An Interactive Knowledge Graph Based Platform for COVID-19 Clinical Research

Juntao Su
sujuntao@gwu.edu
George Washington University

Edward T. Dougherty
edougherty@rwu.edu
Roger Williams University

Shuang Jiang, Fang Jin
sjiang97,fangjin@gwu.edu
George Washington University

ABSTRACT

Since the first identified case of COVID-19 in December 2019, a plethora of pharmaceuticals and therapeutics have been tested for COVID-19 treatment. While medical advancements and breakthroughs are well underway, the sheer number of studies, treatments, and associated reports makes it extremely challenging to keep track of the rapidly growing COVID-19 research landscape. While existing scientific literature search systems provide basic document retrieval, they fundamentally lack the ability to explore data, and in addition, do not help develop a deeper understanding of COVID-19 related clinical experiments and findings. As research expands, results do so as well, resulting in a position that is complicated and overwhelming. To address this issue, we present a named entity recognition based framework that accurately extracts COVID-19 related information from clinical test results articles, and generates an efficient and interactive visual knowledge graph. This knowledge graph platform is user friendly, and provides intuitive and convenient tools to explore and analyze COVID-19 research data and results including medicinal performances, side effects and target populations.

CCS CONCEPTS

• **Computing methodologies** → **Knowledge representation and reasoning**.

KEYWORDS

knowledge graph, named entity recognition, COVID-19, clinical results

ACM Reference Format:
Juntao Su, Edward T. Dougherty, and Shuang Jiang, Fang Jin. 2022. An Interactive Knowledge Graph Based Platform for COVID-19 Clinical Research. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining (WSDM '22), February 21–25, 2022, Tempe, AZ, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3488560.3502193

1 INTRODUCTION

COVID-19 has now spread around the entire world. Individuals affected by this global pandemic have demonstrated a tremendously wide range of symptoms, extending from mild effects to more severe and life-threatening illnesses. A host of medications are still being clinically evaluated for approval to treat this coronavirus disease. While these treatments are still under clinical evaluation, both these as well as those already approved by the FDA for COVID-19 treatment have contributed to an immense amount of articles and reports that document these treatments and their results; not surprisingly, the efficacy of these treatments vary tremendously. While these documents provide a comprehensive collection of treatment-effect data, a major issue that has percolated is the inability for researchers to efficiently and accurately mine these data with the goal of extracting desired information that will facilitate both treatment efficacy as well as therapeutic advancements. To support the biomedical research community's capability to explore these existing results, we have constructed a named entity recognition (NER) based knowledge graph generation tool to efficiently and accurately query, extract, present and analyze these data within an intuitive, comprehensive, and interactive platform.

Knowledge graphs [14] were first introduced by Google in 2012 to interlink data and assist search queries. In such a graph, the nodes represent general entities, for instance, objects, concepts or even forms of data. Node edges on the other hand represent semantic relations between the nodes. In our case, the nodes depict clinical problems, tests and treatments, and edges are the interconnections among these data. In the approaches and tools presented in this work, both the nodes and relations are automatically extracted from COVID-19 clinical documents and reports, and also retain the source of the information. Thus, if a researcher is interested in the performance of a particular pharmaceutical to treat a patient with severe COVID-19 symptoms, for example, the interconnected relations between one node representing that drug and another node representing patients with severe symptoms could inform the researcher of (i) current and anticipated treatment results and (ii) information about the source article of these data.

A fundamental aspect in the construction of a knowledge graph is the assurance that knowledge extraction is relevant, accurate and efficient. For example, many clinical articles tend to describe the background information, study design, and/or statistical methods in detail, however, this is not the information that we seek in this work. Rather, we want to facilitate the querying and exploring of current research approaches, progress and results; to accomplish this, the extracted knowledge will therefore be used as an index for the article. In addition, many current knowledge graph generation tools simply extract information aimlessly which does not necessarily fit the needs of the users of the knowledge graph. Thus, in this work we synthesize the text summarization technique with a clinical named entity recognition (NER) method to efficiently and accurately extract relevant knowledge from clinical articles.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
WSDM '22, February 21–25, 2022, Tempe, AZ, USA.
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9132-0/22/02…$15.00
https://doi.org/10.1145/3488560.3502193

