Title: Investigating COVID-19 Analytics and Research with Clinical Knowledge Organization Systems

Aims

The study investigated the utility of the open-source data analytics, reporting, and integration software KNIME[1] (Berthold et al., 2009), for mapping knowledge organization systems (KOS) dedicated to COVID-19. The research questions of a) how the data analytics and extract transform load (ETL) tools in the software support the task of clinical coding scheme mapping and b) how the output produced from the mapping could be used to annotate clinical trial documents were addressed.

Methods

This study modeled the research design to match the cognitive processes involved in the Design Science Research Cycle, which is ideal for guiding the development and evaluation of artifacts as objects of research where the artifact is used to solve identified problems (Hevner et al., 2004). Three ontologies, the Coronavirus Infectious Disease Ontology (CIDO)[2] (He et al., 2020), the COVID-19 ontology[3](Sargsyan et al., 2020), and Coronavirus Vocabulary (COVOC)[4], with a combined total of 10691 terms were selected and mapped to each other using a variety of machine learning and natural language processing methods. In the string-based matching portion of the workflow, KNIME join nodes were used to identify matching URIs. Additionally, the string distance, similarity search, and similarity learner nodes were configured and used to locate occurrences of terms and identify lexically similar entity names (Aho & Corasick, 1975; Cohen et al., 2003; Monge & Elkan, 1997). In the second stage, sense-based matching is done by analyzing the meaning of the concepts rather than the labels utilizing the document similarity learner node and the concept definitions in the ontologies (Giunchiglia et al., 2004). Finally, the rule-based matching segment of the workflow involved creating a customized node. This node uses Python script and ScispaCy[5] models for processing biomedical text to identify words and phrases within the ontology documents and assess their relationships. Mappings between RDFS, OWL, and SKOS ontological classes, properties, concepts from concept schemes, or transitive super properties were identified. The integrated set of mapped concepts obtained from these workflow segments was used to create a dictionary for the NLP pipeline where annotation of clinical trials was performed with various NLP nodes. The research output is a new knowledge contribution (Gregor & Hevner, 2013) in the form of a reproducible and shareable artifact for mapping disparate terminologies and document annotation that can be utilized beyond the domain of biomedicine.

Background

Since early 2020 the world has been in crisis mode dealing with the COVID-19 pandemic, revealing several challenges and opportunities for data analytics, semantic interoperability, and decision-making. Sharing COVID-19 data is critical for clinical research, drug testing, and therapeutic strategies and for developing public health policies for intervention, control, and management of the disease. However, these rely on data and methodology integrated from multiple disciplines such as biology, medicine, public health, geography, and social science.

---

[1] KNIME 4.4.1. August 2021. KNIME AG. Zurich, Switzerland. https://www.knime.com/.
[2] https://bioportal.bioontology.org/ontologies/CIDO
[3] https://bioportal.bioontology.org/ontologies/COVID-19
[4] https://www.ebi.ac.uk/ols/ontologies/covoc
[5] https://allenai.github.io/scispacy/

In response, various Knowledge Organization Systems (KOS) - encompassing classification schemes such as the International Classification of Diseases (ICD), terminologies such as the Systematized Nomenclature of Medicine -- Clinical Terms (SNOMED-CT), and ontologies like the Coronavirus Infectious Disease Ontology (CIDO) have updated their concept schemes, adding terms specific to COVID-19. Furthermore, researchers have developed new KOS to directly respond to the pandemic, for example, the COVID-19 Ontology. These KOS are critical for defining and structuring concepts and terms in healthcare, enabling information exchange, and ensuring consistent use by stakeholders.

Knowledge Organization Systems (KOS) facilitate the meaningful and accurate utilization of the information exchanged at the syntactic level of interoperability. They further act as a method for information enrichment and analytics processes in addition to supporting comparative, translational, and prognostic research. Used strategically, KOS can reduce the time and cognitive loads associated with combining and linking datasets and are critical to the infrastructure needed for enabling the proper functioning of healthcare and data-driven biomedical research discoveries. However, there is a need for a "common, uniform, and comprehensive approach" to clinical knowledge representation (de Quiros et al., 2018). Taking the best advantage of this data requires mapping terms across vocabularies and annotating text with terminologies and ontologies (Tchechmedjiev et al., 2018).

