CyNER: Cybersecurity Named Entity Recognition System for Efficient Intelligence Analysis

Shota FUJII, Tetsuro Kito, Tomohiro Shigemoto, and Yasuhiro Fujii
Hitachi, Ltd. Research & Development Group, Japan

HITACHI Inspire the Next

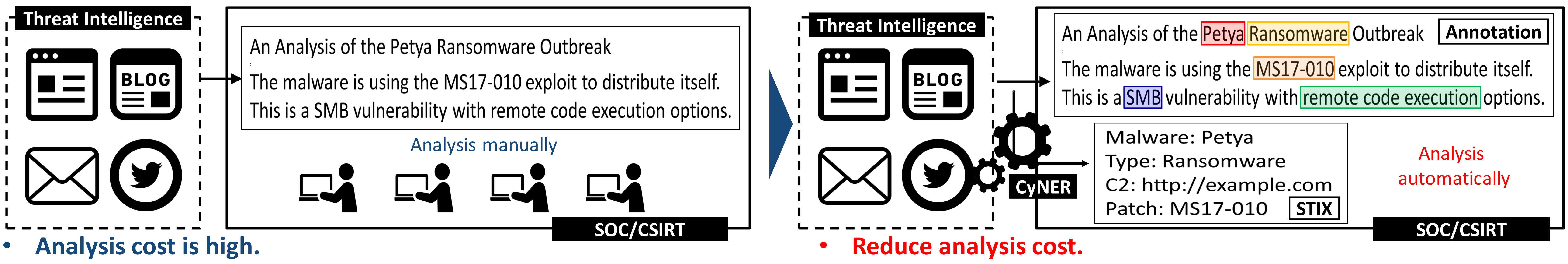
Abstract

- Threat intelligence is effective for obtaining up-to-date security threats and dealing with them.
- However, a number of threat intelligence are written by natural language; therefore, its analysis cost is high and automatic processing (e.g., register suspicious URL to BlackList) is difficult

We propose a CyNER, the Cybersecurity Named Entity Recognition System for Efficient Intelligence Analysis.
The CyNER extracts Named Entities (NEs) and Indicator Of Compromise (IOC) by combining Named Entity Recognition (NER) with regular expression from threat intelligence written by natural language (e.g., blog, mailing list, SNS, etc.).
We find that the CyNER can extract NEs more high accuracy than existing method (f-measure: 0.78). The evaluation also shows the CyNER can extract IOCs more large quantity than simple method (+44.8%).

1. Objectives of This Research

To reduce cost of threat intelligence-related SOC/CSIRT tasks, our research aims to structure threat intelligence as follows:

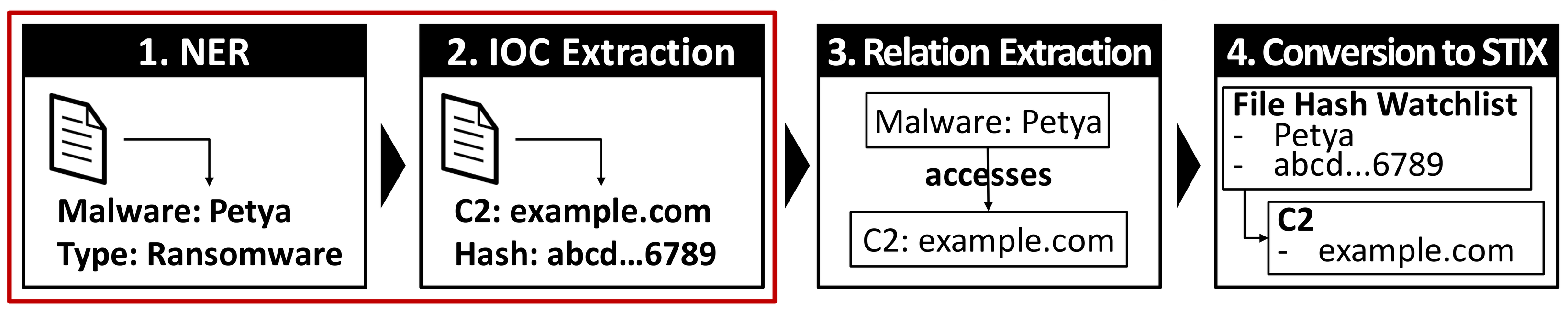


- **Analysis cost is high.**
- **Highly depends on expert's skill and most of procedure is manual.**

- **Reduce analysis cost.**
- **Make automatic processing possible.**

2. Approach to Threat Intelligence Construction

To construct structured threat intelligence, following steps are needed:



This poster describes implementation of 1. NER and 2. IOC extraction.

