Descriptive Knowledge Graph in Biomedical Domain

Kerui Zhu **Jie Huang**† **Kevin Chen-Chuan Chang**
University of Illinois at Urbana-Champaign, USA
{keruiz2, jeffhj, kcchang}@illinois.edu

Abstract

We present a novel system that automatically extracts and generates informative and descriptive sentences from the biomedical corpus and facilitates the efficient search for relational knowledge. Unlike previous search engines or exploration systems that retrieve unconnected passages, our system organizes descriptive sentences as a relational graph, enabling researchers to explore closely related biomedical entities (e.g., diseases treated by a chemical) or indirectly connected entities (e.g., potential drugs for treating a disease). Our system also uses ChatGPT and a fine-tuned relation synthesis model to generate concise and reliable descriptive sentences from retrieved information, reducing the need for extensive human reading effort. With our system, researchers can easily obtain both high-level knowledge and detailed references and interactively steer to the information of interest. We spotlight the application of our system in COVID-19 research, illustrating its utility in areas such as drug repurposing and literature curation.[1]

1 Introduction

Efficiently extracting knowledge from the vast and ever-growing corpus of literature is crucial for researchers to keep up with the latest discoveries and trends in the field. The COVID-19 pandemic has highlighted this need, with thousands of related studies being published in a short period when a new disease emerges. However, surveying the latest findings requires significant effort, and researchers may struggle to see the big picture, leading to duplicated work and delaying the development of treatments (Wang et al., 2021). Hence, an exploration system that can effectively retrieve comprehensive information from the latest literature corpus is important.

Existing exploration systems manage information in generally three granularities: documents, sentences, and knowledge facts. Document-level retrieval usually takes keyphrases (Shen et al., 2018) or questions (Voorhees et al., 2021; Levy et al., 2021) as queries and finds relevant documents. Using such systems, researchers need to read the retrieved documents to find relevant information, which is still time-consuming (Wang et al., 2020b). Sentence-level (Wang and Lo, 2021; Wang et al., 2020b; Lahav et al., 2022) retrieval usually takes entities, entity types, or sentences as queries and finds sentences that contain the entities and entity types or are semantically similar to the input sentence. The retrieved sentences require less reading effort, but they are retrieved as independent text pieces, which don't provide a general overview of the knowledge. Knowledge facts, on the other hand, are usually (head, relation, tail) triples extracted from the corpus and stored as a knowledge graph, which concisely reveals the connection between entities. However, systems that retrieve knowledge facts (Chung et al.; Wang et al., 2021) usually allow queries for entities and entity types only, and the retrieved knowledge facts can only cover the relations in a fixed pre-defined set.

To overcome the limitations mentioned above, we develop an exploration system that manages corpus sentences as a *descriptive knowledge graph* (Huang et al., 2022b). The Descriptive knowledge graph for Explaining Entity Relationships (DEER) is a special knowledge graph where each edge is not a relation label but a set of relational sentences describing the relationship between a pair of entities (Handler and O'Connor, 2018; Huang et al., 2022a; Huang and Chang, 2022; Liu et al., 2023; Huang et al., 2023). We collect entities and sentences from the biomedical corpus to build a domain-specific DEER and provide useful tools for users to effectively query and explore the graph.

Our system allows users with little prior knowl-

[1] Demo video: https://www.youtube.com/watch?v=VvWs9JEP8ro System website: https://zhukerui.github.io/CovidDEER/. †Corresponding author.

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 462–470
December 6-10, 2023 ©2023 Association for Computational Linguistics

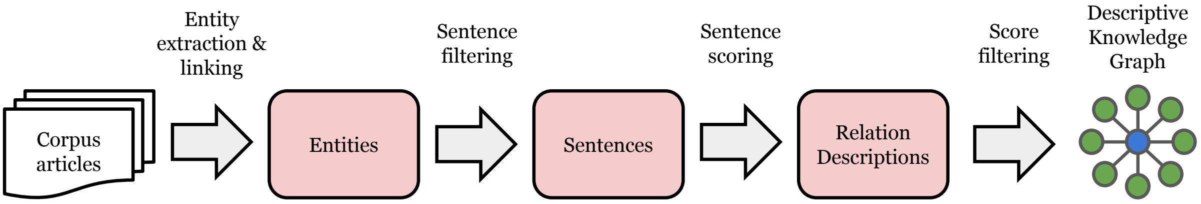


Figure 1: Data pipeline for descriptive knowledge graph construction: 1. Extract entities from corpus articles. 2. Remove sentences with missing subject or object entities. 3. Score sentences as relation descriptions with the RDS. 4. Filter low-score relation descriptions (score < 0.7) and build the graph.

edge to interactively retrieve up-to-date, comprehensive, and easily understandable relational sentences, and explore relationships between entities in one-hop or multi-hop connections. Additionally, we use ChatGPT and a fine-tuned relation synthesis model to generate succinct relation descriptions for entity pairs based on the retrieved sentences to aid users' reading. It is worth mentioning that our system is automatically built without any supervised training or hand-crafted rules, making it seamlessly adaptable to any biomedical corpus with ease, and it can serve as a frontrunner for collecting knowledge in any future emergency.

