COVID-19 Literature Knowledge Graph Construction and Drug Repurposing Report Generation

Qingyun Wang[1], Manling Li[1], Xuan Wang[1], Nikolaus Parulian[1], Guangxing Han[2], Jiawei Ma[2], Jingxuan Tu[3], Ying Lin[1], Haoran Zhang[1], Weili Liu[1], Aabhas Chauhan[1], Yingjun Guan[1], Bangzheng Li[1], Ruisong Li[1], Xiangchen Song[1], Yi R. Fung[1], Heng Ji[1], Jiawei Han[1], Shih-Fu Chang[2], James Pustejovsky[3], Jasmine Rah[4], David Liem[5], Ahmed Elsayed[6], Martha Palmer[6], Clare Voss[7], Cynthia Schneider[8], Boyan Onyshkevych[9]

[1]University of Illinois at Urbana-Champaign [2]Columbia University [3]Brandeis University
[4]University of Washington [5]University of California, Los Angeles [6]Colorado University
[7]Army Research Lab [8]QS2 [9]Department of Defense

hengji@illinois.edu, hanj@illinois.edu, sc250@columbia.edu

Abstract

To combat COVID-19, both clinicians and scientists need to digest vast amounts of relevant biomedical knowledge in scientific literature to understand the disease mechanism and related biological functions. We have developed a novel and comprehensive knowledge discovery framework, **COVID-KG** to extract fine-grained multimedia knowledge elements (entities and their visual chemical structures, relations and events) from scientific literature. We then exploit the constructed multimedia knowledge graphs (KGs) for question answering and report generation, using drug repurposing as a case study. Our framework also provides detailed contextual sentences, subfigures, and knowledge subgraphs as evidence. All of the data, KGs, reports[1], resources, and shared services are publicly available[2].

1 Introduction

Practical progress at combating COVID-19 relies heavily on effective search, discovery, assessment, and extension of scientific research results. However, clinicians and scientists are facing two unique barriers in digesting these research papers.

The first challenge is *quantity*. Such a bottleneck in knowledge access is exacerbated during a pandemic when increased investment in relevant research leads to even faster growth of literature than usual. For example, as of April 28, 2020, at PubMed[3] there were 19,443 papers related to coronavirus; as of June 13, 2020, there were 140K+ related papers, *nearly 2.7K new papers per day* (see Figure 1). The resulting knowledge bottleneck contributes to significant delays in the development of vaccines and drugs for COVID-19. More intelligent knowledge discovery technologies need to be developed to enable researchers to more quickly and accurately access and digest relevant knowledge from the literature.

The second challenge is *quality*. Many research results about coronavirus from different research labs and sources are redundant, complementary, or even conflicting with each other, while some false information has been promoted in both formal publication venues as well as social media platforms such as Twitter. As a result, some of the public policy responses to the virus, and public perception of it, have been based on misleading, and at times erroneous claims. The relative isolation of these knowledge resources makes it hard, if not impossible, for researchers to connect the dots that exist in separate resources to gain new insights.

Let us consider drug repurposing as a case study.[4] Besides the long process of clinical trials and biomedical experiments, another major cause of the lengthy discovery phase is the complexity of the problem involved and the difficulty in drug discovery in general. The current clinical trials for drug repurposing rely mainly on reported symptoms in considering drugs that can treat diseases with similar symptoms. However, there are too many drug candidates and too much misinformation published in multiple sources. The clinicians and scientists thus urgently need assistance in obtaining a reliable ranked list of drugs with detailed evidence, and also in gaining new insights into the underlying molecular cellular mechanisms on COVID-19 and the pre-existing conditions that may affect the mortality and severity of this disease.

To tackle these two challenges we propose a new

[1]Demo video: http://159.89.180.81/demo/covid/Covid-KG_DemoVideo.mp4

[2]Project website: http://blender.cs.illinois.edu/covid19/

[3]https://www.ncbi.nlm.nih.gov/pubmed/

[4]This is a *pre-clinical phase of biomedical research* to discover new uses of existing, approved drugs that have already been tested in humans and so detailed information is available on their pharmacology, formulation, and potential toxicity.

framework, **COVID-KG**, to accelerate scientific discovery and build a bridge between the research scientists making use of our framework and clinicians who will ultimately conduct the tests, as illustrated in Figure 2. **COVID-KG** starts by reading existing papers to build multimedia knowledge graphs (KGs), in which nodes are entities/concepts and edges represent relations and events involving these entities, as extracted from both text and images. Given the KGs enriched with path ranking and evidence mining, **COVID-KG** answers natural language questions effectively. With drug repurposing as a case study, we focus on 11 typical questions that human experts pose and integrate our techniques to generate a comprehensive report for each candidate drug.

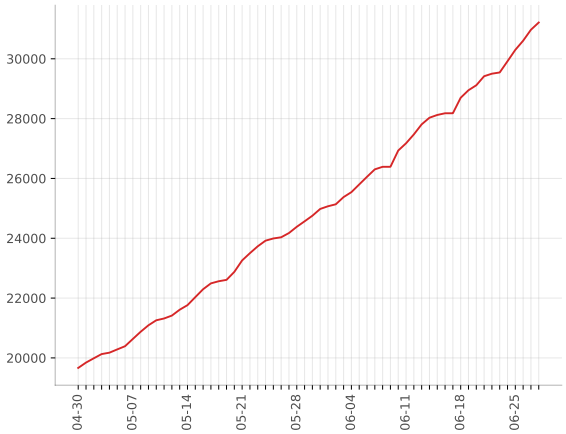

Figure 1: Increasing numbers of COVID-19 papers over time in PubMed website

2 Multimedia Knowledge Graph Construction

2.1 Coarse-grained Text Knowledge Extraction

Our coarse-grained Information Extraction (IE) system consists of three components: (1) coarse-grained entity extraction (Wang et al., 2019a) and entity linking (Zheng et al., 2015) for four entity types: *Gene nodes, Disease nodes, Chemical nodes, and Organism.* We follow the entity ontology defined in the Comparative Toxicogenomics Database (CTD) (Davis et al., 2016), and obtain a Medical Subject Headings (MeSH) Unique ID for each mention. (2) Based on the MeSH Unique IDs, we further link all entities to the CTD and extract 133 subtypes of relations such as *Gene–Chemical–Interaction Relationships, Chemical–Disease Associations, Gene–Disease Associations, Chemical–GO Enrichment Associations and Chemical–Pathway Enrichment Associations.* (3) Event extraction (Li et al., 2019): we extract 13 Event types and the roles of entities involved in these events as defined in (Nédellec et al., 2013), including *Gene expression, Transcription, Localization, Protein catabolism, Binding, Protein modification, Phosphorylation, Ubiquitination, Acetylation, Deacetylation, Regulation, Positive regulation, and Negative regulation.* Figure 3 shows an example of the constructed KG from multiple papers. Experiments on 186 documents with 12,916 sentences manually annotated by domain experts show that our method achieves 83.6% F-score on node extraction and 78.1% F-score on link extraction.

