CDTier:A Chinese Dataset of Threat Intelligence Entity Relationships

Yinghai Zhou, Yitong Ren, Ming Yi, Yanjun Xiao, Zhiyuan Tan, Nour Moustafa, *Senior Member, IEEE*, Zhihong Tian, *Senior Member, IEEE*

Abstract—Cyber Threat Intelligence (CTI), which is knowledge of cyberspace threats gathered from security data, is critical in defending against cyberattacks.However, there is no open-source CTI dataset for security researchers to effectively apply enormous CTI information for security analysis in the field of threat intelligence, particularly in the field of Chinese threat intelligence. As a result, for network security research and development, this paper constructed a Chinese CTI entity relationship dataset–CDTier, which includes: 1) A threat entity extraction dataset composed of 100 CTI reports, 3744 threat sentences and 4259 threat knowledge objects; 2) A dataset for entity relation extraction including 100 CTI reports, 2598 threat sentences and 2562 knowledge object relations. CDTier is, as far as we know, the first CTI dataset. On the CDTier, we trained 4 models for threat entity extraction and relation extraction using well-established and widely used deep learning methods in the NLP. The results showed that the model trained on CDTier extracts knowledge objects and their relationships described in threat intelligence more accurately. This significantly minimizes threat intelligence analysts' work while assessing threat intelligence. The CDTier may be found at https://github.com/MuYu-z/CDTier.

Index Terms—Cyber Threat Intelligence, Threat Entity Extraction, Entity Relation Extraction, NLP, Information Extraction

1 Introduction

The essence of cybersecurity is confrontation. The attacker can typically gain a time and resource advantage over the defender, placing the defender in a passive position [1]. In other words, due to the unequal information and resources of the attacker and the defender, it is difficult for the defender to obtain any information about the attacker before the attacker implements the attack [2]. But the attacker will start the attack after collecting the relevant information of the defender. Additionally, new types of attacks represented by Advanced Persistent Threats (APT) have garnered considerable attention as a result of the enormous evolution of data collection and transmission [3]. Therefore, Cyber Threat Intelligence (CTI) is proposed, which is the information and details of security incidents recorded by security researchers. Gartner defined CTI as knowledge based on evidence that can be used to help security researchers' decision to respond to that threat [4].

However, as more and more enterprises or organizations are taking cybersecurity seriously, the world of regulated and unregulated threat data is expanding at a rapid clip, and this has three major negative effects on CTI. First of all, CTI data sources are numerous and exhausting for analysts. Second, there are many types of CTI and complex application scenarios. Finally, in the Internet era, information generation is fast and threat intelligence is updated quickly [5]. Security researchers and communities from around the world are working to build CTI-sharing standards so that CTIs can be generated and shared quickly. The development of shared standards has largely solved the problem of cumbersome and messy CTI, such as Structured Threat Information Expression(STIX) [6], Trusted Automated eXchange of Indicator Information(TAXII) [7] and Cyber Observable eXpression(CybOX) [8].

In the field of threat intelligence analysis, it is an important task to extract pre-defined threat entity relationships from massive amounts of open-source unstructured text. The relationship between threat entities can be described as a relational triad$< Te_1, Tr, Te_2 >$. Te_1 and Te_2 denotes the threat entities, while Tr is the relationship set $TR< Tr_1, Tr_2, Tr_3, ..., Tr_i >$ between the entities [9]. The goal of threat entity relationship extraction is to extract a triad of threat entities and relationships from open-source heterogeneous CTI texts to improve the quality of valid threat knowledge. As shown in Figure 1, in the Chinese CTI text, the sentence: "UAC-0056组织针对乌克兰传播OutSteel和SaintBot" can be described as <UAC-0056, utilize, OutSteel>, <UAC-0056, utilize, SaintBot> and <UAC-0056, target, Ukraine>.

There are three limitations in dataset regarding Named Entity Recognition (NER) and Relation Extraction (RE). First, because of the unique characteristics of the cybersecurity industry, which are influenced by various policies and regulations regarding confidentiality, there are very few

- *Corresponding author: Zhihong Tian (tianzhihong@gzhu.edu.cn). E-mail: see http://www.michaelshell.org/contact.html*
- *Yinghai Zhou, Yitong Ren and Zhihong Tian are with Cyberspace Institute of Advanced Technology, Guangzhou University, Guangzhou, 510006, China. E- mail: zyh@e.gzhu.edu.cn, renyitong@e.gzhu.edu.cn, tianzhihong@gzhu.edu.cn*
- *Ming Yi is with Institute of Computer Application, China Academy of Engineering Physics,MianYang, 621050, China. E-mail: ridyi@foxmail.com*
- *Yanjun Xiao is with PINGXING Lab (Nsfocus Technology Group Company), Guangzhou, 510663, China. E- mail: xiaoyanjun@nsfocus.com*
- *Zhiyuan Tan is with School of Computing, Engineering and the Building Environment, Edinburgh Napier University, UK. E-mail: z.tan@napier.ac.uk*
- *Nour Moustafa is Postgraduate Discipline Coordinator (Cyber) and Senior Lecturer in Cyber Security & Computing at the School of Engineering and Information Technology (SEIT), University of New South Wales (UNSW)'s UNSW Canberra, Australia. E-mail: nour.moustafa@unsw.edu.au*

Manuscripts received on February 22, 2022; The modified in August

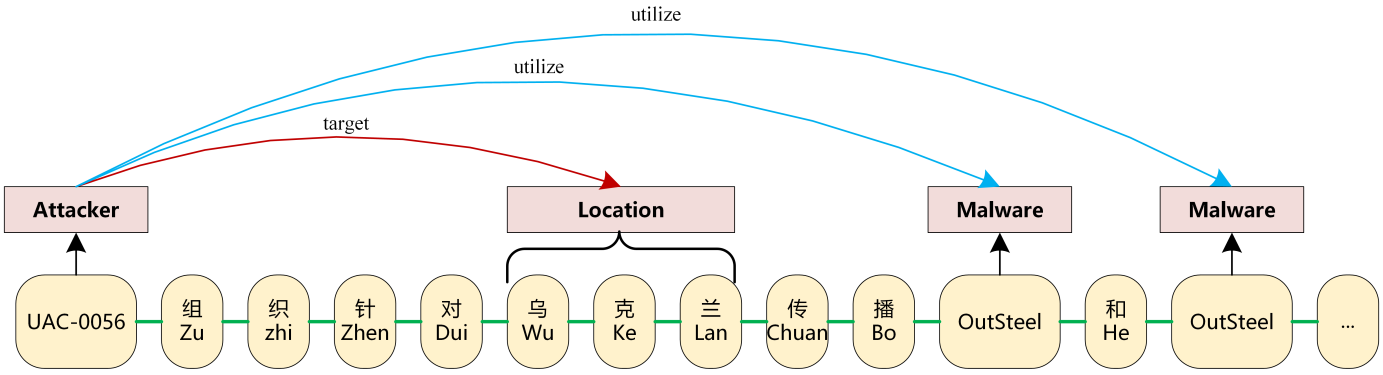


Fig. 1: Examples of Entity Relations in Chinese CTI Text

open source dataset in the field of threat intelligence, and the existing dataset focuses primarily on log-based temporal event relationships [10] [11]. Second, the lack of a Chinese CTI corpus. CTI in Chinese is much less than English CTI, and there is no open-source dataset, either in NER or RE. Thirdly, the processing of Chinese CTI is more complex than English CTI due to the lack of explicit separators between words. As a result, many character-based approaches have been developed [12] [13], but it still needs to be supplemented with word-level information.