2 Graph Construction

DEER (Huang et al., 2022b) is a form of knowledge representation that balances the openness and informativeness of free text and the structured representation of the knowledge graph. In this graph, nodes are entities, and edges are sentences describing the relationship between the two nodes, called *relation descriptions*, pointing from the subject node to the object node in the sentences. The previous DEER graph (Huang et al., 2022b) was built upon Wikipedia. Due to the limitations of the corpus, it does not contain much biomedical domain knowledge. In this section, we will introduce techniques for building a descriptive knowledge graph in the biomedical domain. Based on that, users could retrieve sentences with efficient graph queries and view the result from a connected perspective to gain a more holistic understanding of the retrieved information.

2.1 Corpus

To efficiently establish the system for retrieving knowledge about a specific topic, we build the DEER on a sub-domain corpus. In this work, we use COVID-19 Open Research Dataset (CORD-19) (Wang et al., 2020a) as a representative corpus in the biomedical domain, which comprises scientific papers related to COVID-19 and other coronaviruses, and note the DEER built from it as *CovidDEER*. For demonstration purposes, we used the snapshot on August 8th, 2020 to simulate a corpus when a new disease outbreak and some clinical experimental results have been published. With this corpus, we demonstrate how our system can retrieve valuable information for disease research and drug repurposing.

# documents	# nodes	# edges
72,014	140,574	863,102

Table 1: Corpus and graph statistics for CovidDEER, collected using the RDS threshold of 0.7.

2.2 Pipeline

To construct *CovidDEER*, our system employs a pipeline that processes the corpus as follows:

Entity Extraction and Linking Initially, we extract biomedical entities from each sentence in the corpus and link them to biomedical ontologies using the NCBI Pubtator API and the SciSpacy library (Neumann et al., 2019). Specifically, we link the extracted entities to Cellosaurus, OMIM, MeSH, Gene, Taxonomy, and UMLS metathesaurus.

Sentence Filtering Next, we use SciSpacy to parse the sentences and remove those which do not have a subject entity or object entity as these entities serve as the head and tail entities in the relation description. Missing head or tail entities indicate that these sentences are not appropriate for describing relationships.

Sentence Scoring Then, we gather the parameters for a scoring function and use it to score the sentences. We use the relation description score (RDS) introduced in Huang et al. (2022b) as the scoring function. This scoring function extracts the dependency path between the head entity, tail entity and other relation-related words in a sentence and generates a score between 0 and 1 to indicate how well this sentence expresses the relationship of the entities. Higher score indicates better the sentence as a relation description. A domain-specific RDS scoring function requires data of dependency path frequency from the domain corpus. Once the scoring function is setup with adequate corpus data, it can be frozen to evaluate any in-domain sentences. In this work, we collect dependency path frequencies from the whole CORD-19 corpus.

For each pair of subject-object entities in a sentence, we apply the scoring function and store the sentence and the RDS score with the corresponding entity pair.

Score Filtering For each entity pair, we filter out the low-quality sentences with the RDS score and assign the rest to the edge from the head entity to the tail entity to construct the DEER. In practice, a sentence with RDS score over 0.7 usually has a good quality.

A visualization of the data pipeline is depicted in Figure 1 and the statistics of the CORD-19 corpus and *CovidDEER* are listed in Table 1. Note that the graph can be easily updated with the latest knowledge by extracting relation descriptions from recent papers.

3 Graph Query

To retrieve sentences from the *CovidDEER*, our system provides a Graph Query module with some auxiliary tools to allow an interactive and flexible search. Below are the queries supported by the graph query module.

Entity-Entity Query Entity-Entity query allows users to retrieve relation descriptions between two entities. This is achieved by extracting sentences that lie on the edge connecting the two entities in the constructed CovidDEER graph. Unlike from systems that retrieve all or random sentences where the entities co-occur, our system focuses on returning sentences that capture the primary relationships between the target entities. This gives users more informative and clearer sentences and saves users from being distracted by meaningless sentences.

Entity-Type Query To obtain a more comprehensive overview of relationships between an entity and an entity type, our system also supports Entity-Type query, which will retrieve sentences from edges between an entity and all its neighbors belonging to an entity type. For instance, users can set the entity to *COVID-19* and the entity type to *Chemicals*. Then the system will return relation descriptions between COVID-19 and all kinds of related chemicals, which provides some insights into the Chemical-Disease interactions related to COVID-19. Our system supports all the entity types in the ontologies mentioned in Section 2.2.

