Extracting a Knowledge Base of Mechanisms from COVID-19 Papers

Aida Amini[*1] **Tom Hope**[*1,2] **David Wadden**[1]

Madeleine van Zuylen[2] **Eric Horvitz**[3] **Roy Schwartz**[2,4] **Hannaneh Hajishirzi**[1,2]

∗ *Tom Hope and Aida Amini contributed equally as first authors*

[1] Paul G. Allen School for Computer Science & Engineering, University of Washington

[2] Allen Institute for Artificial Intelligence [3] Microsoft Research [4] The Hebrew University of Jerusalem, Israel

{tomh,roys,hannah}@allenai.org {amini91,dwadden}@cs.washington.edu

Abstract

The urgency of mitigating COVID-19 has spawned a large and diverse body of scientific literature that is challenging for researchers to navigate. This explosion of information has stimulated interest in automated tools to help identify useful knowledge. We have pursued the use of methods for extracting diverse forms of *mechanism relations* from the natural language of scientific papers. We seek to identify concepts in COVID-19 and related literature which represent activities, functions, associations and causal relations, ranging from cellular processes to economic impacts. We formulate a broad, coarse-grained schema targeting mechanism relations between *open, free-form* entities. Our approach strikes a balance between expressivity and breadth that supports generalization across diverse concepts. We curate a dataset[1] of scientific papers annotated according to our novel schema. Using an information extraction model trained on this new corpus, we construct a knowledge base (KB) of 2M mechanism relations, which we make publicly available. Our model is able to extract relations at an F1 at least twice that of baselines such as open IE or related scientific IE systems. We conduct experiments examining the ability of our system to retrieve relevant information on viral mechanisms of action, and on applications of AI to COVID-19 research. In both cases, our system identifies relevant information from our automatically-constructed knowledge base with high precision.

1 Introduction

The global effort to understand the SARS-CoV-2 virus and to mitigate the COVID-19 pandemic is an interdisciplinary endeavor with an intensity the world has rarely seen (Apuzzo and Kirkpatrick 2020). Scientists from many areas, ranging from microbiology to AI, are working to understand the disease, adding to a rapidly expanding body of literature which encompasses both past work on viruses and findings on the novel coronavirus (Wang et al. 2020b). As a recent example, a diverse group of scientists called attention to the airborne transmissibility of the virus based on research spanning virology, aerosol physics, flow dynamics, epidemiology, medicine and building engineering, stating, "expertise in many science and engineering areas enables us to understand the *mechanisms* behind generation of respiratory microdroplets, how viruses survive in microdroplets, and how airflow patterns carry microdroplets in buildings" (Queensland University of Technology 2020). In this paper, our overarching goal is to build a knowledge base (KB) that scientists can use to search and explore **diverse interdisciplinary mechanisms** in literature related to COVID-19.

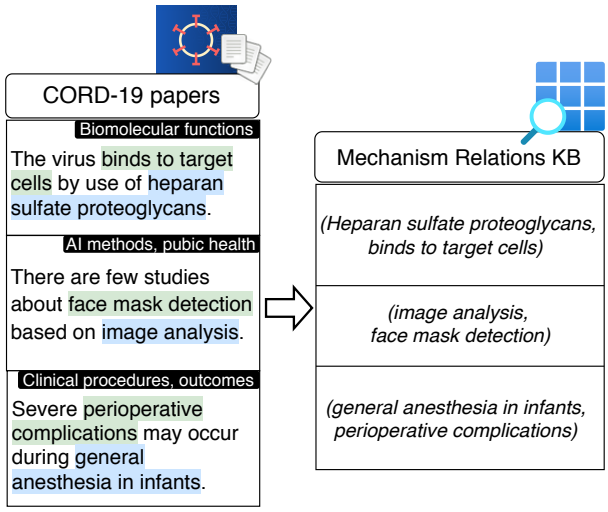

Figure 1: Our knowledge base of mechanism relations spans a wide range of activities, functions, and influences extracted from CORD-19, a corpus of papers related to COVID-19.

Figure 1 shows examples of the types of mechanisms we focus on. These include mentions of mechanistic *activities* (e.g., viral binding), of *functions* that natural or artificial entities serve (e.g., a protein used for binding, or image analysis used in public health), and also more indirect *influences and associations* (such as possible complications associated with a medical procedure). These relationships cover a wide range of domain-specific concepts in scientific papers, providing a **unified language** which can be used for domain-agnostic extraction and scientific search (with results as seen in Figure 1). More broadly, a KB of mechanisms across

∗ *Equal contribution, listed in alphabetical order.*

[1] Data and models are made available at https://git.io/JUhv7.

science could enable the transfer of ideas across disparate areas (Hope et al. 2017; Kittur et al. 2019), and assist in literature-based discovery (Swanson and Smalheiser 1996; Spangler et al. 2014; Nordon et al. 2019) by finding cross-document causal links (Swanson and Smalheiser 1996).

In biomedicine, information extraction (IE) has been used to extract mentions of pinpointed entities such as proteins or chemicals and their relations, including recently from coronavirus-related papers (Ilievski et al. 2020; Ahamed and Samad 2020; Hope et al. 2020). Some of these relations correspond to mechanisms (e.g., chemical-protein regulation, or drug-drug interactions), but capture only a fraction of the full breadth and depth of mechanisms in the literature. It is challenging to formulate comprehensive fine-grained schemas for diverse domains; on the other extreme, Open IE approaches (Etzioni et al. 2008; Stanovsky et al. 2018; Zhan and Zhao 2020) focus on general-purpose, schema-free extraction of relations, but many of the relations are generic and uninformative for scientific applications

In this work, we use *open, free-form* entities with a broad class of relations centered around mechanisms, to strike a balance between expressivity and breadth. Our unified view of mechanisms is designed to help generalize and scale the study of these important relations in the context of the COVID-19 emergency and more broadly. We lay the foundations for the framework, which we hope will open new avenues for boosting knowledge discovery across the sciences.

