COVIDScholar: AI-powered rapid data gathering, analysis, and dissemination

John Dagdelen, Amalie Trewartha, Haoyan Huo, Tanjin He, Kevin Cruse, Zheren Wang, Yuxing Fei, Akshay Subramanian, Kristin Persson, Gerbrand Ceder

There is a critical need for tools that can help the COVID-19 research community stay on top of the emerging literature. Results are being generated faster than researchers and medical professionals can read them and the pace of knowledge creation around the virus is only accelerating. As a result, we are almost certainly overlooking critical connections between ideas and observations (both new and old) that could be the key to developing effective vaccines and therapies for COVID-19. Our team at Lawrence Berkeley National Laboratory has started to address this problem by building covidscholar.org, a knowledge portal tailored specifically for COVID-19 research that leverages natural language processing (NLP) techniques to synthesize the information spread across nearly 100,000 emergent research articles, patents, and clinical trials into actionable insights and new knowledge.

Traditionally, recognizing non-typical connections in the literature and integrating knowledge from different fields requires a large time investment from researchers. However, recent advances in natural language processing (NLP) are enabling automated literature analysis that can bring latent knowledge from unstructured text to light. Our previous work in materials science text mining[1] has shown that word embeddings are capable of capturing 'latent' scientific knowledge from text, such as suggesting novel materials that had been overlooked by human researchers. We are now applying similar techniques to scour the COVID-19 literature (and related research) for drug re-purposing candidates, disease-gene relations, and virus-host interactions.

Our knowledge discovery platform, covidscholar.org, is powered by an automated system that scrapes research documents from dozens of sources across the internet, cleans/repairs metadata as necessary, and analyzes the text with a number of NLP models for classification, information extraction, and scientific language modeling. We then integrate this information with specialized knowledge graphs being developed by researchers from the DOE Systems Biology KBase, which has the potential to give users unparalleled insight into the complex interactions that govern the transmission of COVID-19, the disease's progression, and potential therapeutic strategies. This approach to combining textual information, such as word embeddings, with ontological knowledge graphs has the potential to improve the performance of machine learning models that operate on these data structures and to enable new ways of exploring literature on emerging subjects by leveraging past knowledge more efficiently.

[1] Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. Unsupervised word embeddings capture latent knowledge from materials science literature. Nature, 571(7763):95, 2019.