COVID Information Commons (CIC) Research Lightning Talk

Transcript of a Presentation by Carlos Badenes-Olmedo (Universidad Politecnica de Madrid), July 26, 2023

Title: Drugs4Covid: Knowledge Graph about Drugs used in the Clinical Control of the Coronavirus

NIH Publication: https://pubmed.ncbi.nlm.nih.gov/37156393/

YouTube Recording with Slides

Summer 2023 CIC Webinar Information

Transcript Editor: Lauren Close

Transcript

Slide 1

Carlos Badenes-Olmedo:

Ok, thank you, Lauren. And thank you, Florence for inviting me. Ok, I'm going to share my screen. So, thank you for inviting me to participate in this webinar. I am Carlos Badenos-Olmedo, a researcher from the Ontology Engineering Group and I am also an assistant profit professor in the Universidad Politecnica de Madrid. I am going to present Drugs4COVID project.

This is a project that we propose to create a knowledge graph. This is trying to represent the drugs that we used during the pandemic to define the clinical control of the coronavirus.

Slide 2

The idea is that during the COVID-19 pandemic, worldwide institutions tried to identify or to create data sets with scientific research articles that are related to coronavirus. This kind of information could be useful for hospital pharmacies, for clinical practitioners. Our research group, the Ontology Engineering Group, was trying to promote how can we retrieve information - retrieve knowledge from this kind of data. The first step was to identify what is the most important dataset that we can use. In this case, from the European Union was provided the COVID-19 European data portal, also the [inaudible] COVID data set. Mainly, the COVID-19 open presets data set was the most important data set that combines information from PubMed BioRxiv, MedRxiv, and arXiv information. It also combines information from the World Health

Organization and provides more than 400,000 articles that we need - we thought that we can leverage to provide knowledge. Combining this kind of knowledge with the self-service from Madrid, the [inaudible]. We provide the mechanisms to create knowledge graphs that we can exploit and provide knowledge from this kind of people.

Slide 3

So, first of all, we define a workflow with different steps that provide not only the step but also the recommendations the steps that you can follow to create a final analysis graph and also facilitate the exploitation. We define a six step workflow. The first step is the harvesting. In this step the idea is that you need to identify the data set that is related to coronavirus, and also to evaluate if the data is completely available or not. The second step is the pre-processing because you need to organize the data that is provided, in this case a data set, because this data is not total because we are working with tests. We need to modelize a certain way of [evaluating] the data. Then, we move to the third step, which is information extraction. In this step we need to discover the main elements of this kind of data that we can use to create the analysis graph. In our case, the main elements are the biomedical concepts (for example, the drugs, the diseases, and the genetic information). Then, we need to define a formal description of the domain, which is the fourth step, which is the semantification. In this case, we need to create an ontology to define the relations between the biomedical concepts and to define all the concepts, all the elements, that finally appear in the knowledge graph. The next step is the knowledge graph generation. In this step we need to define the rules to create the instances in the knowledge graph. And finally we can provide the mechanisms to exploit - to facilitate the use of the information that the knowledge graph contains.

Slide 4

So we focus on the first step. The objective is to identify the relevant data sources and also to evaluate the availability of the data. Our proposal is that you need to perform a systematic literature review, taking into account the main concepts of the coronavirus, and then define the data set. For example, from digital repositories, PubMed, BioRxiv, and so on, but also combining with other sources, for example, clinical collections from the coordinating corpus, the lead COVID dataset and also additional resources. For example, patents, encyclopedic articles from Wikipedia. And all this data is organized in this first step.

Slide 5

In the next step, we need to transfer this unstructured data, which are text, into tables, which are structured data. Then, the methodology that we propose is to identify the minimal information you need. The most easy way to transform text into a structured way is to define the full text of the article as the data. In our opinion, this is not the best way to do that and our proposal is to define as minimal information as you need: the paragraph of the articles. In that area you can discover all the references or the relation between the biomedical concepts.

Slide 6

Then the next step is the informational structure. In this case, the idea is to create annotations based on that paragraphs that discover drugs, diseases, and genetic information. In our experience, we fine-tune different language models for each different biomedical concept. The idea is that you need to define a specific language model to identify drugs and also to normalize the drugs according to different vocabularies because in different countries we use different standard codes.

Slide 7

Once we have the annotations with the entities and also the codes we can define the formal space to describe all this information. This is the step that we need to create an ontology. In the biomedical domain there exist a lot of ontologies so the idea is not to create from scratch an ontology. The idea is to reduce system ontologies and to provide the missing information in the new ontology. In our case the ontology was EBOCA and the missing information was to provide the evidence that supports the relationships between drugs, between diseases, and between genetic information. In our case we used the unified medical language system and also the DISNET platform. All this information is combined. We also provide, this is the purple area, the information about the evidence. Which is the evidence? The evidence is the paragraph where the relation between these elements is reported in the scientific article.

Slide 8

Once we have defined the formal domain, the ontology, we need to identify the instances, the claims, the statements retrieved from the scientific articles to create instances in the knowledge graph. This is the knowledge graph generation step so our methodology is proposed to create rules to identify using the previous model language the entities and also the relations between them. Finally, once we have the ontology, we have the instances, we are able to identify the nodes in the graph. For example, the blue ones are the elements, the orange ones are the relations between them, and the purple ones are the evidence that supports these kinds of relations. The evidence is the minimal information unit which is the paragraph and also the articles.

Slide 9

Once we have the knowledge graph, finally, we are available to facilitate the exploitation of the information. The best - the first step is of course we can use SPARQL queries - this is a specific language model that you can create queries to exploit the language, the analysis graph. This requires an expert in this kind of domain.

Slide 10

Our second methodology is to create a question-answer interface that provides the information not only from the knowledge graph but also combining from external sources, from others and knowledge graphs from others, document connections and then support questions in natural language to provide answers, also in natural language. So our platform is for a COVID platform.

Slide 11

All this information is policy available: the knowledge graphs, the models, the data sets, and also the services are - these resources are completely free, are completely public. These are available from these URLs. Thanks for your attention and I'm able to answer any questions that you have.