Finding Novel Links in COVID-19 Knowledge Graph

Drahomira Herrmannova*, Ramakrishnan Kannan, Seung-Hwan Lim, and Thomas E. Potok

Computer Science and Mathematics Division
Oak Ridge National Laboratory

February 22, 2021

Abstract

The scientific literature is expanding at incredible rates, which were recently estimated to be in the millions of new articles per year. Extracting information from such vast stores of knowledge is an urgent need, as exemplified by the recent open release of materials relevant to the current SARS-CoV-2 pandemic. In this context, this challenge seeks to develop algorithms for the analysis and mining of knowledge graphs. The main task in this challenge is to leverage a graph of biomedical concepts related to COVID-19 and the relations between them to try to discover novel, plausible relations between concepts. For this challenge, the participants will be provided with a graph dataset of biomedical concepts and relations between them extracted from scientific literature, along with all-pairs shortest path information between the concepts in the graph. They will be asked to analyze the data and use it to predict which concepts will form direct novel relations in the future. In addition, they will be asked to rank the predicted links according to the predicted importance of each relation.

1 Introduction

The scientific literature is expanding at incredible rates, which were recently estimated to be in the millions of new articles per year [3]. Extracting information from such vast stores of knowledge is an urgent need, as exemplified by the recent open release of materials relevant to the current SARS-CoV-2 pandemic [4]. Given that the volume of information is easily beyond the capacity of any one person, analysts have been strongly motivated to develop automated knowledge-mining methods and extraction tools [10, 8, 1].

*Corresponding author: herrmannovad@ornl.gov

In this context, this challenge seeks to develop algorithms for the analysis and mining of *knowledge graphs*. More specifically, this challenge is based on the process of literature-based discovery [6]. It has been shown that previously unknown relationships exist in the scientific literature that can be uncovered by finding concepts that link disconnected entities [6, 7, 9]. This process, called *Swanson Linking*, is based on the discovery of hidden relations between concepts A and C via intermediate B-terms: if there is no known direct relation A-C, but there are published relations A-B and B-C one can hypothesize that there is a plausible, novel, yet unpublished indirect relation A-C. In this case the B-terms take the role of bridging concepts. For instance, in 1986, Swanson applied this concept to propose a connection between dietary fish oil (A) and Raynaud's disease (C) through high blood viscosity (B), which fish oil reduces [5]. This connection was validated in a clinical trial three years later.

2 Task and Data

The main task in this challenge is to leverage a graph of biomedical concepts related to COVID-19 and the relations between them to try to discover novel, plausible relations between concepts.

2.1 Data

The participants will be provided with the following data:

- **Training data:**
  - Graph representing biomedical concepts and relations between them constructed from PubMed[1], Semantic MEDLINE[2], and CORD-19[3]. This graph will represent a historic snapshot of the data (e.g., a version of the knowledge graph built from papers published up until June 2020). An explanation of the graph is shown in Figure 1.
  - Data representing the shortest path between all pairs of concepts in the above graph.
- **Validation data:**
  - A version of the above graph and shortest path information that includes all available data (e.g., all data up until February 2021).
  - List of concept pairs representing novel relations that have formed since the year captured in the above graph, e.g., all novel relations between the concepts in the graph that have formed between June 2020 and February 2021.

[1] `https://pubmed.ncbi.nlm.nih.gov/`
[2] `https://skr3.nlm.nih.gov/SemMedDB/`
[3] `https://www.semanticscholar.org/cord19`

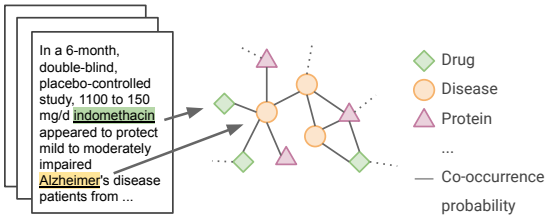

Figure 1: Depiction of the challenge data.