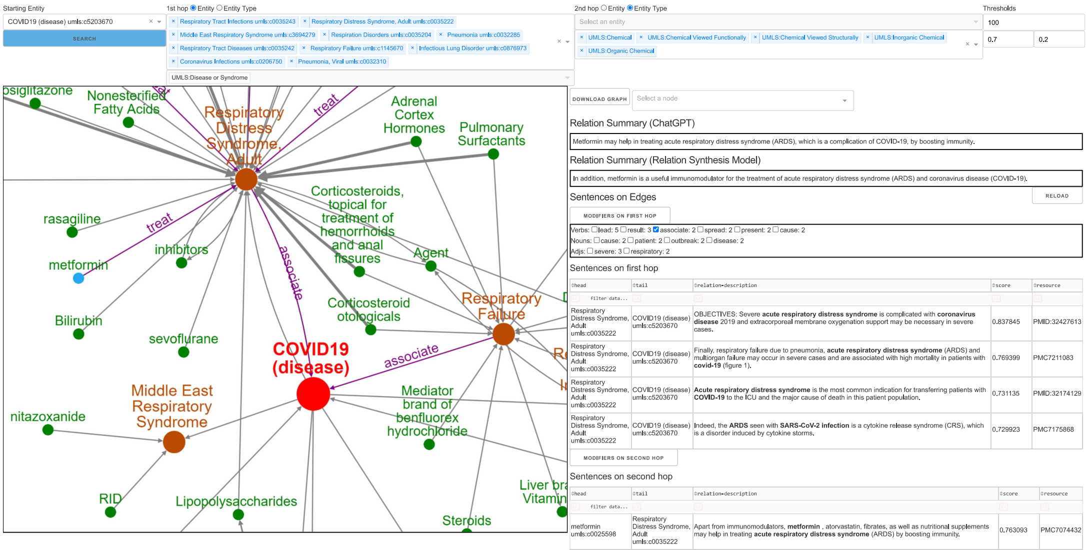
Type	Frequent Modifiers
Nouns	treatment (14), chloroquine (6), efficacy (4), hydroxychloroquine (4), therapy (2), option (2), patient (2)
Verbs	show (7), treat (5), use (5), propose (2), include (2)
Adjs	apparent (4), antiviral (3), effective (3), safe (2), severe (2), antimalarial (2)

Table 2: Frequent Modifiers between Chemicals and COVID-19.

Multi-hop Query Besides finding direct neighbors of an entity, users can also query multi-hop neighbors to explore more indirect connections. By specifying the entities or entity types at each hop, users can retrieve sentences for multi-hop inference. For example, a user may begin with *COVID-19*, set *Symptom* as the first-hop entity type and *Chemical* as the second-hop entity type to explore drugs that can treat COVID-19 related symptoms, and thus, could be used for COVID-19 treatment. With this tool, our system could beat traditional knowledge graphs by providing the contextualized knowledge, and beat the text-based search engines by allowing multi-hop retrieval with one query.

Modifier Filtering When querying a popular entity, *CovidDEER* may return too many edges, which may distract users from catching the general relationships. To alleviate this, we define the words in a sentence that convey the relation information as the *modifiers* and allow users to locate interesting edges using the modifiers. We extract the noun phrases, verbs, and adjectives on the dependency path between the two entities as the modifiers. For example, Table 2 shows the frequent modifiers collected between *COVID-19* and its *Chemical* neighbors. These modifiers provide insights into the *COVID-19-Chemical* relationships and users can click the modifiers to highlight the edges where they occur. This tool could also help users perform a more fine-grained query to reduce unwanted results.

Figure 2 shows an example interface of our system, where the retrieved results are displayed as a



Starting Entity

COVID19 (disease) umls:c5203670

1st hop ● Entity ○ Entity Type

× Respiratory Tract Infections umls:c0035243 × Respiratory Distress Syndrome, Adult umls:c0035222 × Middle East Respiratory Syndrome umls:c3694279 × Respiration Disorders umls:c0035204 × Pneumonia umls:c0032285 × Respiratory Tract Diseases umls:c0035242 × Respiratory Failure umls:c1145670 × Infectious Lung Disorder umls:c0876973 × Coronavirus Infections umls:c0206750 × Pneumonia, Viral umls:c0032310

UMLS:Disease or Syndrome

2nd hop ○ Entity ● Entity Type

Select an entity

× UMLS:Chemical × UMLS:Chemical Viewed Functionally × UMLS:Chemical Viewed Structurally × UMLS:Inorganic Chemical × UMLS:Organic Chemical

Thresholds

100

0.7 0.2

Relation Summary (ChatGPT)

Metformin may help in treating acute respiratory distress syndrome (ARDS), which is a complication of COVID-19, by boosting immunity.

Relation Summary (Relation Synthesis Model)

In addition, metformin is a useful immunomodulator for the treatment of acute respiratory distress syndrome (ARDS) and coronavirus disease (COVID-19).