The techniques, methods, and tools in this work provide an accurate, efficient and seamless framework that addresses the problem of biomedical and clinical researches having an inability to mine COVID-19 treatment approaches and results with efficacy. Our contributions are:

- constructing a tool for generating knowledge graph from COVID-19 articles that integrates the text summarization technique and named entity recognition methods,
- designing a knowledge graph with the extracted knowledge from COVID-19 clinical articles which is easy to search and explore for medical treatments and related knowledge, and
- building a COVID-19 tracking platform that presents an intuitive and interactive knowledge graph with many useful tools for researchers to analyze their data.

2 RELATED WORK

2.1 Knowledge Graph Generation

A knowledge graph is an information representation that uses a graph-structured topological model to interconnect data. A fundamental aspect of generating knowledge graphs is the identification of entities and relations. In practice, there are countless methods to generate knowledge graphs from a knowledge base or source. Exner and Nugues [7] integrated Named Entity Recognition combined with Semantic Role Labeling techniques to assign the named entities as either subject or object via verb role identification. Further, the T2KG tool [10] generated knowledge graphs utilizing a hybrid of a rule-based approach with a vector-based similarity metric to identify similar mappings.

2.2 Biomedical Named Entity Recognition

Entity recognition is the foundational phase within knowledge graph generation. The NER method, in particular, recognizes mention spans of a particular entity type, e.g. Person or Organization, in the input sentence. As its name implies, Biomedical Named Entity Recognition (BioNER) is the NER method trained with a biomedical based dataset. This method can be utilized to recognize biomedical entities such as illness or medicine with higher accuracy [8]. Within BioNER, modeling methods can be divided into four categories: Rule-based, Dictionary-based, Machine Learning based, and Hybrid models, each with distinct benefits dependent on application and dataset [6]. In recent years, research focus has heavily shifted towards either pure machine learning approaches or hybrid techniques that combine rules and dictionaries with machine learning methods [11].

2.3 Relation extraction

After BioNER is applied, the identification of relations between the biomedical named entities follows. For establishing such associations, the majority of studies use approaches that are either co-occurrence based, rule-set based, or machine learning based [15]. In a rule-based approach, for example, the relationships extracted depend highly on the syntactic and semantic analysis of sentences. On the other hand, these dependencies diminish with more dynamic approaches from, for example, machine learning based techniques. In practice, the dependency parser of some common Natural Language Processing tool can be used to extract the relations between recognized entities.

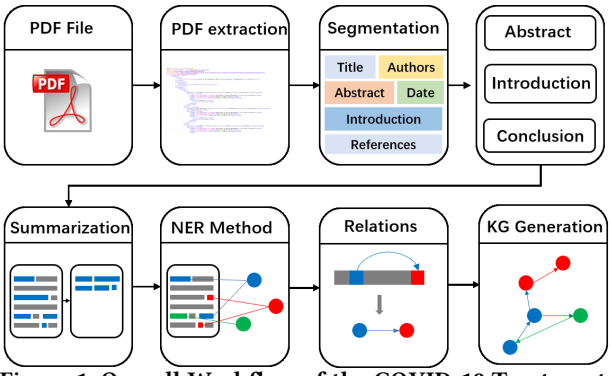


Figure 1: Overall Workflow of the COVID-19 Treatment Information System

3 METHODS

3.1 Overall Workflow

A workflow diagram of our COVID-19 Treatment information system is shown in Fig 1. To begin, we collect raw documents from the aforementioned COVID-19 clinical dataset. Second, we extract the structured XML/TEI encoded documents and delete redundant parts. Third, we implement the text summarizing method to obtain a summary, and extract the entities and relationships with the clinical NER model from this summary and disambiguate entities and relations. The extracted entities and relationships are then saved to specially formatted files and imported into the Neo4j non-relational database; it is worth noting that NoSQL databases, like the Neo4j database, have flexibility, scalability, and efficiency advantages, and in addition, possess the capacity to handle large amounts of unstructured, semi-structured, and structured data, as compared to their traditional SQL database counterparts. And finally, we complete the storage of knowledge, optimize the graph structure, and upload it to our online platform.

