Estimating Cyber Kill Chain Phases from Unstructured Technical Reports

THEIN THIN THARAPHE[1] EZAWA YUKI[1] NAKAGAWA SHUNTA[1]
FURUMOTO KEISUKE[2] SHIRAISHI YOSHIAKI[1,3] NAKAMURA TORU[3]
HASHIMOTO MASAYUKI[3] MOHRI MASAMI[4] MORII MASAKATU[1]

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.

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 use sophisticated intrusion and attack routes to evade detection. Once intruded, it conducts long-term attack activities with a potentially destructive consequence. To counter this, many have been paid attention to the field of Threat Intelligence, 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 different 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 organization often share attacks information in the form of security reports. In order to make a datasets for cyber threat analysis such as Ref. [1], we propose an approach for modeling threat information compiled in various formats and an analysis system based on this method. This paper treats security reports on a paragraph basis considering that one event is described in one paragraph and estimates phases in the cyber kill chain and extract 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. 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, APT goes through seven phases: reconnaissance, weaponization, delivery, exploitation, installation, command and control (C2), and actions on objectives.

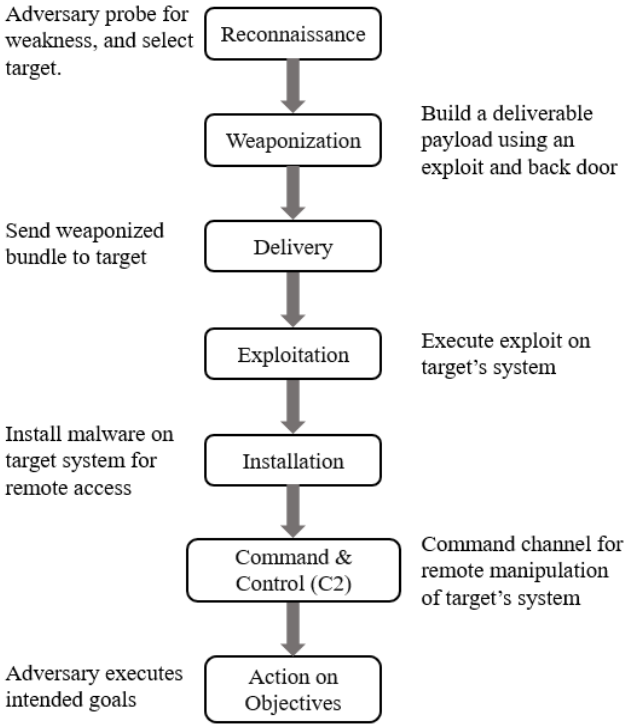

Figure 1. Cyber Kill Chain

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 Cyber Kill Chain model. The Diamond Model as shown in Figure 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 the attackers or organization utilizing capability (tools and techniques) against the victim in order to reach the desired goal. The infrastructure is the logical or physical communication system used by adversary to deliver capability, maintain control and gain benefits from victim.

1 Kobe University
2 National Institute of Information and Communications Technology
3 Advanced Telecommunications Research Institute International
4 Gifu University

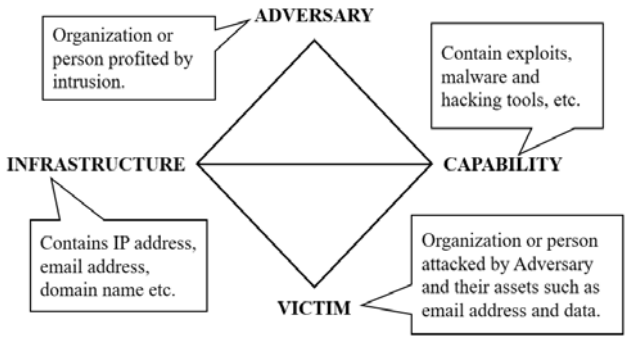

Figure 2. Diamond Model

Infrastructure could be e-mail addresses, domain names, and IP addresses, etc. The victim is always the target of the adversary.