In summary, this paper constructed a Chinese Dataset of Threat intelligence entity relationships (CDTier) to address the above issues and provided better support for the construction of knowledge graphs, the attribution of cyber threats, and other related research and applications. Firstly, CDTier summarized 5 types of threat entities and 11 types of threat entity relationships based on standards such as STIX and the applications of CTI in the real world. Then, the Chinese CTI text data was labeled according to a well-designed scheme. The team included two cybersecurity graduate students and a professor, in order to minimise errors. Finally, in order to verify the practical value of CDTier, we used mainstream natural language processing models to perform entity extraction and relationship extraction for the Chinese threat intelligence text in a pipeline manner. The remaining parts of the essay are arranged as follows:

- Threat intelligence research in several domains is presented in the Section 2.
- The preliminary work, including the defining of entities and relationships, is introduced in the Section 3
- The standard strategy is described in the Section 4, along with the data collection, annotation strategy and annotation method.
- The Section 5 presents the performance of a variety of typical natural language processing models in the CDTier.
- The Section 6 is the conclusion.

2 Related Works

Entity extraction and entity relation extraction based on CTI have been studied in many fields, such as malware, Indicators of Compromise (IOC), threat intelligence behavior analysis and threat entity relation extraction.

Specifically, in malware: Manikandan et al [14]. used CNN-CRF (Convolutional Neural Networks-Conditional Random Field, CNN-CRF) model to identify entities related to malicious programs. Yansi Keim et al. [15] proposed a cyber threat intelligence framework that used the Elastic search–Logstash–Kibana (ELK) stack to provide detailed reports based on input and preprocessing, which effectively detected a new generation of malware. Nidhi Rastogi et al [16]. developed a malware ontology model: MALOnt. MALOnt instantiated a knowledge graph from the CTI corpus containing hundreds of annotated malware, which can be used to structure extraction information and generate knowledge graphs. Haoxi Tan et al [17]. designed ColdPress, an extensible malware analysis platform, which could effectively extracted threat sample information from malware-related threat intelligence. By combining machine learning techniques with feature selection algorithms like CorrACC and CorrAUC, [18] and [19] proposed CorrACC and CorrAUC to identify malicious traffic in the Internet of Things network.

In IOC, Liao et al [20]. developed a system called iACE, the task of extracting IOC from CTI is modeled as a graph similarity problem, which enables iACE to automatically obtain the input-output control system from the CTI and capture its contextual relationship. Long et al [21]. applied neural network patterns to the identification of IOCs based on threat intelligence, so that IOC could be automatically identified from CTI. TIMiner [22] combined regular expressions, NER and syntactic dependencies in the field of cybersecurity to extract IOCs in CTI. HINCTI [23] proposed a heterogeneous graph convolutional network method based on MIIS metric to identify the threat types of infrastructure nodes in CTI to complement the relevant information of IOC. iGen [24]extracted IOCs in STIX-standard threat intelligence from sandbox results.

In threatening behavior analysis, TTPDrill [25] utilized Natural Language Processing techniques and Information Retrieval methods to extract threat actions from unstructured CTI text. Yan et al [26]. used BERT-BiGRU to classify attack behaviors and attack strategies described in threat intelligence, thereby calculating the possibility of attacks and the degree of harm of attacks, especially for unstructured threat intelligence analysis of IIoT (Industrial Internet of Things). Zongxun L et al [27]. proposed a BERT-BiLSTM-CRF-based model to automatically extract threat actions and generate tactics, techniques and procedures (TTPs) from

TABLE 1: CDTier's treatment of eighteen STIX knowledge objects

Treatment	Object in STIX2.1	Sketch
Extraction based on ATT&CK	Attack Pattern	Describes the pattern the attacker attempts to sabotage the target
Corresponding entity **Campaign**	Campaign	Indicates a specific attack
Extraction based on ATT&CK	Course of Action	The act of preventing or responding to an attack
Rarely described in Chinese CTI	Grouping	Data generated during analysis and investigation
Corresponding entity **Industry**	Identity	Represent a particular person, organization, or group
Extracted by regular expression	Indicator	Indicator of threat characteristics
Extracted by regular expression	Infrastructure	Systems, software services and other physical or virtual resources
Corresponding entity **Attacker**	Intrusion Set	A collection of malicious acts and resources used by an organization
Corresponding entity **Location**	Location	Location
Corresponding entity **Tools**	Malware	Programs or codes implanted in the system for destruction
Extract with third-party sandbox	Malware Analysis	The result of an analysis of a malware instance or family
Rarely described in Chinese CTI	Note	Additional information that does not exist in other objects
Extracted by regular expression	Observed Data	Description of observable behavior on the network
Rarely described in Chinese CTI	Opinion	Evaluation of information correctness in STIX objects
Can be directly obtained	Report	CTI
Corresponding entity **Attacker**	Threat Actor	Person, group, or organization with malicious intent
Corresponding entity **Tools**	Tool	Legitimate software that an attacker can use to execute an attack
Extracted by regular expression	Vulnerability	A vulnerability that can be exploited by an attacker

APT reports.

In threat entity relation extraction, Dionsio et al [28]. built a Named Entity Recognition Model using Natural Language Processing technology and Deep Learning algorithms to identify named entities from tweets related to cybersecurity published by Twitter. Gasmi et al [29]. used LSTM (Long Short Term Memory) and CRF-Conditional Random Field) models to extract cybersecurity entities and their relationships from cybersecurity texts. CASIE [30] combined Attention Mechanism and BiLSTM model to extract cybersecurity events from CTI texts. Li T et al [31]. proposed a BiLSTM-CRF model based on Attention Mechanism for named entity recognition of web-safe text. Extractor [32] utilized NLP to extract entity attack behaviors automatically from CTI texts, and used "semantic role annotation" for semantic analysis to understand the relationships between attack entity behaviors, thus transforming unstructured text into graph-structured knowledge.