Then, we have to solve following 3 problems:

Problem 1 In security field, unknown words tend to be generated and its recognition is relatively difficult.

Problem 2 Some IOCs are in images; thus, cannot be extracted simply.

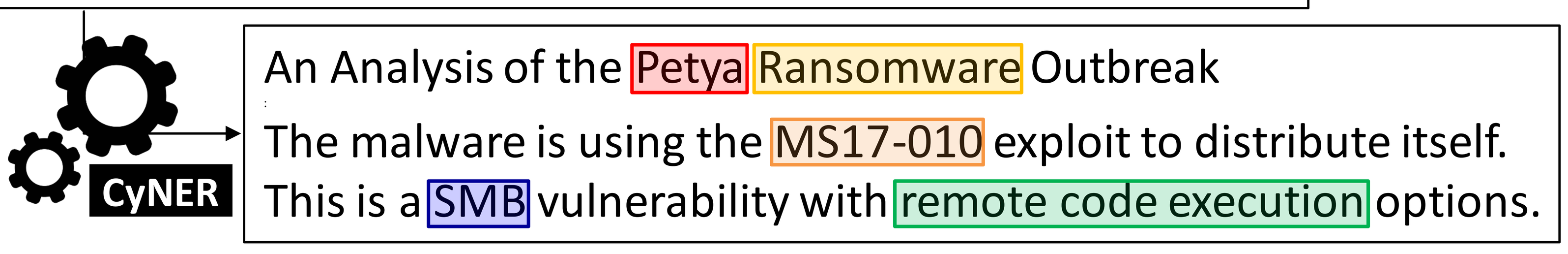
Problem 3 Some IOCs are defanged; thus, cannot be extracted simply.

To achieve NER and IOC extraction with solving above problems, the CyNER carries out them by following steps:

1. NER with Picking up Unknown Words

Threat Intelligence Sample 1

An Analysis of the Petya Ransomware Outbreak
The malware is using the MS17-010 exploit to distribute itself.
This is a SMB vulnerability with remote code execution options.



Malware Name, Malware Type, Patch ID, Protocol, Vulnerability

STEP 1: Named Entity Recognition
By using state-of-the-art NER method [1], extract NEs from threat intelligence (e.g., malware name, malware type, vulnerability, etc.).

STEP 2: Picking up Missed NEs (to solve Problem 1)
In generally, 1st candidate is extracted as NE from the result of NER. In contrast, the CyNER picks up relatively high possible NE from 2nd candidate when the 1st candidate is "not NE".

[1] Ma, X. and Hovy, E.: End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016), pp. 1064–1074 (2016).

