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: Julie Meunier

Transcript

Slide 1

D'accord, merci Lauren. Et merci Florence de m'avoir invité. Je vais partager mon écran. Merci de m'avoir invité à participer à ce webinaire. Je suis Carlos Badenos-Olmedo, chercheur au sein de l'Ontology Engineering Group et je suis également professeur assistant à l'Universidad Politecnica de Madrid. Je vais présenter le projet Drugs4COVID.

Slide 2

L'idée est que pendant la pandémie de COVID-19, les institutions mondiales ont essayé d'identifier ou de créer des ensembles de données avec des articles de recherche scientifique liés au coronavirus. Ce type d'information pourrait être utile aux pharmacies des hôpitaux et aux praticiens cliniques. Notre groupe de recherche, l'Ontology Engineering Group, a essayé de promouvoir la manière dont nous pouvons extraire des informations - extraire des connaissances à partir de ce type de données. La première étape a consisté à identifier l'ensemble de données le plus important que nous puissions utiliser. Dans ce cas, l'Union européenne a fourni le portail de données européen COVID-19, ainsi que l'ensemble de données COVID [inaudible]. L'ensemble de données COVID-19 open presets est l'ensemble de données le plus important, qui combine des informations provenant de PubMed BioRxiv, MedRxiv et arXiv. Il combine également des informations de l'Organisation mondiale de la santé et fournit plus de 400 000 articles dont nous avons besoin - nous avons pensé que nous pouvions les exploiter pour fournir des connaissances. En combinant ce type de connaissances avec le libre-service de Madrid, le [inaudible]. Nous fournissons les mécanismes permettant de créer des graphes de connaissances que nous pouvons exploiter et fournir des connaissances à partir de ce type de personnes.

Slide 3

Ainsi, tout d'abord, nous définissons un flux de travail avec différentes étapes qui fournissent non seulement l'étape mais aussi les recommandations que vous pouvez suivre pour créer un graphique d'analyse final et faciliter l'exploitation. Nous définissons un flux de travail en six étapes. La première étape est la récolte. Dans cette étape, l'idée est d'identifier l'ensemble des données relatives au coronavirus et d'évaluer si les données sont entièrement disponibles ou non. La deuxième étape est le prétraitement, car il faut organiser les données fournies, dans ce cas un ensemble de données, car ces données ne sont pas totales puisque nous travaillons avec des tests. Nous devons modéliser une certaine façon d'[évaluer] les données. Ensuite, nous passons à la troisième étape, qui est l'extraction d'informations. Au cours de cette étape, nous devons découvrir les principaux éléments de ce type de données que nous pouvons utiliser pour créer le graphique d'analyse. Dans notre cas, les principaux éléments sont les concepts biomédicaux (par exemple, les médicaments, les maladies et les informations génétiques). Ensuite, nous devons définir une description formelle du domaine, ce qui constitue la quatrième étape, à savoir la sémantisation. Dans ce cas, nous devons créer une ontologie pour définir les relations entre les concepts biomédicaux et pour définir tous les concepts, tous les éléments, qui apparaissent finalement dans le graphe de connaissances. L'étape suivante est la génération du graphe de connaissances. Dans cette étape, nous devons définir les règles pour créer les instances dans le graphe de connaissances. Enfin, nous pouvons fournir les mécanismes d'exploitation - pour faciliter l'utilisation des informations contenues dans le graphe de connaissances.

Slide 4

Nous nous concentrons donc sur la première étape. L'objectif est d'identifier les sources de données pertinentes et d'évaluer la disponibilité des données. Nous proposons de procéder à une analyse systématique de la littérature, en tenant compte des principaux concepts du coronavirus, puis de définir l'ensemble des données. Par exemple, à partir de dépôts numériques, PubMed, BioRxiv, etc., mais aussi en combinant avec d'autres sources, par exemple, les collections cliniques du corpus de coordination, l'ensemble de données principal COVID et également des ressources supplémentaires. Par exemple, des brevets, des articles encyclopédiques de Wikipedia. Toutes ces données sont organisées au cours de cette première étape.

