Technical Note

Paragraph-based Estimation of Cyber Kill Chain Phase from Threat Intelligence Reports

THIN THARAPHE THEIN¹,a), YUKI EZAWA¹, YOSHIKI SHIRAISHI¹, MASAMI MOHRI³, SHUNTA NAKAGAWA¹, YASUHIRO TAKANO¹, KEISUKE FURUMOTO², YOSHIaki SHIRAISHI¹, MASAMI MOHRI³, YASUHIRO TAKANO¹, MASAKATU MORII¹

Received: March 9, 2020, Accepted: September 10, 2020

Abstract: In order to keep up with the increasing number of cyberattacks, the defense tactics require timely and accurate understanding of the threats and corresponding risks. We propose a scheme for modeling threat information to extract event information from security reports on a paragraph basis and then estimate their kill chain phases. The experimental results show that the model got an average F1-score of 0.67, the average accuracy of 65% of the cyber kill chain phases and 86% of core features can be extracted by using this method.

Keywords: cyber kill chain, diamond model, threat intelligence, security report

1. Introduction

A large number of network attacks, including Advanced Persistent Threat (APT), have been targeting various organizations in recent years. Most APT attacks evade detection and conduct potentially destructive long-term attack activities using sophisticated intrusion routes. To counter this, many security operators, engineers and researchers have been paying attention to Threat Intelligence field, which involves collecting vulnerability and threat information, analyzing and organizing so that they can be easily accessible. By utilizing Threat Intelligence, it is expected to be able to predict future attacks from existing ones and estimate the relevance actions between various attacks. It is therefore necessary to analyze multiple pieces of threat information in an integrated manner.

With the goal of increasing cyber security awareness, various organizations often share analysis of attacks information in the form of security reports. We have proposed a modeling procedure for the integrated analysis of threat information contained in security reports [1]. To use threat information promptly, a challenge is to establish an automatically modeling procedure. This paper treats security reports on a paragraph basis considering that one event is described in one paragraph and then estimates phases in the cyber kill chain and extracts event information.

2. Background

2.1 Cyber Kill Chain

Cyber Kill Chain [2] is an intelligence-driven model for intrusion detection analysis of attack activities with seven stages as shown in Fig. 1. These stages assist the security analysts in having a practical understanding of an adversary’s tactics, techniques, and procedures. The adversary must go through these series of stages (chain) to accomplish the intended goals and breaking of any of these steps will interrupt the entire attack process.

Generally, an APT goes through seven phases: Reconnaissance, Weaponization, Delivery, Exploitation, Installation, Command and Control (C2), and Actions on Objectives.

2.2 Diamond Model

The Diamond Model [3] has been proposed to integrate the series of attack activities by adversary and is typically used in conjunction with the Cyber Kill Chain model. The Diamond Model as shown in Fig. 2 consists of four elements: adversary,
infrastructure, capability, and victim. These processes are called events. In this model, an event, which is a minimum unit of the chain, is represented by a diamond shape and the four elements are located at each vertex of the diamond. These elements are called core features.

The adversary is an attacker or organization utilizing capability (tools and techniques) against victim in order to reach the desired goal. The infrastructure is the logical or physical communication system used by the adversary to deliver capability, maintain control and gain benefits from the victim. The infrastructure could be e-mail addresses, domain names, and IP addresses, etc. The victim is always the target of the adversary.

### 2.3 Activity Thread

A chain of events contained in one attack activity, which have a causal relationship for the purpose of the attack, can be represented by a directed graph called an Activity Thread. In order to represent the order of events in an attack activity, the transitions between the event source and the destination are connected by an arrow. By doing so, the attack activity can be expressed as a form of Activity Thread as shown in Fig. 3 and the causality of the events can be clarified by analyzing the Activity Thread. Once an activity thread based on diamond events has been made, it can identify each event using the Kill Chain model.

### 2.4 Existing Research on Threat Information Extraction

Research on modeling techniques that can automatically extract useful threat information from online security forums, blogs and threat reports has been a focus of interest. Hutchins et al. [2] introduced a method which categorizes APT attacks to kill chain phases in order to have a better understanding of attacker actions, steps, and motives. Huseri et al. [4] proposed a technique that extracts threat actions from unstructured text of security reports based on a semantic relationship. Each threat action is then mapped to appropriate tactics, techniques, and a kill chain phase and generates STIX (Structured Threat Information eXpression) standard formatted reports.

### 2.5 Outline of ChainSmith Model

Zhu et al. [5] proposed a system called ChainSmith that can automatically extract the Indicators of Compromise (IoCs) from security articles and categorized them with their corresponding kill chain phases. The key intuition behind this system is that the context words in adjacent sentences in security articles indicate a kill chain phase, and the context words that directly relate to the IOC determine its level of maliciousness. Moreover, to learn the semantics similarity among words, ChainSmith utilizes a dependency-based wording embedding [6] that uses words dependencies instead of just context words. In this approach, six types of IoCs named entities: URL, IP, hash, malware family, Exploit Kit and CVE, are extracted and classified their kill chain phases by training the neural networks.