3.2 Extraction, Segmentation and Summarization

As shown in Fig 1, the raw PDF documents need to be extracted and segmented in the first step. To accomplish this, we use GROBID [1], which is a machine learning library for extracting, parsing and re-structuring raw PDF documents; GROBID is designed with a particular focus on technical and scientific publications, making it a natural choice for this workflow. With GROBID, we extract raw documents into structuredXML/TEI encoded documents, and then delete the less relevant parts including references and acknowledgements sections.

The next step is to summarize the most informative sentences with text summarization methods. The spaCy library [9], which encapsulates advanced Natural Language Processing methods in Python and Cython, was used for this task. spaCy was also used to clean the text, which includes removing stop words and punctuation

Figure 2: Example of Stanza Clinical NER.

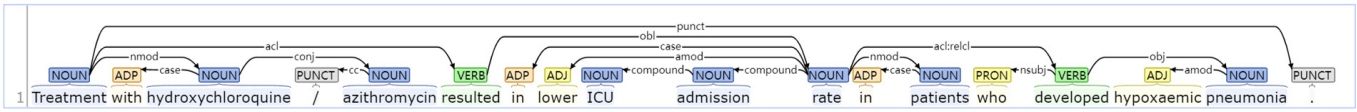


Figure 3: Example of Stanza Universal Dependencies.

marks, and changing all text to lower case. Then with spaCy's tokenizer, we implemented the sequence analysis to finally construct the summarization.

3.3 Clinical Named Entity Recognition

For the COVID-19 clinical test results articles, the most important elements in each sentence for this work are the medical issues to be solved and the corresponding treatments. To do this, we use Stanza's [12] named entity recognition component to extract this information. Stanza NER is adopted from the contextualized string representation-based sequence tagger. In practice, for each domain, Stanza has been trained with both a forward and backward LSTM character-level language model (CharLM) to supplement the word representation in each sentence. At tagging time, it concatenates the representations from these CharLMs at each word position with a word embedding, and feeds the results into a standard one-layer Bi-LSTM sequence tagger with a conditional random field (CRF)-based decoder.

The resulting pretrained CharLMs provide rich domain-specific representations that notably improve the accuracy of the NER models. One of the domain-specific NER models of Stanza is the i2b2 clinical model [16]. It is this model in particular that is utilized for extracting problem, test and treatment entities from various types of clinical notes, reports, and articles. As seen in Fig 2, ''hydroxychloroquine' is recognized as a treatment entity while both 'malaria' and 'mosquito bites' are recognized as problem entities.

3.4 Relation Extraction Based on Dependencies

To analyze the syntactic structure of each sentence, Stanza parses them into the Universal Dependencies (UD) format, where each word in the sentence is assigned a syntactic head that is either another word in the sentence, or in the case of the root word, an artificial root symbol. The dependency parser in Stanza is a variant of the Bi-LSTM-based deep biaffine neural dependency parser. We use the rule-based approach to extract the relations by scanning for verbs and prepositions that correlate two or more nouns or phrases serving as named entities. Fig 3 shows the conclusion of Dubernet et al [5] in Stanza's dependency parser, which builds a tree structure of words from the input sentence and represents the syntactic dependency relations between words. For example, the relation between entity 'hydroxychloroquine/azithromycin' and 'lower ICU admission rate' is 'resulted' (Fig 3).

4 EXPERIMENTS

4.1 Dataset

The dataset of potential pharmaceuticals for COVID-19 treatment compiled by the AIM (Artificial Intelligence in Medicine) [2] tool was used. The use of artificial intelligence for this task makes processing an enormous amount of scientific data possible. The AIM-API gives access to information in a format suitable for processing in JSON format, and for a particular COVID-19 related article, we reliably obtain title, impact factor, publication date, as well as the URL to the article. By sorting and viewing these articles, we translate these raw documents into the knowledge graphs.

