AFRL-AFOSR-JP-TR-2022-0016

A Study on constructing knowledge graph and Graph-based Deep Learning for prediction and ranking problems in cybersecurity

Minh Le Nguyen
JAPAN ADVANCED INSTITUTE OF SCIENCE AND TECHNOLOGY
1-1, ASAHIDAI
NOMI, ISHIKAWA, , 923-1211
JP

03/31/2022
Final Technical Report

Air Force Research Laboratory
Air Force Office of Scientific Research
Asian Office of Aerospace Research and Development
Unit 45002, APO AP 96338-5002

REPORT DOCUMENTATION PAGE

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
20220331	Final	START DATE 20190823	END DATE 20210822

4. TITLE AND SUBTITLE
A Study on constructing knowledge graph and Graph-based Deep Learning for prediction and ranking problems in cybersecurity

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
FA2386-19-1-4041		
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**

6. AUTHOR(S)
Minh Le Nguyen

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
JAPAN ADVANCED INSTITUTE OF SCIENCE AND TECHNOLOGY
1-1, ASAHIDAI
NOMI, ISHIKAWA 923-1211
JP

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AOARD UNIT 45002 APO AP 96338-5002	AFRL/AFOSR IOA	AFRL-AFOSR-JP-TR-2022-0016

12. DISTRIBUTION/AVAILABILITY STATEMENT
A Distribution Unlimited: PB Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This research aims at studying how to build knowledge graph representation for cybersecurity by performing open information extraction techniques on a large scale of text documents. After that, we will investigate how deep learning can be applied for knowledge graph representation in cybersecurity. We will consider the use of attention mechanism in graph deep learning model and compare various deep learning models which can be used for learning from graph representation data. Besides, we will study semi-supervised learning frameworks for utilizing unlabeled data for improving the performance of prediction and ranking problems. The proposed method will be applied for detecting malware and code flaws in cybersecurity.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT U	b. ABSTRACT U	c. THIS PAGE U	SAR	15

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
ALAN LIN	227-7009

AFOSR Deliverable Reporting Format:
Final and Interim Reports

This document has been developed to provide Principal Investigators (PIs), co-PIs, and research organizations with:

- A listing of the questions that will be asked in the new AFOSR project reporting format;
- Assistance in planning for the submission of the report

Overview: There are two main sections of the AFOSR Deliverable Report. Section 1 is filled out in Qualtrics, and Section 2 is uploaded by PDF.

- Section 1: Structured Survey Questions in Qualtrics
  - This section captures information in a structured survey format as required by the Research Performance Progress Report Format (RPPR) guidance. Information in this section will include publications, participants, and other intellectual property questions. All questions in this section will be asked within Qualtrics.

- Section 2: Technical Report PDF
  - This section captures unstructured technical information not captured in the above section. PI's will upload PDF reports that contain information on awards, changes to scope, and other technical updates. This PDF upload will be very similar to previous AFOSR report uploads. Please contact your individual program officer if you have further questions about what should be contained in this report.

***Note: The information being asked in this deliverable report is explicitly defined by the official RPPR guidance which can be found here: http://www.nsf.gov/bfa/dias/policy/rppr/index.jsp. We have automated several questions to make the reporting less burdensome on the principal investigator. Your report link, found in the deliverable reminder email, contains individualized information that is specific to your deliverable report. You may edit this information if anything appears to be incorrect.

Section 1: Structured Survey Questions

Award Information

You will be asked for the following information in the survey. The majority of these fields will be sent pre-populated, but can be edited if necessary. You will be able to manually edit any of these fields if information needs to be updated.

Award Number (FA2386-19-1-4041)

- Report Type
- Principal Investigator: Nguyen Le Minh
- Principal Investigator Email: nguyenml@jaist.ac.jp
- Principal Investigator Phone: (+81) 080-6364-1960

- Project Title: ***A Study on constructing knowledge graph and Graph-based Deep Learning for prediction and ranking problems*** **in cybersecurity**

- Recipient Organization: Japan Advanced Institute of Science and Technology (JAIST)
- Business Office Email: nguyenml@jaist.ac.jp
- Report Due Date: 9/11/2021
- Report Period Start Date: 03/08/2019
- Report Period End Date: 03/08/2021
- Current Program Officer
- Please list any other Co-Program Officers (if applicable)

Publications

[KongICTAI2021] Wei Kun Kong, Teeradaj Racharak, Yiming Cao, Cheng Peng, Minh Le Nguyen, KGWE: A Knowledge-guided Word Embedding Fine-tuning Model, *Proceedings of the 33rd IEEE International Conference on Tools with Artificial Intelligence (ICTAI),* 2021