2. IOC Extraction with OCR and Refang

Threat Intelligence Sample 2

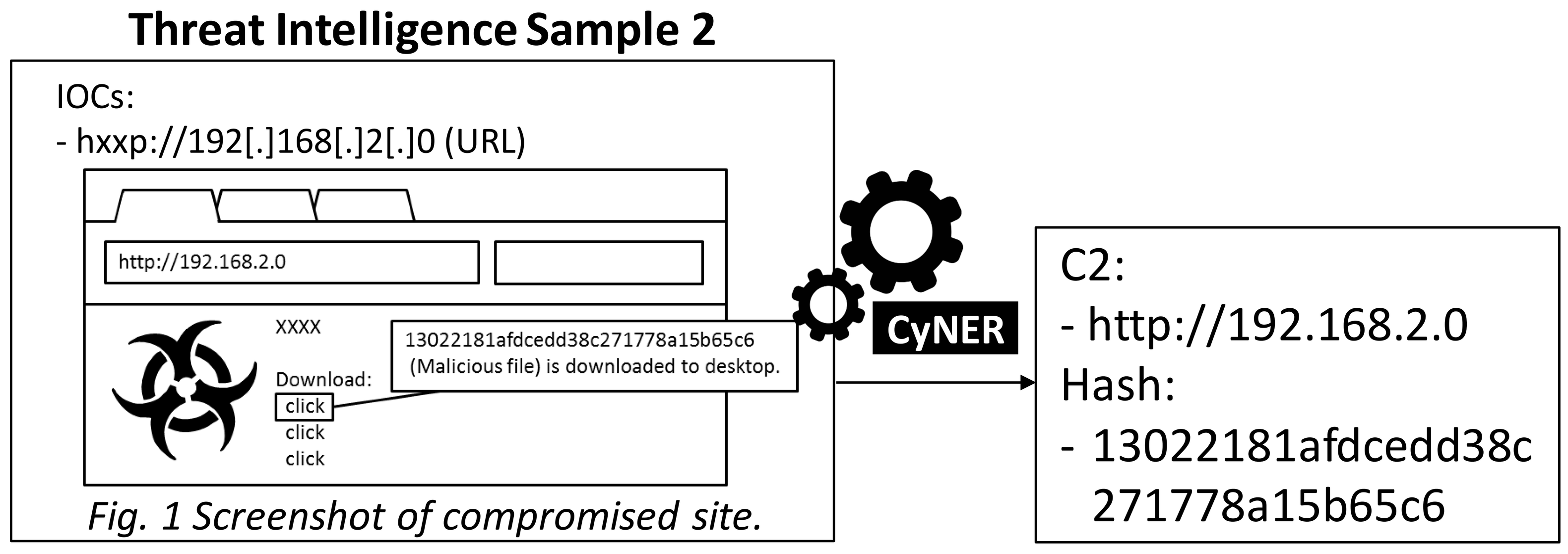


Fig. 1 Screenshot of compromised site.

STEP 1: Convert Image to Text (to solve Problem 2)
To extract IOCs from figure, convert figure to text by using OCR.

STEP 2: Refang (to solve Problem 3)
To extract defanged IOCs, refang these IOCs.
e.g., example[.]com -> example.com

STEP 3: Extract IOC by Regular Expression
Finally, by using regular expression, extract IOCs (e.g., ip address, URL, hash value, etc.).

3. Preliminary Experiment

To verify the effective of the CyNER, we evaluate recognition accuracy of NEs (evaluation 1) and quantity of IOCs (evaluation 2) as follows:

Evaluation 1: accuracy of NEs

Experimental Setup

Dataset: ICS-CERT alerts [2] (123 articles, 4,779sentences)
- ✓ Train: 83 articles (about 70%; July 13, 2011 - Oct. 29, 2013)
- ✓ Verify: 40 articles (about 30%; Oct. 30, 2013 - Jan. 11, 2018)

Result

	precision	recall	F1-measure
Baseline [1]	0.75	0.74	0.73
CyNER	**0.77**	**0.80**	**0.78**

has higher score than baseline method in all evaluation index.

[2] ICS-CERT Alerts, available from <https://ics-cert.us-cert.gov/alerts>.

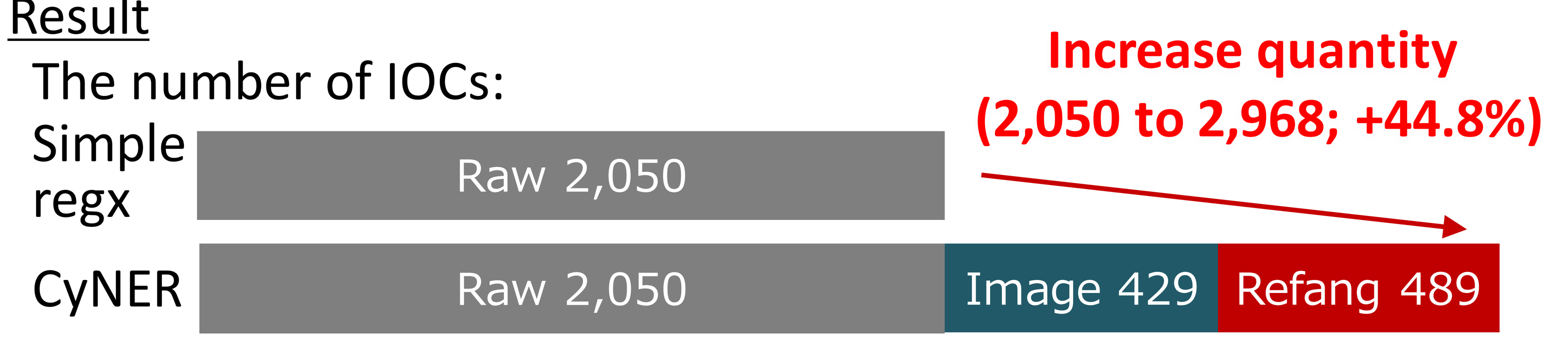
Evaluation 2: quantity of IOCs

Experimental Setup

Dataset: FireEye Threat Research Blog [3] (100 articles)
- ✓ Feb. 22, 2017 - Oct. 11, 2018

Result

The number of IOCs:



increase quantity of IOCs by image2text and refang.

[3] FireEye: Threat Research Blog, available from <https://www.fireeye.com/blog/threat-research.html>.

4. Future Work

- implement relation extraction (e.g., <MALWARE> accesses <C2>) and convert threat intelligence to STIX using by extracted relation.
- evaluate the CyNER with more large datasets and operate the CyNER in SOC/CSIRT and evaluate its practicability.