POLITECNICO MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE E DELL'INFORMAZIONE

Executive Summary of the Thesis

Exploring and searching the CORD-19 big data corpus for supporting COVID-19 research

Laurea Magistrale in Computer Science and Engineering - Ingegneria Informatica

Author: Francesco Invernici

Advisor: Prof. Stefano Ceri

Co-advisor: Dr. Anna Bernasconi

Academic year: 2021-2022

---

1. Introduction

On 11 March 2020, the World Health Organization declared the outbreak of coronavirus disease 2019. (COVID-19) a global pandemic COVID-19 is a contagious disease caused by a virus, the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). As a countermeasure, the governments of many countries in the world had to introduce drastic preventive measures, such as lockdowns and mandatory vaccination campaigns. Thus, the pandemic has had unprecedented effects in modern history on both a social and scientific level. Specifically, the scientific community has joined its efforts and has produced an extraordinary amount of publications concerning the disease, which has never been seen for any other infectious disease. Even though the high volume of publications may represent such a positive effort against the disease, we should not consider only the advantages, but also the questions that it entails. The research against a disease in a pandemic climate is a hard challenge of its own against time and resources, therefore, we need to think about making it effective, efficient, and safe.

Motivation. In order to accelerate the research against the disease, private and public institutions made available large corpora of scientific publications about COVID-19 as open datasets. The largest of these datasets is the COVID-19 Open Research Dataset (CORD-19)[6], curated by the Allen Institute for AI, which in its final release included entries for more than one million papers of which 370 thousand with full-text content. These data present all the 5 V characteristics of Big Data: Volume, Variety, Veracity, Value, and Velocity. In particular, in addition to the high volume, the variety produces strong technological challenges, due to the unstructured nature of content in natural languages. In Health and Life Sciences, veracity is always questioned, since the intrinsic multidisciplinarity of the domain translates to many specific representations that always have to be integrated. The value is a characteristic that we have to preserve from the sources to the results, indeed, we have to guarantee that even rare pieces of information are reported.

Objectives. For this work, we define two main objectives: making scientific information related to COVID-19 easily accessible by users that are experts in the biomedical domain, and enabling a process of knowledge discovery, built from the information available in the publications.

2. CORD-19 Analysis and Information Extraction

At the beginning of this work, we discuss and analyse the largest corpus of COVID-19 related scientific publications available, CORD-19. Specifically, it includes papers and preprints that were collected from several archives, like PubMed, bioRxiv, and arXiv.org. Their metadata were harmonized and deduplicated, and the documents were processed to extract the full text. The dataset consists of a metadata table with all the papers information, such as title and abstract, a compressed archive with the full-text contents, and an embeddings table with a high dimensional vectorial representation based on the content of each paper (SPECTER).

Exploratory data analysis (EDA) of the metadata. The EDA on the metadata has revealed to us the characteristics and the potential issues we have to take care of while handling this dataset, such as the distribution of duplicates, the missing values, and the availability of abstracts and full texts. From this analysis, we decided to consider only the abstracts of the papers as representative of their content, preferring a high number of different papers to a reduced dataset of full-texts.

Content analysis. With this analysis, we aimed to understand the typology of the content of the dataset and its value. The analysis was performed in two phases: K-means clustering of the embeddings table, and visualization of the most frequent words of each cluster in a word cloud figure, after excluding the words that were common in each cluster. Finally, we discovered that CORD-19 could be partitioned into five topic classes: Biology, Epidemiology, Therapy and Treatment, Society and Environment, and Psychology, which are shown in Figure 1.

How to make CORD-19 information accessible.

We bundled the results of the previous analyses into a `Jupyter` notebook. In this notebook, we

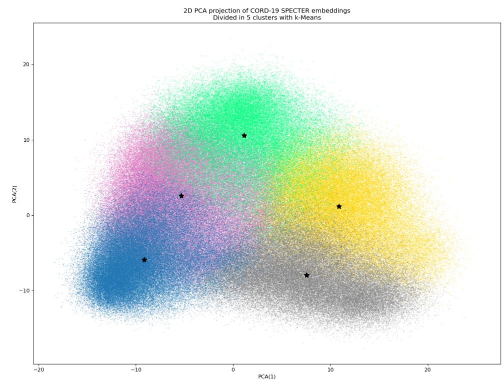

Figure 1: Visualization of CORD-19 clustering partition, projected on the subspace defined by PCA over the SPECTER vectorial embeddings.

also implemented a simple but effective keyword search with the `whoosh` Python library. In this way, we obtained a proof of concept that verified the possibility of searching CORD-19 through keywords and embeddings.