### 3. Proposed Scheme

We aim to establish a modeling method for classifying cyber kill chain phases for threat information described in security reports. We assumed that events are described in each paragraph unit of security reports. In this paper, we propose a model for analyzing security reports in paragraphs, which consists of extracting event information and estimating kill chain phases. Figure 4 shows the flow of the proposed method. It estimates the phase and then extracts informative words for core features of the diamond model from each paragraph. ATT&CK [7] provides lists of attackers and malware, so we use ATT&CK as words list for pattern matching in core feature extraction.
3.1 Word Embedding

In order to understand the semantic similarity among words, the state-of-art word2vec [8] algorithm is used in the proposed method to parse words semantically. The word2vec processes the text corpus as an input and outputs the vectors that are distributed numerical representations of word features. One of the key features of word2vec is that the word vectors generated take up much lesser space than one hot encoded vector. Another is that it holds the semantic meaning of the word since similar words are grouped in a vector space.

3.2 Paragraph-based Estimation of the Cyber Kill Chain Phase

Since the security report hardly describes the Reconnaissance and Weaponization phase of the cyber kill chain, the remaining five phases: Delivery, Exploitation, Installation, Command and Control, and Action on Objectives are only considered in this paper. Figure 5 shows the procedures for estimating the cyber kill chain phase. In our model, we use five binary classifiers using neural networks that are constructed in the same way as in Ref. [5]. In Ref. [5], the binary classifiers are used to predict which phases are represented by each sentence. This is based on a motivation that different IoCs may appear in each sentence and they want to extract them as much as possible to structure the threat information. This paper differs from Ref. [5] in that it focuses on extracting event information from security reports. Therefore, instead of predicting the phases of each sentence, we extract approximate event information by predicting the phases of each paragraph. The classifier is trained by the example sentences from ATT&CK for Enterprise. ATT&CK is a knowledge base managed by MITRE corporation that categorizes the behavior of the attacker in terms of Technique and Tactics. The kill chain phases are then predicted by inputting the security reports into the trained neural networks in a paragraph unit.

3.3 Classification Model

Firstly, preprocessing is done on input text corpus with off-the-shelf NLP techniques. In this step, lowercase conversion, removal of stop-words, punctuation, and special characters are performed. Next, each sentence is tokenized into words and lemmatization is applied to each word. After this process, we parse each word by using word2vec, in which semantically similar words will be in a close position in the vector space. The word vector is trained with the embedding dimension of 100.

In the next step, the kill chain phases are estimated by five binary classifiers of neural networks. The classifier is designed with input, output, and one hidden layer with 50 nodes. In this step, the features to be fed into the classifier are identified. Firstly, informative words are calculated by using the following equation [5]:

$$\text{Score}(w) = \max_{k \in K} \frac{p(w | k)}{p(w)}$$

where the word $w$ represents one of the words appearing in the all documents, $p(w)$ represents the probability of occurrence of the word $w$, and $k$ represents one of the set $K$ of all kill chain phases, and $p(w | k)$ is the probability of occurrence of the word $w$ in the all documents describing $k$. These probabilities are used to calculate a score $\text{Score}(w)$, which represents a degree to which the word $w$ is specific to a certain kill chain phase. We consider informative words to be words with a high score and a high occurrence.

Next, the context words for each sentence that will be fed into the classifiers are calculated. The context words are determined from two statements; informative words of the current sentence and informative words of previous sentences if no informative words are found in current sentences. The average word embedding of the context words, the words used in the title, and the number of IoCs in each paragraph are then passed into the neural networks as the input features. The classifiers are then trained to determine whether each paragraph unit of the security report falls into any of five phases.

3.4 Core Features Extraction from Paragraph mainly with ATT&CK

The task of this paper is to extract core features words consisting in the Diamond Model regardless of the types such as Adversary, Victim, Capability, Infrastructure toward the goal to extract event related words and to classify them into 4 types of core features. In cases where classification is easy, core feature classification is also conducted.

Three types of core features: Adversary, Infrastructure, and Capability are extracted in this paper. And the uncategorized words that could also be considered as the core features are extracted as Candidate Words. A total of 4 core features are extracted. Firstly, word lists of the following four statements are made.