2.2 Fine-grained Text Entity Extraction

However, questions from experts often involve fine-grained knowledge elements, such as *"Which **animo acids** in glycoprotein are most related to Glycan (CHEMICAL)?"*. To answer these questions, we apply our fine-grained entity extraction system CORD-NER (Wang et al., 2020c) to extract 75 types of entities to enrich the KG, including many COVID-19 specific new entity types (e.g., *coronaviruses, viral proteins, evolution, materials, substrates, and immune responses*). CORD-NER relies on distantly- and weakly-supervised methods (Wang et al., 2019b; Shang et al., 2018), with no need for expensive human annotation. Its entity annotation quality surpasses SciSpacy (up to 93.95% F-score, over 10% higher on the F1 score based on a sample set of documents), a fully supervised BioNER tool. See Figure 4 for results on part of a COVID-19 paper (Zhang et al., 2020).

2.3 Image Processing and Cross-media Entity Grounding

Figures in biomedical papers may contain different types of visual information, for example, displaying molecular structures, microscopic images, dosage response curves, relational diagrams, and other unique visual content. We have developed a visual IE subsystem to extract the visual information from figures to enrich the KG. We start by designing a pipeline and automatic tools shown in Figure 5 to extract figures from papers in the CORD-19 dataset and segment figures into nearly half a million isolated subfigures. In the end, we perform cross-modal entity grounding, i.e., associating visual objects identified in these subfigures with entities mentioned in their captions or refer-

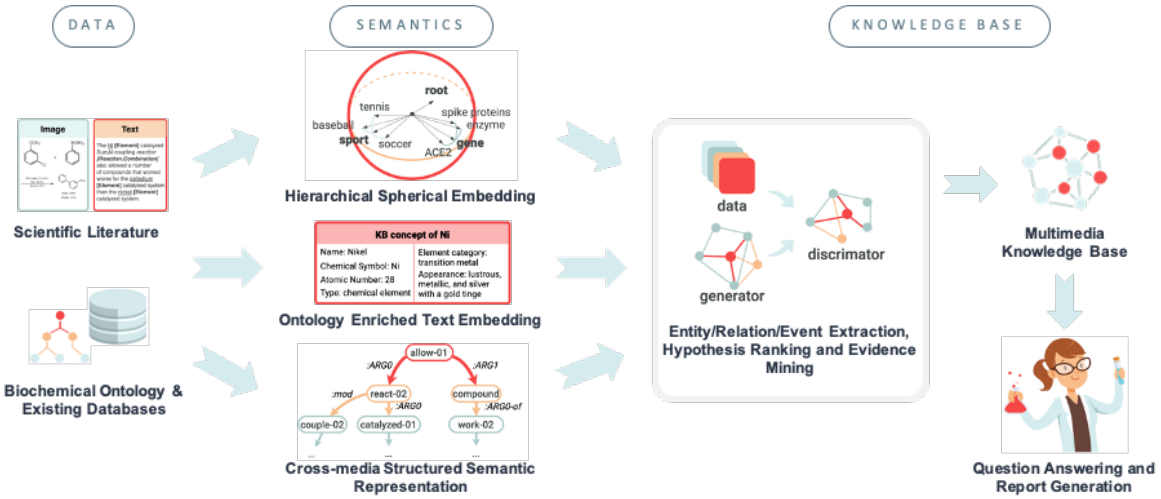

Figure 2: COVID-KG Overview: From Data to Semantics to Knowledge

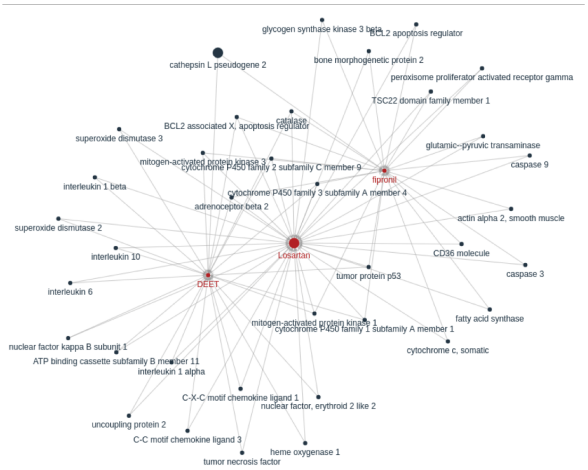

Figure 3: Constructed KG Connecting Losartan (candidate drug in COVID-19) and cathepsin L pseudogene 2 (gene related to coronavirus), where red nodes represent chemicals, grey nodes represents genes, and edges represents gene-chemical relations.

Angiotensin-converting enzyme 2 **GENE_OR_GENOME** (ACE2 **GENE_OR_GENOME**) as a SARS-CoV-2 **CORONAVIRUS** receptor: molecular mechanisms and potential therapeutic target. SARS-CoV-2 **CORONAVIRUS** has been sequenced [3]. A phylogenetic **EVOLUTION** analysis [3, 4] found a bat **WILDLIFE** origin for the SARS-CoV-2 **CORONAVIRUS**. There is a diversity of possible intermediate hosts for SARS-CoV-2 **CORONAVIRUS**, including pangolins **WILDLIFE**, but not mice **EUKARYOTE** and rats **EUKARYOTE** [5]. There are many similarities of SARS-CoV-2 **CORONAVIRUS** with the original SARS-CoV **CORONAVIRUS**. Using computer modeling, Xu *et al.* [6] found that the spike proteins **GENE_OR_GENOME** of SARS-CoV-2 **CORONAVIRUS** and SARS-CoV **CORONAVIRUS** have almost identical 3-D structures in the receptor binding domain that maintains Van der Waals forces **PHYSICAL_SCIENCE**. SARS-CoV spike proteins **GENE_OR_GENOME** has a strong binding affinity to human ACE2 **GENE_OR_GENOME**, based on biochemical interaction studies and crystal structure analysis [7]. SARS-CoV-2 **CORONAVIRUS** and SARS-CoV spike proteins **GENE_OR_GENOME** share identity in amino acid sequences and .....