Background and state of the art

Starting from the successful experience with the proof of concept, we framed the problem of exploring and searching the literature of CORD-19 in the information retrieval (IR) framework. In this framework, an IR model is defined as the quadruple of documents, queries, a retrieval strategy, and a ranking function.

Keyword search. This methodology consists of searching and retrieving scientific documents that are relevant to an unstructured textual query, which is formed by keywords, i.e. concepts and expressions of interest. This technique derives from the probabilistic relevance framework, which is based on a probabilistic estimation of the relevance of each document for a given query. In this model, the collection of documents is represented as an index that maps the terms that occur in the text with the documents containing them. In this work, we consider as ranking function, which estimates the relevance, the Okapi's Best Matching 25 (BM25) [5] and its multi-field variant, BM25F.

Similarity-based search. We define this methodology as the retrieval from a corpus of

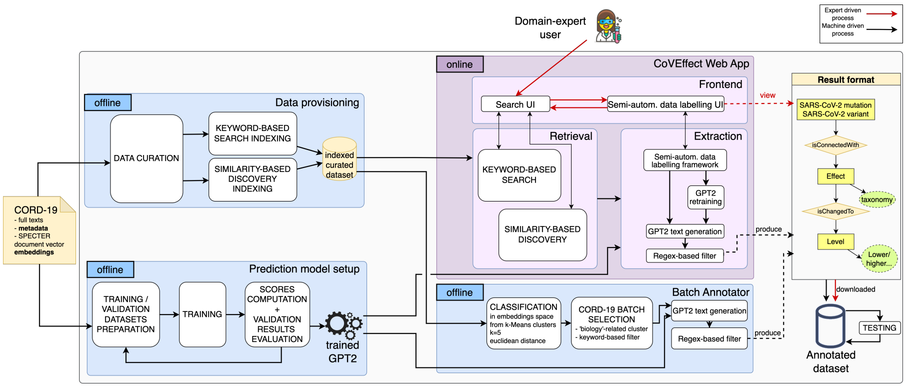

Figure 2: Overview of the CoVEffect application.

scientific publications that are the most similar to one that is chosen by a user, in terms of content. We frame this technique in the vector IR models class, where both queries and documents are represented by vectors, and their similarity is measured as the angle between those vectors through the cosine similarity. Instead of using the basic term-space representation, we adopted the numerical vectorial embeddings representations generated by deep learning models, which are the current state of art in natural language processing. Specifically, we exploited the SPECTER [2] embeddings available for each paper of CORD-19, which are document-level citation-informed representations.

Graph-driven search. With this denomination, we refer to the search and retrieval of scientific literature through the querying of a semantic network, augmented with a medical ontology, built from the information available in CORD-19, with small graphs created by the users as queries. For this work, we considered several precedent experiences of knowledge graphs for COVID-19-related knowledge management, such as the work by Logette *et al.* [4] and we aim to create a user-friendly graph-driven literature search process. In particular, we introduce the *graphical abstract*, a small graph formed by concepts and relations, which the users can use to express research hypotheses and query our semantic network.

3. CoVEffect: Keyword-based and similarity-based extraction of mutations' effects from CORD-19

Since the beginning of the pandemic, a very high variety and volume of data about COVID-19 and SARS-CoV-2 has been made available, therefore, it has been necessary to create a unique abstract model to harmonize them all, as described by Alfonsi *et al.* in [1]. While some of these data was available from sources in structured and semi-structured formats, the mutations' and variants' effects of COVID-19 were only available scattered in the scientific literature. Thus, with the CoVEffect application, we offer a framework for semi-automated labelling and extraction of mutations' effects, in order to enable the training of deep learning predictive models for this task, while overcoming the total lack of specific pre-annotated datasetscommonly necessary to train those models. In this way, we aim to extract quadruples like (type, entity, effect, level), starting from the raw data of CORD-19 and a generic pre-trained language model.

Overview. The framework provided by the CoVEffect application, as shown in Figure 2 consists of two main phases: a retrieval stage, where the users can search and retrieve abstracts of CORD-19's documents, which contain information on COVID-19 mutations' effects; an extraction stage, where the language model provides the users the set of quadruples predicted from the abstracts retrieved in the first phase. After the extraction, the users verify the correctness and completeness of the quadruples, and if the model misses something, they can manually provide the correct set of annotations for that abstract. In addition to the online application, some support processes are provided within CoVEffect: the data provision pipeline that feeds the retrieval stage; the prediction model setup, necessary to bootstrap the prediction model for this extraction task; the batch annotator, to run automatically the predictive model on large portions of CORD-19 and extract mutations' effects in bulk.