Sentences on Edges

Verbs: □ lead: 5 □ result: 3 ☑ associate: 2 □ spread: 2 □ present: 2 □ cause: 2

Nouns: □ cause: 2 □ patient: 2 □ outbreak: 2 □ disease: 2

Adjs: □ severe: 3 □ respiratory: 2

Sentences on first hop

head	tail	relation-description	score	resource
filter data...				
Respiratory Distress Syndrome, Adult umls:c0035222	COVID19 (disease) umls:c5203670	OBJECTIVES: Severe **acute respiratory distress syndrome** is complicated with **coronavirus disease** 2019 and extracorporeal membrane oxygenation support may be necessary in severe cases.	0.837845	PMID:32427613
Respiratory Distress Syndrome, Adult umls:c0035222	COVID19 (disease) umls:c5203670	Finally, respiratory failure due to pneumonia, **acute respiratory distress syndrome** (ARDS) and multiorgan failure may occur in severe cases and are associated with high mortality in patients with **covid-19** (figure 1).	0.769399	PMC7211083
Respiratory Distress Syndrome, Adult umls:c0035222	COVID19 (disease) umls:c5203670	**Acute respiratory distress syndrome** is the most common indication for transferring patients with **COVID-19** to the ICU and the major cause of death in this patient population.	0.731135	PMID:32174129
Respiratory Distress Syndrome, Adult umls:c0035222	COVID19 (disease) umls:c5203670	Indeed, the **ARDS** seen with **SARS-CoV-2 infection** is a cytokine release syndrome (CRS), which is a disorder induced by cytokine storms.	0.729923	PMC7175868

Sentences on second hop

head	tail	relation-description	score	resource
filter data...				
metformin umls:c0025598	Respiratory Distress Syndrome, Adult umls:c0035222	Apart from immunomodulators, **metformin** , atorvastatin, fibrates, as well as nutritional supplements may help in treating **acute respiratory distress syndrome** (ARDS) by boosting immunity.	0.763093	PMC7074432

Figure 2: The web interface of *CovidDEER*. The interface shows a graph retrieved by a two-hop query: *"COVID19"* – 10 *"Disease or Syndrome"* entities – 5 *"Pharmacologic Substance"* related entity types. The *metformin* is selected (in blue) and a directed path, *COVID19* → *Respiratory Distress Syndrome, Adult* → *metformin*, is used for relation summary by ChatGPT and relation synthesis model.

graph, and users can checkout sentences by clicking the edges.

4 Relation Synthesis Model

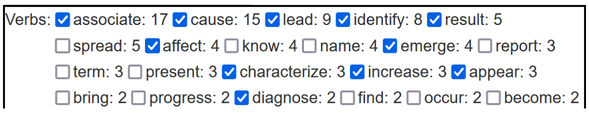
Although *CovidDEER* displays the relational sentences in a graph view to reveal the connections between entities, considerable manual effort is still required to read and digest the information on the edges. Performing multi-hop logic inference is even harder as users need to find associated sentences across different edges. To reduce users' reading effort, we trained a relation synthesis model (Huang et al., 2022b), which is based on a Fusion-in-Decoder model (Izacard and Grave, 2020) trained to take sequences of relation descriptions from the multi-hop paths between two entities in DEER and generate one single relation description for the entities. Each training data is collected by selecting the highest RDS-scored sentences on each edge in the multi-hop paths between a target entity pair as the input and the highest RDS-scored sentence on the one-hop path between the target pair as the output. In order to allow summarizing relation descriptions on each edge, we also add the lower-scored sentences on the one-hop path into the input. Since large language models have demonstrated strong capabilities through simple prompting (OpenAI, 2023; Anil et al., 2023; Qin et al., 2023; Bubeck et al., 2023; Huang and Chang, 2023), in addition to the fine-tuned model, we also prompt ChatGPT (OpenAI, 2022) to generate a short passage to summarize the relationship from the retrieved sentences. Detailed steps for fine-tuning and the prompt for ChatGPT can be found in Appendix B & C. By reading the generated relation descriptions first, users can get a general idea of the relation between the entities and then decide whether to read the retrieved sentences or not.

5 System Demonstration & Evaluation

In this section, we first evaluate the relation synthesis model by assessing the faithfulness of the generation with respect to the input relation descriptions. Then we demonstrate our system's capacity in discovering unknown knowledge and locating information of interest with a drug repurposing task and a literature curation task respectively.

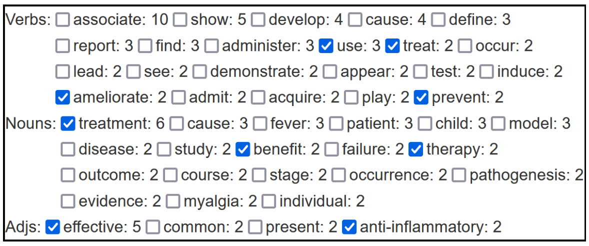
5.1 Relation Synthesis Model Evaluation

Huang et al. (2022b) have demonstrated the capability of the relation synthesis model to generate easily understandable relation descriptions. However, in the biomedical domain, it is crucial for the model to generate truthful sentences and not mislead the reader with erroneous information. Table 4 provides an example of the model's generation, where the extracted relation descriptions for (*COVID-19*, *Pneumonia*) and (*Pneumonia*, *Vaccines*) are the inputs to the model. The 1-hop relation summary is the summarized relation description over the sen-



Verbs: ☑ associate: 17 ☑ cause: 15 ☑ lead: 9 ☑ identify: 8 ☑ result: 5
□ spread: 5 ☑ affect: 4 □ know: 4 □ name: 4 ☑ emerge: 4 □ report: 3
□ term: 3 □ present: 3 ☑ characterize: 3 ☑ increase: 3 ☑ appear: 3
□ bring: 2 □ progress: 2 ☑ diagnose: 2 □ find: 2 □ occur: 2 □ become: 2

Figure 3: Verb Modifiers between *COVID-19* and *Disease* or *Symptom*. "Correlation" related modifiers are checked.



