AUTOMATIC MAINTENANCE OF COVID-19 RELATED KNOWLEDGE GRAPHS BASED ON LARGE-SCALE INFORMATION EXTRACTION ON SCIENTIFIC LITERATURE

by
Argyro (Iro) Sfoungari 6528015
MSc Computing Science, Utrecht University, 2020

Submitted to Utrecht University's Graduate School of Natural Sciences
in partial fulfilment of the requirements for the
Master's degree in Computing Science

Graduate Program in Computing Science
Utrecht University
2020

Approved by
Velegrakis Yannnis - Project Supervisor
Dalamagas Theodore - Host Organisation Supervisor
Chekol Mel - Second Supervisor

Graduate Program in Computing Science
Utrecht University
2020

Contents

1 Abstract

2 Introduction

In December 2019, the first cases of the novel coronavirus disease (COVID-19) appeared in Wuhan City, Hubei Province, China [1]. Coronavirus, which causes severe acute respiratory syndrome, had such strenth that very quickly spread beyond Hubei province, to all of China at first and eventually to the entire world. Both mortality rates and rapid transmission speed were so alarming that on March 11, 2020, the World Health Organization characterized the situation as a global pandemic [2]. Since the beginning of this pandemic, a plethora of publications have appeared and research towards a cure or vaccination will be prevalent globally for some time to come. In fact, according to statistics [3], it turns out that global research has focused extensively on novel coronavirus as scientific literature over this topic is steadily increasing, for instance the week between August 24^{th} and August 30^{th}, the number of publications reached up to 4000. Given the urgency of the situation globally, there is a need for tools that locate with accuracy all related bibliographic data.

The basic assumption before conducting any type of research is that the researchers involved are up to date with developments in their particular field. Then, they need to be aware of the recent progress on a global level which is resulted from the related studies of their colleagues. The remarkable pace of advancement in science requires the involvement and vigilance of experienced researchers who will at the same time be in constant communication with one another for the benefit of all. It is often difficult to examine carefully the limitless amount of information and it is not uncommon for researchers to be disoriented reading research findings that are irrelevant or inaccurate. The document-based form of scientific representation is the only available form of scholarly communication and this seems to be the base of the above-mentioned restriction. Scientific literature involves unstructured text which is difficult for researchers to evaluate and assess. The scientific community is looking for automatic solutions to facilitate scientific text analysis, a difficult task indeed since the form of scientific literature contains terminology, large sentences, many abbreviations, and formulas. Therefore, this kind of text is not machine-understandable and so cannot be processed automatically.

An innovative idea that can supply scientific literature with structure is its representation using Knowledge Graphs (KGs). KGs have been a source of great interest, in recent years, for many researchers working both in academia and industry [4]. Until 2012, the term KG was not particularly popular, at least outside of the scientific circles. The popularization of KGs is principally attributed to Google [5]. The momentum of Google's idea to transfer general knowledge was so powerful [6] that the term KG since then has been widely used while lacking clear definition [4]. Google engineers, by strongly believing that a new era of search has arrived, decided that the web is such a capable

tool that can respond to users' requests with human characteristics. By way of explanation, Google employed KGs to produce information about natural world connections. Processing knowledge in this way not only helps the search engine to better understand the meaning behind the user's queries but also encourages researchers to attain more consistent results [7]. Since 2012, many companies that dominate the specific market use data represented by and processed as graphs. KG in collaboration with other cutting-edge regions (linked data, cloud computing etc.) has generated numerous interesting and innovative applications.

This project inspired by knowledge graphs will attempt to provide assistance to people involved with research regarding COVID-19 which is today's major global health concern. The proposed method involves the employment of knowledge graphs. Since COVID-19 pandemic began to spread, a plethora of publications have appeared and research towards a cure will be prevalent globally for some time to come. In other words, we will attempt to design a system which will explore knowledge graphs (browse and search), so as to assist researchers in finding papers relevant to their interests based on topics, quality, and connections with other papers. The main idea is to turn each paper into a graph using the paper's metadata (e.g. DOI, author list, publication year, venue etc.). The union of all the graphs will result in a big graph the so-called "knowledge graph". Therefore, each paper will be represented as a set of nodes and edges of that graph. The proposed graph will be continuously updated as new publications become available. For each paper, it would be also interesting to do some analysis in order to detect topics (orientation towards NLP, traversal semantics, and data exploration). After all, a system to explore (browse and search) the knowledge graph will be designed in order to support knowledge workers or researchers to retrieve papers relevant to a specific field of research, based on topics, quality, connections (through the path queries) with other papers.