Software architecture. CoVEffect combines a data production access layer, a predictor based on machine learning, and a dedicated interface for expert users; my role has been to build the data provision pipeline in the access layer, which enables both the index and similarity-based search as described in Section 2, and the software components that expose the search interfaces. Specifically, we describe the RESTful APIs `/search` and `/similar` that were designed and implemented to expose the two search methodologies, which were implemented following the microservice paradigm with the Flask Python framework. The implementation of the keyword search is based on the `whoosh` Python library for indexed textual retrieval systems. The implementation of the similarity-based search is based on the `annoy` library, to exploit the index available for the approximate nearest neighbours search algorithm.

Semi-automated data labelling framework. To conclude, we focus on the process behind the machine-human hybrid data labelling framework, which allows the users to train complex multi-task predictive models providing only a minimal set of data, based on the weaknesses found in the iterative training process. In this way, the users may reduce the time-consuming effort that is necessary to manually annotate COVID-19 mutations' effects from the scientific literature. In addition, we show how the *saliency maps*, which are an explainable AI visualization technique, can provide powerful and informative hints about what the model is considering for its predictions.

4. AGAVE: Graph-driven extraction of publications from CORD-19 supporting clinical hypotheses

With AGAVE, a Graphical Abstract Verification Engine, we aim to design and show a graph-driven search system for scientific literature, starting from the challenges described by Kejriwal in [3] in terms of usability and user-friendliness for the applications of knowledge graphs in the biomedical domain. In particular, we aim to make graphs as query objects and not just a data structure that resides in the backend.

Graphical abstracts. For this reason, taking inspiration from many successful examples of the biomedical and clinical literature, where graphical abstracts are exploited to visualize complex research hypotheses in an easy and compact way, we propose a formalization of them, which will be exploited to define a query language for AGAVE. As shown in Figure 3, graphical abstracts are small graphs composed of nodes, which represent concepts, and undirected links, which represent the co-occurrence of the two linked concepts. After the analysis of real graphical abstracts, we could assume that these elements are expressive enough to convey the semantic of a research hypothesis.

The semantic network. After describing the query language, we have to characterize the semantic network, which is a unique graph that represents all the information extracted from the abstracts of the publications of CORD-19. The extraction process is implemented in the data provision pipeline and includes techniques from the NLP and the text-mining domains, such as entity recognition, linking the entities to a medical ontology (UMLS), and statistical evaluation of the co-occurrence relations between entities. Thus, the semantic network, as the graphical abstracts, is composed of nodes that represent entities and links that represent co-occurrence. The resulting network consists of a unique connected component of about 124 thousand concepts and 44 million relations, which were mined from 650 thousand publications.

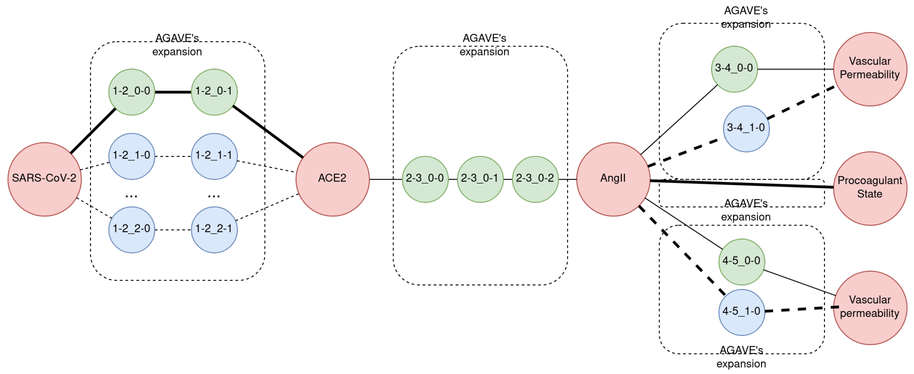

Figure 3: Graphical abstract matching. The nodes provided by the user are in red, while in blue and green there are the paths provided by the system to match the original graphical abstract with the semantic network. This graphical abstract is an excerpt from the one provided by Logette *et al.* in [4].