Our main contributions include:

- We curate and distribute an annotated dataset of mechanisms in COVID-19 papers to help accelerate discovery of functional relations and research in this area. We release a state-of-art IE model (Wadden et al. 2019) trained on our data, and a knowledge base of 2M mechanism relations extracted from the literature. These relations include *direct mechanisms* (mentions of mechanistic activities and functions) and *indirect mechanisms* (influences and associations without explicit mention of the process involved).
- We show 2X improvement in F1 over baselines including a recent scientific IE approach (Luan et al. 2018), Open IE, and semantic role labeling (SRL (Shi and Lin 2019)). These results demonstrate the inability of existing approaches to capture an important class of relations and the utility of our curated dataset.
- In experiments with human evaluators, we reach high precision in two search tasks using our KB: studying mechanisms of the SARS-CoV-2 virus, and exploring applications of AI in the CORD-19 corpus.

2 Background: Mechanisms in Science

The concept of *mechanisms*, also referred to as *functional relations*, is fundamental in biomedical ontologies (Burek et al. 2006; Röhl 2012; Keeling et al. 2019), engineering (Hirtz et al. 2002), and across science. Mechanisms can be natural (e.g., the mechanism by which amylase in saliva breaks down starch into sugar), artificial (electronic devices), non-physical constructs (algorithms, economic policies), and very often a blend (a pacemaker regulating the beating of a heart through electricity and AI algorithms).

In our work we aim to achieve broad coverage of mechanism relations, extending to a wide range of entities and topics observed in COVID-19 papers. For example, in addition to areas such as medicine, microbiology, genetics, proteomics, zoology and virology, topics we cover in our mechanism annotations include computer science, public policies, flow dynamics, building engineering, macroeconomic impacts and international relations. In *Homo Deus: A Brief History of Tomorrow* (Harari 2016), the author writes: "While some experts are familiar with one field, such as AI, nanotechnology, big data or genetics [...] no one is capable of connecting all the dots and seeing [...] how breakthroughs in AI might impact nanotechnology, or vice versa." By building a KB with diverse, domain-agnostic mechanisms, we aim to make progress toward connecting those dots.

Exact definitions of mechanisms are subject to debate in the philosophy of science (Röhl 2012; Keeling et al. 2019). A dictionary definition of mechanisms refers to *a natural or established process by which something takes place or is brought about*. More intricate definitions discuss "complex systems producing a behavior", "entities and activities productive of regular changes", "a structure performing a function in virtue of its parts and operations", or the distinction between "correlative property changes" and "activity determining how a correlative change is achieved" (Röhl 2012). The schema we propose in this work (see Section 3, Figures 1,2) draws inspiration from these existing definitions. We extract activities and functions, and also more general influences and associations. Our work is also related to a large body of literature on extracting information from biomedical papers. This information often corresponds to very *specific* types of mechanisms such as chemical-protein regulation and drug-drug interactions (Li et al. 2016; Segura Bedmar, Martínez, and Herrero Zazo 2013). In the CHEMPROT dataset (Li et al. 2016) for example, texts are annotated for relations capturing interactions between chemicals and proteins (e.g., up/down regulation). The annotation guidelines for CHEMPROT distinguish between *direct* and *indirect* interactions, between relations *explicitly* and *implicitly* referred to in the text, and between texts where "mechanistic information is available" and those where the nature of an interaction is more vague. A semantic predication schema akin to Semantic Role Labeling, with predicates such as `X treats Y` or `X induces Y`, has also been proposed (Kilicoglu et al. 2011). Concepts and relations in that work were also limited to a relatively narrow set of biomedical sub-domains and entities aligned with the UMLS biomedical onotolgy (Bodenreider 2004) such as names of drugs and diseases (see Section 5.1 for more details). Recent work has applied such tools to extract information from the CORD-19 corpus, for constructing KBs (Wise et al. 2020; Wang et al. 2020a) and visualizations (Hope et al. 2020). Unfortunately, biomedical ontologies suffer from cultural differences between disciplines that lead to a lack of a unified language (Wang et al. 2018) and many fragmented classes (Salvadores et al. 2013) – with only a small a fraction at the focus of mainstream biomedical IE. In the next section, we present our schema for unified extraction of mechanisms – in biomedicine, and beyond.

3 Task and Data

3.1 Relation Schema

Our goal is to extract information expressing the important notion of mechanisms. As discussed in Section 2, this seemingly intuitive concept is subject to debate, and an absolute definition is illusive. We opt for a practical approach that is simple enough for annotators and models, inspired by the definitions and schema discussed in Section 2.

Within the concept of mechanisms, we include *activities* (e.g., binding) or explicit mentions of *functions* (e.g., a use for treating), and also *influences* or associations of a more *indirect* nature (such as describing observed effects, without describing the process involved). We further break the concept of mechanisms into relations of the form `( subject, object, class)` with two coarse-grained classes. The first, which we call `direct mechanisms`, includes mechanistic activities, and reference to specific functions. The second, `indirect mechanisms`, includes influences or associations without explicit mechanistic information or mention of a function, and relations that are expressed more *implicitly* in the text.

Indirect mechanisms correspond to texts indicating "input-output correlations" (Röhl 2012), such as indicating that COVID-19 may lead to certain symptoms but not *how*, or mentioning a general association between two proteins. Direct mechanisms describe "inner workings" – revealing more of the intermediate states that lead from initial conditions (COVID-19) to final states (symptoms) (Röhl 2012), or describing explicitly the function served by an entity (whether natural or human-made). This distinction is inspired by the direct and indirect types of relations in the CHEMPROT chemical-protein regulation schema, but covers a much broader set of concepts and domains.

