J. Indian Inst. Sci.
A Multidisciplinary Reviews Journal
ISSN: 0970-4140 Coden-JIISAD

REVIEW ARTICLE

A Web Resource for Exploring the CORD-19 Dataset Using Root- and Rule-Based Phrases

Jacob Collard[1], Talapady Bhat[1], Eswaran Subrahmanian[1,2], Ira Monarch[1], Jonah Tash[1], Ram Sriram[1] and John Elliot[1]*

Abstract | This short paper describes a web resource—the NIST CORD-19 Web Resource—for community explorations of the COVID-19 Open Research Dataset (CORD-19). The tools for exploration in the web resource make use of the NIST-developed Root- and Rule-based method, which exploits underlying linguistic structures to create terms that represent phrases in a corpus. The method allows for auto-suggesting-related terms to discover terms to refine the search of a COVID-19 heterogenous document base. The method also produces taxonomic structures in the target domain as well as providing semantic information about the relationships between terms. This term structure can serve as a basis for creating topic modeling and trend analysis tools. In this paper, we describe use of a novel search engine to demonstrate some of the capabilities above.

Keywords: *Root- and rule-based method, CORD-19 dataset, Auto-suggest search*

1 Introduction

The NIST Root- and Rule-based method (R&R)[1,2] is a framework built around linguistic structures to identify and index key phrases. R&R defines how individual natural language words ("roots") can be combined into structured terms. These structured terms represent natural language phrases in accordance with their linguistic structure, allowing for relationships between the individual words or between complex phrases to be identified. These terms both disambiguate and normalize the natural language text according to their linguistic structure ("rules"). The R&R method draws insight from noun compounds in Sanskrit, German, Latin, and other languages.

The R&R method can be used for simple single term and advanced searches that disambiguate a user's search query to increase the relevance of retrieved documents and to normalize the query to retrieve a wider set of relevant documents. This method can be applied to a variety of different domains without modifying the overall framework; practical differences in vocabulary, language use, and abbreviations can be accounted for by modifying a few simple parameters of the framework.

2 The NIST CORD-19 Web Resource

The COVID-19 Open Research Dataset (CORD-19) provides a collection of articles related to SARS-CoV-2, COVID-19, and related viruses and diseases.[3] The corpus is obtained from PubMed, the WHO, bioRxiv, and medRxiv, and is updated daily with new relevant articles. This large, open-access corpus makes it possible for researchers to track and analyze new developments related to COVID-19. However, due to its size and breadth, answering specific research questions about the data can be difficult when only a subset of the corpus is needed—extracting the relevant subset is a challenging problem. Identifying trends in research also requires the analysis of the relationships between concepts represented in the CORD-19 textual data repository. Both of these tasks can be facilitated through the R&R method—normalized indices can be used to improve search and the structure of the terms can be used to facilitate other analyses.

[1] *National Institute of Standards and Technology, Gaithersburg, MD, USA.*
[2] *Carnegie Mellon University, Pittsburgh, USA.*
**es3e@andrew.cmu.edu*

The CORD-19 Web Resource aims to provide tools that increase the accessibility of CORD-19 for research and other use-cases. The web resource consists of several components including the COVID-19 Data Curation System (CDCS), which is a data registry and a document repository (https://covid19-data.nist.gov/) and the COVID-19 document search portal (https://randr19.nist.gov/) that is described in this paper. The CDCS repository incorporates the R&R method internally for managing search needs over the registry and repository.

The document search portal uses the R&R method to index and query a mirror of the full-text portion of CORD-19. The system also provides automatically extracted key phrases for each document in the corpus.

2.1 Data and Preprocessing

CORD-19 is drawn from three separate data sources:

- PubMed's PMC corpus. CORD-19 contains all sources that match the following 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"
- The WHO corpus of global research on coronavirus disease
- bioRxiv and medRxiv pre-prints that match the same query in PubMed.

CORD-19 also contains data from PubMed, Microsoft Academic, and the WHO COVID-19 database of publications, though not all of these are included in the NIST CORD-19 Web Resource, since not all documents from these sources include open-access full text. At the time of this writing, CORD-19 contains over 82,000 documents with full text.

