Research Paper Review – Knowledge Graphs

Gerald T. Rigdon

March 4, 2025

The paper "Constructing Knowledge Graphs from Cyber Threat Intelligence Using Large Language Model" (Liu, J., Zhan, J., 2023) makes the case for the use of knowledge graphs to automate the extraction of important information from Cyber Threat Intelligence (CTI) reports with the assistance of a Large Language Model (LLM), namely ChatGPT. A knowledge graph is a data structure that organizes information into a graph format. It consists of nodes or entities and edges which are relationships that represent knowledge. While visually akin to call graphs that represent the behavior and internal workings of software systems by displaying interconnecting nodes to model relationships, knowledge graphs focus on the retrieval of targeted salient information from text by leveraging the neural network power of LLMs.

Research Problem

The paper acknowledges the rapid evolution of cybersecurity threats which results in the need to compile volumes of CTI reports in an attempt to keep pace. While these reports contain valuable insights into attack patterns, strategies, and Indicators of Compromise (IoC), it is too time consuming to extract the information manually. While acknowledging the use of Natural Language Processing (NLP) techniques such as Named Entity Recognition (NER) and Relation Extraction (RE), this typically results in the creation of specialized or custom fine-tuned models to generate useful graphs that capture domain specific concerns. Hence, this approach could be just as burdensome or even more so than manual data extraction. To solve this dilemma, the paper proposes a practical solution by employing easily accessible models like ChatGPT to build an effective knowledge extraction pipeline. The key to the research was the development of CTI-specific ontology that allows models like ChatGPT to extract well formatted metadata from CTI reports for the generation of Subject-Predicate-Object (SPO) triplets. In these triplets, the subject and object are entities, and the predicate is the relationship between them. The catalyst was that through zero shot learning or the ability to use a model that has not been fine turned, an LLM can leverage its pre-trained "Off The Shelf" neural network to extract relevant information.

Research Questions

Although not stated outright, the paper answers research questions such as: 1) How can information be automatically extracted from CTI reports? 2) Can a knowledge graph be constructed without the need for LLM fine-tuning? 3) How does the research proposal compare to other approaches such as AttacKG and REBEL? 4) Can the proposed method be effective for other use cases outside the cybersecurity domain?

Deriving these questions is apparent since the paper clearly identifies its contributions. First, it proposes a pipeline for construction of a CTI knowledge graph based on specific ontology requirements. Second, it demonstrates that ChatGPT can be effectively used to simplify the process of building structured knowledge. Third, it boasts that the research is extensible by offering up a re-usable prompt template and processing pipeline.

Hypothesis and Methodology

While there is no specifically stated hypothesis, one could conclude that the research is presented to demonstrate that it is possible to create a pipeline that extracts precise information from CTI reports by generating accurate SPO triplets to create a knowledge graph that can in turn be leveraged to allow one to identify and understand key patterns, relationships, and most importantly, cybersecurity threats.

The research method was conducted using the following approaches employing only ChatGPT, without the need for customized or fined-tuned LLMs.

- Ontology: The ontology, which is a way to create a conceptual or generalized view of data, employed STIX v2.1 Relationship Object (SRO) as the foundational component. From that a customized ontology was built, one that imposed constraints on entity and relation types from CTI reports that enabled a first pass generation of SPO triplet candidates that were subsequently fed back into ChatGPT along with the text from the CTI report. The LLM response produces a final set of SPO triplets providing the information necessary to create a satisfactory knowledge graph. The governing reason for this approach is to provide a way in which humans and machines can easily interpret data and perform actions on them.
- Prompt Engineering: The research designed a prompt, which is a set of instructions that assist an LLM in executing a task and producing a response that meets the user expectations. The prompt is built around three components: Role, Task, and Format. It includes a metadata template in JSON format so that the LLM will respond with data packaged accordingly.
- Metadata Extraction: ChatGPT was used to complete a JSON object representing the metadata of a CTI report where the internal schema was based on the overarching ontology. When an LLM chunks raw text to produce output based on similarity-search algorithms it may miss important information such as an IoC within a CTI report. To shore up this gap, an additional step using a traditional regular expression search was employed to make for a more complete processing step such that the metadata generated would initially produce better candidate SPO triplets.
- SPO Triplet Generation: This approach is the heart of what makes the research proposal innovative. Using the capabilities of ChatGPT, the first pass generated candidate SPO triplets are fed back to

ChatGPT so that it can discern the correct final SPO triplets. As part of this final selection process, ChatGPT is presented with the candidate SPO triplets along with the original CTI report text in the prompt, as context.