Figure 4: Example of Fine-grained Entity Extraction

ring text. To start, since most figures are embedded as part of PDF files, we run Deepfigures (Siegel et al., 2018) to automatically detect and extract figures from each PDF document. Then each figure is associated with text in its caption or referring

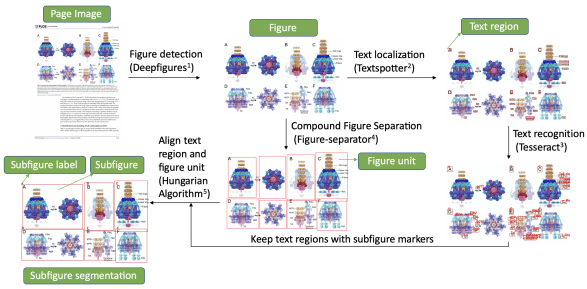

Figure 5: System Pipeline for Automatic Figure Extraction and Subfigure Segmentation. The figure image shown here is from (Kizziah et al., 2020)

context (main body text referring to the figure). In this way, a figure can be attached, at a coarse level, to a KG entity if that entity is mentioned in the associated text.

To further delineate semantic and visual information contained within each subfigure, we have developed a pipeline to segment individual subfigures and then align each subfigure with its corresponding subcaption. We run Figure-separator (Tsutsui and Crandall, 2017) to detect and separate all non-overlapping image regions. On occasion, subfigures within a figure may also be marked with alphabetical letters (e.g., A, B, C, etc). We use deep neural networks (Zhou et al., 2017) to detect text within figures and then apply OCR tools (Smith, 2007) to automatically recognize text content within each figure. To identify *subfigure marker text* and *text labels* for analyzing figure content, we rely on the distance between text labels and subfigures to locate subfigure text markers. Location information of such text markers can also be used to merge multiple image regions into a single subfigure. In

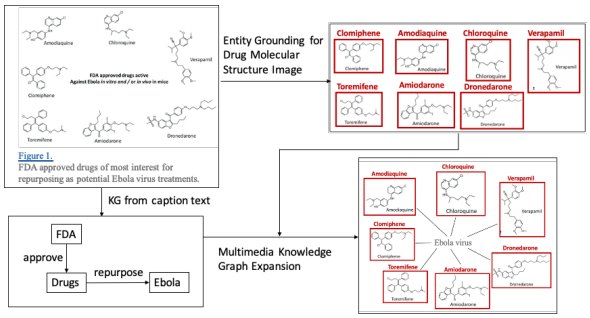

Figure 6: Expanding KG through Subfigure Segmentation and Cross-modal Entity Grounding. The figure image shown here is from (Ekins and Coffee, 2015)

the end, each subfigure is segmented, and associated with its corresponding subcaption and referring context. The segmented subfigures and associated text labels provide rich information that can expand the KG constructed from text captions. For example, as shown in Figure 6, we apply a classifier to detect subfigures containing molecular structures. Then by linking the specific drug names extracted from within-figure text to corresponding drug entities in the coarse KG constructed from the caption text, an expanded cross-modal KG can be constructed that then links images with specific molecular structures to their drug entities in the KG.

2.4 Knowledge Graph Semantic Visualization

In order to enhance the exploration and discovery of the information mined from the COVID-19 literature through the algorithms discussed in previous sections, we create semantic visualizations over large complex networks of biomedical relations using the techniques proposed by Tu et al. (2020). Semantic visualization allows for the visualization of user-defined subsets of these relations interactively through semantically typed tag clouds and heat maps. This allows researchers to get a global view of selected relation subtypes drawn from hundreds or thousands of papers at a single glance. This in turn allows for the ready identification of novel relations that would typically be missed by directed keyword searches or simple unigram word cloud or heatmap displays.[5]

We first build a data index from the knowledge elements in the constructed KGs, and then create a Kibana dashboard[6] out of the generated data indices. Each Kibana dashboard has a collection of visualizations that are designed to interact with each other. Dashboards are implemented as web applications. The navigation of a dashboard is mainly through clicking and searching. By clicking the protein keyword `EIF2AK2` in the tag cloud named *"Enzyme proteins participating Modification relations"*, a constraint on the type of proteins in modifications is added. Correspondingly, all the other visualizations will be changed.

One unique feature of the semantic visualization is the creation of *dense tag clouds* and *dense heatmaps*, through a process of parameter reduction over relations, allowing for the visualization of relation sets as tag clouds and multiple chained relations as heatmaps. Figure 7 illustrates such a dense heatmap that contains relations between proteins and implicated diseases (e.g., *"those proteins that are down-regulators of TNF which are implicated in obesity"*), along with their type information[7].

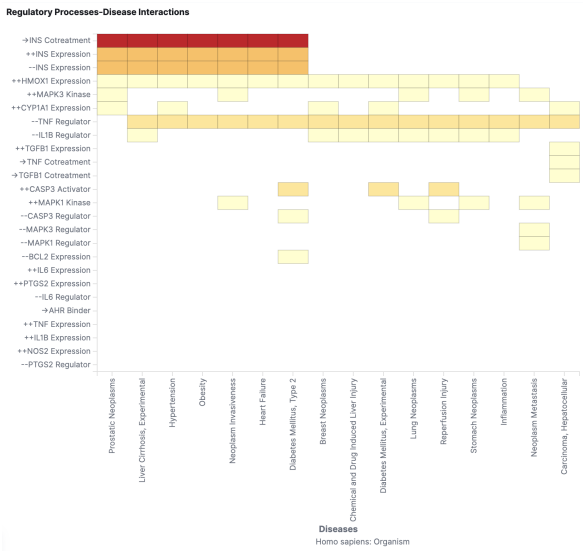

Figure 7: Regulatory Processes-Disease Interactions Heatmap

3 Knowledge-driven Question Answering

In contrast to most current question-answering (QA) methods which target single documents, we have developed a QA component based on a combination of KG matching and distributional semantic matching across documents. We build KG indexing and searching functions to facilitate effective and

[5] https://www.semviz.org/

[6] https://github.com/elastic/kibana

[7] We use the following symbols to indicate the "action" involved in each protein: "++" = increase, "−−" = decrease, "→" = affect.

efficient search when users pose their questions. We also support extended semantic matching from the constructed KGs and related texts by accepting multi-hop queries.