Figure 1 shows some examples, such as SARS-CoV-2 binding to target cells (`direct mechanism`), the use of image analysis for face mask detection (`direct mechanism`), and complications generally associated with a medical procedure (`indirect mechanism`). Our annotation guideline, available in the supplement, shows many more instantiations of these relations.

Finally, to be able to more directly interpret mechanism relations beyond the coarse-grained categorization, we also experimented with granular relations of the form `subject-predicate-object`, where predicates represent a specific type of a mechanism relation explicitly mentioned in the text (e.g., *binds, causes, reduces*; see Figure 2). While more granular, these relations are also less general – as the natural language of scientific papers describing mechanisms often does not conform to this more rigid structure (in Table 1, there are 400 coarse relations that could not be converted to granular form by an annotator). In our experiments, we also train a model that infers the predicates (a list of frequent predicates is available in the supplement).

3.2 Dataset

We construct a dataset (called `COFIE`: COVID-19 Open Functional Information Extraction) by annotating abstracts from the CORD-19 corpus (Wang et al. 2020b), numbering over 200K abstracts and papers relating to past and present coronaviruses and other broadly related literature. We collect 250 abstracts, similar in size to related scientific IE datasets (Luan et al. 2018) which share similar challenges in collecting expert annotations of complex or ambiguous concepts. We use a relatively low-resource, generalizable annotation approach for a rapid response to COVID-19.

Sentences	Granular relations
Viral infections probably **initiate** a large percentage of childhood and adult asthmatic attacks based on a history of preceding ' cold '.	(Viral infections, **initiate,** childhood and adult asthmatic attacks)
PUVA is useful to **treat** human platelet (PTL) concentrates in order to **eliminate** Leishmania spp.	(PUVA, **treat,** human platelet (PTL) concentrates) (PUVA, **eliminate,** Leishmania spp)

Figure 2: Examples of granular relations.

Annotation process To obtain high-quality annotations of CORD-19 abstracts, we follow a three-stage process of (1) annotating entities and relations using biomedical experts, (2) unifying span boundaries with an NLP expert, and (3) verifying annotations with a bio-NLP expert.

In the first stage, 5 annotators with biomedical background annotate all relations reflecting mechanisms as defined in Section 3.1 (the full annotation guidelines can be found in the supplement). Annotators were given examples and had a one-hour training session using Prodigy, a platform with a GUI for rapid annotations (Montani and Honnibal 2018). Entities are annotated only when involved in a relation with another. Following (Luan et al. 2018), annotators perform a greedy annotation preferring longer spans whenever ambiguity occurs as to span boundaries.

We initially observed large variation between annotators and low agreement as measured with strict, exact matching criteria between relations. A deeper look revealed much of the disagreement was due to variations in annotation style rather than meaning. In particular, the largest reason for disagreement was differences in span boundaries, likely stemming from the challenging nature of our task with abstract, soft definitions of relations between free-form spans.

In the second stage of annotation, an NLP expert annotator carried out a round of style unification between annotators to enhance dataset quality. The NLP expert unified entity annotations by adjusting span boundaries while preserving the original meaning. In the last stage, a bio-NLP expert with experience in annotating scientific papers verified the annotations and corrected them as needed. We observe that the bio-NLP expert accepted 81% of the annotations from the second stage without modification, confirming the high quality of the annotated data.

Dataset	Num. Rels	Num. Sents	Avg. span length	Avg. span distance
COFIE	2370	994	4.00	11.40
COFIE-G	1966	867	3.45	11.46

Table 1: Dataset summary. COFIE has coarse relations (*subject, object, class*), class $\in$ {DIRECT, INDIRECT}. COFIE-G has granular relations (*subject, predicate, object*).

3.3 Task Definition

Given an input document $\mathcal{D}$ represented as a sequence of input tokens $\{w_1, \ldots, w_n\}$, the task is to identify all mentions of mechanism relations in $\mathcal{D}$ – including the entities participating in those relations.

Entities In COFIE, we only annotate entity mentions that participate in one of our two relation categories (direct/indirect). These mentions all share a single common entity type. The mention $e = (e_{\text{start}}, e_{\text{end}})$ is represented by the indices of its start and end tokens in $\mathcal{D}$.

Relations A *coarse relation* is represented as a tuple $r_c = (s, o, y)$, where s and o are the subject and object entities. The relation label is given by $y \in \{\text{DIRECT}, \text{INDIRECT}\}$. A *granular relation* is represented as a tuple $r_g = (s, p, o)$. The s and o slots are the same as in coarse relations. p represents a specific type of mechanism relation (which may be direct or indirect). For simplicity, we constrain the predicate p to consist of a single token (usually a verb); p is therefore represented by its token index in $\mathcal{D}$.

3.4 Evaluation metrics

We evaluate our performance in the tasks of entity identification and relation extraction, defined as follows.

Entity identification Given a boolean span matching function $\text{m}(s_1, s_2) = \mathbb{1}(s_1 \text{ matches } s_2)$, a predicted entity mention $\hat{e}$ is correctly *identified* if there exists some gold mention e^* in $\mathcal{D}$ such that $\text{m}(\hat{e}, e^*) = 1$[2].