Verbs: □ associate: 10 □ show: 5 □ develop: 4 □ cause: 4 □ define: 3
□ report: 3 □ find: 3 □ administer: 3 ☑ use: 3 ☑ treat: 2 □ occur: 2
□ lead: 2 □ see: 2 □ demonstrate: 2 □ appear: 2 □ test: 2 □ induce: 2
☑ ameliorate: 2 □ admit: 2 □ acquire: 2 □ play: 2 ☑ prevent: 2
Nouns: ☑ treatment: 6 □ cause: 3 □ fever: 3 □ patient: 3 □ child: 3 □ model: 3
□ disease: 2 □ study: 2 ☑ benefit: 2 □ failure: 2 ☑ therapy: 2
□ outcome: 2 □ course: 2 □ stage: 2 □ occurrence: 2 □ pathogenesis: 2
□ evidence: 2 □ myalgia: 2 □ individual: 2
Adjs: ☑ effective: 5 □ common: 2 □ present: 2 ☑ anti-inflammatory: 2

Figure 4: Modifiers between COVID-19 related *Disease* or *Symptom* and *Chemicals* or *Drugs*. "Treatment" related modifiers are checked.

tences of one pair of entities, and the 2-hop relation summary is the synthesized relation description for (*COVID-19*, *Vaccines*) through aggregating the 2-hop path (*COVID-19*, *Pneumonia*, *Vaccines*).

To evaluate the model's faithfulness, the authors of this work used the model to generate relation descriptions for 20 randomly selected samples from the test dataset, tried to find supporting evidence from the input and gave a score from 1 to 5 for each generation to indicate its faithfulness to the input. The final average score for the 20 samples is 4.10, indicating that the generation is generally supported by the input. However, there is still a gap before we can fully trust it and we suggest users read the retrieved sentences to acquire reliable knowledge and only use the generated relation description as a reference.

5.2 Case Study 1: Drug Repurposing

Drug repurposing intends to identify new uses for drugs that were originally used to treat other diseases. *CovidDEER* can aid researchers in identifying candidate drugs through the following steps:

- Set the target disease as the starting node.
- Search the first-hop neighborhood for diseases and symptoms related to the target disease.
- Search the second-hop neighborhood for drugs used to treat those related diseases and symptoms.

Suppose a researcher wants to discover the candidate drugs for COVID-19. By setting *COVID-19* as the starting node and the *Diseases and Symptoms* entity type as the first-hop neighbors, the system retrieved a set of disease or symptom entities and the frequent modifiers. We select several verb modifiers that might indicate a "correlation" relationship between the entities and COVID-19. Then, we update the first-hop neighbors to 10 of these "correlated" entities and set 5 *Pharmacologic substance* related entity types as the second-hop neighbors. The retrieved two-hop graph can be seen in Figure 2. Similarly, we select several modifiers that might indicate a "treatment" relation to find candidate drugs. The selected modifiers are shown in Figures 3 and 4 and the collected candidate drugs are listed in Table 3.

Candidate drugs
nitric oxide, lamb preparation, beta-Lactams, Leukotriene B4, sphingosine 1-phosphate, amoxicillin, Macrolide Antibiotics, Macrolides, beta-Lactams, rifampin, Hydroxymethylglutaryl-CoA Reductase Inhibitors, methylprednisolone, trivalent influenza vaccine, Fibrates, lipid modifying drugs, plain, Corticosteroid ophthalmologic and otologic preparations, metformin, inhibitors, Corticosteroid otologicals, Bilirubin, Fibrates, nitazoxanide, atorvastatin, Artemisinins, antagonists

Table 3: Collected candidate drugs for COVID-19 treatment.

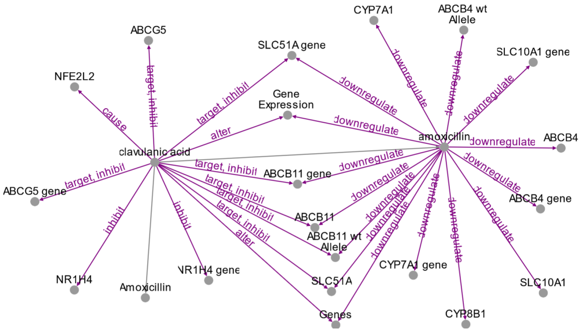


Figure 5: The local graph built from the passage in PMID: 34767876 with *Chemicals* and *Gene* entities.

