Knowledge Graph Construction for Detecting Cybersecurity Attacks

Shraddha Mukesh Makwana
University of Alberta
CCID:smakwana
smakwana@ualberta.ca

Pranjal Dilip Naringrekar
University of Alberta
CCID:naringre
naringre@ualberta.ca

1 Motivation

In this project, we aim to develop a cybersecurity knowledge base system that can be queried by Security Analysts working in a Security Operations Center to detect cyber-attacks. Cyber-attacks target individuals, small and medium enterprises with an aim to compromise confidentiality, integrity, and availability of information. Every year cyber defense professionals and researchers find millions of attack and malware variants. Security Analysts hence, have to be up to date with these attacks. Therefore, this idea arises from the fact that a Security Analyst's background knowledge is sometimes insufficient for detecting evolving attacks as they are complex and varied. In order to help them with a decision about cyber-attack, all open source threat intelligence sources need to be stored in a structured way in a cybersecurity knowledge graph that can be queried (Sarhan and Spruit, 2021).

The major motivation of developing this system comes from a news article called 'Attack Samba Server' which specifies that an attack can be caused due to some vulnerabilities exposed on port 445. The attackers usually take advantage of the Log4j vulnerability and port 445 to get access to the network. The analysts need to process the information of these attacks with respect to their local defensive setup. Therefore, our system would take the two entities as input and then derive the relationship between them to aid the security analyst with knowledge of possible attacks. For instance, two entities like Log4j and port 445 are given as input and an exploit relationship is derived and a knowledge graph with other relationships is also formed.

2 Related Work

Jia et al. (2018) firstly, constructed a knowledge graph for cybersecurity by collecting and analyzing structured and unstructured data. Secondly, they used Named Entity Recognizer (NER) to extract the entities and to train the extractor. Thirdly, they constructed the ontology according to the information that has been obtained and then generated the cybersecurity knowledge base. They deduced rules based on a quintuple model. New attributes were deduced using the attribute value prediction formula and new relationships between instances were obtained using the relational reasoning predictive formula and the path-ranking algorithm.

A research for improving the cybersecurity knowledge graph was performed by Pingle et al. (2019). RelExt made use of a semantic triple generation technique that consisted of two cybersecurity entities and the relationship between the entities. They made use of deep learning approaches to extract possible relationships and evaluated their technique by asserting the entity relationship generated by RelExt in the Knowledge Graph to get information about various cybersecurity threats.

Piplai et al. (2020) describes a system that makes use of After Action Reports to extract cyber-knowledge. The entities are extracted using a customized extractor called the Malware Entity Extractor. Later, they construct a neural network to predict how pairs of 'malware entities' are related to each other. Furthermore, similar entities are fused to improve CKG that would aid the security analyst to execute queries to retrieve better answers.

3 Proposal

The main focus of our project is to develop a system with plenty of knowledge about the cybersecurity domain that is a knowledge graph and would assist the Security Analyst to make informed decisions about an attack by querying the system.

As shown in Figure 1, our project is divided into three sub-categories. In the first part, we would gather data and perform filtering process based on

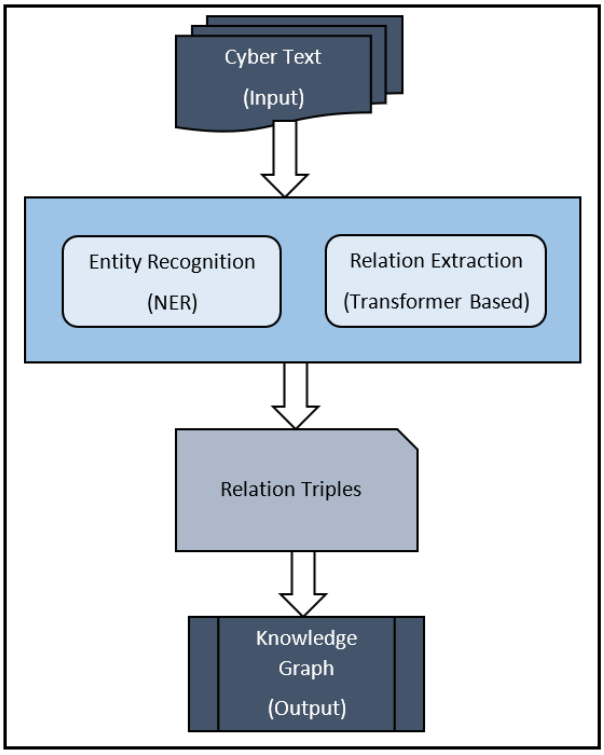

Figure 1: System Architecture

the required cybersecurity text, in second part we will perform cybersecurity entity recognition using Named Entity Recognition and extract the relationship amongst these entities using Transformer Based Model to create a semantic triplet and lastly we would construct a knowledge graph that can be queried. Our aim is to build a relation extraction model which will be a classifier that predicts a relation 'r' for a given pair of entities e1, e2.

For instance, 'cache' and 'Out-of-order execution' are related to one another resulting in attacks like Spectre and Meltdown. If these entities are searched on google it would not be able to relate it to an attack, however, our relation extractor will make sure to find the relation 'combined with' between these two entities.

4 Experimental Evaluation

We will make use of two main indicators: precision and recall to determine how accurately the relationship (eg. 'hasVulnerability') class has occurred. To comprehensively evaluate the system we will also make use of the F1-score parameter. It is calculated as the harmonic mean of precision and recall, as follows:

$$F1 = 2 \times (\frac{(precision \times recall)}{(precision + recall)})$$

Along with this, we will assert the semantic triplets in the knowledge graph to ensure correctness.

5 Expectations

The foremost expectation out of this project is that our system should perform the tasks of cybersecurity entity recognition, entity relationship extraction and knowledge graph construction effectively. Along with this, we also expect that the system is able to help the Security Analyst to make informed decisions about any new cybersecurity attack by answering their queries appropriately.

6 Detailed Timeline

In the given time frame, we aim to identify the cybersecurity entities by mid of February. Following that, we aim to achieve the task of finding relationships between these cybersecurity entities using transformers by the end of February. Later, until the end of March we will focus on constructing the knowledge graph for the derived information. The future scope of this project is to build an Analyst Augmentation Systems that would combine multiple cybersecurity knowledge graphs to provide more accurate data.

7 Github Repository URL

The url to track our project work can be found at

`https://github.com/pranjal080598/KG_Cybersecurity_Attack`

References

Yan Jia, Yulu Qi, Huaijun Shang, Rong Jiang, and Aiping Li. 2018. A practical approach to constructing a knowledge graph for cybersecurity. *Engineering*, 4(1):53–60. Cybersecurity.

Aditya Pingle, Aritran Piplai, Sudip Mittal, Anupam Joshi, James Holt, and Richard Zak. 2019. Relext: Relation extraction using deep learning approaches for cybersecurity knowledge graph improvement.

Aritran Piplai, Sudip Mittal, Anupam Joshi, Tim Finin, James Holt, and Richard Zak. 2020. Creating cybersecurity knowledge graphs from malware after action reports. *IEEE Access*, 8:211691–211703.

Injy Sarhan and Marco Spruit. 2021. Open-cykg: An open cyber threat intelligence knowledge graph. *Knowledge-Based Systems*, 233:107524.