Extracting a knowledge base of biomedical mechanisms from COVID-19 scientific literature

Aida Amini[1], Tom Hope[2], David Wadden[1,2], Roy Schwartz[2], and Hannaneh Hajishirzi[1,2]

[1]University of Washington
[2]Allen Institute for AI
{*amini91, dwadden, hannaneh*}@*cs.washington.edu*
{*tomh, roys*}@*allenai.org*

Abstract

The COVID-19 pandemic has sparked an influx of research by scientists worldwide, leading to a rapidly evolving and highly interdisciplinary corpus of papers. In this emergency scenario, there is a need for automatic information extraction (IE) systems to help scientists stay current on the latest findings and accelerate scientific discovery.

In this work, we describe our ongoing efforts to automatically extract structured representations of biomedical systems from the scientific papers contained in the COVID-19 Open Research Dataset (CORD-19). At the time of this writing the dataset has accumulated over 128K relevant papers, both historical and cutting-edge, to corona-viruses dominating areas of virology, epidemiology and biology.

Most previous work on biomedical information extraction has focused on identifying a set of relations specifically found in that scientific field(e.g., interactions between drugs and proteins). In contrast, we focus on a lightweight and salable approach by identifying general scientific relations, across a variety of fields, from the mechanism by which a virus gains entry into a host cell, to the effects of psychological counseling on the mental health of health care providers.

We extract relations capturing a broad notion of *mechanisms* and *effect* in COVID-19 papers, spanning a diverse range including psychological intervention techniques, computational algorithms, and molecular mechanisms of viral entry into cells. This unified view of natural and artificial mechanisms can help generalize across the COVID-19 corpus and is designed to help scientists study fundamental questions: What things do – and what effects they have.

We collect a set of *mechanisms* and *effect* annotations from domain experts, allowing them to choose any range of text spans from the provided paragraph for marking the relations. We further define a mapping schema for previously introduced scientific datasets to unify the notations. These datasets cover a various scientific domains such as computer science, chemistry and biology.

We construct a dataset for research on this task by combining annotations from two sources: (1) We collect expert annotations of *mechanisms* and *effect* relations found in the CORD-19 corpus, and (2) We merge data from existing scientific IE datasets ranging from computer science to molecular biology, converting from the specific relation types in these data to our general relations. We train an information extraction model on this dataset, and used the model to construct a knowledge graph of mechanisms and effects from the entirety of the CORD-19 corpus. Our curated data, models, and extracted knowledge graph will be made publicly available to facilitate future research efforts.