Findings

The study findings support using ETL tools such as KNIME for terminology mapping and document annotation without requiring high complexity manual or automated procedures and potentially reducing the cognitive load of any human experts involved. For example, lexical algorithms identified 602 of the 666 mappings between the CIDO/COVID19 ontology and 294 matches that were not present in the Gold Standard set of terms. The workflow consistently identifies lexically similar terms and similar URIs across ontologies. However, human effort or more advanced algorithms are needed to assess the types of semantic relations. Results from the evaluation demonstrate moderate to high precision in the mapping results compared to a gold standard in both string-based matching algorithms and those that rely on semantics.

Additionally, the model scorer for document annotation achieved high precision and recall (99%), indicating that the annotation workflow identifies more relevant results than irrelevant results and that most dictionary terms are identified. The mapping output innovatively supports semantic enrichment of words by providing a list of tailored terms that can be used to train the annotation model to identify the desired entity types within the unstructured text. The resulting workflow tool facilitates easy loading and analysis of datasets, data cleaning and transformation, reductions in operating, product, personnel, and project-related costs, and insights into community-based development that anyone, expert and non-expert alike, can use.

Relevance to Themes of the Workshop

The study investigates and discusses how KOS are used for computational reasoning, enhancing analysis and discoverability, and providing a common terminology for discipline-neutral dialog, supporting COVID-19 research. Using a few good KOS, it also explores the annotation and production of appropriate metadata and a summary of findings from clinical trial data. This is relevant to the Workshop themes regarding Health-related KOS issues and contributions, Terminology/Vocabulary Development, and KOS Mapping.

References:

Aho, A. V., & Corasick, M. J. (1975). Efficient string matching: An aid to bibliographic search. Communications of the ACM, 18(6), 333–340. https://doi.org/10/crw9b4

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., ... & Wiswedel, B. (2009). KNIME-the Konstanz information miner: version 2.0 and beyond. *AcM SIGKDD explorations Newsletter*, *11*(1), 26-31. Available from https://www.knime.com/downloads

Cohen, W. W., Ravikumar, P., & Fienberg, S. E. (2003). A Comparison of string distance metrics for name-matching tasks. In web, 2003, 73–78.

De Quiros, F. G. B., Otero, C., & Luna, D. (2018). Terminology services: Standard terminologies to control health vocabulary: Experience at the Hospital Italiano de Buenos Aires. Yearbook of Medical Informatics, 27(01), 227–233. https://doi.org/10.1055/s-0038-1641200

Giunchiglia, F., Shvaiko, P., & Yatskevich, M. (2004). S-Match: An algorithm and an implementation of semantic matching. In C. J. Bussler, J. Davies, D. Fensel, & R. Studer (Eds.), The Semantic Web: Research and Applications (pp. 61–75). Springer. https://doi.org/10/b4t35k

He, Y., Yu, H., Ong, E., Wang, Y., Liu, Y., Huffman, A., ... & Smith, B. (2020). CIDO, a community-based ontology for coronavirus disease knowledge and data integration, sharing, and analysis. Scientific data, 7(1), 1-5. https://doi.org/10.1038/s41597-020-0523-6

Hevner, A., & Chatterjee, S. (2010b). Design science research in information systems. In A. Hevner & S. Chatterjee (Eds.), Design Research in Information Systems: Theory and Practice (pp. 9–22). Springer US. https://doi.org/10.1007/978-1-4419-5653-8_2

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. MIS Quarterly, 28(1), 75–105. https://doi.org/10.2307/25148625

Monge, A., & Elkan, C. (1997). An efficient domain-independent algorithm for detecting approximately duplicate database records.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2008). A design science research methodology for information systems research. Journal of Management Information Systems, 24(3), 45–77. https://doi.org/10/cxnmc8

Sargsyan, A., Kodamullil, A. T., Baksi, S., Darms, J., Madan, S., Gebel, S., ... & Hofmann-Apitius, M. (2020). The COVID-19 ontology. Bioinformatics, 36(24), 5703-5705. https://doi.org/10.1093/bioinformatics/btaa1057

Tchechmedjiev, A., Abdaoui, A., Emonet, V., Melzi, S., Jonnagaddala, J., & Jonquet, C. (2018). Enhanced functionalities for annotating and indexing clinical text with the NCBO Annotator+. Bioinformatics, 34(11), 1962–1965. https://doi.org/10/gdk4vz