[NguyenICTAI2021] Chau Nguyen, Vu Tran and Minh Le Nguyen, Enrichment of Features for Malware-Related Sentence Classification using External Knowledge, *Proceedings of the 33rd IEEE International Conference on Tools with Artificial Intelligence (ICTAI), 2021*

[KongKSE2020] Wei Kun Kong, Teeradaj Racharak, Minh Le Nguyen, Can Knowledge Enhance Reading Comprehension? An Integrated Approach with Semantic Lexicon, *Proceedings of the 11th IEEE International Conference on Knowledge and Systems Engineering (KSE)*, Can Tho, Vietnam, 2020

[ZinICART2021] May Myo Zin, Teeradaj Racharak, Minh Le Nguyen, Construct-Extract: An Effective Model for Building Bilingual Corpus to Improve English-Myanmar Machine Translation, *In Proceedings of 13th International Conference on Agents and Artificial Intelligence (ICAART)*, 2021

[Le&ICP2021] Tung Le, Huy Nguyen, Minh Le Nguyen, Vision And Text Transformer For Predicting Answerability On Visual Question Answering, *In Proceedings of IEEE International Conference on Image Processing (ICIP)*, 934-938, 2021

[Le&Journal2021] Tung Le, Huy Tien Nguyen, Minh Le Nguyen: **Multi visual and textual embedding on visual question answering for blind people.** Neurocomputing 465: 451-464 (2021)

[Vu-AI&Law2021] Vu, S.T., Le Nguyen, M. & Satoh, K. Abstract meaning representation for legal documents: an empirical research on a human-annotated dataset. *Artif Intell Law* (2021). https://doi.org/10.1007/s10506-021-09292-6

[Tran and Nguyen-2021] XC Tran, LM Nguyen, ReLink: Open information extraction by linking phrases and its applications. International Conference on Distributed Computing and Internet Technology, 44-62, 2021, https://doi.org/10.1007/978-3-030-65621-8_3 (invited paper)

[Tran&NguyenCL2021] Van-Khanh Tran, Le-Minh Nguyen, Variational model for low-resource natural language generation in spoken dialogue systems, *Computer Speech & Language*, Volume 65, 2021,101120, https://doi.org/10.1016/j.csl.2020.101120.

[Chau2021] Nguyen Minh Chau. A study on graph neural networks and pretrained models for analyzing cybersecurity texts, *Master thesis, Information School, JAIST* (Sep 2021)

[KienColing2020] Phi Manh Kien, Ha-Thanh Nguyen, Ngo Xuan Bach, Vu Tran, Minh Le Nguyen, Tu Minh Phuong: **Answering Legal Questions by Learning Neural Attentive Text Representation.** COLING2020: 988-998

[NHT2021]Nguyen Ha Thanh, Bui Minh Quan, Dang Tran Binh, Nguyen Le Minh, Evaluate and Visualize Legal Embeddings for Explanation Purpose, *In Proceedings KSE 2021*

[BMQ2021] Bui Minh Quan, Vu Tran, Nguyen Ha Thanh, Dang Tran Binh, Nguyen Le Minh, How Curriculum Learning Performs on AMR Parsing, *In Proceedings KSE 2021*

[Chau&SCIDOCA2021]. Chau Nguyen, Minh-Phuong Nguyen, Tung Le and Le-Minh Nguyen. Cybersecurity Text Analysis: Identification of Token Labels in Cybersecurity Texts, Presentation at *SCIDOCA 2021*

[Nguyen&COLIEE2020] Ha-Thanh Nguyen, Hai-Yen Thi Vuong, Phuong Minh Nguyen , Binh Tran Dang, Quan Minh Bui, Sinh Trong Vu , Chau Minh Nguyen, Vu Tran, Ken Satoh, and Minh Le Nguyen, JNLP Group: Deep Learning for Legal Processing in COLIEE 2020, *In COLIEE 2020 competition (the best system)*

Participants

You will be asked to provide the information below for: (1) PIs; and (2) each person who worked on and was funded by this grant during the current reporting period. Please include all participants including yourself.