A common category of queries is the connections between two entities. Given two entities in a query, we generate a subgraph covering salient paths between them to show how they are connected through other entities. Figure 3 is an example subgraph summarizing the connections between *Losartan* and *cathepsin L pseudogene 2*. The paths are generated by traversing the constructed KG, and are ranked by the number of papers covering the knowledge elements in each path in the KG. Each edge is assigned a salience score by aggregating the scores of paths passing through it. In addition to knowledge elements, we also present related sentences and source information as evidence. We use BioBert (Lee et al., 2020), a pre-trained language model to represent each sentence along with its left and right neighboring sentences as local contexts. Using the same architecture computed on all respective sentences and the user query, we aggregate the sequence embedding layer, the last hidden layer in the BERT architecture with average pooling (Reimers and Gurevych, 2019). We use the similarity between the embedding representations of each sentence and each query to identify and extract the most relevant sentences as evidence.

Another common category of queries includes entity types, rather than entity instances, and requires extracting evidence sentences based on type or pattern matching. We have developed EVIDENCEMINER (Wang et al., 2020a,b), a web-based system that allows for the user's query as a natural language statement or an inquiry about a relationship at the meta-symbol level (e.g., CHEMICAL, PROTEIN) and then automatically retrieves textual evidence from a background corpora of COVID-19.

4 A case study on Drug Repurposing Report Generation

4.1 Task and Data

A human-written report about drug repurposing usually answers the following typical questions.

1. Current indication: what is the drug class? What is it currently approved to treat?
2. Molecular structure (symbols desired, but a pointer to a reference is also useful)
3. Mechanism of action i.e., inhibits viral entry, replication, etc. (w/ a pointer to data)
4. Was the drug identified by manual or computation screen?
5. Who is studying the drug? (Source/lab name)
6. In vitro Data available (cell line used, assays run, viral strain used, cytopathic effects, toxicity, LD50, dosage response curve, etc.)
7. Animal Data Available (what animal model, LD50, dosage response curve, etc.)
8. Clinical trials on going (what phase, facility, target population, dosing, intervention etc.)
9. Funding source
10. Has the drug shown evidence of systemic toxicity?
11. List of relevant sources to pull data from.

The answers to questions #5 and #11 are extracted based on the meta-data sections of research papers in scientific literature, including the author affiliation and acknowledgement sections. The answers for other questions are all extracted based on the knowledge graphs constructed and knowledge-driven question-answering method described above.

As in our case studies, DARPA biologists inquired about three drugs, Benazepril, Losartan, and Amodiaquine, and their links to COVID-19 related chemicals/genes as shown in Figure 8:

BM1_00870 BM1_06175 BM1_16375 BM1_17125 BM1_22385 BM1_30360 BM1_33735 BM1_56245 BM1_56735 BM1_00870 BM1_06175 BM1_16375 BM1_17125 BM1_22385 BM1_30360 BM1_33735 BM1_56245 BM1_56735 CATB-10270 CATB-1418 CATB-1674 CATB-16A CATB-16D2 CATB-1852 CATB-1874 CATB-2744 CATB-3098 CATB-348 CATB-3483 CATB-5880 CATB-84 CATB-912 CATD CATHY CATK CATL CATL-LIKE CTS12 CTS3 CTS6 CTS7 CTS7-PS CTS8 CTS8L1 CTS8-PS CTSA CTSA.L CTSB CTSBA CTSBB CTSB.L CTSB-PS CTSB.S CTSC CTSC.L CTSC.S CTSD CTSD2 CTSD.S CTSE CTSEAL CTSE.L CTSE.S CTSF CTSF.L CTSG CTSH CTSH.L CTSH-PS CTSJ CTSK CTSK1 CTSK.L CTSL CTSL.1 CTSL3 CTSL3P CTSLA CTSLB CTSLL CTSL.L CTSLL3 CTSLP1 CTSLP2 CTSLP3 CTSLP4 CTSLP6 CTSLP8 CTSM CTSM-PS CTSM-PS2 CTSO CTSO.L CTSQ CTSQL2 CTSR CTSS CTSS1 CTSS.2 CTSS2.1 CTSS2.2 CTSSL CTSS.L CTSS.S CTSV CTSV.L CTSW CTSW.L CTSZ CTSZ.L CTSZ.S LOAG_18685 SMP_013040.1 SMP_034410.1 SMP_067050 SMP_067060 SMP_085010 SMP_085180 SMP_103610 SMP_105370 SMP_158410 SMP_158420 SMP_179950 TSP_01409 TSP_02382 TSP_02383 TSP_03306 TSP_07747 TSP_10129 TSP_10493 TSP_11596 LMAN1 LMAN1L LMAN1.L LMAN1.S LMAN2 LMAN2L MBL1P MBL2 ACE2 FURIN TMPRSS2

Figure 8: COVID-19 related chemicals/genes.

Our KG results for many other drugs are visualized at our website[8]. We download new COVID-19 papers from three Application Programming Interfaces (APIs): NCBI PMC API, NCBI Pubtator API, and CORD-19 archive. We provide incremental updates including new papers, removed papers and updated papers, and their metadata information at our website[9].

[8] http://blender.cs.illinois.edu/covid19/visualization.html

[9] http://blender.cs.illinois.edu/covid19/

4.2 Results

As of June 14, 2020 we collected 140K papers. We selected 25,534 peer-reviewed papers and constructed the KG that includes 7,230 Diseases, 9,123 Chemicals and 50,864 Genes, with 1,725,518 Chemical-Gene links, 5,556,670 Chemical-Disease links, and 7,7844,574 Gene-Disease links. The KG has received more than 1,000+ downloads. Our final generated reports[10] are shared publicly. For each question, our framework provides answers along with detailed evidence, knowledge subgraphs, image segmentation and analysis results. Table 1 shows some example answers.