We experiment with three different span matching functions. The most conservative is m_{exact}, which is true if two spans have the same start and end tokens. Given the heterogeneous nature of the spans present in our dataset, this metric is overly stringent. Therefore, following common practice in work on Open IE (Stanovsky et al. 2018), we also report results using two more lenient matching functions. The similarity function $\text{m}_{\text{rouge}}(s_1, s_2)$ is true if ROUGE-L$(s_1, s_2) > 0.5$ (Lin 2004), and the function $\text{m}_{\text{subset}}(s_1, s_2)$ true if s_1 is contained in s_2 or vice versa.

Relation identification / classification Given a boolean span matching function, a predicted coarse-grained relation $\hat{r} = (\hat{s}, \hat{o}, \hat{y})$ is correctly *identified* if there exists some gold relation $r^* = (s^*, o^*, y^*)$ in $\mathcal{D}$ such that $\text{m}(\hat{s}, s^*) = 1$ and $\text{m}(\hat{o}, o^*) = 1$. It is properly *classified* if, in addition, $\hat{y} = y^*$. Granular relations are evaluated in the same fashion as coarse-grained relations, with the additional requirement that the predicted predicate token $\hat{p}$ must match the gold p^*.

Relation identification measures the model's ability to identify relations of any type - direct or indirect - while relation classification aims to discriminate between direct and indirect types of mechanism mentions in the text.

[2] Since there is only one entity type in COFIE, an entity mention is correctly classified as long as its span is correctly identified.

4 Knowledge Base Construction

Our goal is to build a KB that can be used by scientists to explore relevant literature in a structured manner, supporting *search for relations* with queries that target specific mechanism relations, as well as more flexible, soft exploration tasks. We train the state-of-the-art DYGIE (Wadden et al. 2019) relation extraction model on our COFIE dataset. Using the trained model, we construct a rich knowledge base to help scientists explore mechanism relations extracted from across many scientific papers. Our approach for constructing a KB consists of the following steps.

1. **Extraction.** We apply our model to approximately 160K abstracts in the CORD-19 corpus, extracting over 2M relations – 1.4M direct mechanism relations, and 630K indirect mechanisms. We select a subset of high confidence relations (with softmax score $\geq 90\%$) and perform standard data cleaning, yielding a KB of 1.5M relations.
2. **Semantic similarity.** Our KB consists of diverse, open spans of text - over 900K unique surface forms after standard string normalization (such as removing punctuation, lemmatizing and lowercasing). In order to support search over our KB, we adapt a recent biomedical language model (LM) trained on over 2.5M papers in PubMed (Gururangan et al. 2020) by fine-tuning it on semantic similarity tasks with the approach in (Reimers and Gurevych 2019) – we fine-tune the LM on the semantic textual similarity (STS) and Stanford Natural Language Inference (SNLI) (Bowman et al. 2015) datasets, to capture a broad notion of similarity. We further tune the LM on the BIOSSES (Soğancıoğlu, Öztürk, and Özgür 2017) dataset, with 100 sentence pairs from biomedical papers annotated for similarity. This approach allows us to capture related concepts (such as *cardiac injury* and *cardiovascular morbidity*), as well as simpler surface matches.
3. **Approximate nearest neighbors search.** Finally, to perform search over this KB in an efficient manner, we employ a recent system (Johnson, Douze, and Jégou 2017) designed for fast similarity-based search over vectors (such as our text embeddings). We create an index of embeddings corresponding to the 900K unique surface forms. In our setting, queries consist of terms representing the subject and/or object of a relation. Queries are embedded using the same language model. Relations are retrieved within 2 seconds on a standard CPU-only laptop.

5 Evaluating Extracted Relations

We evaluate the extracted entities and relations on the three sub-tasks introduced in Section 3.4: relation classification, relation identification, and entity identification.

5.1 Setup

Implementation Details We use the DYGIE library[3], with SciBERT (Beltagy, Lo, and Cohan 2019) token embeddings finetuned on our task. We employ minimal random hyperparameter search to select the best-performing model on the development set. Full details and code are in the supplement.

Baselines We compare our method with the baselines below. Some of these baselines involve training a model on an existing dataset. In these cases, we preprocess the dataset by "mapping" all relation types to the `direct mechanism` or `indirect mechanism` relations in `COFIE`. This mapping was performed by a bio-NLP expert annotator.

SemRep. We train DYGIE on the SemRep dataset (Kilicoglu et al. 2011), consisting of 500 sentences from MEDLINE abstracts and annotated for semantic predication. Concepts and relations in this dataset relate to clinical medicine, substance interactions, genetic etiology of disease and pharmacogenomics. Concepts are tied to the UMLS biomedical ontology (Bodenreider 2004) and focused on pinpointed entities as in most biomedical IE resources. Some of the relations correspond to mechanisms (such as X TREATS Y or X CAUSES Y); other relations are even broader, such as PART-OF or IS-A – we do not attempt to capture these categories as they often do not reflect a functional relation.

SciERC We train DYGIE on the SciERC dataset (Luan et al. 2018), consisting of 500 abstracts from computer science papers that are annotated for a set of relations, including for USED-FOR relations between methods and tasks. We naturally map this relation to our `MECHANISM` label and discard other relation types.

SRL Our task consists of functional relations between open, flexible spans. A natural baseline to try is thus Semantic Role Labeleing (SRL). Using a pre-trained SRL model (Shi and Lin 2019), we select relations of the form (`Arg0, verb, Arg1`), and evaluate using our partial metrics applied to `Arg0` and `Arg1` respectively.

SRL-Bio predicates. We adapt the SRL baseline by filtering predicates down to a list of 80 biomedical verbs that are publicly available from a biomedical proposition bank named BioPro (Chou et al. 2006).

SRL-Mechanism Continuing the above baseline, we task a biomedical annotator with mapping each predicate verb to either `DIRECT MECHANISM` or `INDIRECT MECHANISM`, using this mapping as SRL's predictions.

