Investigating Ontology Use in Artificial Intelligence and Machine Learning for Biomedical Research - A Preliminary Report from A Literature Review

Asiyah Yu Lin*[1], Andrey Ibrahim Seleznev[2], Tianming "Danny" Ning[3], Paulene Grier[4,5], Lalisa "Mariam" Lin[6], Christopher Travieso[7], Ansu Chatterjee[8], Jaleal Sanjak[9]

[1] *National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA*
[2] *Walter Johnson High School, Bethesda, MD 20817, USA*
[3] *Winston Churchill High school, Potomac, MD 20854, USA*
[4] *Thomas Stone High School, Waldorf, MD 20601, USA*
[5] *College of Southern Maryland, La Plata, MD 20646, USA*
[6] *Walt Whitman High School, Bethesda, MD 20817, USA*
[7] *Our Lady Of Good Counsel High School, Onley, MD 20832, USA*
[8] *Office of Director, National Institutes of Health, Bethesda, MD 20852, USA*
[9] *National Center for Advancing Translational Sciences, National Institutes of Health, Rockville, MD 20850, USA*

Abstract

In this report, the authors conducted a comprehensive literature review to answer a question: how ontologies are being used in the AI/ML approaches to solve biomedical research problems? A selection of 107 papers were reviewed and data were extracted to answer question regarding how, what, who and where the ontology-aware AI/ML approach were applied in biomedical domain, as well as the mechanics of ontology use in AI/ML framework. The ontologies either was used as categories of data or used to compute the knowledge. Among many other ontologies, the Gene Ontology dominated the use of ontologies in AI/ML based biomedical problem solving. Lack of collaborations were observed via the co-authorship network analysis.

Keywords

Ontology, Artificial Intelligence, Machine Learning, literature review,

1. Introduction

As a form of knowledge representation, ontologies organize the knowledge and data hierarchically ("tree-like") and horizontally ("network-like" or "graph-like") using semantic relations, such as "is-a" or "part-of". Artificial Intelligence (AI) and/or Machine Learning (ML) often apply mathematical models that require numeric data as input. The fast growing and big volume of biomedical data has benefited the fast-advancing AI/ML algorithms and frameworks. However, leveraging the non-numerical, semantic, and hierarchical relations from an ontology remains a challenge in AI/ML [1]. In this report, the authors conducted a literature review to answer a question: how ontologies are being used in the AI/ML approaches to solve biomedical research problems?

2. Method

On the date of Sep.4, 2022, a total of 503 papers were retrieved from PubMed Central® (PMC) archive using keywords appeared in title and abstract: ontology, artificial intelligence, machine learning, deep learning, neural network, and embedding within 5 years' range, from 2017 to 2022. Out of the 503 papers, the authors selected 250 papers highly relevant papers to screen due to the time constrain. In total, 107 papers were selected for this report based on the

ICBO 2022, September 25-28, 2022, Ann Arbor, MI, USA
EMAIL: asiyah.lin@nih.gov (A. 1);
ORCID: 0000-0003-2620-0345 (A. 1);

eligibility criteria of a research paper solving a biomedical scientific problem. Excluded papers (n= 143) are review or comment papers, papers that do not solve a biomedical scientific problem, rather an engineer problem such as Natural Language Processing (NLP) problems or using AI/ML to develop ontology (*e.g.* predict new relations or new classes), and irrelevant papers. To facilitate the information extraction process and make the user interface easy and intuitive to use, a Google Form was designed for extracting the related text from the papers. The senior reviewer (AYL) then reviewed all the 250 papers' screening and 107 papers' information extraction to cross check the results. The raw dataset of reviewers' response was deposited to the Zenodo repository. A DOI id (10.5281/zenodo.7769984) was reserved for this dataset. Co-authorship network analysis were conducted using Gephi software (https://gephi.org/).

3. Results

After reviewing the abstracts of 250 papers, authors identified four major categories of ontology use in AI/ML: 1. Use the whole ontology or ontology terms as data labels to be the training datasets; 2. Transform the ontological representation into numerical data representation that will be used in the downstream AI/ML, which includes calculate term's semantic similarities, construct concepts association matrix, and use word embedding algorithms, and etc.; 3. The ontology as a graph structure or network structure used as a part of neural network architecture; 4. The ontology classification is the target of the AI/ML classifier.

What follows are the specific questions being answered via this exercise of literature review.

1. What biomedical problems are solved using ontology aware-AI/ML?

The biomedical problems that were being solved are mostly focused on gene function prediction (25 papers), or ontology annotation (14 papers). 7 papers using ontology-aware AI/ML to perform protein/gene interaction prediction, and 6 papers predict disease gene or protein or variant prediction. Other topics including drug-drug interaction, drug-drug interaction, drug repurpose, drug target, drug toxicity, pathway membership prediction. In the clinical area, a few papers focus on clinical outcomes prediction from EHR, anatomical site prediction from radiology report, image, or pathology report, and predicting patient similarity from clinical trial. Interestingly, there are papers using ontology and AI/ML to mine the social media data for sentiment prediction and drug off-label use prediction.