- Provide the name and identify the role the person played in this project.
- **PIs: Nguyen Le Minh (Professor) : Manage all projects**
- **RA: Nguyen Minh Chau (15/10/2019-22/9/2021): working on cybersecurity text (working time: 20months)**
- RA: Dang Hoang Anh (19/04/2021-22/9/2021): Graph deep learning (working time: 06 months)
- RA: Bui Minh Quan (1/08/2021-20/08/2021): Data Aannotation and AMR parsing for extracting graph (working time: 1 month)
- RA: Nguyen Minh Phuong (Ph.D. student) (1/8/2021-20/8/2021): data annotation (working time: 1month)
- RA: Le Thanh Tung (Ph.D. student) (19/04/2021-30/09/2021) : Graph deep learning and VQA (working time: 5 months)
- RA: Kong Wei Kun (Ph.D student) (19/04/2021-22/9/2021): graph deep learning + data annotation (working time: 1month)
- RA: Peng Cheng (1/8/2021-20/8/2021): data annotation (working time: 1month)
- RA: Nguyen Ha Thanh (Ph.D student) (1/8/2021-20/8/2021): data annotation
- RA: Dang Tran Binh (Ph.D student) (19/04/2021-22/09/2021): data annotation
- **Researcher: Dr. Tran Duc Vu (April 1, 2020 – Sep 30, 2020): working on graph deep learning and its application (working time: 06 months)**

Other Partners or Collaborators

We have a colloboration with Prof Tomoko Matsui and Assistant Professor. Tran Duc Vu (ISM). Dr. Tran Duc Vu was our formal postdotral researcher and he has just promoted to ISM as an assistant professor. In JAIST, we have our collaborations with Prof Satoshi Tojo, and Prof Mizuhito Ogawa.

Section 2: Technical Report

Abstract

This research aims at studying how to build knowledge graph representation for cybersecurity by performing open information extraction techniques on a large scale of text documents. After that, we will investigate how deep learning can be applied for knowledge graph representation in cybersecurity. We will consider the use of attention mechanism in graph deep learning model and compare various deep learning models which can be used for learning from graph representation data. Besides, we will study semi-supervised learning frameworks for utilizing unlabeled data for improving the performance of prediction and ranking problems. The proposed method will be applied for detecting malware and code flaws in cybersecurity.

Accomplishments

Research Objective

Our goal is to develop a method which can automatically extract a knowledge graph by reading text documents in cybersecurity. The machine reading techniques including open information extraction and textual entailment recognition are investigated to obtain knowledge graph representation. Then, we would like to develop a *graph-deep learning model that can utilize the knowledge graph to make the learning model more robust. We will consider two learning problems including classification and ranking in cybersecurity*. The classification task aims at determining the type of malware or code flaws. The ranking

task aims at ranking the effect of malware or code flaws. We will focus on using *unlabelled data for improving our learning framework*. In addition, our learning method will address "how the autonomy adapts to changes to the underlying data (e.g. concept drift) or functions with limited annotated training data." We will consider both graph-deep learning for classification and ranking problems with unlabled data.

Major activites

In this research project, we develop a method for building graph knowledge from cybersecurity text and its application to the problem of classification and ranking. The main activities are to study a core technology for graph extraction and its application in machine learning. The major activities are as follows: (1) Developing open information extraction tools for cybersecurity text (2) Exploring the use knowledge graph representation and its application for classification and ranking. (3) For the education activities, the research project has been supported by many students, including master and doctoral students. Mr. Nguyen Minh Chau has finished his master thesis on knowledge graphs and cybersecurity extraction [Chau2021]. (4) For international colloprations, we have organized an internal workshop between the Explainable AI center and UET and Le Quy Don Technical University related to AI and cybersecurity for research activities. We presented our works on cybersecurity at the workshops and published our results in many international conferences and journals. We attended the conferences for presenting our papers in ICTAI 2021 and KSE 2020, KSE 2021, SCIDOCA 2021.

Specific objectives

The first objective of the result is to study how to extract knowledge from cybersecurity text. The second objective is to use the knowledge graph for enhancing the performance of classification and ranking problems. In the first object, we would like to study our technique for extracting knowledge graphs from cybersecurity text. In the application, we

apply external knowledge (i.e., knowledge graph) to enhance ranking and classification performance. Another objective is to study a method that utilizes the knowledge from entity embeddings learned from any knowledge graph embedding model and shows to fine-tuning the pre-trained models and word embedding. We also consider various works on the classification and ranking problems using external knowledge for improving deep learning models.

Signifcant results

In the first objective, we would like to study our technique for extracting knowledge graphs from cybersecurity text. As a result, we develop our method using open information extraction with considering the chunking information for extracting the triple relations, including entity, relation, and entity. After that, we can build a knowledge graph by considering entity as a node and relation as links between two nodes. We published this work in the journal paper [Tran&Nguyen-2021].