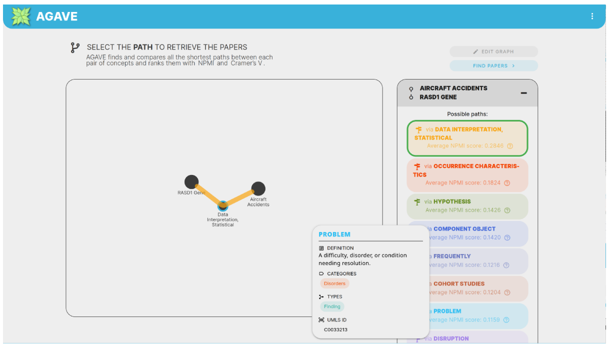

Figure 4: AGAVE web interface at the path selection page. The relation between the two entities of the original graphical abstract has been expanded with another entity in the middle. On the right, there is the list to choose the best expansion.

Graphical abstract matching. A considerable part of this work is dedicated to the description of the retrieval strategy that we designed for AGAVE. Briefly, given a graphical abstract produced by the user, we have to compare it with the semantic network of the system, in order to be able to retrieve the papers that explain the relations of the graphical abstract. We defined this process as *matching*. As shown in Figure 3, the system proposes to the user a set of paths of its semantic network for each relation of the original graphical abstract. These paths, which are highlighted by the dashed areas between the original (red) concepts, are called *expansions*, as the system has to expand the original immediate links with longer paths to match the original nodes with its own topology. The users may choose the one path from each expansion that best suits their needs, or they can trust the ranking computed by the system, which is based on an estimation of the mutual information. Once the graphical abstract is matched with the semantic network, the system retrieves the papers from the database and ranks them in function of the number of relations of the expanded graphical abstract they contain.

Software architecture. The AGAVE application service consists of three software components (i.e. backend, frontend, and data provision pipeline) and two databases (i.e. a Neo4J graph database for the semantic network, a relational database for the indexes and the metadata). The backend exposes the algorithm logic through RESTful interfaces, which are implemented with `connexion`, a Flask-based Python framework for openAPI specifications, and supports concurrent sessions. The frontend, which is shown in Figure 4, follows likewise the process of the retrieval strategy and is implemented with the Vue.js framework and the D3.js library for the graph visualizations. The application is distributed as a set of Docker images and the deployment contemplates the execution of those images as containers.

Evaluation. To evaluate the system, we relied on the opinion of a biomedical expert, who is part of the GeCo team. In this process, we chose several research hypotheses, recreated them using both the keyword search system and AGAVE, and retrieved the results from the systems. Then, we showed the results to the expert, who gave us several comments. For instance, AGAVE resulted very effective in searching wide research hypotheses with many concepts, because of its capability of managing and ranking results with different topics, while the keyword search provided poor results with that kind of hypothesis. In conclusion, AGAVE has proven itself as a powerful tool for knowledge discovery and shown that it is possible to make the user participate in a graph-driven search process in an intuitive manner.

5. Conclusions and Future Work

With the aim of making scientific information related to COVID-19 accessible and the related knowledge discoverable, we implemented three information retrieval strategies and we designed and developed two applications: CoVEffect and AGAVE. The retrieval stage of CoVEffect was successfully able to support the users in the training process of the model. AGAVE turned out to be a sound starting point for the scientific literature search with a graph-centric user experience. As the user from the biomedical field reported in the evaluation process, the system is valid and reflects the intentions for which it was created.

Future Work. CoVEffect may be easily repurposed for other labelling tasks, both from the domain of CORD-19 and others. AGAVE, likewise, can be reapplied to other contexts and datasets, and with the addition of other ontologies in the semantic network.

6. Acknowledgements

I dedicate this space to the advisor and the coadvisor of this work, Prof. Stefano Ceri and Dr. Anna Bernasconi, without whom all this would have been unimaginable. I also thank the GeCo team members for their collaboration on these projects.

References

[1] Tommaso Alfonsi, Ruba Al Khalaf, Stefano Ceri, and Anna Bernasconi. Cov2k model, a comprehensive representation of sars-cov-2 knowledge and data interplay. *Scientific Data*, 9(1):1–12, 2022.

[2] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. Specter: Document-level representation learning using citation-informed transformers. *arXiv preprint arXiv:2004.07180*, 2020.

[3] Mayank Kejriwal. Knowledge graphs and covid-19: opportunities, challenges, and implementation. 2020.

[4] Emmanuelle Logette, Charlotte Lorin, Cyrille Favreau, Eugenia Oshurko, Jay S Coggan, Francesco Casalegno, Mohameth François Sy, Caitlin Monney, Marine Bertschy, Emilie Delattre, et al. A machine-generated view of the role of blood glucose levels in the severity of covid-19. *Frontiers in Public Health*, page 1068, 2021.

[5] Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.

[6] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, et al. Cord-19: The covid-19 open research dataset. *ArXiv*, 2020.