Although entity and entity relationship extraction based on threat intelligence has been studied in numerous fields, they either focused on English texts or considered only data-normative cyber threat entities. At the same time, irrelevant information such as advertisements, product descriptions and message boards in threat intelligence texts can significantly degrade the quality of text analysis models. In addition, it's more challenging for Chinese CTI text because there are a large number of English organization names and jargon in Chinese CTI that need vectorized representation.

3 PRELIMINARY WORK

It is essential to extract entities and define the relationship between entities based on the content and criteria in CTI during the process of entity and entity relationship extraction based on CTI. This will ensure that accurate results are obtained. However, there is not a single Chinese annotated dataset that takes into consideration this issue.

As the first Chinese annotated dataset in the field of CTI, CDTier reasonably selects 5 entities, including attacker, tool, industry, region and campaign, and 11 relationship types to describe threat entity behavior with reference to the ATT&CK [33], STIX [6] and actual business requirements.

3.1 The Definition of Five Entities

STIX [6], a recognized threat intelligence sharing standard in the cybersecurity field, has gone through two versions since it was proposed. There were 8 knowledge objects proposed in STIX1.x. STIX2.0 expanded knowledge objects to 12. STIX2.1 proposed in September 2021 is an extension of STIX2.0, expanding the knowledge objects in STIX 2.1 to 18. From the earliest version of STIX1.x to today's STIX2.1, STIX has gradually developed from practical to intelligent by designing the structured and systematic representation and description of threat information, so that threat information can be presented to security analysts more directly. And store it in JSON for faster machine reading.

However, the challenge of CTI is that security vendors are difficult to obtain enough information to accurately perceive all threat details. Therefore, there is very little CTI that fully meets the STIX standard in real-world scenarios. From the perspective of practical application, CDTier summarizes the meaning of 18 knowledge objects in STIX and the processing methods of CDTier, and abstracts five key knowledge objects as entities: attacker, tool, industry, region and campaign, as shown in Table 1.

Therefore, CDTier defined five key knowledge objects as follows:

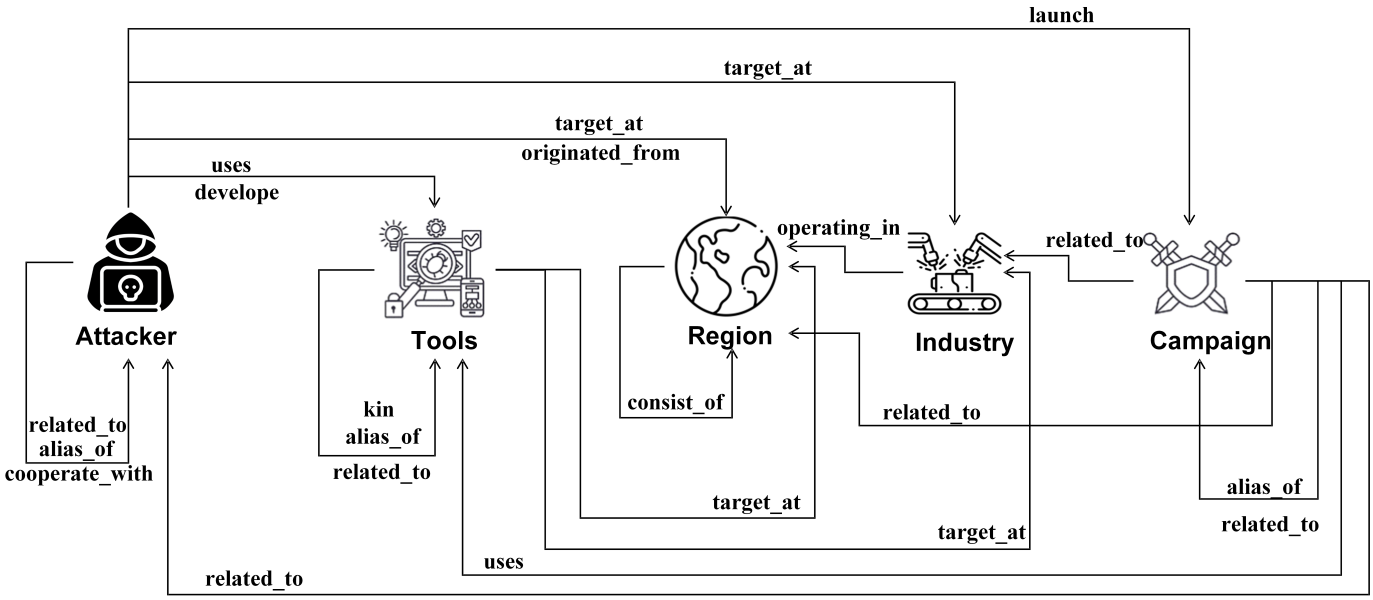


Fig. 2: Eleven relationship types corresponding to five key knowledge objects

- Attacker: the person, group, or organization that carried out a specific malicious attack in a threat intelligence description.
- Tools: Malware, legitimate software, or self-developed attack tools targeting a specific domain used by attackers in the threat intelligence description.
- Industry: The target industry of the attacker in the campaign.
- Region: The region targeted by the attacker or the region the attacker belongs to in the threat intelligence description.
- Campaign: Indicates a specific attack activity initiated by an attacker.

3.2 Eleven Entity Relation Types

Considering the particularity of CTI texts and disclosure standards, CDTier selects 11 relationship types to describe threat entity relationship types for the five key knowledge objects described in 3.1. In fact, the 11 relationship types were chosen because they are mostly described in Chinese CTI texts, and there are only a few other relationship types involved. In order to facilitate the understanding of the relationship types involved in CDTier, this paper presents 11 relation types corresponding to the five key knowledge objects in Figure 2. Include:

- alias_of: The same attacker and tool described in different threat intelligence. This relationship can be expressed in three triples.

$$< Attacker, alias\_of, Attacker >$$
$$< Campaign, alias\_of, Campaign >$$
$$< Tools, alias\_of, Tools >$$

- related_to: a knowledge object that is related to a certain extent but still unable to determine what the relationship is. This relationship can be expressed in six triples.

$$< Attacker, related\_to, Campaign >$$
$$< Campaign, related\_to, Campaign >$$
$$< Campaign, related\_to, Region >$$
$$< Campaign, related\_to, Industry >$$
$$< Tools, related\_to, Tools >$$

- uses: between two knowledge objects with a utilization relationship. This relationship can be expressed in two triples.

$$< Attacker, uses, Tools >$$
$$< Campaign, uses, Tools >$$

- target_at: The threat agent (attacker, malware) launched an attack on a specific target. This relationship can be expressed in four triples.

$$< Attacker, target\_at, Industry >$$
$$< Attacker, target\_at, Region >$$
$$< Tools, target\_at, Region >$$
$$< Tools, target\_at, Region >$$