The remaining of this paper is organized as follows: Section 2 discusses the related work; Section 3 presents the problem statement.....

3 Related Work

In recent years, generating KGs from unstructured text attracts significant attention within the research community. Information Extraction (IE) techniques are intended to produce a structured version of the information presented in the text so as to make it computer-understandable. Text representation through KGs provides opportunities for simple and easy content navigation and surpasses conventional keyword search, enabling the production of direct answers to complex queries.

3.1 General Knowledge Extraction

This paper assumes that KG generation from unstructured text is a 4-step process: Preprocessing, Named Entity Recognition (NER), Relationship Identification (RI) and Enrichment as illustrated in Figure 1. NER and RI are necessary steps and always required, whereas Preprocessing and Enrichment are optional.

Step 1*: Text preprocessing* is a widely used NLP task and is mainly applied to simplify text for future use. Tasks such as co-reference resolution, abbreviation resolution, sentence simplification, tokenization, stop words removal, etc. may differ depending on the data quality and the nature of the project concerned. Even though text preprocessing is a crucial step, it is frequently omitted. This happens either because scientists rely on usual datasets that have been already preprocessed, or because there are advanced tools using ultra-modern technologies (e.g. neural networks) that have decreased the necessity for preprocessing [7].

Step 2*: Named Entity Recognition* follows preprocessing and is still another important NLP task. NER identifies named entities (i.e. natural world objects) mentioned in the text and then classifies these entities into predetermined categories.

Step 3*: Relationship Identification.* Once NER step has been completed, the next step is to identify relationships between these entities and describe them as triples. Relation triples represent text in the sequence of subject, predicate, and object. Subject and object represent the extracted entities of Step 2, while predicate is the relation that connects these entities. Triple extraction can be achieved through multiple information extraction techniques.

Step 4*: Enrichment.* A KG can either be enriched or rely on preexisting semantic data models, that represent knowledge about a specific domain i.e. ontologies. An ontology uses a standard vocabulary to describe the main concepts of a domain as well as the relations among the concepts themselves. The need to develop a universal vocabulary shared by a community is crucial as it promotes interoperability, which is the main idea for KG development. Although some projects do not include an ontology, some others use it either to enhance relation and triple extraction or to enrich the KG.

What follows is a presentation of related projects that have been categorized according to the methods Steps 2 and 3 are performed but are also according to other parameters.

3.2 Methodologies

3.2.1 Named Entity Recognition and Semantic Annotations

Named Entity Recognition (NER), is considered a fundamental process in KG generation, as it identifies and disambiguates name entities from unstructured text, which then can be linked to relevant concepts or in other words, can be semantically annotated [8]. DepressionKG [9] is a disease-centric knowledge graph generated by the integration of multiple heterogeneous resources, whose