2.3 Activity Threat

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

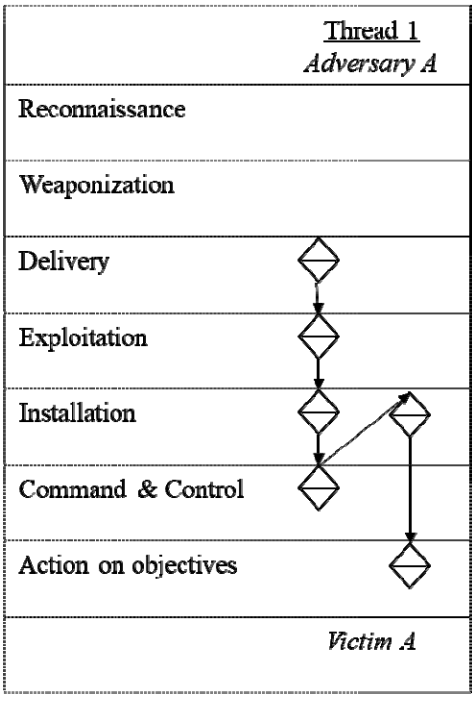

Figure 3. Activity Thread

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 ATP 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 semantic relationship. Each threat action is then mapped to appropriate tactics, techniques, and kill chain phase and generate STIX (Structured Threat Information eXpression) standard formatted reports.

3. 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 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 dependency-based wording embedding [6] that uses words dependencies instead of just context words. In this approach, six types of IoC named entity: URL, IP, hash, malware family, Exploit Kit and CVE, are extracted and classified their kill chain phases by training the neural networks.

4. Proposed Scheme

We aim to establish a modeling method for classifying cyber kill chain phases for threat information described in security reports. The first requirement for this approach is to extract event information from security reports. We assumed that events are described in each paragraph unit of security reports. In this paper, we propose a method for analyzing security reports in paragraphs, which includes extracting event information and estimating kill chain phases. Fig. 4 shows the flow of the proposed method. Firstly, it estimates phase and extracts informative words for core features of diamond model from each paragraph.

4.1 Word Embedding

In order to understand semantic similarity among words, the state-of-art word2vec [7] algorithm is used to parse words semantically. The word2vec processes text corpus as an input and outputs the vectors that are distributed numerical representations of word features. The key features of word2vec is that the word vectors generated take up much lesser space than one hot encoded vector. And also, it holds semantic meaning of the word since similar words are grouped in vector space.
Eg. Vec[King] – Vec[Man] + Vec[Woman] = Vec[Queen]

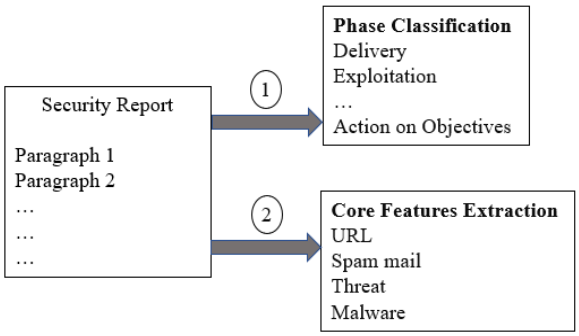

Figure 4. Overview of Proposed Method

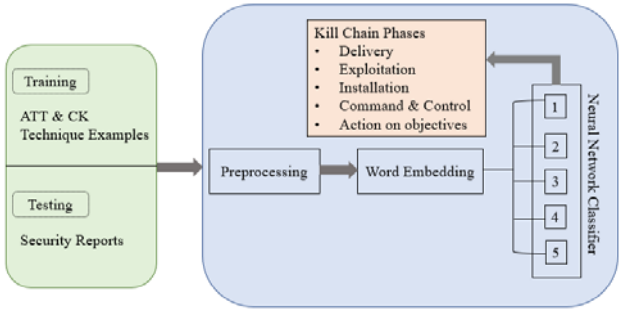

Figure 5. Kill Chain Phase Classifier

4.2 Paragraph-based Estimation of Cyber Kill Chain Phase

Since the security report hardly describes Reconnaissance phase and Weaponization phase of the cyber kill chain, the only remaining five phases: Delivery, Exploitation, Installation, Command and Control, and Action on Objectives are considered in here. Figure 5 shows the procedure for estimating the cyber kill chain phase. In this method, five binary classifiers of neural networks are used. These models have the same configuration, but train with different data. The five classifiers correspond to Delivery, Exploitation, Installation, Command and Control, and Action on Objectives respectively. These classifiers are used to estimate whether the contents of security reports belong to the phase described. The classifier is trained by the dataset, which includes the example sentences from ATT&CK [8] for Enterprise and manually labeled their phases. ATT&CK is a knowledge base managed by MITER 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 report into the trained neural networks in a paragraph unit.

