Building an AI-powered Literature Review for COVID-19

Jan Bremer[1], Maikel Boot[2], Lucas Buyon[3], Paul Mooney[4], Tayab Waseem[5]

[1] Department of Neurophysiology and Pathophysiology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany
[2] Department of Microbial Pathogenesis, Yale University, New Haven CT, USA
[3] Department of Immunology and Infectious Diseases, Harvard TH Chan School of Public Health, Boston, MA, USA
[4] Kaggle, Boulder, CO, USA
[5] Wagner Macula & Retina Center, Norfolk, VA, USA

The COVID-19 pandemic has accelerated the pace at which scientific manuscripts are published. Thousands of papers have been published as pre-print articles on open access platforms such as bioRxiv or medRxiv, but existing methods to rapidly summarize the research are lacking. This affects the rate at which scientists can write literature reviews or grants on emerging scientific topics. Here, we report on efforts to combine custom search tools, text extraction algorithms, and expert curation to rapidly summarize the COVID-19 literature.

Methods and organizational details:

A team of over 150 medical and scientific volunteers from global institutes has created live literature reviews of over 88 COVID-19 related research questions. These questions were drawn from the NASEM's SCIED (National Academies of Sciences, Engineering, and Medicine's Standing Committee on Emerging Infectious Diseases and 21st Century Health Threats) research topics, the World Health Organization's R&D Blueprint for COVID-19, the New England Journal of Medicine, and other sources. The relevant publications for each question were summarized in a tabular format. Search tools used for relevant literature reviews included: PubMed, Google, and COVID-19 specific search engines such as covid19-research-explorer.appspot and covidex. Our literature review tables currently cover 6.2% (1128/18103) of published papers published since February 2020.

Project goals:

To evaluate whether custom search tools and text extraction algorithms can speed up literature reviews, we are conducting a utility study in collaboration with several US institutes, across various disciplines. Briefly, a research team is split into two groups, both teams create a literature review, but only one of them gets an article summary table that provides important text extractions for every relevant publication.

In addition, we challenged the Kaggle community of machine learning developers (nearly 5 million users) to create an AI-based literature review tool based on golden standard tables from our dataset. User submissions will be in the form of Python or R notebooks and the submissions will be evaluated for the completeness of the text extractions, the accuracy at which they identify in-scope papers, and the clarity of the documentation of their approach. The code created in this challenge is open source and falls under an Apache 2.0 license, which will serve as the base for further development of this technology.