5.3 Case Study 2: Literature Curation

Literature curation (Wiegers et al., 2009) is the process of identifying documents relevant to a task or topic and locating and annotating the content of interest in these documents. The latter requires the curator to read through the whole document, which can be tedious and time-consuming. Our system provides an interface where users can run the pipeline in Figure 1 to build a DEER on any article indexed in PubMed and use this graph to locate the information relevant to the curation target.

Suppose a curation task is to collect Drug-Target interactions and a curator is assigned a relevant article. The curator first submits the article's PubMed reference number (PMID) or PMC Identifier (PM-

	(COVID-19, Pneumonia)	(Pneumonia, Vaccines)
Extracted relation descriptions	**Coronavirus disease 2019** (COVID-19) is a novel type of highly contagious **pneumonia** caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Conversely, SARS-CoV, MERS-CoV, and **COVID-19** may initially present asymptomatically, but can progress to **pneumonia**, shortness of breath, renal insufficiency and, in some cases, death.	Despite the availability of safe and effective antibiotics and **vaccines** for treatment and prevention, **pneumonia** is a leading cause of death worldwide and the leading infectious disease killer. Despite advances in managerial practices, **vaccines**, and clinical therapies, **pneumonia** remains a widespread problem and methods to enhance host resistance to pathogen colonization and pneumonia are needed.
1-hop relation summary	**COVID-19** is a highly contagious **pneumonia** caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2).	Despite the availability of safe and effective antibiotics and **vaccines** for treatment and prevention, **pneumonia** remains a major cause of death worldwide.
2-hop relation summary (COVID-19, Vaccines)	**COVID-19** is a major cause of death worldwide, despite the availability of safe and effective antibiotics and **vaccines** for treatment and prevention of pneumonia.	

Table 4: Example of relation description extracted or generated by the relation synthesis model.

CID) to the interface. The interface will return a DEER built from the article and a list of entity types appear in the DEER. Then the curator can read the entity types and select the types that are relevant to the task, for instance, *Chemicals* and *Gene*, and refresh the DEER with only entities of the selected types remaining. An example DEER built from PMID 34767876 with the above three types of entities remaining and some modifiers on the edges is shown in Figure 5. From the graph, the curator can easily see the possible Drug-Target interactions in the passage and click the edges to verify the information. The content of PMID 34767876 is placed in Appendix A.

6 Related Work

Exploration System Exploration systems aim to help users learn the content in the data sources through simple queries (Wang and Lo, 2021). Some systems are designed to retrieve sentence-level text pieces for a specific need. Wang et al. (2020b) retrieve textual evidence that semantically matches the queried statement. Lahav et al. (2022) build a set of scientific challenges and directions extracted from a corpus and retrieves challenges and directions through entity co-occurrence. Taub-Tabib et al. (2020) develop a lightweight query language to retrieve sentences that syntactically match an example sentence. In contrast, our system collects relational sentences into a graph structure and displays the retrieved sentences in a graphic view that shows the connection between the text pieces, which is not seen in previous works.

Literature-Based Discovery Literature-based discovery (LBD) tools aim to discover unknown knowledge and generate new hypotheses by connecting current knowledge scattered in different literature together (Swanson, 2008), which is commonly used in biomedical tasks like drug repurposing and interaction prediction. Early LBD tools (Swanson, 1986; Smalheiser and Swanson, 1996) require manual effort in organizing information from the passages. Recent studies (Pu et al., 2023) approach LBD as a link prediction task over knowledge bases, where new knowledge is discovered as predicting new links between concepts. Our system is more like the early LBD tools. With the high-quality relational sentences and the advance of LLM in language understanding, our system greatly alleviates the user's reading workload and allows a rough verification of the generated hypothesis based on the retrieved sentences.

7 Conclusion

In this work, we developed an exploration system in the biomedical domain that operates on a COVID-related corpus facilitating efficient retrieval of relational knowledge and enabling tasks such as drug repurposing and literature curation. We demonstrate the advantages of managing a raw text corpus in a descriptive knowledge graph, including streamlined management, support for multi-hop reasoning across sentences from various articles, and comprehensive visualization of entity connections in the domain. Additionally, we equipped users with a modifier filtering module and a relation synthesis model that offer an overview of the relations on the edge before reading. In future work, we aim to enhance the accuracy and reliability of the relation descriptions generated for user reference.

Limitations

Our system currently only support at most 2-hop query, since the number of entities in the graph will grow exponentially as the path gets longer, which will cause difficulty in reading the graph and reasoning along the path. This hinders our system from studies of more complex network like biochemical pathways, which involves several steps of reaction, and limits the possible knowledge that could be discovered by the system.

The relation synthesis model is trained to generate a single relation description, which is sometimes incapable to cover all the necessary information about the relationship of the target entities. ChatGPT could generate a short paragraph with more details included, but it is more costly than running a local fine-tuned model.