Slide 5

Dans l'étape suivante, nous devons transférer ces données non structurées, qui sont du texte, dans des tableaux, qui sont des données structurées. La méthodologie que nous proposons consiste à identifier les informations minimales dont vous avez besoin. La façon la plus simple de transformer du texte en données structurées est de définir le texte intégral de l'article comme données. À notre avis, ce n'est pas la meilleure façon de procéder et notre proposition consiste à définir l'information minimale dont vous avez besoin : le paragraphe de l'article. Dans cette zone, vous pouvez découvrir toutes les références ou la relation entre les concepts biomédicaux.

Slide 6

L'étape suivante est la structure informationnelle. Dans ce cas, l'idée est de créer des annotations basées sur ces paragraphes qui permettent de découvrir des médicaments, des maladies et des informations génétiques. D'après notre expérience, nous affinons différents modèles de langage pour chaque concept

biomédical. L'idée est qu'il faut définir un modèle de langage spécifique pour identifier les médicaments et aussi pour normaliser les médicaments en fonction de différents vocabulaires, car dans les différents pays, nous utilisons des codes standard différents.

Slide 7

Une fois que nous avons les annotations avec les entités et les codes, nous pouvons définir l'espace formel pour décrire toutes ces informations. C'est à cette étape que nous devons créer une ontologie. Dans le domaine biomédical, il existe de nombreuses ontologies. L'idée n'est donc pas de créer une ontologie à partir de zéro. L'idée est de réduire les ontologies du système et de fournir les informations manquantes dans la nouvelle ontologie. Dans notre cas, l'ontologie était EBOCA et l'information manquante consistait à fournir les preuves des relations entre les médicaments, entre les maladies et entre les informations génétiques. Dans notre cas, nous avons utilisé le système de langage médical unifié et la plate-forme DISNET. Toutes ces informations sont combinées. Nous fournissons également, dans la zone violette, des informations sur les preuves. Qu'est-ce que la preuve ? La preuve est le paragraphe où la relation entre ces éléments est rapportée dans l'article scientifique.

Slide 8

Une fois que nous avons défini le domaine formel, l'ontologie, nous devons identifier les instances, les affirmations, les déclarations extraites des articles scientifiques pour créer des instances dans le graphe de connaissances. Il s'agit de l'étape de génération du graphe de connaissances. Notre méthodologie est donc proposée pour créer des règles permettant d'identifier, à l'aide du langage du modèle précédent, les entités et les relations entre elles. Enfin, une fois que nous avons l'ontologie et les instances, nous sommes en mesure d'identifier les nœuds du graphe. Par exemple, les nœuds bleus sont les éléments, les nœuds orange sont les relations entre eux et les nœuds violets sont les preuves qui soutiennent ces types de relations. La preuve est l'unité d'information minimale qui est le paragraphe et les articles.

Slide 9

Une fois que nous disposons du graphe de connaissances, nous sommes enfin en mesure de faciliter l'exploitation des informations. La meilleure - la première étape est bien sûr d'utiliser les requêtes SPARQL - il s'agit d'un modèle de langage spécifique qui permet de créer des requêtes pour exploiter le langage, le graphe d'analyse. Cela nécessite un expert dans ce type de domaine.

Slide 10

Notre deuxième méthodologie consiste à créer une interface question-réponse qui fournit des informations non seulement à partir du graphe de connaissances, mais aussi en combinant des sources externes, d'autres personnes et des graphes de connaissances d'autres personnes, en documentant les connexions, puis en prenant en charge les questions en langage naturel pour fournir des réponses, également en langage naturel. Notre plateforme est donc une plateforme COVID.

Slide 11

Toutes ces informations sont disponibles : les graphes de connaissances, les modèles, les ensembles de données et les services sont - ces ressources sont entièrement gratuites et publiques. Elles sont disponibles à partir de ces URL. Je vous remercie de votre attention et je suis en mesure de répondre à vos questions.