OpenIE Finally, we also experiment with the supervised Open Information Extraction model of (Stanovsky et al. 2018), similar in nature and in motivation to SRL.

5.2 Automatic evaluations

Relation Prediction Table 2 reports the performance of the DYGIE model trained on `COFIE` and of the different baselines. Our model outperforms baselines by a factor of 2X or more, primarily showing the inability of existing frameworks to capture an important class of relations. Table 3 shows a comparison of precision, precision@K and recall of the DYGIE model trained on our data and on SciERC. The model trained on our data achieves 90% P@50, indicating that correct predictions are assigned higher confidence scores. Figure 3 (left) shows precision scores for top K predicted relations, sorted by prediction confidence. The DYGIE model trained on `COFIE` maintains a high precision score ($\geq 70\%$) within top-20% predictions. As discussed in Section 4, we construct a KB by filtering for high-confidence relations, thus having high P@K is important.

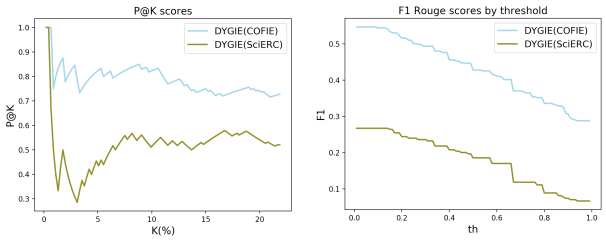

Figure 3: (Left) Precision@K of our model compared the pre-trained SciERC baseline. P@K for our model is high in absolute numbers. (Right) F1 as function of the Rouge-L span matching threshold. Our default threshold is 0.5.

In Figure 3 (right) we show the relation identification $F1$ score for different thresholds of the Rouge-L matching metric. Our default threshold is 0.5. We conduct this analysis primarily to make sure results are reasonably robust in a local neighborhood around 0.5. As expected, we observe a steady decline in F1 as the threshold increases, however the curve declines moderately, and even with the most conservative threshold of 1.0 (i.e. an exact match) F1 is substantially higher than our best performing baseline (SciERC).

Granular relation prediction In addition to coarse-grained relation prediction, we also train a model on `COFIE-G` and measure the prediction quality. Our evaluation shows that the model trained to predict `(s, predicate, o)` triples achieves $F1$ scores of 43.2 and 24.4 using the substring and exact match metrics, respectively. When predicting relations without trigger labels (i.e., `(s, o)`), the model achieves $F1$ scores of 51.7 and 27.6 on the same two metrics.

5.3 Human evaluation of predicted relations

To complement the automatic evaluation metrics in section 2, we conduct an additional human evaluation to measure the quality of the predictions in our KB – do automated metrics capture the true quality of predictions, or are they an under/overestimation as measured by human judgments?

We employ two annotators with biomedical and computer science background, and show them predicted relations for sentences selected randomly from our test data so that we can compare to our automated metrics over ground truth annotation. In particular, we show each annotator 200 relations and the sentences from which they were extracted: 100 predicted by our approach (DYGIE trained on `COFIE`), and 100 using the pre-trained semantic role labeling (SRL) baseline in Section 2. Annotators are asked to evaluate relations in

[3] https://github.com/dwadden/dygiepp

	Relation Classification			Relation Identification			Entity Identification		
Model	Substr	Rouge	Exact	Substr	Rouge	Exact	Substr	Rouge	Exact
OpenIE	-	-	-	24.2	15.5	0.6	71.4	25.6	7.8
SRL	-	-	-	32.7	24.5	1.0	72.5	27.7	6.9
SRL-Bio predicates	-	-	-	13.1	7.5	0.4	70.5	25.6	6.8
SRL-Mechanism	10.4	4.7	0.4	13.1	7.5	0.4	70.5	25.6	6.8
DYGIE(SemRep)	11.4	6.8	3.0	17.2	8.3	3.3	68.1	32.5	22.1
DYGIE(SciERC)	24.9	18.6	6.7	27.9	20.4	7.4	73.0	39.2	23.4
DYGIE(COFIE)	**49.6**	**42.8**	**28.8**	**53.4**	**45.6**	**30.0**	**82.4**	**50.2**	**38.8**

Table 2: F1 scores of partial and exact matching metrics. Relations from SRL and OpenIE do not map directly to DIRECT MECHANISM and INDIRECT MECHANISM classes, and do not have relation classification scores. We also explore mapping SRL predicates to these two classes.

		Relation Classification				Relation Identification				Entity Identification		
Model	metric	P	P@50	R	F1	P	P@50	R	F1	P	R	F1
DYGIE(COFIE)	substr	59.7	90.0	42.4	49.6	63.9	94.0	45.8	53.4	95.7	72.4	82.4
DYGIE(COFIE)	rouge	49.9	80.0	37.5	42.8	53.1	84.0	40.0	45.6	64.1	41.3	50.2
DYGIE(COFIE)	exact	31.3	52.0	26.6	28.8	32.6	56.0	27.8	30.0	48.4	32.4	38.8
DYGIE(SciERC)	substr	70.4	70.0	15.1	24.9	78.6	84.0	16.9	27.9	95.5	59.1	73.0
DYGIE(SciERC)	rouge	52.0	54.0	11.3	18.6	57.1	64.0	12.4	20.4	58.0	29.6	39.2
DYGIE(SciERC)	exact	18.4	26.0	4.1	6.7	20.4	30.0	4.5	7.4	34.7	17.7	23.4

Table 3: Precision, recall and P@K of DYGIE(COFIE) in comparison to the DYGIE(SciERC) baseline.

similar fashion to our annotation guidelines, tagging relations as positive if they reflect a mechanism described in the text, and if they consist of coherent argument spans that capture essential information for the relation to hold, but not redundant or irrelevant information.

