Extracting a knowledge base of mechanisms from COVID-19 papers

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

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

Abstract

The COVID-19 pandemic has sparked an influx of research by scientists worldwide, leading to a rapidly evolving corpus of interdisciplinary papers. At the time of this writing the COVID-19 Open Research Dataset (CORD-19) has amassed over 128K relevant papers, both historical and cutting-edge. In this emergency scenario, there is a need for automatic information extraction (IE) to provide scientists with structured knowledge, and to accelerate exploration and discovery.

In this work we extract relations capturing a broad notion of *mechanisms* in CORD-19 papers – spanning a range of mechanisms as diverse as psychological intervention techniques, computational algorithms, and molecular mechanisms of viral cell entry. This unified view of natural and artificial mechanisms can help generalize across the CORD-19 corpus and is designed to help scale the study of the many different types of processes, activities and functions described in the dataset.

We collect a set of annotations from domain experts for *direct mechanisms* (operations and functions explicitly described in the text) and *indirect mechanisms* (observed effects and interactions without an explicit description of a direct functional relation). For example, descriptions of the mechanism by which the SARS-CoV-2 virus binds to cells, or of a diagnostic procedure based on computer vision – are considered direct mechanisms. Conversely, descriptions of indirect mechanisms can for example be of observed links between COVID-19 and certain symptoms, with no explicit mention of the functional process leading from the disease to the symptoms. This distinction between direct and indirect relations is inspired by a review of biomedical and scientific ontologies (e.g., direct and indirect regulation of proteins by chemicals).

We allow annotators to select free-form text spans as the arguments in our mechanism relations; this is in contrast to many existing datasets of annotated scientific relations which are often entity-centric (e.g., protein-chemical interactions). We do so in order to capture the complexity and diversity of the many concepts and ideas described in the corpus, in a scalable approach. To address the challenging nature of the annotation task with multiple "correct" annotations of complex and diverse spans, we conduct a multi-round annotation process with final adjudication by a domain expert experienced in bioNLP annotations.

Our annotations are used in combination with existing datasets from different domains to train a relation extraction model, using a mapping schema for previously introduced scientific datasets, selecting only direct and indirect mechanisms (e.g., *DIRECT UP-REGULATION* in the chemprot dataset) and unifying relation labels into our typology using a domain expert. Our results indicate we outperform baselines including openIE and SRL, and also supervised models trained on related science IE datasets in the biomedical and computer science domains. We use a biomedical language model that we fine-tune to capture semantic similarity, build a graph of similar mechanisms and induce concepts by finding cliques. To support search over our KB, we use the same language model for retrieving relations similar to the query. To help boost community efforts we release our curated data and models as well as a large-scale knowledge graph of extracted mechanisms.

*. Equal contribution.