Several clinicians and medical school students in our team have manually reviewed the drug repurposing reports for three drugs, and also the KGs connecting 41 drugs and COVID-19 related chemicals/genes. In checking the evidence sentences and reading the original articles, they reported that most of our output is informative and valid. For instance, after the coronavirus enters the cell in the lungs, it can cause a severe disease called Acute Respiratory Distress Syndrome. This condition causes the release of inflammatory molecules in the body named cytokines such as Interleukin-2, Interleukin-6, Tumor Necrosis Factor, and Interleukin-10. We see all of these connections in our results, such as the examples shown in Figure 3 and Figure 9. With further checks on these results, the scientists also indicated that many results were worth further investigation. For example, in Figure 3 we can see that Lusartan is connected to tumor protein p53 which is related to lung cancer.

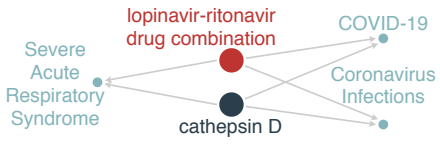

Figure 9: Connections Involving Coronavirus Related Diseases

Question	Example Answers	
Q1	Drug Class	angiotensin-converting enzyme (ACE) inhibitors
	Disease	hypertension
	Evidence Sentences	[PMID:32314699 (PMC7253125)] Past medical history was significant for hypertension, treated with amlodipine and benazepril, and chronic back pain.
		[PMID:32081428 (PMC7092824)] On the other hand, many ACE inhibitors are currently used to treat hypertension and other cardiovascular diseases. Among them are captopril, perindopril, ramipril, lisinopril, benazepril, and moexipril.
Q4	Disease	COVID-19
	Evidence Sentences	[PMID:32081428 (PMC7092824)] By using a molecular docking approach, an earlier study identified N-(2-aminoethyl)-1 aziridine-ethanamine as a novel ACE2 inhibitor that effectively blocks the SARS-CoV RBD-mediated cell fusion. This has provided a potential candidate and lead compound for further therapeutic drug development. Meanwhile, biochemical and cell-based assays can be established to screen chemical compound libraries to identify novel inhibitors.
Q6	Disease	cardiovascular disease
	Evidence Sentences	[PMID:22800722 (PMC7102827)] The in vitro half-maximal inhibitory concentration (IC50) values of food-derived ACE inhibitory peptides are about 1000 fold higher than that of synthetic captopril but they have higher in vivo activities than would be expected from their in vitro activities.....
Q8	Disease	COVID-19
	Evidence Sentences	[PMID:32336612 (PMC7167588)] Two trials of losartan as additional treatment for SARS-CoV-2 infection in hospitalized (NCT04312009) or not hospitalized (NCT04311177) patients have been announced, supported by the background of the huge adverse impact of the ACE Angiotensin II AT1 receptor axis over-activity in these patients.
		[PMID:32350632 (PMC7189178)] To address the role of angiotensin in lung injury, there is an ongoing clinical trial to examine whether losartan treatment affects outcomes in COVID-19 associated ARDS (NCT04312009).
		[PMID:32439915 (PMC7242178)] Losartan was also the molecule chosen in two trials recently started in the United States by the University of Minnesota to treat patients with COVID-19 (clinical trials.gov NCT04311177 and NCT 104312009).

Table 1: Example Answers for Questions in Drug Repurposing Reports

5 Related Work

Extensive prior research work has focused on extracting biomedical entities (Zheng et al., 2014; Habibi et al., 2017; Crichton et al., 2017; Wang et al., 2018; Beltagy et al., 2019; Alsentzer et al., 2019; Wei et al., 2019; Wang et al., 2020c), relations (Uzuner et al., 2011; Krallinger et al., 2011; Manandhar and Yuret, 2013; Bui et al., 2014; Peng et al., 2016; Wei et al., 2015; Peng et al., 2017; Luo et al., 2017; Wei et al., 2019; Li and Ji, 2019; Peng et al., 2019, 2020), and events (Ananiadou et al., 2010; Van Landeghem et al., 2013; Nédellec et al., 2013; Deléger et al., 2016; Wei et al., 2019; Li et al., 2019; ShafieiBavani et al., 2020) from biomedical literature, with the most recent work focused on COVID-19 literature (Hope et al., 2020; Ilievski et al., 2020; Wolinski, 2020; Ahamed and Samad, 2020).

Most of the recent biomedical QA work (Yang et al., 2015, 2016; Chandu et al., 2017; Kraus et al., 2017) is driven by the BioASQ initiative (Tsatsaronis et al., 2015), and many live QA systems, including COVIDASK[11] and AUEB[12], and search en-

[10] http://blender.cs.illinois.edu/covid19/DrugRe-purposingReport_V2.0.docx

[11] https://covidask.korea.ac.kr/

[12] http://cslab241.cs.aueb.gr:5000/

gines (Kricka et al., 2020; Esteva et al., 2020; Hope et al., 2020; Taub Tabib et al., 2020) have been developed. Our work is an application and extension of our recently developed multimedia knowledge extraction system for the news domain (Li et al., 2020a,b). Similar to the news domain, the knowledge elements extracted from text and images in literature are complementary. Our framework advances state-of-the-art by extending the knowledge elements to more fine-grained types, incorporating image analysis and cross-media knowledge grounding, and KG matching into QA.

6 Conclusions and Future Work

We have developed a novel framework, **COVID-KG**, that automatically transforms a massive scientific literature corpus into organized, structured, and actionable KGs, and uses it to answer questions in drug repurposing reporting. With **COVID-KG**, researchers and clinicians are able to obtain informative answers from scientific literature, and thus focus on more important hypothesis testing, and prioritize the analysis efforts for candidate exploration directions. In our ongoing work, we have created a new ontology that includes 77 entity subtypes and 58 event subtypes, and we are building a neural IE system following this new ontology. In the future, we plan to extend **COVID-KG** to automate the creation of new hypotheses by predicting new links. We will also create a multimedia common semantic space (Li et al., 2020a,b) for literature and apply it to improve cross-media knowledge grounding and inference.