Along with the technique using open information extraction, we consider a study on abstract meaning representation and apply it to cybersecurity. We published our work on legal text parsing [Vu-AI&Law2021] and adapted this method to cybersecurity. In the application, we apply our technique to identifying malware-related sentences. This binary classification task is even challenging for neural network approaches as the performance of those approaches on the task is far from perfect. The previous approaches focused on using only annotated data to tackle the task, which may limit the performance of the classifiers. In this work, we propose to leverage external knowledge to enrich the features of the sentences. The experimental results demonstrate that, with the enriched features, a support vector machine (SVM) model gains about 9% on the F1 score compared to the model's performance without the enriched features. We also achieved the best F1 score on the task of identifying malware-related sentences. The results of this work are published in [NguyenICTAI2021][Chau2021]. We also work on identifying token labels in malware

sentences which is presented in SCIDOCA 2021 [Chau&SCIDOCA2021]. As a result, our method combining BERT and sequence learning framework attained the promising results. Another work is to study a method that utilizes the knowledge from entity embeddings learned from any knowledge graph embedding model and shows to aid in fine-tuning the pre-trained models and word embedding [KongKSE2020][KongICTAI2021]. This work supports enhancing the performance of deep learning with the support of external knowledge. We also consider various works on ranking problems with considering the use of external knowledge for transformer models. During the study on deep learning and application, we also develop our method of using VAE for natural language generation. This work is published in the high-rank journal – Computer Speech and Language [TranNguyenCL2021]. The main contribution and significant of this work are to present a Variational-based NLG (VNLG) framework tackling the NLG issues of having a low-resource setting data. Based on this framework, we first propose a novel adversarial VNLG which consists of two critics which are Domain and Text similarity critics in an adversarial training procedure, solving the first issue in domain adaptation. For the second issue of having lim- ited in-domain data, we propose a dual variational model which is a combination of a variational-based generator and a variational CNN-DCNN. We extensively conducted the experiments of both proposed models in various training scenarios, such as domain adaptation and training models from scratch, with varied proportion of training data, across four different domains. The experimental results show that, while the former generator could perform acceptably well in a new, unseen domain using a limited amount of target domain data, the latter model shows its ability to work well when the training in- domain data is scarce. We perform a ranking method for ranking question and answering system [KienColing2020] and present a new method for the visual question and answering (VQA) with pre-trained models [Le&Journal2021] [Le&ICP2021].

In conclusion, there are some significant achievements we obtained during the project as

follows. (1) We have developed a method for extracting knowledge from text. We perform a technique based on open information extraction – ReLink which use linguistic structure as chunking for extraction the triples (entity, relation, entity) and apply cybersecurity text. As a result, we published our work in the conference [Tran&Nguyen2021]. (2) We also use our technique for making a knowledge graph extraction on COVID-19's scientific data. As a result, we published our system available on the website (**https://www.jaist.ac.jp/is/labs/nguyen-lab/systems/covrelex/home/**) (3)We have investigated the use of external knowledge via graph deep learning for improving the performance of machine learning in the problem of classifying malware-related documents. Another work is to purpose Knowledge Graph for Word Embedding (KGWE) that utilizes the knowledge from entity embeddings learned from any knowledge graph embedding model and shows to aid in fine-tuning the static word embeddings. We evaluate the proposed model using the word similarity task with various benchmarks, and the results demonstrate that the word embeddings fine-tuned by the KGWE with a BoW-based encoder can significantly outperform the baseline word embeddings.

Impacts

Development of the principal discipline(s) of the project

This research shows a method for constructing knowledge graph representation for cybersecurity by performing open information extraction techniques and abstract meaning representation parsing on a large scale of text documents. In addition, we investigated how deep learning can be applied for knowledge graph representation in cybersecurity. The results on extracting cybersecurity text can be adapted to other domains such as scientific data and biomedical documents. We have also employed our work to develop a COVID-19's scientific support system which is made available online in our website. Besides, our

semi-supervised learning frameworks and pre-trained models could utilize unlabeled data for improving the performance of prediction and ranking problems other than cybersecurity text. The use of knowlege graph can help to improve pre-trained models and support enhancing the performance of deep learning models in some important applications. Integrating deep learning with knowledge graph can help the learning framework enhance the interpretability.

