Actionable Cyber Threat Intelligence using Knowledge Graphs and Large Language Models

DECISION: Accept

AUTHORS: Romy Fieblinger, Md Tanvirul Alam and Nidhi Rastogi.

Summary of Reviews

- Review 1: 1 (3)
- Review 2: 1 (3)

Reviews

Review 1

Total score: 1

Overall evaluation: 1

Reviewer's confidence: 3

The study tests different open-source LLMs and various techniques (prompt engineering, guidance framework and fine-tuning, etc.) to automate the extraction of knowledge graphs (KG), which can yield actionable cyber threat intelligence (CTI). The main conclusion, however, is that using LLMs, regardless of on the type of techniques (although some outperforms others) on large datasets to extract KG for CTI, yields limited results.

Strength of the paper:

- Does the work that many analysts need to do: test different models and techniques to see which one outperforms others for CTI

- Highlights the impact of various extraction methodologies.

- Can become a reference for other researchers working on similar topics (use of LLM for CTI)

- The paper is well written and clear, it can be published as is (one typo first page, second column, second paragraph: "and and")

There are two main weaknesses to the paper:

1) The results are limited since the models do not perform well. At this point, the paper can be used as a reference for further improvement (as argued by the authors).

2) The paper does not go beyond technical results for fine-tuning LLMs. Why and how is this relevant for the WACCO venue?

Areas of improvement:

- The authors could better explain why KG are key to CTI (why choose this format over another?)

- Figures of triples as opposed to text would lighten the text

- The parameters set for each LLM type and technique could be added in an annex for research reproducibility and clarity purposes (temperature, max token, etc.)

Review 2

Total score: 1

Overall evaluation: 1

Reviewer's confidence: 3

Overall, I greatly enjoy reading the paper contribution and the authors experimental setup to critically evaluate the possibilities that LLMs offer for CTI processing. I do find some minor concerns with the methodology that could be better substantiated.

- When it comes to the engineering of the prompts, how did the process actually work? Did you follow some guidelines like the ones provided by OpenAI (https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-the-openai-api)

- The choice of ROUGE-N as metric to evaluate the performance is fully justified. Why not using ROUGE-BE which extends ROUGE by incorporating semantic role labeling and entity recognition for a more comprehensive evaluation of text summarization systems?

- The paper's methods are evaluated on CTI extraction from Android malware reports, yet the reliance on only 36 malware types raises concerns about external validity. Have you considered the potential bias and limited generalizability of findings? Could expanding the dataset to encompass a broader range of malware types address these concerns and provide more robust insights into real-world CTI scenarios?

- While your use of the BRAT annotation tool to manually annotate reports sounds doable for the sample size at hand, I'm curious about the manpower involved. How many annotators were engaged in this process, and was there any inter-annotator agreement checking to ensure consistency and reliability across annotations?

- The paper underscores the fine-tuned 7B model's optimal performance on test data, supported by ROUGE scores and human evaluations. However, its application to a large-scale dataset for KG generation yielded noisy triples, indicating scalability challenges from curated test datasets to

unprocessed data. This raises questions about strategies to mitigate noise in large-scale KG generation, the efficacy of post-processing techniques, and the scalability of fine-tuning approaches. What strategies could be explored to mitigate noise in large-scale KG generation? How effective are current post-processing techniques in improving KG quality? What scalability issues might arise with fine-tuning approaches, and how can they be addressed?

- While the paper compares prediction efficacy between LADDER-generated triples and those from the fine-tuned 7B chat model, some aspects could be improved (or further clarified). The discrepancy in overall scores compared to previous research (e.g., Alam et al.) raises questions about the generalizability of the proposed methodology across different datasets and domains. Moreover, exploring the interpretability and explainability of predicted relationships could provide valuable insights into the underlying mechanisms driving link prediction in unstructured CTI knowledge graphs. How can the proposed methodology be adapted to incorporate domain-specific knowledge mappings to improve prediction performance and applicability in real-world scenarios? What strategies can be employed to enhance the interpretability and explainability of predicted relationships within the CTI knowledge graph?