1. Computer related words described in Wikipedia [9]
2. Software names such as malware and tools etc., described in ATT&CK
3. Group names of attack activities described in ATT&CK
4. New General Service List (NGSL) [10]

First, words to be stored in the core features are extracted. The IP address, URL, e-mail address, file name, and CVE are extracted using the IoCs extraction tool called Cyobstract [11]. The extracted IP addresses, URLs, and e-mail addresses are classified into Infrastructure. The file name and CVE are classified into Capability. The extracted words are deleted from the text.

Next, we check whether the words in the list (1), (2), and (3) are described in the text. In that case, they are extracted and deleted from the text. Words derived from (1), (2) and (3)
are classified into Candidate Words, Capability and Adversary respectively. After that, we check whether the words in the list (4) are described in the text. In that case, they are deleted from the list of words. Then, words that have not been deleted in the previous removal of the NGSL word list are classified into Candidate list of words. Then, words that have not been deleted in the previous removal of the NGSL word list are classified into Candidate Words.

4. Experiment

In this section, the effectiveness of the proposed method described in the previous section is calculated. The purpose of this proposed method is to correctly classify the cyber kill chain phase as much as possible. The experiment is conducted with emphasis on this point.

4.1 Dataset

Two datasets are used for training and estimation of cyber kill chain phases. There are no open source datasets with a labeled phase for security datasets. Thus, we use ATT&CK for the training dataset because Tactics and Technique of ATT&CK are related to the Cyber Kill Chain Phase. For training data, a total of 3101 example sentences for each technique of ATT&CK for Enterprise are collected and we manually labeled their kill chain phases. For testing the model, four security reports published between November 2018 and December 2018 by Trend Micro [12] and McAfee [13] are collected. Each paragraph unit from the collected security reports is labeled with the kill chain phase, and candidate words to be used in the core features are extracted. These datasets are used as a ground truth for testing the effectiveness of the proposed model.

4.2 Results

The proposed method is applied to the collected security report for evaluation. The cyber kill chain phase and core features of the events extracted are compared with the manually annotated ground truth data. The following two items are evaluated:

- Whether the proposed model can correctly estimate the cyber kill chain phase for each paragraph of the security report by accuracy and F1-score.
- Whether core features words are extracted regardless of the types by recall, not whether core features words are classified into 4 types.

For the evaluation of the kill chain phase classification, the proposed model can correctly estimate with an average accuracy of 65% and a 0.67 F1-score as shown in Table 1. Though the estimation of the C2 phase resulted in a relatively low accuracy, the other four phases performed pretty well, notably the action on objectives phase estimates with approximately 80% accuracy.

| Kill chain phase        | Accuracy | F1-score |
|-------------------------|----------|----------|
| Delivery                | 0.63     | 0.72     |
| Exploitation            | 0.70     | 0.80     |
| Installation            | 0.62     | 0.70     |
| Command & Control (C2)  | 0.51     | 0.45     |
| Action on Objectives    | 0.79     | 0.70     |
| **Average**             | **0.65** | **0.67** |

Table 1 Kill chain phase classification result.

5. Conclusion

In this paper, we proposed a method of estimating the cyber kill chain phases and extracting event information in a paragraph-based analysis of security reports which include summarization of threat information. We trained the classification model with ATT&CK in experiments. As a result, the proposed model can estimate the cyber kill chain phase with an average F1-score of 0.67 and an average accuracy of 65%, and the core features can be extracted with a 86% recall by using this method. As for future work, we plan to test the proposed model on the dataset with several technical reports collected from various reliable security blogs or forums. We will also challenge to classify candidate words into 4 types of core features including the most difficult task of classification of Victim.

Acknowledgments These research results were obtained from the commissioned research by National Institute of Information and Communications Technology (NICT), JAPAN.

References

[1] Ito, D., Nomura, K., Kamizono, M., Shiraishi, Y., Takano, Y., Mohri, M. and Morit, M.: Modeling Attack Activity for Integrated Analysis of Threat Information, IEICE Trans. Information and Systems, Vol.E101-D, No.11, pp.2658–2664 (2018).
[2] Hutchins, E.M., Cloppert, M.J. and Amin, R.M.: Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains, Leading Issues in Information Warfare & Security Research, Vol.1, pp.80–106 (2011).
[3] Caltagirone, S., Andrew, P. and Christopher, B.: The Diamond Model of Intrusion Analysis, Center for Cyber Threat Intelligence and Threat Research, Hanover, MD (2013).
[4] Husari, G., Al-Shaer, E., Ahmed, M., Chu, B. and Niu, X.: TTPDrill: Automatic and Accurate Extraction of Threat Actions from Unstructured Text of CTI Sources, Proc. 33rd Annual Computer Security Applications Conference, pp.103–115 (2017).
[5] Zhu, Z. and Dumitras, T.: ChainSmith: Automatically Learning the Semantics of Malicious Campaigns by Mining Threat Intelligence Reports, 2018 IEEE European Symposium on Security and Privacy (EuroSeP), pp.458–472 (2018).
[6] Levy, O. and Goldberg, Y.: Dependency-based word embeddings, ACL, pp.302–303 (2014).
[7] MITRE ATT&CK, available from (https://attack.mitre.org/).
[8] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J.: Distributed Representations of Words and Phrases and their Compositionality, Advances in Neural Information Processing Systems, pp.3111–3119 (2013).
[9] Wikipedia: List of words about computers, available from (https://simple.wikipedia.org/wiki/List_of_words_about_computers).
[10] Browne, C., Culligan, B. and Phillips, J.: The New General Service List, available from (http://www.newgeneralservicelist.com) (2013).
[11] Cybstract, available from (https://github.com/cmu-se/cybstract).
[12] Trend Micro: Security Intelligence Blog, available from (https://blog.trendmicro.com/trendlabs-security-intelligence).
[13] McAfee, McAfee Labs, available from (https://www.mcafee.com/blogs/other-blogs/mcafee-labs).
Thin Tharaphe Thein received her B.E. degree from Yangon Technological University, Myanmar in 2019. She is currently pursuing Master’s degree at Kobe University. Her research focuses on threat intelligence.