Other disciplines:

One of the good issues is that our integrating knowledge graphs and deep learning would contribute to other fields like the bioinformatics domain. The machine learning models will have incorporated rich information from knowledge graphs that can also enhance interpretability. It can contribute to the explainable AI field or make the machine learning model is more robust. Otherwise, the supporting system for reading using knowledge graphs can help researchers save more time accessing the new knowledge. It might boost other scientific research.

Describe the impact in this reporting period on the development of human resources

The project's impact is that we support provided opportunities for research and teaching in knowledge graph extraction, deep learning, and information extraction. One example is the master students who studied in this project and got their master's degree. After getting the master's degree, the students continued to study a Ph.D. program at our laboratory. The researcher who participated in the project has been promoted to Assistant Professor at the other university in Japan.

The impact on teaching and educational experiences

The impact on teaching and educational experiences are vast. We have introduced the results of using deep graph learning for malware analysis in our group seminar and as an

example in the NLP lecture at JAIST. We also develop an online resource on cybersecurity and knowledge graph searching for COVID-19. The system for extracting knowledge graphs on scientific papers on COVID-19 is indicated on the website.
https://www.jaist.ac.jp/is/labs/nguyen-lab/systems/covrelex/home/

The impact in this reporting period on physical, institutional, and information resources that form infrastructure.

We have developed a collaboration in research with Vietnamese Universities in the study of malware detection. A collaboration among Professors in JAIST within our project was also developed. We share our data and tools related to the knowledge graph and deep learning utilizing the knowledge graph. We also develop training data for AMR parsing in cybersecurity and machine learning techniques for analyzing cybersecurity text.

Impact on society beyond science and technology:

The impact on society beyond science and technology can be understood that if our technique can support users for reading the reports of cybersecurity text quicly. It can be also applied for other domains such as legal domain and scientific documents.

Technical Updates

This section will show some additional information regarding to the work we conducted during the project.

Update 1: Technical updated for our framework of malware-related sentence prediction – we published in ICTAI 2021 ([NguyenICTAI2021])

The framework showed our results when we used Graph Attention Network by utlizing Atribute Reference Guide to enrich the features of sentences via the two methods:

- Produce weak labels as features

- Produces attribute label weights via heuristically determining how likely a sentence is classified in each of 444 malware attribute labels

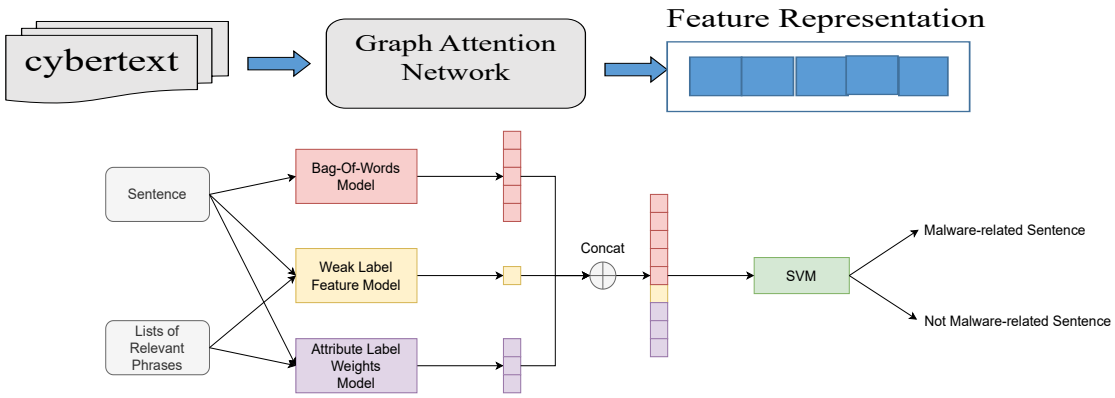

Figure 1. The proposed method for malware-related sentence prediction

Update 2: The technical summary of the SCODICA2021 [Cybersecurity Text Analysis: Identification of Token Labels in Cybersecurity Texts]

We present a follow-up research on the task of cybersecurity text analysis (which contains four subtasks). In previous research, we achieve state-of-the-art performance on the task (task 1): identify malwarerelated sentences. In this paper, we proceed with tackling another task (task 2), which is to identify token labels for cybersecurity sentences. Specifically, we formulate this task as a sequence labelling task. While previous work focused on exploiting the ability of Conditional Random Fields (CRFs) for sequence labelling, we further leverage the language understanding ability of pretrained language model to tackle this subtask. Additionally, we also leverage our model for task 1 to help produce the predictions for task 2, results in the state-of-the-art performance on the MalwareTextDBv2.0 dataset (the largest dataset for cybersecurity text analysis) with an improvement of 3.70% on relaxed F1 score comparing to previous approaches.