4.3 Classification Model

Firstly, preprocessing is done on input text corpus with off-the-shelf NLP techniques. In this step, lowercase conversion, removal of stopwords, punctuation and special characters are performed. Next, each sentence is tokenized into words and lemmatization is applied to each word. After this step, we parse each word by using word2vec, in which semantically similar words will be in a close position in vector space which is trained by maximizing the probability of a word given the words around it. The word vector is trained with the embedding dimension of 100. In next step, the kill chain phases are estimated by utilizing 5 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 feed into the classifier are identified. Firstly, Informative words are calculated by using the following equation [5]:

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

where $p(w)$ is the probability of word w, $p(w|k)$ is the probability of word w for describing kill chain k, and K is total kill chain phases.

Next, the context word for each sentence that will be feed into the classifier are calculated. The context is determined from two statements; informative words of current sentences and informative words of previous sentences if no informative words are found in current sentences. The average word embedding of the context word are then passed into the neural networks. We train the classifier to estimate whether each paragraph unit of security report falls into any of 5 phases.

4.4 Core Features Extraction from Paragraph

The purpose of this section is to extract the candidate words that should be included in the core features of diamond model. 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 name such as malware and tools etc., described in ATT&CK
(3) Group name 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 Indicator of Compromise (IoC) 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. If they are, extract them and delete them 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. If they are, delete them 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 Words.

5. Experiment

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

5.1 Dataset

Two datasets are used for training and estimation of cyber kill chain phases. For training data, the example sentences for each technique of ATT&CK for Enterprise are collected. There are 3101 example sentences, which are partially described in Figure 6. These example sentences are manually labeled with their corresponding kill chain phases. As for testing the model, security reports published in TrendLabs Security Intelligence Blog [12] by Trend Micro and McAfee Labs Category blog [13] by McAfee are gathered. From these sources, four reports between November 2018 to December 2018 are collected. Each paragraph units from the collected security reports are attached with a phase label and, candidate words to be used in the core features are extracted. These data are used as a ground truth.

5.2 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:

(1) Whether the proposed model can correctly estimate the cyber kill chain phase for each paragraph of security report

(2) Correctly extract the core features from the paragraphs of security reports or not.

Table 1. Phase Classification Result

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

6. Conclusion

In this paper, we proposed a method of estimating cyber kill chain phases and extracting event information in a paragraph-based analysis of security reports which include summarization threat information. The experiment 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.

Reference

[1] D. Ito, K. Nomura, M. K. Kamizono, Y. Shiraishi, Y. Takano, M. Mohri, M. Morii, “Modeling Attack Activity for Integrated Analysis of Threat Information”, IEICE Trans. on Information and Systems, vol.E101-D, no.11, pp.2658-2664 (2018).

[2] E.M. Hutchins, M.J. Cloppert, and R.M. Amin, “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] S. Caltagirone, P. Andrew, and B. Christopher, “The Diamond Model of Intrusion Analysis,” Center for Cyber Threat Intelligence and Threat Research, Hanover, MD, 2013.

[4] Husari, Ghaith, Ehab Al-Shaer, Mohiuddin Ahmed, Bill Chu, and Xi Niu. “TTPDrill: Automatic and Accurate Extraction of Threat Actions from Unstructured Text of CTI Sources,” in Proceedings of the 33rd Annual Computer Security Applications Conference, pp. 103-115. 2017.

[5] Z. Zhu, T. Dumitras, “Chainsmith: Automatically learning the semantics of malicious campaigns by mining threat intelligence reports,” in 2018 IEEE European Symposium on Security and Privacy (EuroS&P), pp. 458-472, 2018.

[6] O. Levy and Y. Goldberg, “Dependency-based word embeddings.” in ACL, pp. 302–30, 2014.

[7] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in Advances in neural information processing systems, pp. 3111–3119, 2013.

[8] MITRE ATT&CK™. https://attack.mitre.org/

[9] Wikipedia, “List of words about computers,” https://simple.wikipedia.org/wiki/List_of_words_about_computers

[10] Browne, C., Culligan, B. & Phillips, J. (2013), “The New General Service List,” http://www.newgeneralservicelist.org.

[11] Cyobstract. https://github.com/cmu-sei/cyobstract

[12] Trend Micro, “Securtiy Intelligence Blog,” https://blog.trendmicro.com/trendlabs-security-intelligence

[13] McAfee, “McAfee Labs,” https://www.mcafee.com/blogs/other-blogs/mcafee-labs