only the title and abstract were available.

2.3 Search Engine Architecture

The information extracted is preprocessed using a standard Natural Language Processing pipeline similar to Covidex [15] and later indexed by Elasticsearch. By choosing Elasticsearch, we utilize a traditional monolithic ranking based architecture which is a combination of Boolean, TF/IDF, and the vector space models. ElasticSearch provides built-in analyzers which divides text into terms on word boundaries, based on Unicode Text Segmentation algorithm, and handles stemming, stopword removal, and tokenization internally. For text search, we use a combination of title, abstract, and available full-texts. These are indexed as separate fields into the Elasticsearch JSON documents. This breaking down the document in smaller atomic units ensures higher recall for the system [8].

The COVIDSeer website is deployed through the Django web framework because of its ease-of-use in serving web files, its Python back-end which easily supports Elasticsearch’s Python library, and existing internal search engine frameworks that we built upon to remove development overhead. The user interface is developed through a hybrid of Django’s template language and Vue.js. Vue is utilized to increase COVIDSeer’s reactivity and browser load times. Pages are able to load much faster if the data is retrieved asynchronously through a REST API tied to the back-end instead of hanging the page while data is retrieved from Elasticsearch. Using our experiences from CiteSeerX [14], we wanted to see how well we could develop a similar search engine for COVID-19 research. With our rich experience in CiteSeerX [14], we were able to deploy the initial version COVIDSeer in 2 weeks. We also wanted to make other data available in a form that can be easily used by researchers. With this objective, we focused on two key areas: first, extracting more information from the dataset by means of abstract and keyphrase extraction and; second, enhancing the users’ search experience by means of faceted search and similar paper recommendations.

Faceted Search. Faceted search presents users with key-value metadata that can be used for query refinement [7]. Narrowing down traditional search results along multiple explicit dimensions allows the classification to be assessed in multiple ways. Therefore, we built and integrated a filtering mechanism for further accessing results of a query of interest. Available facets include authors, source, journal, publication year, presence/absence of full-text/abstracts. Sources are archival services or online repositories like bioRxiv, medRxiv, and others preprint services. This allows users to select filters from one or multiple categories and the intersection of all is presented in the search results.

[7] https://fatcat.wiki/

[8] This study was based on GROBID 0.4 and PDFBox 1.8.6. However, for the latest releases no comparison exists.

Paper Recommendations. We also provide a list of top-10 similar papers for each paper in the dataset. We do this by first obtaining the top-10 candidate similar papers using tf-idf representation and cosine similarity of the abstracts and titles [10]. Second, we rerank these top-10 papers using SciBERT [1] embeddings of the abstracts and titles. To represent a paper, we average the contextual word-level embeddings for each abstract and title. Once we have these vectors we find the cosine similarity between the paper and the candidates returned by tf-idf similarity in the first step. It should be noted that our paper recommendation method is based on semantic similarity and recommended papers are within the CORD-19 corpus. A similar approach has been demonstrated in the work Citeomatic [2], where content-based recommendations were shown to give better results than citation graph-based approaches.

3 Discussion

COVIDSeer was launched online on March 30th, 2020. Since it was available online, we have received approximately 2,100 queries from 1,003 unique IP addresses. A systematic evaluation requires sufficient labeled and/or user-click data, which is is being collected. We target to study how keyphrases assist users to more efficiently navigate to desired papers. This work was done on a fixed snapshot of dataset released by CORD-19 and IA. With new weekly releases of CORD-19 dataset it is hard to keep our crawl repository and search engine updated, as crawling the actual PDFs for the articles will require sophisticated crawlers which we plan to build in the future. Nevertheless, we believe the enhancements and enrichment of the original dataset proposed and demonstrated in this paper will be valuable to the research community.

4 Conclusion and Future Directions

Employing a amalgamation of information retrieval and machine learning techniques, we have enhanced the CORD-19 dataset through the addition of abstracts and principal keyphrases in the available literature. With the evolving nature of the CORD-19 dataset, we learned that building up a pipeline that updates the dataset used by the search engine is important. Furthermore, we use a 2 step model to recommend similar papers to a given paper with the aim of helping a researcher in his goal of finding relevant related work.

In addition to the information retrieval tasks, the proposed dataset can be coupled with machine learning to perform tasks such as Image Captioning, Table and Visual Question Answering, and Document Figure Classification.

Future work could be to implement author name disambiguation so as to correctly associate every author to their research paper. Another task would be creating a set of question templates for each of the found clusters and annotating figures and tables associated with them.

5 Acknowledgements

We gratefully acknowledge partial support from the National Science Foundation.

References

[1] 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)*. 3606–3611.

[2] Chandra Bhagavatula, Sergey Feldman, Russell Power, and Waleed Ammar. 2018. Content-Based Citation Recommendation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 238–251.

[3] Cornelia Caragea, Florin Adrian Bulgarov, Andreea Godea, and Sujatha Das Gollapalli. 2014. Citation-Enhanced Keyphrase Extraction from Research Papers: A Supervised Approach. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1435–1446. https://doi.org/10.3115/v1/D14-1150

[4] Hung-Hsuan Chen, Jian Wu, and C Lee Giles. 2017. Compiling Keyphrase Candidates for Scientific Literature Based on Wikipedia.

[5] Christopher Clark and Santosh Divvala. 2015. Looking Beyond Text: Extracting Figures, Tables, and Captions from Computer Science Paper. (2015).

[6] Emanuele Guidotti and David Ardia. 2020. COVID-19 data hub. (2020).

[7] Jonathan Koren, Yi Zhang, and Xue Liu. 2008. Personalized Interactive Faceted Search. Association for Computing Machinery, New York, NY, USA.

[8] Jimmy Lin. 2009. Is searching full text more effective than searching abstracts? *BMC bioinformatics* 10, 1 (2009), 46.

[9] Patrice Lopez. 2009. GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications. In *ECDL*.

[10] Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage Re-ranking with BERT. *arXiv preprint arXiv:1901.04085* (2019).

[11] Pradeep B Teregowda, Isaac G Councill, Juan Pablo Fernández Ramírez, Madian Khabsa, Shuyi Zheng, and C Lee Giles. 2010. SeerSuite: Developing a Scalable and Reliable Application Framework for Building Digital Libraries by Crawling the Web. *WebApps* 10 (2010), 14–14.

[12] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Michael Kinney, Ziyang Liu, William. Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The Covid-19 Open Research Dataset. *ArXiv* abs/2004.10706 (2020).

[13] Jian Wu, Jason Killian, Huaiyu Yang, Kyle Williams, Sagnik Ray Choudhury, Suppawong Tuarob, Cornelia Caragea, and C. Lee Giles. 2015. PDFMEF: A Multi-Entity Knowledge Extraction Framework for Scholarly Documents and Semantic Search. Association for Computing Machinery, New York, NY, USA.

[14] Jian Wu, Kunho Kim, and C. Lee Giles. 2019. CiteSeerX: 20 years of service to scholarly big data. *Proceedings of the Conference on Artificial Intelligence for Data Discovery and Reuse* (2019).

[15] 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. arXiv:2004.05125 [cs.CL]