Acknowledgements

This material is based upon work supported by the National Science Foundation IIS 16-19302 and IIS 16-33755, Zhejiang University ZJU Research 083650, IBM-Illinois Center for Cognitive Computing Systems Research (C3SR) and IBM-Illinois Discovery Accelerator Institute (IIDAI), grants from eBay and Microsoft Azure, UIUC OVCR CCIL Planning Grant 434S34, UIUC CSBS Small Grant 434C8U, and UIUC New Frontiers Initiative. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the funding agencies.

References

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Meng-Han Chung, Jun Zhou, Xiaodong Pang, Yuchuan Tao, and Jinfeng Zhang. BioKDE: a Deep Learning Powered Search Engine and Biomedical Knowledge Discovery Platform.

Abram Handler and Brendan O'Connor. 2018. Relational summarization for corpus analysis. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1760–1769.

Jie Huang, Kevin Chang, Jinjun Xiong, and Wen-mei Hwu. 2022a. Open relation modeling: Learning to define relations between entities. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 297–308, Dublin, Ireland. Association for Computational Linguistics.

Jie Huang and Kevin Chen-Chuan Chang. 2022. Ver: Learning natural language representations for verbalizing entities and relations. *ArXiv preprint*, abs/2211.11093.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics.

Jie Huang, Yifan Gao, Zheng Li, Jingfeng Yang, Yangqiu Song, Chao Zhang, Zining Zhu, Haoming Jiang, Kevin Chen-Chuan Chang, and Bing Yin. 2023. Ccgen: Explainable complementary concept generation in e-commerce. *arXiv preprint arXiv:2305.11480*.

Jie Huang, Kerui Zhu, Kevin Chen-Chuan Chang, Jinjun Xiong, and Wen-mei Hwu. 2022b. DEER: Descriptive knowledge graph for explaining entity relationships. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6686–6698, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Gautier Izacard and Edouard Grave. 2020. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, pages 874–880.

Dan Lahav, Jon Saad Falcon, Bailey Kuehl, Sophie Johnson, Sravanthi Parasa, Noam Shomron, Duen Horng Chau, Diyi Yang, Eric Horvitz, Daniel S. Weld, and Tom Hope. 2022. A Search Engine for Discovery of Scientific Challenges and Directions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):11982–11990.

Sharon Levy, Kevin Mo, Wenhan Xiong, and William Yang Wang. 2021. Open-Domain Question-Answering for COVID-19 and Other Emergent Domains. *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 259–266.

Chenzhengyi Liu, Jie Huang, Kerui Zhu, and Kevin Chen-Chuan Chang. 2023. DimonGen: Diversified generative commonsense reasoning for explaining concept relationships. In *Proceedings of the 61st Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 4719–4731, Toronto, Canada. Association for Computational Linguistics.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and robust models for biomedical natural language processing. In *BioNLP 2019 - SIGBioMed Workshop on Biomedical Natural Language Processing, Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327. Association for Computational Linguistics (ACL).

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. *OpenAI*.

OpenAI. 2023. Gpt-4 technical report.

Yiyuan Pu, Daniel Beck, and Karin Verspoor. 2023. Graph embedding-based link prediction for literature-based discovery in alzheimer's disease. *Journal of Biomedical Informatics*, page 104464.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Jiaming Shen, Jinfeng Xiao, Xinwei He, Jingbo Shang, Saurabh Sinha, and Jiawei Han. 2018. Entity Set Search of Scientific Literature: An Unsupervised Ranking Approach. *41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018*, pages 565–574.

Neil R Smalheiser and Don R Swanson. 1996. Indomethacin and alzheimer's disease. *Neurology*, 46(2):583–583.

D. R. Swanson. 2008. *Literature-Based Discovery? The Very Idea*, pages 3–11. Springer Berlin Heidelberg, Berlin, Heidelberg.

Don R Swanson. 1986. Fish oil, raynaud's syndrome, and undiscovered public knowledge. *Perspectives in biology and medicine*, 30(1):7–18.

Hillel Taub-Tabib, Micah Shlain, Shoval Sadde, Dan Lahav, Matan Eyal, Yaara Cohen, and Yoav Goldberg. 2020. Interactive Extractive Search over Biomedical Corpora. *BioNLP*, pages 28–37.

Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R. Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2021. Trec-covid: Constructing a pandemic information retrieval test collection. *SIGIR Forum*, 54(1).

Lucy Lu Wang and Kyle Lo. 2021. Text mining approaches for dealing with the rapidly expanding literature on COVID-19. *Briefings in Bioinformatics*, 22(2):781–799.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Nancy Xin Ru Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020a. CORD-19: The COVID-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Qingyun Wang, Manling Li, Xuan Wang, Nikolaus Parulian, Guangxing Han, Jiawei Ma, Jingxuan Tu, Ying Lin, Ranran Haoran Zhang, Weili Liu, Aabhas Chauhan, Yingjun Guan, Bangzheng Li, Ruisong Li, Xiangchen Song, Yi Fung, Heng Ji, Jiawei Han, Shih-Fu Chang, James Pustejovsky, Jasmine Rah, David Liem, Ahmed ELsayed, Martha Palmer, Clare Voss, Cynthia Schneider, and Boyan Onyshkevych. 2021. COVID-19 literature knowledge graph construction and drug repurposing report generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 66–77, Online. Association for Computational Linguistics.