- originated_from: Which region the attacker came from. This relationship can be represented by a triple.

$$< Attacker, originated\_from, Region >$$

- launch: The attacker launches a campaign. This relationship can be represented by a triple.

$$< Attacker, originated\_from, Region >$$

- kin: Malware belonging to the same family. This relationship can be represented by a triple.

$$< Tools, kin, Tools >$$

- consist_of: Describes the dependencies between regions. For example, South Korea consists of Seoul, Busan, Daegu, Incheon, Gwangju, Daejeon and Ulsan. This relationship can be represented by a triple.

$$< Region, consist\_of, Region >$$

- operating_in: Describes an industry that belongs to a certain region. This relationship can be represented by a triple.

$$< Industry, operating\_in, Region >$$

- develop: Describes the development of a tool, malware, or script by an attacker. This relationship can be represented by a triple.

$$< Attacker, develop, Tools >$$

4 Annotation Strategy

4.1 Data Collection

There were a total of 200 Open-Source Chinese Threat Intelligence pieces included in CDTier, which were contributed by thirteen different security firms. Table 2 provided a concise summary of the essential facts that were included in the CDTier, which included a total of 6342 threat sentences and 493304 Chinese characters. There are a total of 2598 threat sentences and 287383 words that are utilized for relation extraction, whereas there are 3744 threat sentences and 205921 words that are used for entity extraction.

TABLE 2: Statistics for CDTier

Data source		Count of CTI	Count of Sentences	Count of Words
entity extraction	Qianxin	34	784	32282
	Threatbook	27	1092	35855
	Sangfor	10	424	34351
	Tencent	6	268	17925
	Gcow	4	101	8208
	360	4	116	14848
	Antiy	3	214	16998
	Venustech	3	349	22199
	Rising	3	218	11888
	Nsfocus	2	55	2972
	DAS-SECURITY	2	77	5162
	Knownsec	2	46	3233
relation extraction	Data gathered by a security enterprise	100	2598	287383
total		200	6342	493304

4.2 Annotation Strategy

Sequence labeling is a problem that cannot be ignored in natural language processing tasks. Its essence is to label each element of a sequence with a label. The standard method is to use BIO tags, marking each element as "B-X", "I-X", or "O". 'B-X' means that the segment to which this element belongs to category 'X' and is at the head of the segment. "I-X" means that the segment to which this element belongs is in category "X" and is in the middle or end of the segment. And "O" means that the element does not belong to any category [5]. For threat objects in the Chinese threat intelligence corpus, the tool YEDDA [34] is used to manually annotate them according to the annotation format of BIO, as shown in Table 3.

To extract the semi-structured information of the corpus in the research of entity relationship extraction, for the sentence-level threat object relationship corpus, the annotation team code a labeling script based on the KMP [35] algorithm to convert the sentence-level Chinese threat intelligence corpus into JSON format to label. Each JSON corresponds to a sentence-level threat relation corpus, which contains four keys: token, h, t and relation. "token" refers to the sentence-level threat relation corpus after tokenization; "h" refers to the head entity and its position in the sentence, which corresponds to Te_1 of the triplet; "t" refers to the tail entity and its position in the sentence, which corresponds to the three Te_2 of the tuples; and "relation" refers to the relationship between "h" and "t". An example is shown as Figure 3.

TABLE 3: Entity Annotation

Entity type	BIO
Attacker	B-Attacker/I-Attacker
Tools	B-Tools/I-Tools
Industry	B-Industry/I-Industry
Region	B-Region/I-Region
Campaign	B-Campaign/I-Campaign
Non entity	O

4.3 Annotation Process

The annotation team consists of two Ph.D. students and a cybersecurity research professor. All team members are native Chinese speakers. The professor serves as the supervisor and is primarily responsible for resolving difficult marking problems and developing annotation standards.The whole process is divided into five stages:

(1) Learn the content and format of threat intelligence, and make tentative annotations (10 threat intelligence documents).

(2) The common parts (60 threat intelligence documents) were marked, and the differences were compared. Supervisors participate in discussions to determine whether revisions are required.

(3) Perform experiments to validate the labeling results, and then adjust the accuracy and breadth of the labeling based on the experimental results.

(4) Label the remaining corpus (30 threat intelligence documents) parts based on the experimental results.

(5) Each annotator cross-validates the annotation results and discusses the case from different viewpoints.

(6) The supervisor makes the final decision on all marked documents. The datasets of CDTier are all generated according to the above five stages, which take about 2 months.

Finally, to eliminate errors in the annotation process, we make the three efforts as follows: (1) revise the annotations according to the experimental results; (2) manually remove the redundant redundant information in the sentences; (3) only consider the threat entities within the sentences relation. In addition, the submitted CDTier is the last version to pass all validations.

5 Experiments

Experiments on entity extraction and relation extraction are carried out in the form of a pipeline to research the downstream tasks provided by CDTier.

```
{
    'token': ['透', '明', '部', '落', '在', '2019', '年', '下', '半', '年', '的',
    '活', '动', '一', '直', '针', '对', '阿', '富', '汗', '地', '区', '，', '在',
    '2020', '年', '开', '始', '再', '次', '转', '为', '针', '对', '印', '度', '地',
    '区', '，', '大', '约', '在', '2020-01', '月', '左', '右', '，', '其', '以',
    '职', '位', '招', '聘', '题', '材', '、', '军', '队', '题', '材', '对', '印',
    '度', '目', '标', '发', '起', '攻', '击'],
    'h': {
        'name': '透明部落',
        'pos': [0, 3]
    },
    't': {
        'name': '阿富汗',
        'pos': [95, 97]
    },
    'relation': 'target_at'
}
```

Fig. 3: An example of CDTier's annotation

5.1 Threat entity extraction

Since it was proposed in 2018, the pre-training model BERT [36] has received a lot of attention and has been used in many fields such as named entity recognition and association extraction. The BERT network architecture essentially utilizes the multi-layer Transformer architecture described in the book "Attention is all you need" [37] and implements the Attention mechanism to convert the distance between two words at any position into one, effectively eliminating the need for Long-term dependency problems in naming and entity recognition tasks. Therefore, this paper trains and tests the CDTier using Google's official Chinese BERT pre-training model and a variety of typical BERT-based entity extraction.

5.1.1 Evaluation

As CDTier annotates multiple knowledge objects, typical evaluation measures such as precision rate P(Formula 1), recall rate R (Formula 2) and F1 value (Formula 3) are implemented. Macro-average$P_{macro(P,R,F1)}$ and micro-average $P_{micro(P,R,F1)}$ are used to evaluate the overall performance of entity extraction, where macro-average $P_{macro(P,R,F1)}$ is the arithmetic mean of each entity's performance indicators and micro-average $P_{micro(P,R,F1)}$ is the arithmetic mean of instance documents' performance indicators.

$$P = \frac{N_c}{N_c + N_d} \tag{1}$$

$$P = \frac{N_c}{N_a} \tag{2}$$

$$F1 = \frac{2PR}{P + R} \tag{3}$$

$N_c$is the number of correctly recognized entities, N_d is the number of incorrectly recognized entities and N_a is the number of all entities.