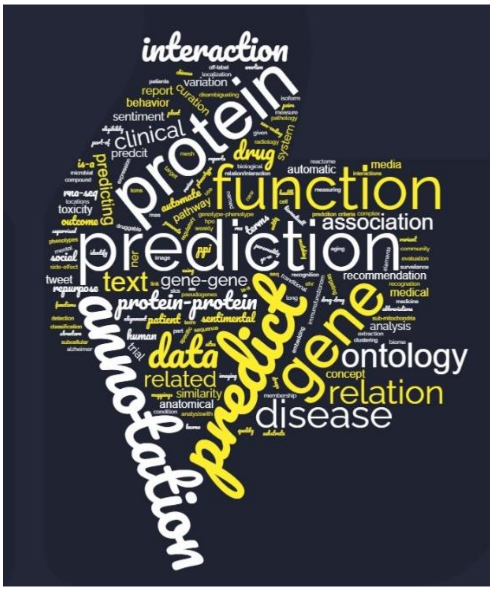

Figure 1: Word cloud of the biomedical problems (generated by https://www.wordclouds.com/)

2. What ontologies are being used?

Besides 7 papers that did not mention the name of ontologies used, 100 papers have specified the ontologies being used. The use of Gene Ontology (GO) is dominant: out of 107 papers, 65 (60.7%) were utilize GO in their AI/ML pipeline or architecture to solve their scientific problems. The next 4 most frequently used ontologies are: SNOMED CT and Human Phenotype Ontology (HPO) (9 papers, 8.4%), UMLS (6 papers, 5.6%) and Disease Ontology (DOID) (5 papers, 4.7%). Besides those, the Infectious Disease Ontology (IDO), ChEBI, FMA and Chinese version MeSH were used more than 1 papers. Many papers develop specific ontologies for their specific task. In addition to the dominate use of GO, 38 (35.5%) ontologies cover topics related to disease, phenotype, or conditions. This result shows the lack of diversity of biomedical ontology use in AI/ML for biomedical research. It also shows the potential benefit of a unified ontology that covers diseases, phenotypes, and conditions.

3. How ontology is being used in the AI/ML algorithm or architecture?

There are two big categories on how ontology is being used in AI/ML algorithms: A) using ontology as categories of data, or B) compute the knowledge. In category A, 42 papers (39%) were using ontologies as training data, and 24 papers (22%) were using ontologies as classifier's target. In category B, the most popular use is to transform the ontology into numeric presentation. 54 papers (50.4%) were using different methodologies, such as embedding, semantic similarity, and information content, to convert a text-based ontology into a matrix table with numbers. Only 12 papers (11.2%) utilized the whole ontology's content and structure as a layer in a neural network architecture.

Out of the 107 papers, 31 papers (29%) applied neural network architecture. Among which, 11 papers used convolutional neural network, 7 papers used deep neural network, 6 papers used long short-term memory network including Bi-LSTM and Bo-LSTM, 3 papers on recurrent neural network, 2 papers on artificial neural network. Deep learning technology were applied in 4 papers. There is a growing practice to use a variety of embedding methods to transform the ontology into a low-dimensional vector space. 6 papers were using Node2Vec, 4 papers using Word2Vec, 2 papers on Doc2Vec, 2 papers on Onto2Vec, and 1 paper on OPA2Vec and DL2Vec. While new methodologies are tested in those papers, traditional classifiers are still being applied: 8 papers applied Support Vector Machine (SVM), 6 papers applied Random Forest, 4 papers used Naive Bayes classifier or k-nearest neighbor and 3 papers used logistic regression techniques. In most of the case, the authors claimed that ontology-aware AI/ML outperforms traditional classifiers.

4. Who and where publish those papers?

The authors also looked at the geographical distribution of the papers that are published. The top 5 countries that publish the most are: USA (33 papers), China (26 papers), UK and Saudi Arabia (10 papers each), France (7 papers), and Germany, Korea, and Portugal (7 papers each). 26 papers have authors across different countries. Out of which, 4 papers produced by China-USA collaborations, and 2 papers produced by France and Lebanon collaboration. The observation of USA publishing dominant maybe biased, because the authors only selected the USA based PMC as the source database to retrieve papers.