- Visualization: The final step of visualization is what brings the knowledge graph to life. In this visual rendering, a human analyst is provided with a graphical view of the interconnections of an Advanced Persistent Threat (APT).

Dataset and Sampling

The dataset for the study included the use of 13 CTI reports. Although this is a limited sample size, there was strong sample diversity in that the reports were collected from real-world public sites and each report chosen represented a different cyber-attack campaign.

Ground Truth, Analysis and Results

The ground truth or baseline test set, which serves as the evaluation criteria, was established from conventional manual approaches. Next, output from these 13 CTI reports was produced using the popular REBEL and AttacKG CTI extraction methods. Finally, the ChatGPT extraction pipeline results were compared against the REBEL and AttacKG results with respect to the ground truth baseline.

The metrics were based on Precision, Recall, and F1 where:

TP = True Positive (Predicted positive and is positive)
TN =True Negative (Predicated negative and is negative)
FP = False Positive (Predicted positive but is negative)
FN = False Negative (Predicted negative but is positive)

Precision = TP / TP + FP
Recall = TP / TP + FN
F1 = (Recall * Precision / Recall + Precision) * 2

When comparing the results on entity identification in threat intelligence, AttacKG demonstrated remarkable results in IoC identification related to things such as IP addresses and executables whereas ChatGPT excelled in recognizing entities in the form of proper nouns. The marked advantage of ChatGPT in this area means that it was able to extract organization names, malware names, and unknown file formats. The average results across the 13 CTI reports showed ChatGPT and AttacKG comparable in Precision (0.76 vs 0.80 respectively), whereas ChatGPT dominated in Recall (0.82 vs .047) and higher in the F1 score (0.78 vs 0.58).

In comparison to the REBEL approach the average results demonstrated ChatGPT superiority where: Precision (0.60 vs 0.25), Recall (0.60 vs .050), and in the F1 score (0.56 vs 0.31).

However, as a word of caution, it was noted that while the ChatGPT results were impressive it sometimes was overly creative resulting in incorrect entity relationships.

Limitation and Gaps

One potential limitation is that the final pipeline stage could be LLM dependent. At this part of the pipeline ChatGPT is given the candidate SPO triplets along with the original CTI report text. There is an implicit assumption here that the chosen LLM can process the report in its context window. Is this a single report? What if the report is hundreds or thousands of pages in length, exceeding LLM context window token limits? Another weakness is that the metrics evaluation was limited to Precision, Recall, and F1 scoring. Adding other metrics such as Accuracy, Specificity, Sensitivity, etc. would be more comprehensive, providing further insights as would the exploration of alternative benchmarking techniques. Finally, the paper did not address the production of and processing of CTI reports in a real-time context. As the rate of cyber-attacks escalate there may be a need to generate intelligence information in real-time which would result in continuous CTI report generation. Hence, the research should consider how well the proposed pipeline could perform with these demands.

Research Extension Possibilities

By parsing these reports into structured JSON files and utilizing key evaluation metrics, the study aims to improve the efficiency and effectiveness of information extraction processes. This approach not only enhances the understanding of cybersecurity threats but also provides a robust framework for future research and development in the field. The paper recognizes the limitations of the current approach where the quality of the SPO triplets is heavily influenced by the quality of the chosen ontology. Hence, it recommends further experiments that could be done simply by repeating the experiment with more advanced LLMs and better future iterations of ChatGPT. Certainly, more work could be done to enhance the LLM pipeline's ability to understand the context in which entities and relationships are mentioned, thereby reducing false positives, and improving the accuracy of the SPO triplets. Moreover, there are many easily accessible models other than ChatGPT which could be used to conduct LLM comparisons using the proposed automated pipeline. Finally, more work could be done to continuously update the ontology to consider new threats and techniques as the cybersecurity landscape changes.

Conclusion

The research paper provides a comprehensive overview of the limitations that exist in processing an abundance of CTI reports and presents an effective alternative. It provides a practical and efficient approach to a complex problem in cybersecurity and leverages the capabilities of easily accessible LLMs, like ChatGPT, to automate a manual process and even outperform other popular methods. The proposed solution

is adaptable and can be configured for different domains, making it versatile. Finally, it cites a number of additional references that provide further context which makes the overall presentation a valuable resource for other researchers in this field of study.

Marymount Honor Pledge

I agree to uphold the principles of honor set forth by this community in the Marymount University mission statement and the Academic Integrity Code and Community Conduct Code, to defend these principles against abuse or misuse, and to abide by the regulations of Marymount University.

References

Liu, J., Zhan, J. (2023). “Constructing Knowledge Graphs from Cyber Threat Intelligence Using Large Language Model.” In Proceedings, 2023 IEEE International Conference on Big Data.