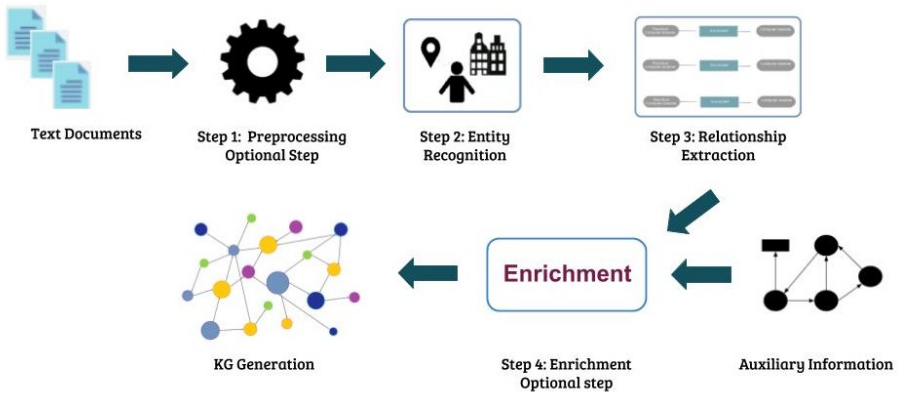

Figure 1: Steps leading to KG generation

content has been semantically annotated with medical terminologies using Xerox's NLP tool XMedlan. Following the same line of reasoning, COVID-KG [10] designers use CORD-NER [11] annotation system, which combines NER methods from four different sources and promotes comprehensive name entity annotation. In the CORD-19 Named Entities Knowledge Graph (NEKG) part of the Covid-on-the-Web [12] project, NEs are identified and then used as a method to build relevant links between articles in the CORD-19 [13] corpus and knowledge bases with entity description such as DBpedia, Wikidata and Bioportal ontologies. Annotation tasks are considered very challenging due to ambiguity, therefore there are tools (e.g. DBpedia Spotlight [14]), aiming to disambiguate natural language references in particular resources.

3.2.2 NLP

Notwithstanding the above, KGs can be generated using NLP or NLP along with another technique. KGen designers [15] do not assume that NER should be completed before RI, so they perform both steps at the same time. The first step in the KGen generation method is the preprocessing of the unstructured text. In so doing, the scientific text is simplified for future use. Then, the preprocessed text undergoes Semantic Role Labeling (SRL) to extract triples in the form of subject, predicate, and object. SRL is a technique that identifies the verb in a sentence as well as the semantic role of the remaining words (agent, patient). For each triple, the verb is assigned to the predicate, while subject and object are assigned to the agent and patient respectively. Concurrently, concepts and entities from the preprocessed text are extracted through tokenization and PoS tagging, which then are mapped to an ontology. The product of the above-mentioned steps is a final set of triples.

3.2.3 Other Methods

As research over KGs is evolving, new projects that do not fall under the aforementioned categories are constantly presented. COVID-19 knowledge Graph (CKG) [16] is an inventive KG that combines scientific metadata (papers, authors, institutions) and information derived from the scientific text (concept, topics). CKG generation is based upon Comprehend Medical (CM) Detect Entities V2, an Amazon Web Service [17] that combines NLP and Machine Learning (ML) and extracts NEs and relationships between them. CM also classifies the extracted entities into entity types or categories and entity attributes. Concerning the topic detection, Latent Dirichlet Allocation (LDA) along with multi-label classification are applied towards assigning topic labels (provided by professionals) to topic models. Argumentative Knowledge Graph (AKG), part of the Covid-on-the-Web project [12] is another exceptional KG. AKG designers employ ACTA [18], a platform that automatically analyzes clinical studies, to generate an argumentative KG. ACTA is based on argumentative mining methods, which identify the argumentative elements inside text (i.e. claims, premises), followed by their relations (i.e. support, attack), resulting in a graph containing the former as nodes and the latter as edges.

3.2.4 Manual Enhancement

As indicated above there are many automatic techniques that lead to a KG generation including metadata and topics extracted from scientific text. However, automation falls short when semantic representation deals with more complex structures. KG based on SemSur ontology [19] is a KG generated with less automation compared to other projects, which however, offers a much more complex structure to represent survey articles descriptively. Experts extract instances manually based on the ontology classes. They then import those instances to the ontology and finally the KG is generated including the communitys contribution.

3.2.5 KGs from Structured Data

There are cases, however, where none of the preceding steps are needed, as data are already in a structured form like RDF, CSV etc. In such cases, there are still obstacles either during data acquisition or during schema mapping. SCM-KG [20] project contains a KG generated from scientific metadata and is based on the idea of a functional framework, which combines data of different structured formats. In such cases converting data in a common model i.e. RDF and mapping them to an ontology, is very challenging as direct mapping is not always possible. Dealing with this issue, SCM-KG pipeline uses Sparqlify-CSV, to map CSV files to an ontology. When this is not feasible ETL component shapes the data in the required format.