$P_{macro(P,R,F1)}$:

$$\begin{cases} P_{macro(P)} = \frac{1}{n}\sum_{i=1}^{n} P_i \\ P_{macro(R)} = \frac{1}{n}\sum_{i=1}^{n} R_i \\ P_{macro(F1)} = \frac{1}{n}\sum_{i=1}^{n} F1_i \end{cases} \tag{4}$$

$P_{micro(P,R,F1)}$:

$$\begin{cases} P_{micro(P)} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FP_i} \\ P_{micro(R)} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FN_i} \\ P_{micro(F1)} = \frac{2 * P_{micro(P)} * P_{micro(F)}}{P_{micro(P)} * P_{micro(F)}} \end{cases} \tag{5}$$

Among them, TP is the number of positive examples that are correctly classified and FP is the number of positive examples that are wrongly classified. FN is the number of misclassified samples.

5.1.2 Data set

The training set, validation set and test set are divided in an 8:1:1 ratio by CDTier. Table 4 shows the distribution of five knowledge objects on the dataset in the experiment: Attacker, Tools, Industry, Region and Campaign.

5.1.3 Comparison and analysis of correlation algorithms

Four typical BERT-based NER models(BERT+LSTM, BERT+BiLSTM, BERT+BiLSTM+CRF and BERT+BiLSTM+GRU+CRF) are trained and tested on CDTier in this paper.

BERT+**LSTM**: The architectural model is shown in Figure 4. The LSTM model is a type of recurrent neural network

TABLE 4: Distribution of annotated data sets

Entity type	Train Set	Validation set	Test Set
Attacker	1289	119	86
Tools	1144	89	89
Industry	672	87	104
Region	845	137	149
Campaign	91	6	9

that has been widely used in NLP tasks due to its superior performance in handling long-distance relationships. When the sequence labeling task is done, the neural network's current output is dependent not only on the current input, but also on the previous output and the LSTM model can well extract the context information of words in the sequence labeling task.

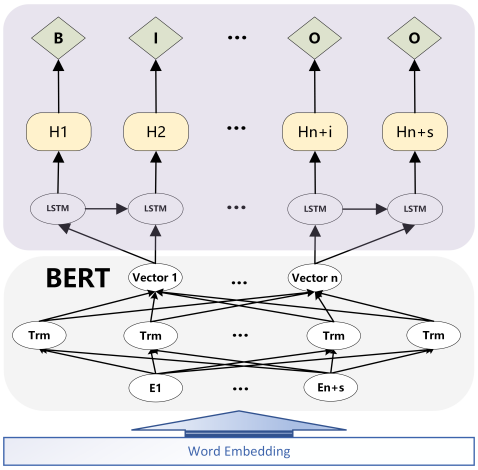


Fig. 4: Model of the BERT+LSTM neural network

BERT+**BiLSTM**: The architectural model is shown in Figure 5. The main limitation of single LSTM model is that it cannot handle contextual signals at the same time. Graves A et al [38]. proposed the BiLSTM model, which is a bidirectional long short-term memory network made up of a forward and a backward LSTM. Compared with single LSTM, BiLSTM can better capture the bidirectional semantic features of the sequence.

BERT+BiLSTM+**CRF**: The architectural model is shown in Figure 6. The CRF [39] model plays a very important role in the sequence labeling task and the CRF model is used to obtain a globally optimal sequence label in the sequence labeling task.

BERT+BiLSTM+**GRU**+CRF: In article [5], we describe and use this model in detail. GRU was proposed by Cho K et al [40]. in 2014, and the complex structure of LSTM is optimized accordingly. Compared with LSTM, using GRU can achieve similar results, and it is also relatively easier to train, so the training effect can be improved to a greater extent.

The overall performance of various models is assessed using evaluation indicators such as macro-average macro and micro-average micro. Table 5 shows the experimental results using the macro-average as the evaluation indicator, and Table 6 shows the experimental results using the micro-average as the evaluation indicator.

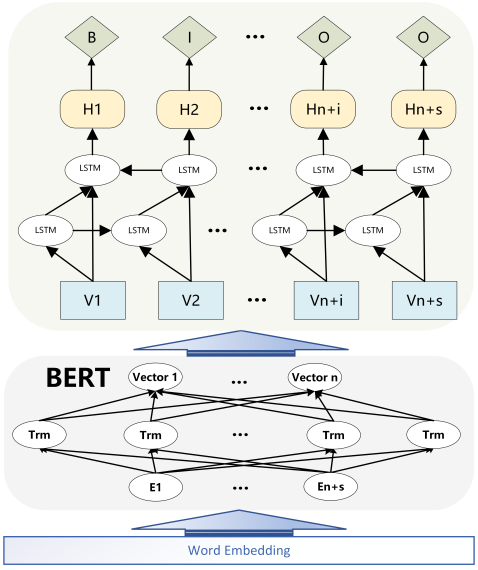


Fig. 5: Model of the BERT+BiLSTM neural network

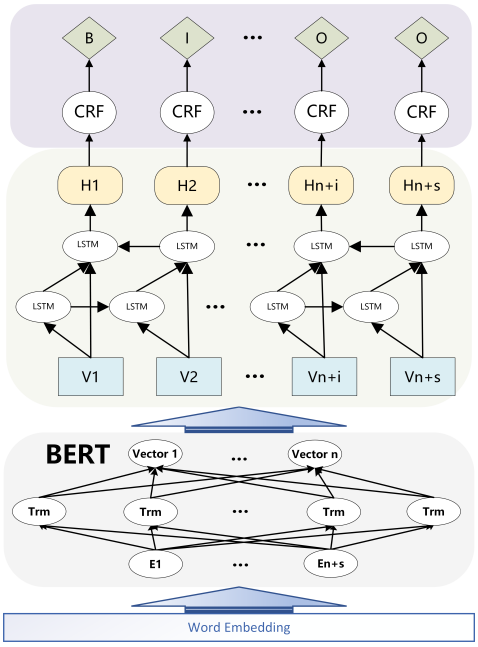


Fig. 6: Model of the BERT+BiLSTM+CRF neural network

BERT+LSTM VS BERT+BiLSTM: Based on BERT, since BiLSTM can use the bidirectional structure to obtain context sequence information, the performance of the BiLSTM model is significantly improved compared with a single LSTM, which is improved by 11.66% in $P_{macro(F1)}$ and 12.41% in $P_{micro(F1)}$.

