From Word Embedding to Cyber-Phrase Embedding: Comparison of Processing Cybersecurity Texts

Cameron Noakes

February 2022

1 Sentence Overview

The researchers demonstrate that using cyber-phrase embedding on cybersecurity text is a promising approach to overcome difficulties of manually providing an exhaustive list of variants of threats and in identifying such cyber intelligence from natural language texts. The researchers also have created an open source project to make their tools and data available to the public. The researchers used 8926 cyber-related articles under categories of computer security to develop their research and proposed model.

2 Introduction

It was outlined by the researchers that previous work has attempted to classify cyber threat information (CTI) in natural language texts but some papers did not conclude on any development or evaluation results, leaving a gap in the market for this complete development and research seen in this paper.

The goal is to identify cyber intelligence in natural language texts that map to similar terms to those in bold in the table provided in the research seen in table 1.

Recognition of a difficult overcoming challenge was displayed of how it is not always possible to enumerate all equivalent expressions of a given cyber concept such as the program not recognising how 'Windows' and 'Operating System' are related.

The referenced research is an open-source cybersecurity embedding model developed by UMBC which was used to identify common ground between studies to better develop a more comprehensive model and research to conduct.

The UMBC model outlined performs better than referenced Google model because it is trained using a cybersecurity text corpus. However, the UMBC model still missed many matches that humans could easily recognize.

The proposed model was explained as a trained cyber-phrase embedding model, which could have the option of training the model with the same text corpus as UMBC's model for performance comparison but unfortunately the researchers outline how the corpus is not publicly available and the UMBC model is the only open source word embedding mode.

3 Cyber Phrase Embedding

The embedding model captures syntagmatic (denoting the relationship between two or more linguistic units) and paradigmatic relationships among words. Observations were conducted by the researchers to identify that in addition to words, phrases with multiple words often form the basic unit for relationships previously mentioned.

A common explained example is that if an attack was about password spraying and the terms 'password' and 'spraying' were identified separately, this could make the attack fall into a whole other completely different category. The researchers use a rule-based crawler to collect cybersecurity related information from cyber threat intelligence reports.

The researchers discussed how they removed any unnecessary data such as advertisements, references, contacts, codes, links, and others using pre-defined rules. They used 8926 cyber-related articles under categories of computer security to develop their research and proposed model.

4 Evaluation

The researchers compare five embedding models (their phrase embedding model (CP-CBOW) and word embedding models using the same corpus as the phrase embedding model. They demonstrate, through examples, that phrase embedding can lead to valuable semantic relationships that cannot be constructed based on a word embedding model.

5 Conclusion

Phrase embedding as an effective text mining approach to aid the understanding of cybersecurity texts. This allows textual descriptions to be used in better understanding cyber texts and provides an approach to gathering more information via textual description mining.

The conducted research demonstrated that phrase embedding outperforms state of the art open source word embedding models. There was acknowledgement of present challenges that remain to improve the performance. The accuracy of the embedding models can often depend on the size of the training corpus and is the main factor at play.

6 What problem does this solve?

Textual descriptions can often have a lot of related information that is not seen in numerical values for various applications such as vulnerability databases, exploits and alike. Due to this, the textual descriptions can provide more information for ML algorithms, mitigations and attacking infrastructure.

7 How do they solve it?

The researchers identified this and took this into account by using phrase embedding models to extract and relate textual descriptions to the ATT&CK tactics and techniques in the framework matrix.

8 How do they assess it?

The provided evaluation demonstrates that the model developed by the researchers outperforms any pre-existing models out there for the same task and

solving the same problem which creates an in-demand aspect to the research conducted and solution developed.