Yuki Ezawa received his B.E. degree from Kobe University, Japan in 2019. He is currently pursuing Master’s degree at Kobe University. His research focuses on information security including blockchain applications.

Shunta Nakagawa received his B.E. degree in electrical engineering from Kobe University, Japan, in 2019. His research interests include machine-learning based cyberattack analysis.

Keisuke Furumoto received his B.E., M.E., and D.E. degrees from Kobe University, Japan, in 2013, 2014 and 2018, respectively. He joined the National Institute of Information and Communications Technology (NICT) in 2018. His current research interests include information security and machine learning.

Yoshiaki Shiraishi received his B.E. and M.E. degrees from Ehime University, Japan, and his Ph.D. degree from the University of Tokushima, Japan, in 1995, 1997, and 2000, respectively. From 2002 to 2006 he was a Lecturer at the Department of Informatics, Kindai University, Japan. From 2006 to 2013 he was an Associate Professor at the Department of Computer Science and Engineering, Nagoya Institute of Technology, Japan. Since 2013, he has been an Associate Professor at the Department of Electrical and Electronic Engineering, Kobe University, Japan. His current research interests include information security, cryptography, computer network, and machine learning based cyberattack analysis. He received the SCIS 20th Anniversary Award and the SCIS Paper Award from ISEC group of IEICE in 2003 and 2006, respectively. He received the SIG-ITS Excellent Paper Award from SIG-ITS of IPSJ in 2015. He is a member of IEEE, ACM, and a senior member of IEICE, IPSJ.

Masami Mohri received her B.E. and M.E. degrees from Ehime University, Japan, in 1993 and 1995 respectively. She received her Ph.D. degree in Engineering from the University of Tokushima, Japan in 2002. From 1995 to 1998 she was an Assistant Professor at the Department of Management and Information Science, Kagawa Junior College, Japan. From 1998 to 2002 she was a Research Associate of the Department of Information Science and Intelligent Systems, the University of Tokushima, Japan. From 2003 to 2007 she was a Lecturer of the same department. From 2007 to 2017, she was an Associate Professor at the Information and Multimedia Center, Gifu University, Japan. Since 2017, she has been an Associate Professor at the Department of Electrical, Electronic and Computer Engineering in the same university. Her research interests are in coding theory, information security, and cryptography. She is a member of IEEE and a senior member of IEICE.

Yasuhiro Takano received his Ph.D. (Info. Sc.) and Dr. Sc. (Tech.) degrees, respectively, from Japan Advanced Institute of Science and Technology (JAIST) and the University of Oulu, Finland, in 2016. He is currently with Kobe University as an Assistant Professor. His research interests include signal processing for communications engineering.

Masakatu Morii received his B.E. degree in electrical engineering and his M.E. degree in electronics engineering from Saga University, Saga, Japan, and his D.E. degree in communication engineering from Osaka University, Osaka, Japan, in 1983, 1985, and 1989, respectively. From 1989 to 1990 he was an Instructor in the Department of Electronics and Information Science, Kyoto Institute of Technology, Japan. From 1990 to 1995 he was an Associate Professor at the Department of Computer Science, Faculty of Engineering, Ehime University, Japan. From 1995 to 2005 he was a Professor at the Department of Intelligent Systems and Information Science, Faculty of Engineering, the University of Tokushima, Japan. Since 2005, he has been a Professor at the Department of Electrical and Electronic Engineering, Faculty of Engineering, Kobe University, Japan. His research interests are in error correcting codes, cryptography, discrete mathematics, computer networks and information security. He is a member of IEEE and a fellow of IEICE.