Ethical Considerations

Required Workflow for Using Our System

Human review required. Our knowledge discovery tool provides investigative leads for pre-clinical research, not final results for clinical use. Currently, biomedical researchers scour the literature to identify candidate drugs, then follow a standard research methodology to investigate their actual utility (involving literature reviews, computer simulations of drug mechanisms and effectiveness, in-vitro studies, cellular in-vivo studies, etc. before moving to clinical studies.). Our tool COVID-KG (and all knowledge discovery tools for biomedical applications) is not meant to be used for direct clinical applications on any human subjects. Rather, our tool aims to highlight unseen relations and patterns in large amounts of scientific textual data that would be too time-consuming for manual human effort. Accordingly, the tool would be useful for stakeholders (e.g., biomedical scientists) to identify specific drug candidates and molecular targets that are relevant in their biomedical and clinical research aims. The use of our knowledge discovery tool allows the researcher to narrow down the set of candidate drugs to investigate rapidly, but then proceed with the usual sequence of steps before kicking off expensive and time-consuming clinical tests. Failure to follow this sequence of events, and use of the system without the required human review, could lead to misguided experimental design wasting time and resources.

Check evidence and source before using our system results. In addition, our tool provides source, confidence values and rich evidence sentences for each node and link in the KG. To curtail potential harms caused by extraction errors, users of the knowledge graphs should double-check the source information and verify the accuracy of the discovered leads before launching expensive experimental studies. We spell out here the positive values, as well as the limitations and possible solutions to address these issues for future improvement. Moreover, any planned investigations involving human subjects should first be approved by the stakeholder's IRB (Institutional Review Board) who will oversee the safety of the proposed studies and the role of COVID-KG before any experimental studies are conducted. COVID-KG is a tool to enhance biomedical and clinical research; it is not a tool for direct clinical application with human subjects.

Limitations of System Performance and Data Collection

System errors. Our system can effectively convert a large amount of scientific papers into knowledge graphs, and can scale as literature volume increases. However, none of our extraction components is perfect, they produce about 6%-22% false alarms and misses as reported in section 2. But as we described in the workflow, all of the connections and answers will be validated by domain experts by checking their corresponding sources before they are included in the drug repurposing report. COVID-KG is developed for pre-clinical research to target down drugs of interest for biomedical scientists. Therefore, no human subjects or specific populations are directly subjected to COVID-

KG unless approved by the stakeholder's IRB who oversees the safety and ethical aspects of the clinical studies in accordance with the Belmont report (https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html). Accordingly, COVID-KG will not impose direct harm to vulnerable human cohorts or populations, unless misused by the stakeholders without IRB approval. With regards to potential harm in preclinical studies, users of COVID-KG are advised to verify the accuracy of the discovered leads in the source information before conducting expensive experimental studies.

Bias in training data. Proper use of the technology requires that input documents are legally and ethically obtained. Regulation and standards (e.g. GDPR[13]) provide a legal framework for ensuring that such data is properly used and that any individual whose data is used has the right to request its removal. In the absence of such regulation, society relies on those who apply technology to ensure that data is used in an ethical way. The input data to our system is peer-reviewed publicly available scientific articles. Additional potential harm could come from the output of the system being used in ways that magnify the system errors or bias in its training data. The various components in our system rely on weak distant supervision based on large-scale external knowledge bases and ontologies that cover a wide range of topics in the biomedical domain. Nevertheless, our system output is intended for human interpretation. We do not endorse incorporating the system's output into an automatic decision-making system without human validation; this fails to meet our recommendations and could yield harmful results. In the cited technical reports for each component in our framework, we have reported detailed error rates for each type of knowledge element from system evaluations and provide detailed qualitative analysis and explanations.

Bias in development data. We also note that the performance of our system components as reported is based on the specific benchmark datasets, which could be affected by such data biases. Thus questions concerning generalizability and fairness should be carefully considered. Within the research community, addressing data bias requires a combination of new data sources, research that mitigates the impact of bias, and, as done in (Mitchell et al., 2019), auditing data and models. Sections 2 and 4.1 cite data sources used for training to support future auditing. A general approach to properly use our system should incorporate ethics considerations as the first-order principles in every step of the system design, maintain a high degree of transparency and interpretability of data, algorithms, models, and functionality throughout the system, make software available as open-source for public verification and auditing, and explore countermeasures to protect vulnerable groups. In our ongoing and future work, we will keep increasing the annotated dataset size, add more rounds of user correction and validation, and iteratively incorporate feedback from domain experts who have used the tool, to create new benchmarks for retraining model and conducting more systematic evaluations. We recommend caution of using our system output until a more complete expert evaluation has occurred.

Bias in source. Furthermore, our system output may include some biases from the sources, by way of biases in the peer-reviewing process. In our previous work (Yu et al., 2014; Ma et al., 2015; Zhi et al., 2015; Zhang et al., 2019), we have aggregated source profile, knowledge graphs, and evidence for fact-checking across sources. We plan to extend our framework to include fact-checking to enable practitioners and researchers to access up-to-the-minute information.

Bias in test queries. Finally, the queries (i.e., the lists of candidate drugs and proteins/genes) are provided by the users who might have biases in their selection. Addressing the user's own biases falls outside the scope of our project, but as we have stated in the previous subsection, we direct users to carefully examine source information (author, publication date, etc.) and detailed evidence (contextual sentences and documents) associated with the extracted connections.

[13]The General Data Protection Regulation of the European Union https://gdpr.eu/what-is-gdpr/.

Acknowledgement

This research is based upon work supported in part by U.S. DARPA KAIROS Program No. FA8750-19-2-1004, U.S. DARPA AIDA Program # FA8750-18-2-0014, .S. DTRA HDTRA I -16-1-0002/Project #1553695, eTASC - Empirical Evidence for a Theoretical Approach to Semantic Components, U.S. NSF No. 1741634, the Office of the Director of National Intelligence (ODNI), and Intelligence Advanced Research Projects Activity (IARPA) via contract FA8650-17-C-9116. The views and conclusions contained herein are

those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Sabber Ahamed and Manar Samad. 2020. Information mining for covid-19 research from a large volume of scientific literature. *Information Retrieval Repository*, arXiv:2004.02085.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Sophia Ananiadou, Sampo Pyysalo, Jun'ichi Tsujii, and Douglas B Kell. 2010. Event extraction for systems biology by text mining the literature. *Trends in biotechnology*, 28(7):381–390.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. 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)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Quoc-Chinh Bui, Peter MA Sloot, Erik M Van Mulligen, and Jan A Kors. 2014. A novel feature-based approach to extract drug–drug interactions from biomedical text. *Bioinformatics*, 30(23):3365–3371.