Full-text documents in CORD-19 are all represented using a JSON format that contains metadata about each article, including the title, authors, unique identifiers (such as PubMed Ids), and the abstract. Each JSON file also contains the sections of each article, divided into paragraphs and labeled.

To process a corpus, R&R requires documents mapped into a flattened structure—i.e., any hierarchical divisions such as sections and subsections will be ignored. To provide a more nuanced search, we choose to treat the paragraph as the core document in the corpus. To accomplish this, we first preprocess the CORD-19 data to produce CSV files where each row represents a single paragraph in the corpus. Each row contains the text of the paragraph as well as metadata including the authors, unique identifiers, and title of the source document.

Once the documents are in the appropriate format, they can be processed using R&R. R&R takes as input a collection of documents—in this case, paragraphs from full-text articles in CORD-19. The output is a collection of structured terms with reference to the original documents (or, in this case, paragraphs).

2.2 R&R Processing

R&R processes each sentence in the document according to the following general set of rules, which produces the list of terms that correspond to the phrases in that sentence[5]. The first rule constructs a syntactic constituency parse of the sentence[2].

Once a constituency tree is constructed, the second rule of the R&R system selects all subtrees which match a particular filter—in this case, all subtrees whose syntactic heads are either nouns or verbs are included. This means that the extracted terms will represent noun or verb phrases, but not other types of phrases such as adjective or prepositional phrases. Adjectives and prepositions may still occur inside of the phrases that are extracted, but do not appear as syntactic heads of any of the terms.

Finally, once phrases are extracted, they are normalized. This is done by reconfiguring each tree, so that the syntactic heads of each subtree occur at the far right of the tree—in other words, each tree is coerced into a head-final configuration. At the same time, stop words such as "of" and "by" are removed, resulting in many similar phrases being represented by the same structures. However, phrases with significantly different meanings will still have different representations, even if the words are the same.

As an example, consider the phrase "Subversion of cellular autophagosomal machinery by RNA viruses". This entire phrase can be represented as a syntactic tree, normalized, and extracted by the R&R system. A number of other phrases can also be extracted from the same sentence, such as "cellular autophagosomal machinery" and "RNA viruses", since these are themselves noun phrases. However, "by RNA viruses" is not used in this particular resource,

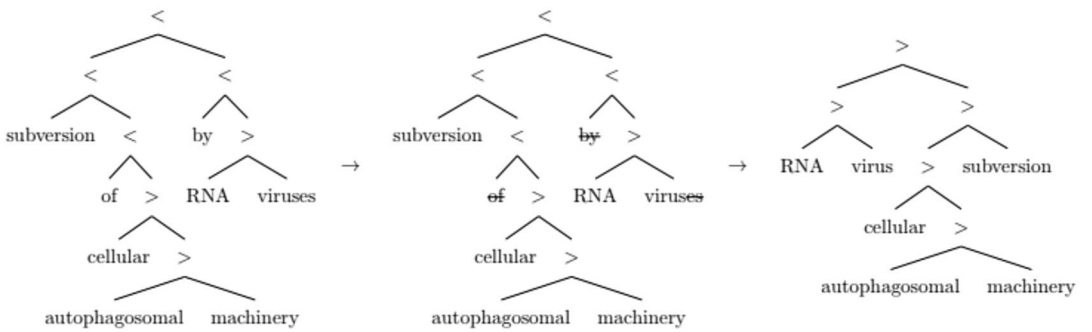

Figure 1: Constructing an R&R term.

since prepositional phrases are not extracted. This example is shown in Fig. 1.

In summary, first, the phrase is parsed into a constituency tree. Then, stop words such as "of" and "by" are removed and individual words are normalized, e.g., by removing plural forms such as "viruses" and replacing them with lemmas, such as "virus". Finally, the tree is modified, so that the head of each branch is the right child of the node.