3.2.6 Related work summary

Project	Input Data	Preprocessing	NER and RI	Ontology
DepressionKG [9]	Medical Data from multiple heterogeneous resources: PubMed, ClinicalTrial, Medical Guidelines, DrugBank, DrugBook, DrugBook, Wikipedia Antidepressant side effects, SIDER, SNOMED CT and Patient Data		Direct entity and concept identification when is feasible, but in case of unstructured text NER and RI are performed using, a semantic annotation tool (XMedlan).	
Multimedia COVID-KG [10]	CORD-19 [13]		i) Text knowledge extraction (NER and entity linking based on the ontology defined in the CTD). ii) Highly detailed entity extraction and semantic annotation using CORD-NER [11]. iii) Additional step: Image processing to extract visual information	Entity linking step is based on the ontology defined in the Comparative Toxicogenomics Database(CTD) [21].
NEKG Covid-on-the-Web [12]	CORD-19 [13]		NER and RI through semantic annotation using DBpedia, Wikidata and Bioportal.	DCMI, Bibliographic Ontology (FaBiO), Bibliographic Ontology, FOAF, Schema.organd Web Annotation Vocabulary have been used to enrich the KG.
KGen [15]	Unstructured text in plain-text format (e.g., a *.txt file)	NLP to simplify sentences: Sentence splitter, Abbreviation resolver, Sentence simplification	i) SRL extracts triples in the form of subject, predicate, object. ii) Entity and concept identification (Tokenization and PoS Tagger), which then are mapped to the ontology recommended by the NCBO bioportal. iii) Combination of the above-mentioned steps (i,ii) to produce the final set of triples.	Step ii: Performs entity identification from sentenses to obtain links to the domain ontology that is recommended by the NCBO bioportal. In this case the ontology is used to enforce entity linking.
AKG Covid-on-the-Web [12]	CORD-19 [13]		Argumentative element identification (i.e. claims, premises) and relationship identification (i.e. support, attack) have been established automatically using ACTA [18], a clinical text analysis tool.	Argument Model Ontology (AMO), the SIOC Argumentation Module (SIOCA) and the Argument Interchange Format have been used to enrich the KG.
CKG [16]	CORD-19 [13]		In this project NER and relationship extraction are performed together using Amazon Comprehend Medical (CM) [17]. For topic detection LDA and multi-label classification have been used.	
KG based on SemSur ontology [19]	Specific research articles (surveys) and the respective surveyed articles		i) SemSur ontology is created, ii) Instances from text are extracted manually based on SemSur ontology, iii) SemSur Ontology + Instances + Community : build the KG	This project is built on the top of SemSur ontology, which uses substes of other existing vocabularies i.e. DC, SWRC, FOAF, MLS, DCMI, DEO, LSC, DOAP.
SCM-KG [20]	Heterogeneous data in different structured formats such as CSV, RDF, web pages etc.		Dealing with structured data from heterogeneous resources → No NER AND RI. i) Data acquisition ii) Creation of an integration ontology (see next column) iii) Data conversion in a common model and ontology mapping. This may be feasible when dealing with RDF data. Though, there are cases where direct mapping of e.g. CSV is not possible. → ETL component to shape the data etc. iv) similarity measures to prevent multiple instances referring to same things.	KG generation is based on its core ontology, which is created using subsets of: SWRC, Dublin Core, and FOAF vocabularies. It is instantiated with data from the sources, and every time a concept is missing, a new relation type is defined.