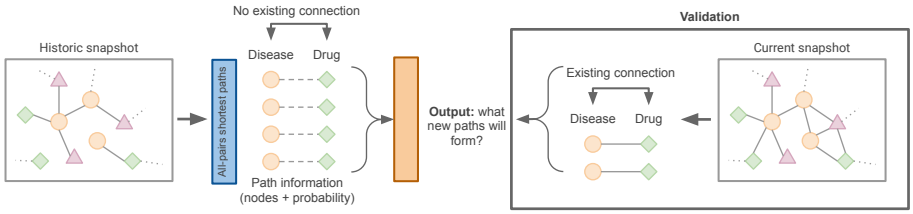

Figure 2: Depiction of the challenge task.

– Additionally, each concept pair will be assigned several ranks representing the importance of the relation. These ranks will include rank according to the moth and year the connection has been formally established (approximated using the date of publication of the associated research article), number of citations received by the relevant article, and other relevant ranks.

The participants will be asked to use the graph and the all-pairs shortest paths information provided (training data) to predict which concepts will form a direct connection in the future (validation data). This process is depicted in Figure 2.

A detailed description of the dataset including the format and the process used to produce the dataset is provided in [2]. The dataset that will be provided for this competition represents an updated version of the dataset[4] described in [2]. The updated version includes concepts that were extracted from articles published since the first release of the dataset.

The participants are allowed to leverage external data, particularly data from PubMed and Semantic MEDLINE that are not included in the provided dataset.

2.2 Tasks

- Analyze and visualize the provided graph and all-pairs shortest paths (APSP) data and provide statistics such as, betweenness centrality, av-

[4] https://dx.doi.org/10.13139/OLCF/1646608

erage path length and frequency of concept occurrence in different paths.

- Compare APSP path statistics with the future connections – are the length of a path between two concepts or other path statistics indicative of whether the concepts will form a connection in the future?
- Develop a classification or a model that uses the APSP data or other relevant data as input and predicts which concepts will form a connection in the future.
- Develop ranking model/function for ranking the predicted links according to their importance.

References

[1] Biomedical Data Translator Consortium et al. "Toward a universal biomedical data translator". In: *Clinical and translational science* 12.2 (2019), p. 86.

[2] Drahomira Herrmannova et al. "Scalable Knowledge-Graph Analytics at 136 Petaflop/s – Data Readme". In: (Aug. 2020). DOI: 10.13139/OLCF/1646608. URL: https://zenodo.org/record/3980252.

[3] Esther Landhuis. "Scientific literature: information overload". In: *Nature* 535.7612 (2016), pp. 457–458.

[4] Office of Science and Technology Policy. *Call to Action to the Tech Community on New Machine Readable COVID-19 Dataset*. Online. Accessed: 2020-04-18. 2020.

[5] Don R Swanson. "Fish oil, Raynaud's syndrome, and undiscovered public knowledge". In: *Perspectives in biology and medicine* 30.1 (1986), pp. 7–18.

[6] Don R. Swanson and Neil R. Smalheiser. "An interactive system for finding complementary literatures: a stimulus to scientific discovery". In: *Artificial Intelligence* 91.2 (Apr. 1997), pp. 183–203.

[7] Don R. Swanson, Neil R. Smalheiser, and Vetle I. Torvik. "Ranking indirect connections in literature-based discovery: The role of medical subject headings". In: *Journal of the American Society for Information Science and Technology* 57.11 (2006), pp. 1427–1439.

[8] Menasha Thilakaratne, Katrina Falkner, and Thushari Atapattu. "A Systematic Review on Literature-based Discovery: General Overview, Methodology, & Statistical Analysis". In: *ACM Computing Surveys (CSUR)* 52.6 (2019), pp. 1–34.

[9] Vahe Tshitoyan et al. "Unsupervised word embeddings capture latent knowledge from materials science literature". In: *Nature* 571.7763 (July 2019), pp. 95–98. DOI: 10.1038/s41586-019-1335-8. URL: https://www.nature.com/articles/s41586-019-1335-8.

[10] Hsih-Te Yang et al. “Literature-based discovery of new candidates for drug repurposing”. In: *Briefings in bioinformatics* 18.3 (2017), pp. 488–497.