There is one last step to the R&R process: the creation of a linearized representation of the tree. The final tree is flattened into a human- and machine-readable form that captures the same information as the tree. This form represents nodes with numerals indicating the distance to the bottom of the tree. Thus, the final term in this example would be "RNA:0:virus:3 :cellular:1:autophagosomal:0:machinery:2:sub version". The logic of the numbering system is that lowest branches of the tree on the right are coded as "autophagosomal:0:machinery" and the next level of the tree the term "cellular: 1: autophagosomal:0:machinery" corresponding to the term "cellular autophagosomal machinery" and so on up the tree. The left side of the tree has two terms at the same level "RNA:0:Virus", but it is at the 3rd level with respect to the term "autophagosomal machinery." This structured term is then tied to metadata, including the paragraph, sentence, and phrases that contain the term. When presented to the end user, the term is usually expressed as one of the canonical natural language phrases that originally generated it—in this case, "subversion of cellular autophagosomal machinery by RNA viruses". This allows the term to be easily understood, while maintaining the structure in the underlying representation. This structure can be used to find related terms and create taxonomies. Because of the structure of the term, other phrases that simply contain the same words are not necessarily identified as similar unless they also have overlapping sub-structures, such as "autophagosomal:0:machinery".

2.3 Constructing the Knowledge Base

Once the set of terms is constructed from the R&R method, the terms and corresponding data are loaded into a relational database. A web interface allows the user to enter search terms to find documents relevant to those search terms. When the user enters a search term, the term can be converted into an R&R term following the same method outlined above. This can be done in real time, to generate potential terms that contain the search term as the user types. Potential terms can then be extended with real terms that occur in the database to provide auto-suggestions of related terms that are present in the corpus. These terms are displayed using canonical phrases, allowing the user to easily understand them.

The data are stored in three layers. The first layer contains the data from CORD-19 in tabular form, including the text of the documents. Each paragraph is its own row, with metadata connecting it to the larger document. This allows for the interface to represent snippets of larger documents and the core metadata, as shown in Fig. 2.

The second layer contains the relationships between terms and documents. Again, terms are related to snippets, not directly to documents, allowing for the relevant portion of the document to be identified, rather than the entire document, which may contain other, irrelevant sections. This is used to generate previews of the data, as shown in Fig. 3.

The third layer contains the terms and relationships between them. This allows for the auto-suggest system to identify terms that

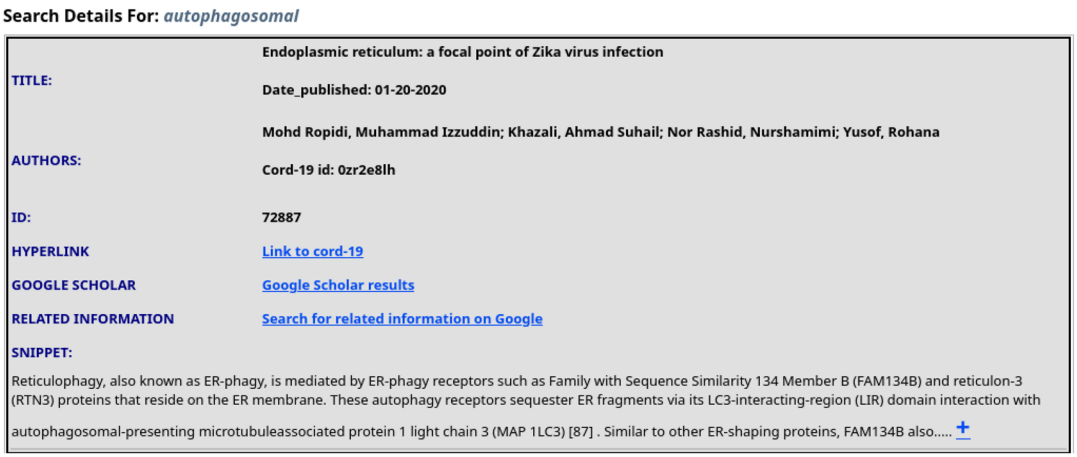

Figure 2: Web interface detailed view.