5. How did authors collaborate in research?

The authors were interested in learning about who are the researchers in this field and how they collaborate. A network analysis was performed based on the co-authorship. The resulted research network shows a lack of collaboration in this research area. Most of the authors are isolated groups (Figure 2A). The hub analysis of the network reveals one active hub center, Dr. Robert Hoehndorf from the King Abdula University of Science and Technology (KAUST) at Saudi. He has many papers published with many authors; however, his co-authorship network is limited between the UK and Saudi Arabia (Figure 2B). Community analysis showed that beside the community formed by the UK and Saudi, a few Chinese researcher forms their own community via co-authorships. This result shows that a lot of collaborative activities, such as focused conference, workshops, meetings, and hackathons are needed to promote creativity and innovation of science. The authors suggested that more workshops such as Role of Ontology in Biomedical AI (ROBI) should be held, and a community of such scientists working in this specific area should be established.

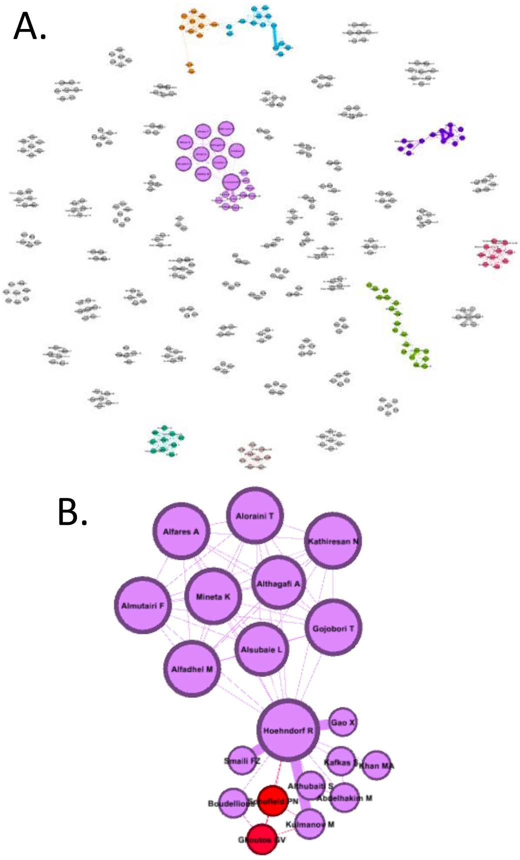

Figure 2A: Network analysis of the co-authorship of the 107 papers. A node denotes an author' Color denotes community; size of the nodes denotes the centrality of an author; the size of

link denotes the counts of co-authorship between authors.

Figure 2B: Hub analysis showed that Dr. Robert Hoehndorf and his group forms an active hub and a small community comprised of Dr. Hoehndorf's collaborators in Saudi Arabia and UK.

4. Conclusion

In conclusion, ontology provides contextually rich data to help the AI/ML to achieve a higher performance compared to the similar methods without ontologies. However, the applications of ontology-aware AI/ML in biomedical domain are still limited to gene or protein function predictions. The lack of cross-discipline collaborations specifically in applications in biomedical domain is alarming. Fundings to support collaborative initiatives and community development are needed in this area. Workshops such as ROBI should be continued and expanded.

Utilizing the graph-structural and semantics within an ontology requires more complex neural network architecture along with many other components such as the neuro-symbolic approach. Explainable AI is an emerging field where the explanatory techniques can explicitly show why a recommendation, or a prediction is made. This literature review is biased by the selection of PMC as the pool to retrieve. Many methodological papers were published as conference proceedings or white papers. Rising topics such as neuro-symbolic, explainable AI were not investigated. The future work includes extending the search to other repositories, such as Europe PMC, IEEE, PMLR, DBLP, arXiv, and to other topics such as neuro-symbolic [2], explainable AI [3] use in biomedical domain. Leveraging an ontology of AI/ML to annotate more details on AI/ML components to allow better analysis is another future direction as well.

5. Acknowledgement

AYL, AC and JS are supported by The Office of Data Science Strategy, NIH, via the Data and Technology Advancement (DATA) National Service Scholar program. AIS, TDN, PG, LL, and CT are the members of a local youth group Biomedical Informatics Research for Youth located in Maryland. This group is founded by AYL. AYL and JS are mentors of this youth group. AYL conceived the idea of the paper, designed the search strategy and survey questions, evaluated the results, conducted analysis, and wrote the manuscript. All other authors contributed to the paper review and data entry. AIS and TDN contributed to data curation for co-authorship network analysis.

6. References

[1] Kulmanov M, Smaili FZ, Gao X, Hoehndorf R: Semantic similarity and machine learning with ontologies. Brief Bioinform 2021, 22(4).

[2] Hassan M, Guan H, Melliou A, Wang Y, Sun Q, Zeng S, Liang W, Zhang Y, Zhang Z, Hu Q: Neuro-Symbolic Learning: Principles and Applications in Ophthalmology. arXiv preprint arXiv:220800374 2022.

[3] Holzinger A, Malle B, Saranti A, Pfeifer B: Towards multi-modal causability with Graph Neural Networks enabling information fusion for explainable AI. Information Fusion 2021, 71:28-37.