Khyathi Chandu, Aakanksha Naik, Aditya Chandrasekar, Zi Yang, Niloy Gupta, and Eric Nyberg. 2017. Tackling biomedical text summarization: OAQA at BioASQ 5B. In *BioNLP 2017*, pages 58–66, Vancouver, Canada,. Association for Computational Linguistics.

Gamal Crichton, Sampo Pyysalo, Billy Chiu, and Anna Korhonen. 2017. A neural network multi-task learning approach to biomedical named entity recognition. *Bioinformatics*, 18(1):368.

Allan Peter Davis, Cynthia J. Grondin, Robin J. Johnson, Daniela Sciaky, Benjamin L. King, Roy McMorran, Jolene Wiegers, Thomas C. Wiegers, and Carolyn J. Mattingly. 2016. The Comparative Toxicogenomics Database: update 2017. *Nucleic Acids Research*, 45(D1):D972–D978.

Louise Deléger, Robert Bossy, Estelle Chaix, Mouhamadou Ba, Arnaud Ferré, Philippe Bessières, and Claire Nédellec. 2016. Overview of the bacteria biotope task at BioNLP shared task 2016. In *Proceedings of the 4th BioNLP Shared Task Workshop*, pages 12–22, Berlin, Germany. Association for Computational Linguistics.

Sean Ekins and Megan Coffee. 2015. Fda approved drugs as potential ebola treatments. *F1000Research*, 4.

Andre Esteva, Anuprit Kale, Romain Paulus, Kazuma Hashimoto, Wenpeng Yin, Dragomir Radev, and Richard Socher. 2020. Co-search: Covid-19 information retrieval with semantic search, question answering, and abstractive summarization. *Information Retrieval Repository*, arXiv:2006.09595.

Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser. 2017. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):37–48.

Tom Hope, Jason Portenoy, Kishore Vasan, Jonathan Borchardt, Eric Horvitz, Daniel S Weld, Marti A Hearst, and Jevin West. 2020. Scisight: Combining faceted navigation and research group detection for covid-19 exploratory scientific search. *Information Retrieval Repository*, arXiv:2005.12668.

Filip Ilievski, Daniel Garijo, Hans Chalupsky, Naren Teja Divvala, Yixiang Yao, Craig Rogers, Ronpeng Li, Jun Liu, Amandeep Singh, Daniel Schwabe, et al. 2020. Kgtk: A toolkit for large knowledge graph manipulation and analysis. *Artificial Intelligence Repository*, arXiv:2006.00088.

James L Kizziah, Keith A Manning, Altaira D Dearborn, and Terje Dokland. 2020. Structure of the host cell recognition and penetration machinery of a staphylococcus aureus bacteriophage. *PLoS pathogens*, 16(2):e1008314.

Martin Krallinger, Miguel Vazquez, Florian Leitner, David Salgado, Andrew Chatr-Aryamontri, Andrew Winter, Livia Perfetto, Leonardo Briganti, Luana Licata, Marta Iannuccelli, et al. 2011. The protein-protein interaction tasks of biocreative iii: classification/ranking of articles and linking bio-ontology concepts to full text. *BMC bioinformatics*, 12(S8):S3.

Milena Kraus, Julian Niedermeier, Marcel Jankrift, Sören Tietböhl, Toni Stachewicz, Hendrik Folkerts, Matthias Uflacker, and Mariana Neves. 2017. Olelo: a web application for intuitive exploration of biomedical literature. *Nucleic acids research*, 45(W1):478–483.

Larry J Kricka, Sergei Polevikov, Jason Y Park, Paolo Fortina, Sergio Bernardini, Daniel Satchkov, Valentin Kolesov, and Maxim Grishkov. 2020. Artificial intelligence-powered search tools and resources in the fight against covid-19. *EJIFCC*, 31(2):106.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Diya Li, Lifu Huang, Heng Ji, and Jiawei Han. 2019. Biomedical event extraction based on knowledge-driven tree-LSTM. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1421–1430, Minneapolis, Minnesota. Association for Computational Linguistics.

Diya Li and Heng Ji. 2019. Syntax-aware multi-task graph convolutional networks for biomedical relation extraction. In *Proc. EMNLP2019 Workshop on Health Text Mining and Information Analysis*.

Manling Li, Alireza Zareian, Ying Lin, Xiaoman Pan, Spencer Whitehead, Brian Chen, Bo Wu, Heng Ji, Shih-Fu Chang, Clare Voss, Daniel Napierski, and Marjorie Freedman. 2020a. GAIA: A fine-grained multimedia knowledge extraction system. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 77–86, Online. Association for Computational Linguistics.

Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, and Shih-Fu Chang. 2020b. Cross-media structured common space for multimedia event extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2557–2568, Online. Association for Computational Linguistics.

Yuan Luo, Özlem Uzuner, and Peter Szolovits. 2017. Bridging semantics and syntax with graph algorithms—state-of-the-art of extracting biomedical relations. *Briefings in bioinformatics*, 18(1):160–178.

Fenglong Ma, Yaliang Li, Qi Li, Minghui Qiu, Jing Gao, Shi Zhi, Lu Su, Bo Zhao, Heng Ji, and Jiawei Han. 2015. Faitcrowd: Fine grained truth discovery for crowdsourced data aggregation. In *Proc. the 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD2015)*.

Suresh Manandhar and Deniz Yuret, editors. 2013. *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Association for Computational Linguistics, Atlanta, Georgia, USA.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229.

Claire Nédellec, Robert Bossy, Jin-Dong Kim, Jungjae Kim, Tomoko Ohta, Sampo Pyysalo, and Pierre Zweigenbaum. 2013. Overview of BioNLP shared task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 1–7, Sofia, Bulgaria. Association for Computational Linguistics.

Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. 2017. Cross-sentence n-ary relation extraction with graph lstms. *Transactions of the Association for Computational Linguistics*, 5:101–115.

Yifan Peng, Qingyu Chen, and Zhiyong Lu. 2020. An empirical study of multi-task learning on BERT for biomedical text mining. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 205–214, Online. Association for Computational Linguistics.

Yifan Peng, Chih-Hsuan Wei, and Zhiyong Lu. 2016. Improving chemical disease relation extraction with rich features and weakly labeled data. *Journal of cheminformatics*, 8(1):53.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 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)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Elaheh ShafieiBavani, Antonio Jimeno Yepes, Xu Zhong, and David Martinez Iraola. 2020. Global locality in biomedical relation and event extraction. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 195–204, Online. Association for Computational Linguistics.