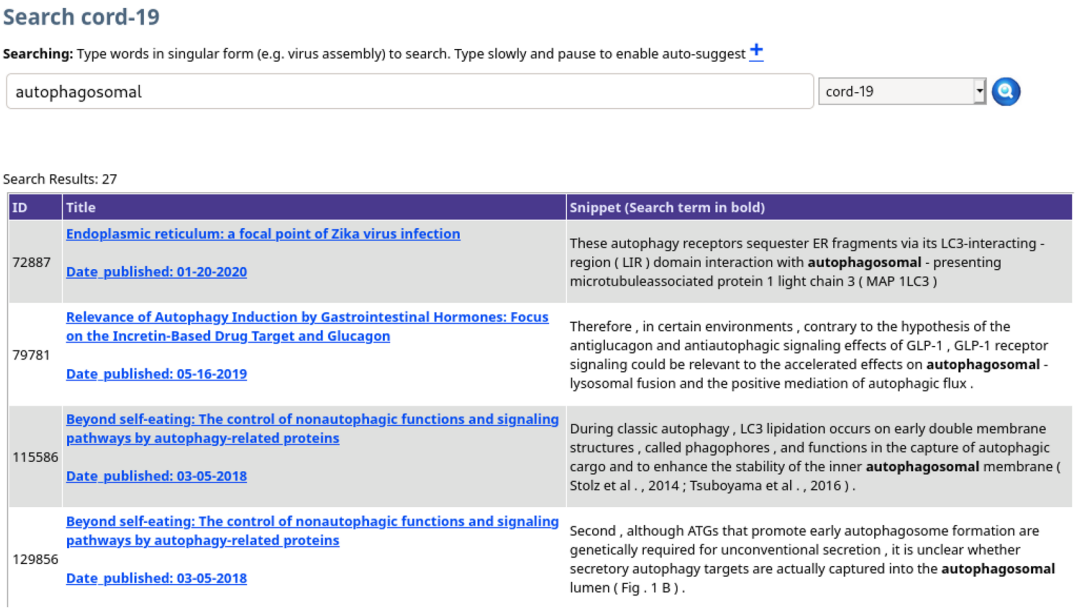

Figure 3: Web interface preview.

closely match the user's input and suggest other related terms.

2.4 Analysis

The R&R system also allows for more nuanced analyses of terms. Though there are many possible uses for the structure of R&R terms, one immediately apparent use is in the construction of a taxonomy. A taxonomy relates terms by identifying hierarchical relationships between terms, e.g., by identifying one term as a hyponym of another. Such a taxonomy can be generated automatically from R&R terms, since closely related terms will always share the same head, which is the right-most root of any given term. In fact, any term which adds roots to the left of another is a hyponym of that term, so long as it does not cross constituent boundaries. As an example, a simple taxonomy can help to identify different kinds of viral particles that are being discussed in the literature, as shown in Fig. 4.

This is only a sample of a much larger taxonomy, which can provide further relationships between important concepts as well as information about the evolution of those relationships over time, when combined with the metadata

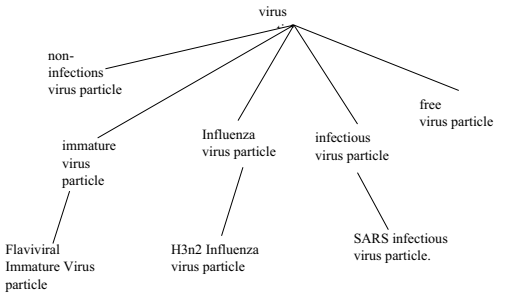

Figure 4: Sample taxonomy of terms.

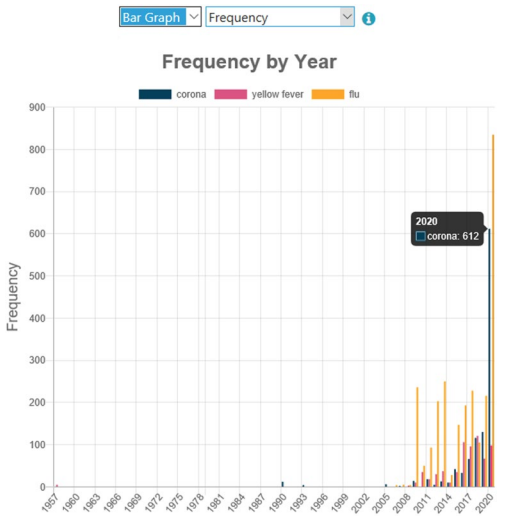

Figure 5: Trending of the terms corona, Yellow Fever, and flu (From CORD-19 dataset).

in the corpus and made available on the web interface.