BERT+BiLSTM VS BERT+BiLSTM+CRF: Comparing the experimental results of BERT+BiLSTM and BERT+BiLSTM, after adding the CRF module, $P_{macro(F1)}$ and $P_{micro(F1)}$ increased by 0% and 6.13% respectively, mainly because the CRF module can It makes good use of the relevance of similar tags to obtain contextual information.

BERT+BiLSTM+CRF VS BERT+BiLSTM+GRU+CRF: A GRU layer is added between the BiLSTM layer and

TABLE 5: Comparison of macro

Models	macro		
	$P_{macro(P)}$	$P_{macro(R)}$	$P_{macro(F1)}$
BERT+**LSTM**	52.51%	61.50%	55.82%
BERT+**BiLSTM**	60.09%	75.79%	66.66%
BERT+BiLSTM+**CRF**	69.99%	57.79%	66.66%
BERT+BiLSTM+**GRU**+CRF	77.03%	79.18%	77.99%

TABLE 6: Comparison of micro

Models	micro		
	$P_{micro(P)}$	$P_{micro(R)}$	$P_{micro(F1)}$
BERT+**LSTM**	51.11%	61.50%	55.82%
BERT+**BiLSTM**	58.83%	75.79%	66.24%
BERT+BiLSTM+**CRF**	69.25%	75.79%	72.37%
BERT+BiLSTM+**GRU**+CRF	76.76%	79.18%	77.95%

the CRF layer in the BERT+BiLSTM+CRF model, makes the BERT+BiLSTM+GRU+CRF model improve by 11.33% in $P_{macro(F1)}$ and 5.58% in $P_{micro(F1)}$ compared with BERT+BiLSTM+CRF. This is because of the multi-layer stacked neural network structure, the model depth is deeper and the extracted features are deeper, resulting in more accurate predictions.

Finally, the performance of extracting entities using different models is compared and the results are shown in Figure 7.

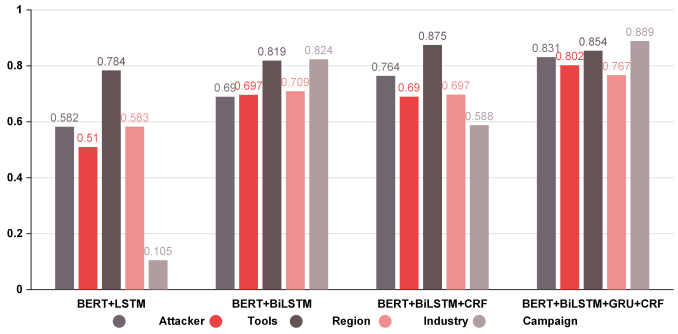


Fig. 7: Experimental comparison of various model effects on various entities

5.2 Relation extraction

Deep learning technology has been widely applied by researchers in the field of entity relation extraction after several years of development [41]. Figure 8 depicts the deep learning process framework for entity relation extraction. The labeled text corpus is first obtained by manually annotating the corpus or automatically aligning the remote knowledge base, and then the word2vec model is applied to the labeled corpus, with the semantic information of the word represented by word vector, position vector and grammatical relationship vector. The feature vector serves as the basis for a neural network. input from the following feature extraction, the semantic features are weighted further using softmax, and the entity-relationship pair is output [42].

This paper used OpenNRE [43] and DeepKE [44] to train four models based on CNN, RNN, GCN and BERT in the

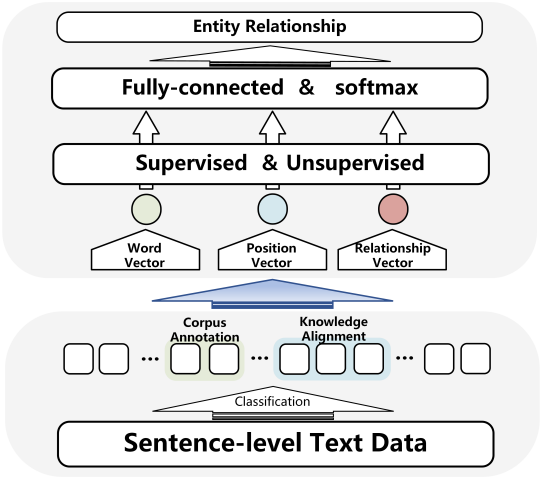


Fig. 8: Deep learning entity relation extraction framework

CDTier in order to evaluate the dataset's validity and the generated threat entity relationship extraction model is used to identify threat entity relationships.

5.2.1 Evaluation

Threat entity relation extraction took the Te_1, $Te_2$and sentence-level corpus in the triplet $< Te_1, Tr, Te_2 >$ as the input of the experiment, predicted the relationship Tr(Equation 6) between Te_1 and Te_2.

$$Acc = N_c/(N_c + N_d) \tag{6}$$

In Formula 6, N_c is the number of relationships with the highest confidence in the correct relationship classification, N_d is the number of relationships with the highest confidence in the correct classification, and N_a is the number of all relationship classifications. The classification confidence calculation formula is as follows(Formula 7):

$$Confidence(E_A \xrightarrow{r}> E_B) = P(A|B) \tag{7}$$

E_A represented the entity A, E_B represented the entity B. It reveals the probability that entity B and entity A have a relationship r when entity A appears.

5.2.2 Data set

CDTier contains 2598 knowledge object relation extraction samples. Due to the small amount of corpus, there is no guarantee that the test set to be submitted is completely identical to the training set. Therefore, no validation set was set up in the experiment and the training set and test set were randomly allocated in a ratio of 4:1. Table 7 shows the overall size of the dataset; Table 8 shows the specific distribution of the 11 threat entity relationships on the dataset.

TABLE 7: Data Set Size

Threat Intelligence	Sentence-level threat corpus	Training Set	Test Set
100	2598	2078	520

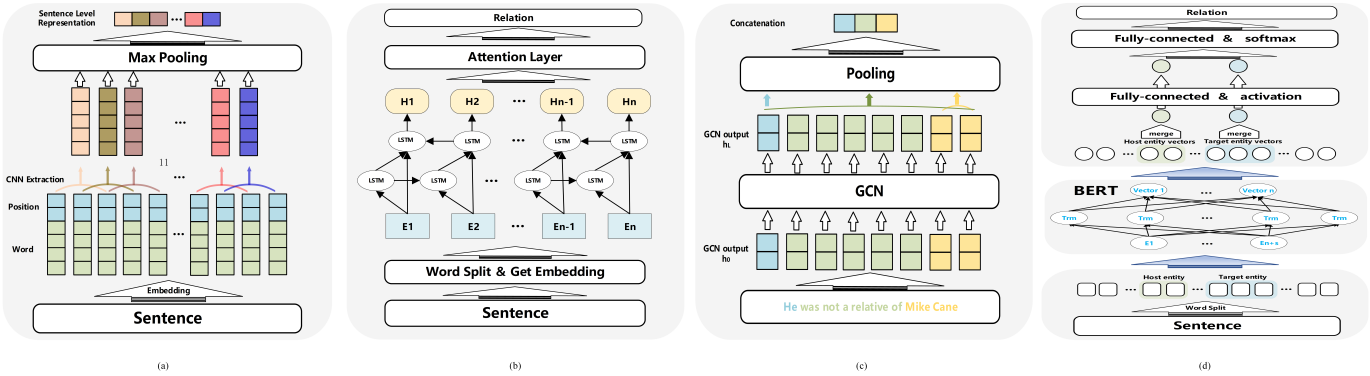


Fig. 9: Framework diagram of four typical relation extraction models: CNN, RNN, GCN and BERT