Table:1 Summarized information information regrading the projects

3.3 Open Science Graphs

Open Science Graphs (OSGs) [22] are scientific KGs that support free and open access to graph representations of scientific metadata, but also promote FAIRness of science, as they can be easily accessed by anyone interested to share or use academic information. In other words, OSGs, are metadata graphs that are continuously updated by researchers whose studies are stored in global repositories. OpenAIRE, is a European project aligned with OSGs. OpenAIRE Research Graph represents metadata and links arising from more than 13,000 data sources globally [22]. After collection, data from heterogeneous sources are subjected to some processing, deduplication, and harmonization with a standard format. The OpenAIRE Research Graph data model [23] is made simple with the intention of promoting universal participation and collaboration. Synergic contribution and collaboration is considered necessary, in order to facilitate the ultimate goal of creating a Global Research KG. At this time, there are many projects like The Open Research Knowledge Graph [24], Research Graph[1] and The OpenCitations[2] graph that share the same principles and exchange metadata and links with the OpenAIRE Research Graph.

4 Problem Statement

Given a selection of scientific documents i.e. scientific articles and their metadata, this project's contribution is:

C1 A proposal for an idea to generate a KG describing entities extracted from data and relations between them. More specifically, the KG will represent in its main nodes' four entities of inherent article information (Publication title, Authors, Publication Year, DOI) and relationships among them. Also it will represent topics extracted from the unstructured text which will be linked to the respective publication as well. Paper entities can connect with either Authors, Publication Year, DOI and Topic entity, or all of them. The structure of a sub-graph is illustrated in Figure 2.

C2 A systematic and semi-automatic method to generate the KG mentioned above. Assuming that the implementation of a completely automated method of generating KG from scientific text is hard to achieve, a systematic and semi-automatic method will be proposed instead. Through this method, multiple information extraction techniques will be employed, using NLP to build the fundamental relationships in the graph, but also to extract topics identified in text and link them to the publication they belong to.

[1] `https://opencitations.net/`
[2] `https://opencitations.net/`

C3 A method for efficient incremental update of the graph, as new publications become available.

C4 Methods for querying the KG, that can derive information from the data and present it to the user. $E \subseteq V \times V$

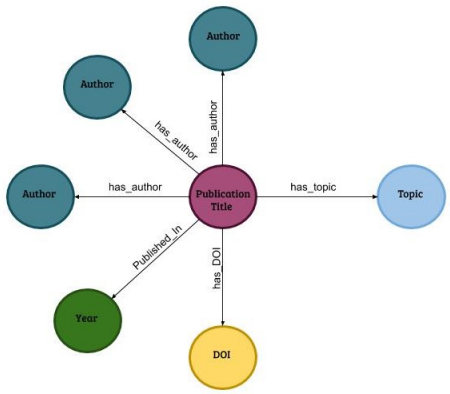

Figure 2: Sub-graph, part of the KG. Publication entities (purple) can connect to Authors (blue), Publication year (green), DOI (yellow), and Topics (light blue).

5 Solution

6 Results

7 Discussion

References

[1] H. Harapan, N. Itoh, A. Yufika, W. Winardi, S. Keam, H. Te, D. Megawati, Z. Hayati, A. L. Wagner, and M. Mudatsir, "Coronavirus disease 2019 (covid-19): A literature review," *Journal of Infection and Public Health*, 2020.

[2] M. L. Ranney, V. Griffeth, and A. K. Jha, "Critical supply shortages—the need for ventilators and personal protective equipment during the covid-19 pandemic," *New England Journal of Medicine*, vol. 382, no. 18, p. e41, 2020.

[3] Q. Chen, A. Allot, and Z. Lu, "Keep up with the latest coronavirus research," *Natur*, vol. 579, no. 7798, pp. 193–193, 2020.

[4] L. Ehrlinger and W. Wöß, "Towards a definition of knowledge graphs.," *SEMANTiCS (Posters, Demos, SuCCESS)*, vol. 48, pp. 1–4, 2016.

[5] A. Singhal, "Introducing the knowledge graph: things, not strings," *Official google blog*, vol. 5, 2012.

[6] H. Paulheim, "Knowledge graph refinement: A survey of approaches and evaluation methods," *Semantic web*, vol. 8, no. 3, pp. 489–508, 2017.

[7] T. Al-Moslmi, M. G. Ocaña, A. L. Opdahl, and C. Veres, "Named entity extraction for knowledge graphs: A literature overview," *IEEE Access*, vol. 8, pp. 32862–32881, 2020.