Table 4 shows a major increase in relation accuracy, as compared to results obtained with automated metrics. In particular we reach average positive rating of over 91%, a high figure in absolute terms, and more than double the rating of SRL (41.7%). Inter-rater agreement was high at 71 by Cohen's Kappa score and 73 by Matthew Correlation Coefficient (MCC) score. Interestingly, we observe that while in absolute numbers the gap between our human-evaluated accuracy and the partial metrics is high, there is strong correlation between them in the overall trend. In particular, accuracy as measured using the rouge-L score to match relations increases more than two-fold for our model in comparison to SRL (from 22.4 to 55.2). A similar trend is seen for the substring-inclusion measure (from 28.7 to 68.6).

We conclude via human judgments that our predicted relations are of overall sufficiently high quality, and that our automated metrics correlate with human judgments.

6 Knowledge Base evaluations

We show how our system can be used to *search for mechanism relations* across a KB of 2M functional relations, and evaluate its utility in two search applications: Studying the viral mechanisms of SARS-CoV-2, and discovering medical applications of AI in the literature.

Targeted biomedical search. This task involves searching for SARS-CoV-2 mechanism relations focused on a specific well-known statement or question regarding the virus (e.g., *SARS-CoV-2 binds ACE2 receptor to gain entry into cells*).

model	Avg. Human	Rouge	Substr
DYGIE(COFIE)	91.1	55.2	68.6
SRL	41.7	22.4	28.7

Table 4: Human evaluation stats for our predictions vs. baseline SRL. We note that human evaluation scores are considerably higher than captured by our automated metrics wrt ground truth annotations, yet still correlated with them - indicating their usefulness in this challenging setting.

In this scenario, we issue queries which specify both the subject and the object of a mechanism relation (e.g., for a given relation (s_1 = SARS-CoV-2, s_2 = binds ACE2 receptor), retrieve relations where s_1^* is relevant/similar to "SARS-CoV-2" and s_2^* relevant to "binds ACE2 receptor"). This task is designed to test our framework's ability to support researchers looking to quickly generate a list of relations pertaining to a specific hypothesis.

Open-ended cross-domain search. This task is focused on discovering diverse ways in which AI research areas or methods are applied in the CORD-19 corpus. Unlike the previous scenario, here evaluators are given queries where only the subject of the relation is specified, s_1 – with queries consisting of popular, leading subfields and methods within AI (e.g., deep reinforcement learning or text analysis). The aim of this task is to evaluate whether we can support exploratory search over relations, potentially surfacing inspirations for new applications of AI against COVID-19, or helping biomedical researchers and practitioners discover where AI methods are being used.

Task Setup and Evaluation In both tasks, our goal is to see if we can retrieve *relevant relations* that expert annotators consider useful and correct. To evaluate these tasks, we re-

Statement/hypothesis/area	Query over relations (s_1,s_2)	Example results
Cardiac injury is common in critical cases of COVID-19. *INDIRECT MECHANISM: Association/correlation*	s_1 ~ 'COVID-19', s_2 ~ 'Cardiac injury'	...a healthy young gentleman with **COVID-19 pneumonia**, who developed **acute ST - segment elevation myocardial infarction**...
SARS-CoV-2 binds ACE2 receptor to gain entry into cells. *DIRECT MECHANISM: Binding function*	s_1 ~ 'SARS-CoV-2', s_2 ~ 'binds ACE2 receptor to gain entry into cells'	...the **spike glycoprotein of SARS-CoV-2** is phylogenetically close to bat coronavirus and strongly **binds with ACE2 receptor protein**...
Open ended: Machine learning uses in the corpus.	s_1 ~ 'Machine learning'	...proposed to **predict tissue outcome in acute stroke patients** using **machine learning methods incorporating multiparametric imaging data** .
Open ended: Reinforcement Learning uses.	s_1 ~ 'Reinforcement Learning'	...shows that **deep reinforcement learning** can be used to learn **mitigation policies in complex epidemiological models**...

Table 5: Queries and example results retrieved (sentences cut to fit). Subject/object (s_1/s_2) of extracted relations appear in bold.

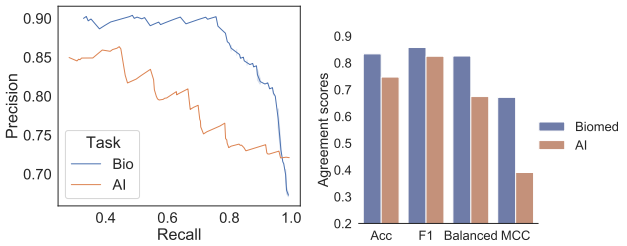

Figure 4: (Left) Precision vs. recall for the search tasks (viral mechanisms, AI methods). Retrieved relations are ranked by embedding-based similarity to a query and compared to gold labels for evaluation. (Right) Average pairwise annotator agreement by several metrics. In the AI task human labels were more diverse but with overall high precision / recall.

cruit 7 annotators with graduate-level education in computer science (AI), medicine, biology and material science.

Table 5 shows examples from the two tasks. In the search focusing on viral mechanisms, 10 claims written by a medical student regarding COVID-19 were taken from a collection of statements prepared in recent automated scientific claim-verification work (Wadden et al. 2020). For example, in Table 5, one such statement regards an association (indirect mechanism) between cardiac injury and COVID-19. We formulate a query for indirect mechanism relations, shown in the second column of the table. In the second task focusing on exploring AI applications, we select a representative list of top methods and areas within AI. Task descriptions, queries and instructions are available in the supplement.