4.2 Knowledge Graph Demonstration

By importing the extracted information to Neo4j, the structured data of the relations between clinical entities is obtained. Figure 4 demonstrates a visualization of the culminating knowledge graph interactive website from our framework on three clinical articles [3][4][13], and shows the clinical results on a certain drug named Hydroxychloroquine (HCQ) sulfate. Here, red nodes represent clinical problems, like coronavirus or renal dysfunction, yellow nodes represent the tests used in the articles, and green nodes represent the treatment, which may include pharmaceuticals and their ingredients.

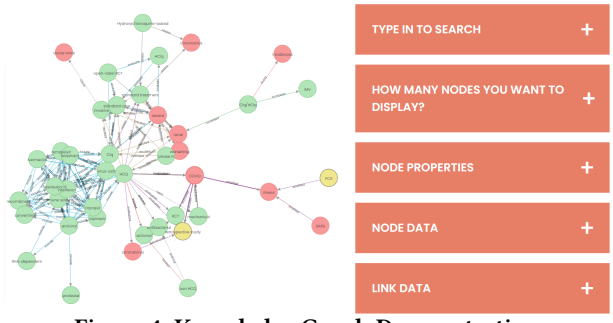


Figure 4: Knowledge Graph Demonstration

4.3 Querying Demonstration

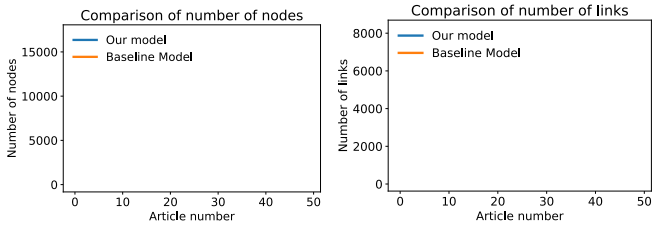
As a next step in analysis, nodes and relations within the knowledge graph can then be queried for a more in-depth investigation. Fig 5 shows the results of querying the HCQ node. The framework provides two methods to show the results, namely a graphical format (Fig 5).The querying results show that the extracted relations contain useful information like 'HCQ resulted in a significant worsening of clinical status', which is precisely the fundamental conclusion of Réa-Neto et al [13]. This example also provides verification that the framework and tool successfully extracted COVID-19 information from the clinical articles.

4.4 Comparison with Baseline tool

A fundamental advantage of our model is the accuracy by which the knowledge graph finds and refines entities and relations. Specifically, integrating clinical NER with text summarization provides

Figure 5: Querying Demo - Graphic Format

a comprehensive knowledge graph with exhaustive and thorough relation generation. To provide a comparison, we also design a baseline model which did not use the clinical NER and the text summarization method. The number of nodes in the baseline tool is much larger than our tool generates. In addition, as the number of nodes increases, the structure of the relations will eventually break down. Further, without a specific goal, the extracted entities and relations may be completely irrelevant to the key information that is desired. These observations are substantiated in Fig 6. Fig 6(a) shows that as more clinical test articles are used, the number of extracted nodes in the baseline tool increases at a rate much faster than our model. The primary reason for why our model does not improperly grow in this fashion is the fact that the clinical NER method ensures that the extracted nodes are in fact clinical terms. Fig 6(b) shows a similar comparison, and results, with relations. Notably, with a significantly lower amount of nodes, the number of relations of our model is consistently less than half of that produced by the baseline model.



(a) Nodes (b) Relations

Figure 6: Comparison with Baseline Model

4.5 COVID-19 Tracking Platform

The knowledge graph is brought online for use by the research community in the form of a web-based COVID-19 Tracking Platform. To provide an efficient and intuitive interface for users to visualize and explore COVID-19 data, the page is divided into two parts. As shown in Fig 4 and Fig 5, the left part contains a canvas to show the knowledge graph, and the right part provides tool boxes to interact with the knowledge graph. Various ways are provided to interact with the graph including hovering, clicking and double clicking. One handy tool is a search bar that allows users to enter keywords, and search results are returned in the graph window. In addition, "Node properties" shows all of the information of the node that is currently focused on. The fully functioning platform is provided to the research community and can be accessed at https://www.covid19kgd3.com/vacc_info_drug.html.