Update technical 3:

We have applied open information extraction to extract the cybersecurity text (Malware

DB 2.0) obtain a knowledge graph for those documents. Figure 2 shows some information about our knowledge graph.

Total Sentences: 11255			
Top K	Number of sufficient triples	Nodes	Edges
1	10594	16488	10418
2	18782	27220	18472
3	24085	33689	23683
all	29504	39951	29030

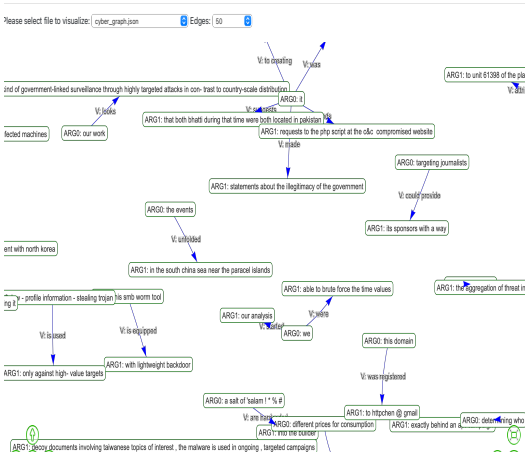

Figure 2. Knowledge graph extraction from cybersecurity text.

Update 4: Our results for graph searching and ranking on COVID-19's scientific data

This system presents CovRelex, a scientific paper retrieval system targeting entities and relations via relation extraction on COVID-19 scientific papers. This work aims at building a system supporting users efficiently in acquiring knowledge across a huge number of COVID-19 scientific papers published rapidly. Our system can be accessed via https://www.jaist.ac.jp/is/labs/ nguyen- lab/systems/covrelex/. Figure 3 illustrated our system for extracting knowledge graph from COVID-19 scientific papers. We obtained approximately **33.3 million relations and 6.4 million entities. This system is available online the relations and entties is ready for scientific research.**

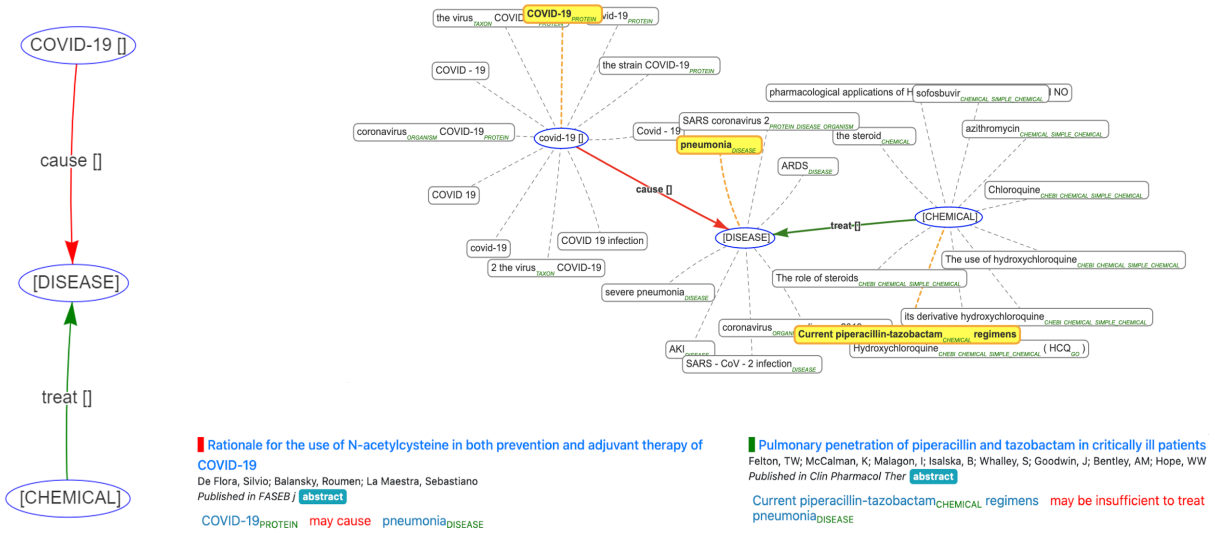

Figure 3. An example of the Covrelex system

Thank you for your support in helping AFOSR discover, shape and champion basic science research that profoundly impacts that future of the Air and Space Forces!