TABLE 8: Distribution of relational data set

	Training set	Test set	Total
alias_of	271	77	348
cooperate_with	27	6	33
related_to	136	27	163
uses	269	58	327
target_at	784	195	943
originated_from	159	34	193
launch	36	4	40
kin	9	1	10
consist_of	54	12	66
operating_in	291	90	381
develop	42	16	58

5.2.3 *Analysis of correlation algorithm*

To train and test the CDTier, the experiments will use four common relation extraction models: CNN, RNN, GCN and BERT. Figure 9 shows the overall architecture of the four extraction models.

Convolutional Neural Networks (CNN) are a type of deep feed forward neural network that includes characteristics such as local connections and weight sharing [45]. The pre-trained or randomly initialized embedding is utilized to turn the vocabulary in the phrase into a word vector, and the displacement vector of the entity word is represented by the corresponding displacement of the entity word in the sentence and its context, as illustrated in Figure 9(a). Following that, the CNN network is applied to gather sentence-level features, and the pooling approach is used to achieve feature vector representation after compression. Finally, the feature vector is input into a completely continuous neural network layer to perform feature relationship categorization in sentences.

Recurrent neural network (RNN) is a kind of recurrent neural network that takes sequence data as input, conducts recursion in the sequence's development direction, and connects all nodes (recurrent units) in a chain [46]. As shown in Figure 9(b), different from the traditional use of the CNN network system to extract sentence features, RNN first uses a bidirectional LSTM network to extract sentence features, and at the same time adds a self-attention mechanism to weigh the output feature vector, and finally get the eigenvector representation with bias. Similarly, the resulting vector is mapped into a fully connected neural network layer, which finally completes the classification of relations.

Graph Convolutional Network(GCN) applies the "convolution" idea to graph data. This approach is often made up of many convolutional layers. Each convolutional layer can transfer node information to a single jumping neighbor and update the nodes using the aggregate function in its own single Information acquired near the leap [47]. Figure 9(c) depicts the use of GCN in the extraction process. The approach utilized in the picture field is followed in this experiment. In the graph convolution approach, the dependency analysis tree of the sentence is turned into a full adjacency matrix, and each word in the clause is a Node that makes graphs. Syntactic information may be recovered in this manner, and then the work of relation classification can be performed using the pooling layer and the fully connected layer.

Bidirectional Encoder Representation from Transformers(BERT) model has shown excellent results in a variety of NLP classification and sequence labeling applications. The fundamental difference between relation classification and other NLP tasks is that it depends on both sentence information and information from the two target entities. To perform the objective of relation extraction, all sentences are directly input into the BERT layer, as illustrated in Figure 9(d), and the result feature vector is directly input into the fully connected layer.

Table 9 shows the effects of CNN, RNN, GCN and BERT on the CDTier. And a comparison of each model's performance during doing relation prediction on a Chinese threat corpus of 55 characters on an Intel(R) Xeon(R) Silver 4210R CPU.

TABLE 9: Model performance comparison

Relation extraction model	The best accuracy in 50 epochs	Prediction time of the same sentence (seconds)
Based on CNN	62.69%	1.1649303
Based on RNN	60.96%	1.0084550000000005
Based on GCN	54.62%	0.9934381999999999
Based on BERT	89.40%	2.6463578939437866

It can be seen that the BERT-based threat entity relation extraction model outperforms other models on our

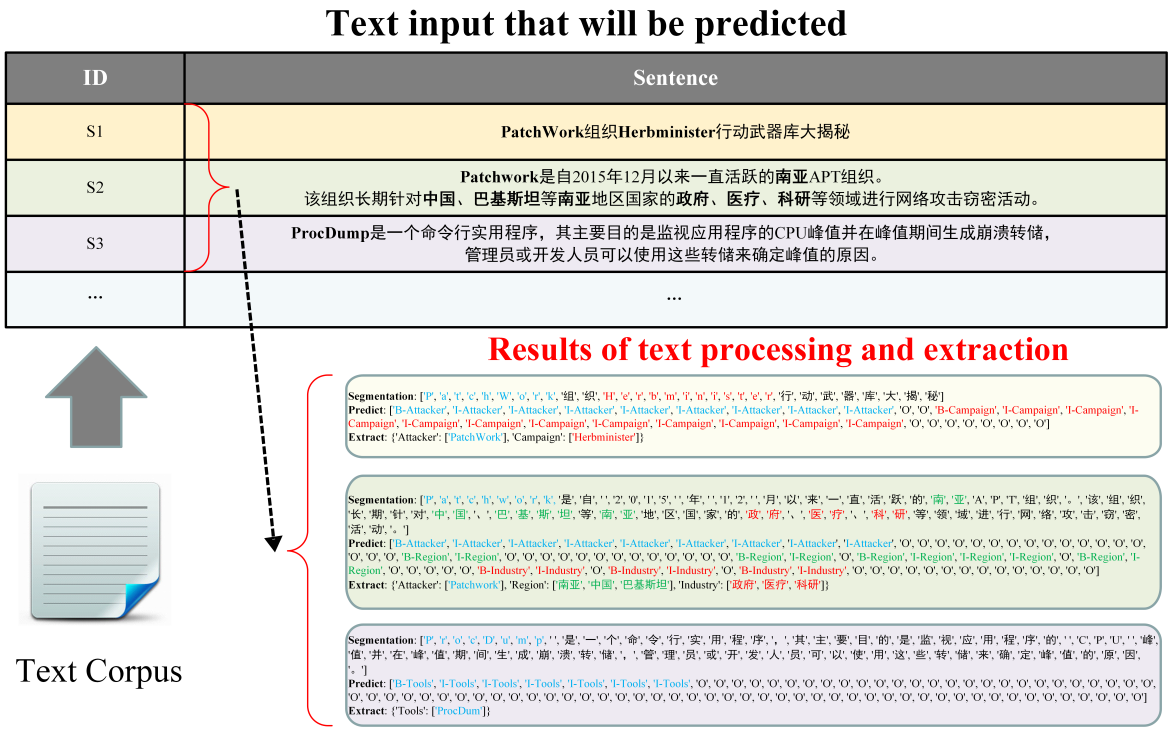


Fig. 10: Example of sentence-level CTI corpus entity extraction

annotated dataset, which further proves the superiority of BERT in natural language processing tasks. On the data set we marked, however, the graph neural network method based on GCN performs the poorest. It can be shown that, when compared to the experiments given in this chapter, the classic deep learning method has a greater overall result. In terms of performance, the BERT-based model's prediction time is significantly longer than other models now that it includes a pre-trained model. The performance of the prediction is generally inversely proportional to the accuracy.