5 CONCLUSION AND DISCUSSION

In this paper, we have presented an NER-based approach for constructing a knowledge graph based tool to support COVID-19 clinical research exploration. The approach consists of four parts: text extraction, text summarization, entity/relation extraction, and knowledge graph generation. The resulting tool provides a simple and seamless way to search and explore COVID-19 research from virtually any kind of clinical document, report, or article. Scenario demonstrations showcase the ability of our model to generate knowledge graphs with more efficient nodes while retaining imperative relations. Finally, we built an online COVID-19 tracking platform that presents an intuitive and interactive knowledge graph with numerous useful mechanisms for researchers to explore, analyze, and synthesize their data. It is our hope that this work facilitates medical related research advancements by providing an approach and platform that helps researchers find, explore, and assess the latest and most up-to-date clinical trial treatments and results.

REFERENCES

[1] 2008–2021. GROBID. https://github.com/kermitt2/grobid. arXiv:1:dir:dab86b296e3c3216e2241968f0d63b68e8209d3c

[2] Project AIM. [n.d.]. Up-to-date mapping of COVID-19 treatment and vaccine development. https://covid19-help.org/. Accessed June 6, 2021.

[3] Saleh Alghamdi et al. 2021. Clinical Efficacy of Hydroxychloroquine in Patients with COVID-19: Findings from an Observational Comparative Study in Saudi Arabia. *Antibiotics* 10, 4 (2021), 365.

[4] Cheng-Pin Chen et al. 2020. A multicenter, randomized, open-label, controlled trial to evaluate the efficacy and tolerability of hydroxychloroquine and a retrospective study in adult patients with mild to moderate coronavirus disease 2019 (COVID-19). *PloS one* 15, 12 (2020), e0242763.

[5] Arthur Dubernet et al. 2020. A comprehensive strategy for the early treatment of COVID-19 with azithromycin/hydroxychloroquine and/or corticosteroids: Results of a retrospective observational study in the French overseas department of Réunion Island. *Journal of Global Antimicrobial Resistance* 23 (2020), 1–3.

[6] Safaa Eltyeb and Naomie Salim. 2014. Chemical named entities recognition: a review on approaches and applications. *Journal of cheminformatics* 6, 1 (2014), 1–12.

[7] Peter Exner and Pierre Nugues. 2012. Entity Extraction: From Unstructured Text to DBpedia RDF triples.. In *WoLE@ ISWC*. 58–69.

[8] Robert Gaizauskas et al. 2003. Protein structures and information extraction from biological texts: the PASTA system. *Bioinformatics* 19, 1 (2003), 135–143.

[9] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. *spaCy: Industrial-strength Natural Language Processing in Python.* https://doi.org/10.5281/zenodo.1212303

[10] Natthawut Kertkeidkachorn and Ryutaro Ichise. 2017. T2kg: An end-to-end system for creating knowledge graph from unstructured text. In *Workshops at the Thirty-First AAAI.*

[11] Nadeesha Perera et al. 2020. Named entity recognition and relation detection for biomedical information extraction. *Frontiers in Cell and Developmental Biology* 8 (2020), 673.

[12] Peng Qi et al. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations.* https://nlp.stanford.edu/pubs/qi2020stanza.pdf

[13] Álvaro Réa-Neto et al. 2021. An open-label randomized controlled trial evaluating the efficacy of chloroquine/hydroxychloroquine in severe COVID-19 patients. *Scientific reports* 11, 1 (2021), 1–10.

[14] Amit Singhal. 2012. Introducing the Knowledge Graph: things, not strings. https://blog.google/products/search/introducing-knowledge-graph-things-not. [Online].

[15] Hui Yang et al. 2011. Mining biomedical text towards building a quantitative food-disease-gene network. In *Learning structure and schemas from documents.* Springer, 205–225.

[16] Yuhao Zhang et al. 2020. Biomedical and Clinical English Model Packages in the Stanza Python NLP Library. *arXiv preprint arXiv:2007.14640* (2020).