Another form of analysis is the study of particular terms that are trending over time. The example below shows the trending of three terms, Corona, Yellow Fever, and flu. As we can see in Fig. 5, the usage of these terms peaks at around the period of the epidemic. Similar analysis can be extended to more complex searches that include more than a single term.

Future research involves analysis of R&R terms for knowledge generation. For example, topic models can group terms together not just by identifying relationships based on their structure, but also on relationships between term co-occurrence in particular documents or sections. This can also be considered both for the whole corpus, for particular subsets of the corpus, or to track the evolution of the corpus over time.

The majority of the examples discussed in this paper are constructed using noun phrases, but verb phrases were extracted as well, according to the same R&R principles. Verb phrases can be included in the taxonomy, but more importantly, they can also be used to identify other relationships between terms. Verbs naturally define relations between zero or more arguments, depending on the transitivity of the verb in question. Because of the way R&R terms are constructed, it is possible to identify these arguments as well as the core relation to produce knowledge graphs that show how all of the terms in a corpus are related. This is a work in progress, but could be combined with the interface presented in this article to explore the new developments and track changes in the field of virus research, or in a particular sub-area. The approach is generalizable to any context of the domain.

Though these innovations have not yet been incorporated into the web interface, the interface as it stands provides linguistically aware access to the CORD-19 full-text collection, allowing for a more accessible, domain-specific search that can correctly disambiguate and identify crucial terms in the corpus.

3 Conclusion

This paper uses the exemplar of CORD-19 to illustrate the implementation of the R&R method in the NIST CORD-19 Web Resource as a tool for researchers in the domain of coronavirus research. The R&R method proceeds from the premise that different scientific domains use different sublanguages. Therefore, for example, the same natural language word is used differently (belongs to different word co-occurrence classes) from one domain or sublanguage to another[4]. Usage is determined by the community of practice. By automatically extracting terms from actual language use in a scientific domain, our approach creates conceptual structures that can provide an evolving standardized vocabulary for search. It, thus, performs very well in contrast to the traditional standardized search vocabularies as it provides knowledge structures that are derived directly from published scientific texts[5]. We have illustrated the advantage of the method to create taxonomies of terms and to provide auto-suggestions based on the usage of terms and phrases in the target domain. The CORD-19 Web Resource illustrates this functionality and provides an ongoing setting for further developments. For example, our evolving data-driven indexing approach is being combined with topic modeling, topic trend analysis, and knowledge graphs to create new research

supporting capabilities. The aim is to facilitate gaining appreciation of a research landscape when no researcher has time to read everything being published in their field. The hope is to help researchers bring together areas of research that only a few or perhaps none has thought of doing before that can lead to new concepts, methods, and discoveries.

Commercial products are identified in this article to adequately specify the procedure. This does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

Publisher's Note

Received: 7 September 2020 Accepted: 7 September 2020
Published online: 29 September 2020

References

1. Bhat TN, Bartolo LM, Kattner UR, Campbell CE, Elliott JT (2015) Strategy for extensible, evolving terminology for the materials genome initiative efforts. J Mater 67:1866–1875
2. Collard J, Bhat TN, Subrahmanian E, Sriram RD, Elliot JT, Kattner UR, Campbell CE, Monarch I (2018) Generating domain terminologies using root-and rule-based terms 1. Wash Acad Sci J Wash Acad Sci 104(4):31–78
3. Wang LL, Lo K, Chandrasekhar Y, Reas R, Yang J, Eide Burdick D, Eide D, Funk K, Katsis Y, Kinney R, Liu Z, Merrill W, Mooney P, Murdick D, Rishi D, Sheehan J, Shen Z, Stilson B, Wade A, Wang K, Wang NX, Wilhelm C, Xie B, Raymond DM, Weld DS, Etzioni O, Kohlmeier S (2020) CORD-19: The Covid-19 Open Research Dataset. From ArXive: https://arxiv.org/pdf/2004.10706.pdf. Accessed 15 Aug 2020
4. Harris Z, Gottfried M, Ryckman Th, Daladier A, Mattick P (1989) The form of information in science: analysis of an immunology sublanguage. In: Boston Studies in the Philosophy and History of Science. Springer Netherlands, Amsterdam
5. Collard J, Bhat TN, Elliott J, Sriram R, Monarch I, Subrahmanian E (2020) Information retrieval with root- and rule-based terms. Available at SSRN: https://ssrn.com/abstract=3565983 or http://dx.doi.org/https://doi.org/10.2139/ssrn.3565983

Jacob Collard is a National Research Council Post-doctoral Fellow at NIST. He received his PhD in computational linguistics in 2020 from Cornell University for work on unsupervised learning to model natural language syntax by incorporating linguistic theory with statistical processing. He has worked with NIST as an independent consultant for 5 years to develop tools for deriving representations of concepts and semantic relations from natural language text.