For both tasks, given a query we retrieve the top 1000 most similar relations from our KB, requiring the cosine distance between the embeddings of each of $s1$ (subject) and $s2$ (object) and the query to be at least 0.75. For queries consisting of both s_1 and s_2 terms, we compute the average distances to their respective query terms, and then select the top and bottom 10 relations (20 per query, 200 per task, and 400 relations in total), shuffle their order, and present to annotators together with the original sentence from which each relation was extracted. We ask evaluators to rate whether the retrieved relations are relevant, as judged by the system's ability to identify $(s1, s2)$ relations such that (1) they are relevant to the query, and (2) the sentence in which $(s1, s2)$ are mentioned expresses a mechanism relation between the two terms, rather than incidentally mentioning them together. This is a challenging task, evaluating both *retrieval* of relations that are semantically similar to the query, and also accurate *extraction* of relations from sentences. In total, we collect 1700 relevance labels across both tasks.

Results Figure 4 (left) shows our results for both tasks. We rank results by their similarity to the query as described, and measure precision and recall. We measure average pairwise annotator agreement with several metrics: accuracy (proportion of matching labels), F1 (taking into account precision and recall symmetrically), balanced accuracy (downweighting the positive ratings to counter their higher proportion), and the Matthew Correlation Coefficient (MCC) score.

In the viral mechanisms task, we achieve high precision of 90% that remains stable for recall values as high as 70%. This reflects our experiment in which annotators viewed relations constrained to be highly or moderately similar to the query, and our ability to retrieve relevant relations. Agreement is relatively high by all metrics. In the AI applications task, our model achieves a precision of 85% at a recall of 40%, but drops more quickly. This is likely due to the more exploratory nature of the task, and the use of concepts from biomedicine and computer science with jargon subtleties not all annotators could precisely understand (e.g., network models vs. *neural* networks). Despite this challenge, overall agreement was high or moderate in the AI task too.

7 Conclusion

In this paper we extract a knowledge base (KB) of mechanism and effect relations from papers relating to COVID-19. Our KB can help scientists search and explore relations spanning viral mechanisms of action, diagnostic algorithms, disease symptoms and many more. We release a dataset an-

notated for mechanism relations, and an IE model trained on our data with an improvement in F1 of at least 2X over baselines. We demonstrate our system's use in searching for biological mechanisms employed by the SARS-CoV-2 virus, and for applications of AI methods in this domain. Our unified view of mechanisms can help generalize and scale the study of the virus and related areas of relevance in the fight against COVID-19. We hope our framework can support research on COVID-19, and boost scientific knowledge discovery more broadly.

References

[Ahamed and Samad 2020] Ahamed, S., and Samad, M. 2020. Information mining for covid-19 research from a large volume of scientific literature. *arXiv preprint arXiv:2004.02085*.

[Apuzzo and Kirkpatrick 2020] Apuzzo, M., and Kirkpatrick, D. D. 2020. Covid-19 changed how the world does science, together. https://www.nytimes.com/2020/04/01/world/europe/coronavirus-science-research-cooperation.html.

[Beltagy, Lo, and Cohan 2019] Beltagy, I.; Lo, K.; and Cohan, A. 2019. Scibert: A pretrained language model for scientific text. In *EMNLP*. Association for Computational Linguistics.

[Bodenreider 2004] Bodenreider, O. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*.

[Bowman et al. 2015] Bowman, S. R.; Angeli, G.; Potts, C.; and Manning, C. D. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

[Burek et al. 2006] Burek, P.; Hoehndorf, R.; Loebe, F.; Visagie, J.; Herre, H.; and Kelso, J. 2006. A top-level ontology of functions and its application in the open biomedical ontologies. *Bioinformatics*.

[Chou et al. 2006] Chou, W.-C.; Tsai, R. T.-H.; Su, Y.-S.; Ku, W.; Sung, T.-Y.; and Hsu, W.-L. 2006. A semi-automatic method for annotating a biomedical proposition bank. In *Proceedings of the workshop on frontiers in linguistically annotated corpora 2006*, 5–12.

[Etzioni et al. 2008] Etzioni, O.; Banko, M.; Soderland, S.; and Weld, D. S. 2008. Open information extraction from the web. *Communications of the ACM* 51(12):68–74.

[Gururangan et al. 2020] Gururangan, S.; Marasović, A.; Swayamdipta, S.; Lo, K.; Beltagy, I.; Downey, D.; and Smith, N. A. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *ArXiv* abs/2004.10964.

[Harari 2016] Harari, Y. N. 2016. *Homo Deus: A brief history of tomorrow*. Random House.

[Hirtz et al. 2002] Hirtz, J.; Stone, R. B.; McAdams, D. A.; Szykman, S.; and Wood, K. L. 2002. A functional basis for engineering design: reconciling and evolving previous efforts. *Research in engineering Design* 13(2):65–82.

[Hope et al. 2017] Hope, T.; Chan, J.; Kittur, A.; and Shahaf, D. 2017. Accelerating innovation through analogy mining. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[Hope et al. 2020] Hope, T.; Portenoy, J.; Vasan, K.; Borchardt, J.; Horvitz, E.; Weld, D. S.; Hearst, M. A.; and West, J. 2020. Scisight: Combining faceted navigation and research group detection for covid-19 exploratory scientific search. *EMNLP*.

[Ilievski et al. 2020] Ilievski, F.; Garijo, D.; Chalupsky, H.; Divvala, N. T.; Yao, Y.; Rogers, C.; Li, R.; Liu, J.; Singh, A.; Schwabe, D.; et al. 2020. Kgtk: A toolkit for large knowledge graph manipulation and analysis. *arXiv preprint arXiv:2006.00088*.

[Johnson, Douze, and Jégou 2017] Johnson, J.; Douze, M.; and Jégou, H. 2017. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*.