Jingbo Shang, Liyuan Liu, Xiaotao Gu, Xiang Ren, Teng Ren, and Jiawei Han. 2018. Learning named entity tagger using domain-specific dictionary. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2054–2064, Brussels, Belgium. Association for Computational Linguistics.

Noah Siegel, Nicholas Lourie, Russell Power, and Waleed Ammar. 2018. Extracting scientific figures with distantly supervised neural networks. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, page 223–232, New York, NY, USA. Association for Computing Machinery.

Ray Smith. 2007. An overview of the tesseract ocr engine. In *Proceedings of the 9th international conference on document analysis and recognition (ICDAR 2007)*, volume 2, pages 629–633.

Hillel Taub Tabib, Micah Shlain, Shoval Sadde, Dan Lahav, Matan Eyal, Yaara Cohen, and Yoav Goldberg. 2020. Interactive extractive search over biomedical corpora. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 28–37, Online. Association for Computational Linguistics.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):138.

Satoshi Tsutsui and David J Crandall. 2017. A data driven approach for compound figure separation using convolutional neural networks. In *Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 533–540.

Jingxuan Tu, M. Verhagen, B. Cochran, and J. Pustejovsky. 2020. Exploration and discovery of the covid-19 literature through semantic visualization. *ArXiv*, abs/2007.01800.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Sofie Van Landeghem, Jari Björne, Chih-Hsuan Wei, Kai Hakala, Sampo Pyysalo, Sophia Ananiadou, Hung-Yu Kao, Zhiyong Lu, Tapio Salakoski, Yves Van de Peer, et al. 2013. Large-scale event extraction from literature with multi-level gene normalization. *PloS one*, 8(4):e55814.

Qingyun Wang, Lifu Huang, Zhiying Jiang, Kevin Knight, Heng Ji, Mohit Bansal, and Yi Luan. 2019a. PaperRobot: Incremental draft generation of scientific ideas. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1980–1991, Florence, Italy. Association for Computational Linguistics.

Xuan Wang, Yingjun Guan, Weili Liu, Aabhas Chauhan, Enyi Jiang, Qi Li, David Liem, Dibakar Sigdel, John Caufield, Peipei Ping, et al. 2020a. Evidenceminer: Textual evidence discovery for life sciences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 56–62.

Xuan Wang, Weili Liu, Aabhas Chauhan, Yingjun Guan, and Jiawei Han. 2020b. Automatic textual evidence mining in covid-19 literature. *Computation and Language Repository*, arXiv:2004.12563.

Xuan Wang, Xiangchen Song, Yingjun Guan, Bangzheng Li, and Jiawei Han. 2020c. Comprehensive named entity recognition on cord-19 with distant or weak supervision. *Computation and Language Repository*, arXiv:2003.12218.

Xuan Wang, Yu Zhang, Qi Li, Xiang Ren, Jingbo Shang, and Jiawei Han. 2019b. Distantly supervised biomedical named entity recognition with dictionary expansion. In *Proceedings of the 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 496–503.

Xuan Wang, Yu Zhang, Xiang Ren, Yuhao Zhang, Marinka Zitnik, Jingbo Shang, Curtis Langlotz, and Jiawei Han. 2018. Cross-type biomedical named entity recognition with deep multi-task learning. *Bioinformatics*, 35(10):1745–1752.

Chih-Hsuan Wei, Alexis Allot, Robert Leaman, and Zhiyong Lu. 2019. PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Research*, 47(W1):587–593.

Chih-Hsuan Wei, Yifan Peng, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Jiao Li, Thomas C Wiegers, and Zhiyong Lu. 2015. Overview of the biocreative v chemical disease relation (cdr) task. In *Proceedings of the 5th BioCreative challenge evaluation workshop*, volume 14.

Francis Wolinski. 2020. Visualization of diseases at risk in the covid-19 literature. *Information Retrieval Repository*, arXiv:2005.00848.

Zi Yang, Niloy Gupta, Xiangyu Sun, Di Xu, Chi Zhang, and Eric Nyberg. 2015. Learning to answer biomedical factoid & list questions: Oaqa at bioasq 3b. *CLEF (Working Notes)*, 1391.

Zi Yang, Yue Zhou, and Eric Nyberg. 2016. Learning to answer biomedical questions: OAQA at BioASQ 4B. In *Proceedings of the Fourth BioASQ workshop*, pages 23–37, Berlin, Germany. Association for Computational Linguistics.

Dian Yu, Hongzhao Huang, Taylor Cassidy, Heng Ji, Chi Wang, Shi Zhi, Jiawei Han, Clare Voss, and Malik Magdon-Ismail. 2014. The wisdom of minority: Unsupervised slot filling validation based on multi-dimensional truth-finding. In *Proc. The 25th International Conference on Computational Linguistics (COLING2014)*.

Haibo Zhang, Josef M Penninger, Yimin Li, Nanshan Zhong, and Arthur S Slutsky. 2020. Angiotensin-converting enzyme 2 (ace2) as a sars-cov-2 receptor: molecular mechanisms and potential therapeutic target. *Intensive care medicine*, 46(4):586–590.

Xiaomei Zhang, Yibo Wu, Lifu Huang, Heng Ji, and Guohong Cao. 2019. Expertise-aware truth analysis and task allocation in mobile crowdsourcing. *IEEE Transactions on Mobile Computing*.

Jin Guang Zheng, Daniel Howsmon, Boliang Zhang, Juergen Hahn, Deborah McGuinness, James Hendler, and Heng Ji. 2014. Entity linking for biomedical literature. In *BMC Medical Informatics and Decision Making*.

Jin Guang Zheng, Daniel Howsmon, Boliang Zhang, Juergen Hahn, Deborah McGuinness, James Hendler, and Heng Ji. 2015. Entity linking for biomedical literature. In *Proceedings of the BMC Medical Informatics and Decision Making*, volume 15.

Shi Zhi, Bo Zhao, Wenzhu Tong, Jing Gao, Dian Yu, Heng Ji, and Jiawei Han. 2015. Modeling truth existence in truth discovery. In *Proc. the 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD2015)*.

Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. 2017. East: an efficient and accurate scene text detector. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 5551–5560.