Xuan Wang, Yingjun Guan, Weili Liu, Aabhas Chauhan, Enyi Jiang, Qi Li, David Liem, Dibakar Sigdel, John Caufield, Peipei Ping, and Jiawei Han. 2020b. 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, Online. Association for Computational Linguistics.

Thomas C. Wiegers, Allan P. Davis, K. Bretonnel Cohen, Lynette Hirschman, and Carolyn J. Mattingly. 2009. Text mining and manual curation of chemical-gene-disease networks for the Comparative Toxicogenomics Database (CTD). *BMC Bioinformatics*, 10(1):326.

A Content of PMID 34767876

> Molecular mechanisms of hepatotoxic cholestasis by clavulanic acid: Role of NRF2 and FXR pathways.
>
> Treatment of beta-lactamase positive bacterial infections with a combination of amoxicillin (AMOX) and clavulanic acid (CLAV) causes idiosyncratic drug-induced liver injury (iDILI) in a relevant number of patients, often with features of intrahepatic cholestasis. This study aims to determine serum bile acid (BA) levels in amoxicillin/clavulanate (A+C)-iDILI patients and to investigate the mechanism of cholestasis by A+C in human in vitro hepatic models. In six A+C-iDILI patients, significant elevations of serum primary conjugated BA definitely demonstrated A+C-induced cholestasis. In cultured human Upcyte hepatocytes and HepG2 cells, CLAV was more cytotoxic than AMOX, and, at subcytotoxic concentrations, it altered the expression of more than 1,300 genes. CLAV, but not AMOX, downregulated the expression of key genes for BA transport (BSEP, NTCP, OSTalpha and MDR2) and synthesis (CYP7A1 and CYP8B1). CLAV also caused early oxidative stress, with reduced GSH/GSSG ratio, along with induction of antioxidant nuclear factor erythroid 2-related factor 2 (NRF2) target genes. Activation of NRF2 by sulforaphane also resulted in downregulation of NTCP, OSTalpha, ABCG5, CYP7A1 and CYP8B1. CLAV also inhibited the BA-sensor farnesoid X receptor (FXR), in agreement with the downregulation of FXR targets BSEP, OSTalpha and ABCG5. We conclude that CLAV, the culprit molecule in A+C, downregulates several key biliary transporters by modulating NRF2 and FXR signaling, thus likely promoting intrahepatic cholestasis. On top of that, increased ROS production and GSH depletion may aggravate the cholestatic injury by A+C.

B Relation Synthesis Model Fine-tuning

As the CORD-19 dataset we use for demonstration is not large enough to train a relation synthesis model, we collected a training, validation, and test dataset from a subset of articles randomly selected from PubMed. All the target sentences have an RDS score greater than 0.75 and all the input sentences have an RDS score greater than 0.7. This resulted in a total of 615,561, 12,824, and 12,825 data in the training, validation, and test dataset respectively, which is comparable in size to the one used in DEER. We trained the model for 20 epochs. Other settings are the same as Huang et al. (2022b). In Section 5, we discuss our manual evaluation of the quality of generation.

C Relation Synthesis with ChatGPT

The input to ChatGPT consists of a prompt and an overall relation context.

- The prompt is “Given the context below, describe the relation between [target_entity_1] and [target_entity_2] in one sentence.” with [target_entity_1] and [target_entity_2] replaced by the target entity pair.
- The overall relation context is formed by concatenating the relation context on each edge with a new line character. The relation context on each edge starts with a description “Relation between [head_entity] and [tail_entity]:
  n” with [head_entity] and [tail_entity] replaced by the head and tail entity of the edge and is followed by the top 5 sentences on the edge sorted by the RDS score and concatenated by the new line character.

Using the sentences in Table 4 as an example, the input for ChatGPT is

> Given the context below, describe the relation between COVID-19 and Vaccines in one sentence.
>
> Relation between COVID-19 and Pneumonia:
> Coronavirus disease 2019 (COVID-19) is a novel type of highly contagious pneumonia caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2).
> Conversely, SARS-CoV, MERS-CoV, and COVID-19 may initially present asymptomatically, but can progress to pneumonia, shortness of breath, renal insufficiency and, in some cases, death.
>
> Relation between Pneumonia and Vaccines:
> Despite the availability of safe and effective antibiotics and vaccines for treatment and prevention, pneumonia is a leading cause of death worldwide and the leading infectious disease killer.
> Despite advances in managerial practices, vaccines, and clinical therapies, pneumonia remains a widespread problem and methods to enhance host resistance to pathogen colonization and pneumonia are needed.