[Keeling et al. 2019] Keeling, D. M.; Garza, P.; Nartey, C. M.; and Carvunis, A.-R. 2019. Philosophy of biology: The meanings of'function'in biology and the problematic case of de novo gene emergence. *Elife* 8:e47014.

[Kilicoglu et al. 2011] Kilicoglu, H.; Rosemblat, G.; Fiszman, M.; and Rindflesch, T. C. 2011. Constructing a semantic predication gold standard from the biomedical literature. *BMC bioinformatics* 12(1):1–17.

[Kittur et al. 2019] Kittur, A.; Yu, L.; Hope, T.; Chan, J.; Lifshitz-Assaf, H.; Gilon, K.; Ng, F.; Kraut, R. E.; and Shahaf, D. 2019. Scaling up analogical innovation with crowds and ai. *Proceedings of the National Academy of Sciences*.

[Li et al. 2016] Li, J.; Sun, Y.; Johnson, R. J.; Sciaky, D.; Wei, C.-H.; Leaman, R.; Davis, A. P.; Mattingly, C. J.; Wiegers, T. C.; and Lu, Z. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database* 2016.

[Lin 2004] Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

[Luan et al. 2018] Luan, Y.; He, L.; Ostendorf, M.; and Hajishirzi, H. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. *arXiv preprint arXiv:1808.09602*.

[Montani and Honnibal 2018] Montani, I., and Honnibal, M. 2018. Prodigy: A new annotation tool for radically efficient machine teaching. *Artificial Intelligence* to appear.

[Nordon et al. 2019] Nordon, G.; Koren, G.; Shalev, V.; Horvitz, E.; and Radinsky, K. 2019. Separating wheat from chaff: Joining biomedical knowledge and patient data for repurposing medications. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 9565–9572.

[Queensland University of Technology 2020] Queensland University of Technology, v. m. 2020. Researchers: Covid-19 spreads ten meters or more by breathing. https://medicalxpress.com/news/2020-07-covid-ten-meters.html.

[Reimers and Gurevych 2019] Reimers, N., and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and*

the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 3973–3983.

[Röhl 2012] Röhl, J. 2012. Mechanisms in biomedical ontology. In *Journal of Biomedical Semantics*, volume 3, 1–14. BioMed Central.

[Salvadores et al. 2013] Salvadores, M.; Alexander, P. R.; Musen, M. A.; and Noy, N. F. 2013. Bioportal as a dataset of linked biomedical ontologies and terminologies in rdf. *Semantic web* 4(3):277–284.

[Segura Bedmar, Martínez, and Herrero Zazo 2013] Segura Bedmar, I.; Martínez, P.; and Herrero Zazo, M. 2013. Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). Association for Computational Linguistics.

[Shi and Lin 2019] Shi, P., and Lin, J. 2019. Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*.

[Soğancıoğlu, Öztürk, and Özgür 2017] Soğancıoğlu, G.; Öztürk, H.; and Özgür, A. 2017. Biosses: a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics* 33(14):i49–i58.

[Spangler et al. 2014] Spangler, S.; Wilkins, A. D.; Bachman, B. J.; Nagarajan, M.; Dayaram, T.; Haas, P.; Regenbogen, S.; Pickering, C. R.; Comer, A.; Myers, J. N.; et al. 2014. Automated hypothesis generation based on mining scientific literature. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1877–1886.

[Stanovsky et al. 2018] Stanovsky, G.; Michael, J.; Zettlemoyer, L.; and Dagan, I. 2018. Supervised open information extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 885–895.

[Swanson and Smalheiser 1996] Swanson, D., and Smalheiser, N. 1996. Undiscovered public knowledge: A ten-year update. In *KDD*.

[Wadden et al. 2019] Wadden, D.; Wennberg, U.; Luan, Y.; and Hajishirzi, H. 2019. Entity, relation, and event extraction with contextualized span representations. *arXiv preprint arXiv:1909.03546*.

[Wadden et al. 2020] Wadden, D.; Lo, K.; Wang, L. L.; Lin, S.; van Zuylen, M.; Cohan, A.; and Hajishirzi, H. 2020. Fact or fiction: Verifying scientific claims. *arXiv preprint arXiv:2004.14974*.

[Wang et al. 2018] Wang, L. L.; Bhagavatula, C.; Neumann, M.; Lo, K.; Wilhelm, C.; and Ammar, W. 2018. Ontology alignment in the biomedical domain using entity definitions and context. *arXiv preprint arXiv:1806.07976*.

[Wang et al. 2020a] Wang, J.; Pham, H. A.; Manion, F.; Rouhizadeh, M.; and Zhang, Y. 2020a. Covid-19 signsym: A fast adaptation of general clinical nlp tools to identify and normalize covid-19 signs and symptoms to omop common data model. *arXiv preprint arXiv:2007.10286*.

[Wang et al. 2020b] Wang, L. L.; Lo, K.; Chandrasekhar, Y.; Reas, R.; Yang, J.; Eide, D.; Funk, K.; Kinney, R.; Liu, Z.; Merrill, W.; et al. 2020b. Cord-19: The covid-19 open research dataset. *arXiv preprint arXiv:2004.10706*.

[Wise et al. 2020] Wise, C.; Ioannidis, V. N.; Calvo, M. R.; Song, X.; Price, G.; Kulkarni, N.; Brand, R.; Bhatia, P.; and Karypis, G. 2020. Covid-19 knowledge graph: Accelerating information retrieval and discovery for scientific literature. *arXiv preprint arXiv:2007.12731*.

[Zhan and Zhao 2020] Zhan, J., and Zhao, H. 2020. Span model for open information extraction on accurate corpus. In *AAAI*.