[8] A. Kiryakov, B. Popov, I. Terziev, D. Manov, and D. Ognyanoff, "Semantic annotation, indexing, and retrieval," *Journal of Web Semantics*, vol. 2, no. 1, pp. 49–79, 2004.

[9] Z. Huang, J. Yang, F. van Harmelen, and Q. Hu, "Constructing knowledge graphs of depression," in *International Conference on Health Information Science*, pp. 149–161, Springer, 2017.

[10] Q. Wang, M. Li, X. Wang, N. Parulian, G. Han, J. Ma, J. Tu, Y. Lin, H. Zhang, W. Liu, *et al.*, "Covid-19 literature knowledge graph construction and drug repurposing report generation," *arXiv preprint arXiv:2007.00576*, 2020.

[11] X. Wang, X. Song, Y. Guan, B. Li, and J. Han, "Comprehensive named entity recognition on cord-19 with distant or weak supervision," *arXiv preprint arXiv:2003.12218*, 2020.

[12] F. Michel, F. Gandon, V. Ah-Kane, A. Bobasheva, E. Cabrio, O. Corby, R. Gazzotti, A. Giboin, S. Marro, T. Mayer, *et al.*, "Covid-on-the-web: Knowledge graph and services to advance covid-19 research," in *International Semantic Web Conference*, 2020.

[13] L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Eide, K. Funk, R. Kinney, Z. Liu, W. Merrill, *et al.*, "Cord-19: The covid-19 open research dataset," *ArXiv*, 2020.

[14] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer, "Dbpedia spotlight: shedding light on the web of documents," in *Proceedings of the 7th international conference on semantic systems*, pp. 1–8, 2011.

[15] A. Rossanez and J. C. dos Reis, "Generating knowledge graphs from scientific literature of degenerative diseases.," in *SEPDA@ ISWC*, pp. 12–23, 2019.

[16] C. Wise, V. N. Ioannidis, M. R. Calvo, X. Song, G. Price, N. Kulkarni, R. Brand, P. Bhatia, and G. Karypis, "Covid-19 knowledge graph: Accelerating information retrieval and discovery for scientific literature," *arXiv preprint arXiv:2007.12731*, 2020.

[17] "Amazon Comprehend Medical." `https://aws.amazon.com/comprehend/medical/`.

[18] T. Mayer, E. Cabrio, and S. Villata, "Acta: a tool for argumentative clinical trial analysis," 2019.

[19] S. Fathalla, S. Vahdati, S. Auer, and C. Lange, "Towards a knowledge graph representing research findings by semantifying survey articles," in *International Conference on Theory and Practice of Digital Libraries*, pp. 315–327, Springer, 2017.

[20] A. Sadeghi, C. Lange, M.-E. Vidal, and S. Auer, "Integration of scholarly communication metadata using knowledge graphs," in *International Conference on Theory and Practice of Digital Libraries*, pp. 328–341, Springer, 2017.

[21] A. P. Davis, C. J. Grondin, R. J. Johnson, D. Sciaky, B. L. King, R. McMorran, J. Wiegers, T. C. Wiegers, and C. J. Mattingly, "The comparative toxicogenomics database: update 2017," *Nucleic acids research*, vol. 45, no. D1, pp. D972–D978, 2017.

[22] A. Aryani, M. Fenner, P. Manghi, A. Mannocci, and M. Stocker, "Open science graphs must interoperate!," in *ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium*, pp. 195–206, Springer, 2020.

[23] P. Manghi, A. Bardi, C. Atzori, M. Baglioni, N. Manola, J. Schirrwagen, and P. Principe, "The openaire research graph data model," Apr. 2019.

[24] M. Y. Jaradeh, A. Oelen, K. E. Farfar, M. Prinz, J. D'Souza, G. Kismihók, M. Stocker, and S. Auer, "Open research knowledge graph: next generation infrastructure for semantic scholarly knowledge," in *Proceedings of the 10th International Conference on Knowledge Capture*, pp. 243–246, 2019.