Talapady Bhat is a project leader at NIST and one of his recent goals is to develop tools and techniques to enable archiving, searching and sharing scientific information. Prior was a co-PI of the RCSB Protein Data Bank project at NIST. Scientific citation index to his publications, as per Google Scholar exceeds 39,000. He got his Ph.D. degree from the Department of Physics, Indian Institute of Science in 1976.

Dr. Eswaran Subrahmanian is a Research Professor at the Institute for Complex Engineered Systems and Engineering and Public Policy at Carnegie Mellon University and Guest Researcher at the National Institute of Standards and Technology, USA. He has held visiting professorships at the Faculty of Technology and Policy Management at TU-Delft (Netherlands), the University of Lyon II; and the National Institute of Standards and Technology. His research is in the areas of Socio-technical systems design, Decision support systems, Engineering informatics, Design theory and methods, and engineering design education. He has three edited books and is a coauthor of a book on Design. His group integrated ethnographic methods in designing design processes and collaborative work support systems with Westinghouse, ABB, Alcoa, Bombardier, Boeing, and Robert Bosch. He is a founding member of a Bangalore-based non-profit working urban design issues using Simulation, Visualization and Gaming startup called Fields of View in India. He is the co-chair of the Special Interest Group on Design theory in the Design Society, Distinguished Scientist of the Association of Computing Machinery and a Fellow of the American Association of Advancement of Science.

Ira Monarch has investigated information design and process issues in large-scale engineering programs, both military and industrial, for over thirty-five years. After retiring as Senior Member of the Technical Staff from the Software Engineering Institute (SEI), he began working with a team at NIST to improve search and mining of NIST databases and other information sources.

Jonah Tash is a student intern working part time with Dr T. N. Bhat at NIST during his studies at University of Maryland.

Ram D. Sriram is currently the chief of the Software and Systems Division, Information Technology Laboratory, at the National Institute of Standards and Technology. Before joining the Software and Systems Division, Sriram was the leader of the Design and Process group in the Manufacturing Systems Integration Division, Manufacturing Engineering Laboratory, where he conducted research on standards for interoperability of computer-aided design systems. Prior to joining NIST, he was on the engineering faculty (1986–1994) at the Massachusetts Institute of Technology (MIT) and was instrumental in setting up the Intelligent Engineering Systems Laboratory. Sriram has co-authored or authored more than 275 publications, including several books. Sriram was a founding co-editor of the International Journal for AI in Engineering. Sriram received several awards including: an NSF's Presidential Young Investigator Award (1989); ASME Design Automation Award (2011); ASME CIE Distinguished Service Award (2014); the Washington Academy of Sciences' Distinguished Career in Engineering Sciences Award (2015); ASME CIE division's Lifetime Achievement Award (2016); CMU CEE Lt. Col. Christopher Raible Distinguished Public Service Award. Sriram is a Fellow of ASME, AAAS, IEEE, Solid Modeling Association, and Washington Academy of Sciences, a Distinguished Member (life) of ACM and Senior Member (life) AAAI. Sriram has a B.Tech. from IIT, Madras, India, and an M.S. and a Ph.D. from Carnegie Mellon University, Pittsburgh, USA.

Dr. John T. Elliott is the group leader of Cell Systems Science Group at NIST. He is currently developing quantitative microscopy techniques for measuring cellular response in a variety of applications. Specific projects have involved development of novel cell stains for automated fluorescence microscopy, fixation techniques to preserve GFP within cells, fluorescence reference materials for intra-laboratory and inter-laboratory standardization of fluorescent microscopes and open source image analysis software to facilitate quantification of 3-color microscopy images.