5.3 Case of extraction

5.3.1 Case of entity extraction

To train the CDTier, the paper used the best-performing BERT+BiLSTM+GRU+CRF algorithm. "PatchWork Herbminister's Operational Arsenal Revealed" [48], an APT threat intelligence report regarding the APT organization Patchwork published on October24, 2022 by KnowSec, was chosen for entity extraction testing. According to the information, KnowSec researchers obtained access to some of the organization's assets, acquired the tools for its bulk attack operation, and named the infiltration effort "Operation Herbminister" based on the names of the tools. The active weapon chain includes 5 forms, and the weapon tool type contains penetration kits (cftmon, Mathworks, htaccess), data stealing kits (CEQ, Isass, DellDrive), remote control Trojans (DellSupportnew, on), isolated network breakthrough kits (MatlabWebManager, wversion), record suite (MSUsoCoreWorker, Servics), and controlled machine operation suite (WeRecUpdater, WinShow). The extraction test results for a few common sentence-level CTI corpora in this intelligence are displayed in Figure 10.

The experimental results in Figure 10 showed that the model trained with CDTier successfully extracts the details of the attackers, battles, tools used, industries involved and regions involved described in sentences S1, S2, S3 and S4, which generally reflects the success of the entity extraction part of CDTier.

5.3.2 Case of relation extraction

NSFOCUS released the threat intelligence "APT organization LOREC53 (Lori Bear) recent large-scale cyber attack activities against Ukraine" [49] on February 16, 2022, which describes the NSFOCUS Fuying lab that captured a large number of phishing attack activities against Ukraine and the associated malicious files, which include pdf, doc, cpl, lnk and other types. After analysis, it was confirmed that this series of phishing activities all came from the APT organization LOREC53. During the period from the end of 2021 to February 2022, the organization used multiple attack techniques to deliver multiple phishing documents to key state institutions in Ukraine, including the Ministry of Defense, the Ministry of Finance, embassies, state enterprises and public medical facilities and conducted cyber-attack activities aimed at gathering information about the organization's personnel.

The analysis was performed for one of the sentence-level threat corpus: "本次发现的多起攻击事件，都是Lorec53(洛瑞熊)在2021年底至2022年2月不同时间段内执行的针对乌克兰政府部门、军队、国企的大规模网络攻击活动的一部分。" . The purpose is to use the BERT-based threat entity relationship extraction model to predict the relationship between entities "Lorec53" and "Lore Bear", entities "Lorec53" and "Ukraine", entities "Government Depart-

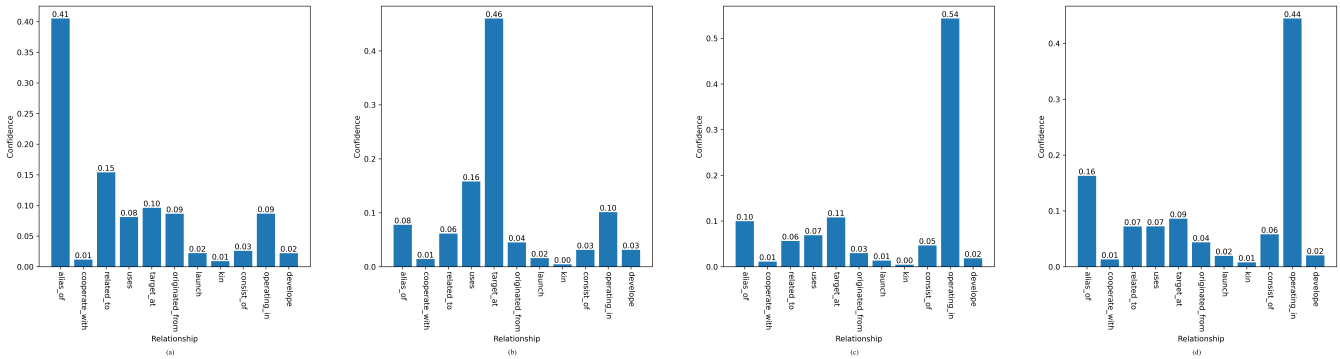


Fig. 11: Example of sentence-level CTI corpus relation extraction

ment" and "Ukraine", entities "Army" and "Ukraine". Figure 11 showed the results of specific prediction.

As shown in Figure 11(a), the relationship between "Lorec53" and "洛瑞熊" with the highest confidence is "alias_of", which means that "Lorec53" and "Lori Bear" have the same name and the prediction is accurate. The most confident relationship between "Lorec53" and "乌克兰," as shown in Figure 11(b), is "target at", demonstrating that "Lorec53" attacked "Ukraine" and that the prediction was correct. According to Figure 11(c), "operating_in", which stands for "政府部门" for "乌克兰", is the relationship with the highest level of confidence between "government department" and "Ukraine", the prediction is true. As shown in Figure 11(d), the relationship with the highest confidence level between "军队" and "乌克兰" is "operating_in", which means "army" of "Ukraine", the prediction is accurate. Therefore, the extraction results generally reflect the validity of the relational extraction of CDTier.

6 Conclusion

The CTI Chinese corpus's lack significantly limits text-based prediction models and their downstream applications. This paper summarized 5 threat entities and 11 threat entity relations according to the STIX2.1 standard and the actual need for threat intelligence and constructed a Chinese threat intelligence dataset (CDTier). CDTier included 200 WeChat intelligence documents from 13 security vendors, including 677 threat sentences, 707,716 Chinese characters, 4,259 threat entities and 2,562 threat entity relations. This is the first Chinese dataset of CTI. Using expert knowledge, we designed a sensible labeling method and conducted an extensive quality evaluation on the CDTier during the annotation process. Furthermore, experiments on named entity recognition and relation extraction are performed on the CDTier using mature and widely used deep learning techniques in the NLP area to validate CDTier's research value.

Acknowledgments

This research was supported by National Natural Science Foundation of China under Grant No. U20B2046, and the National Key Research and Development Program of China under Grant 2021YFB2012402, Guangdong Province Universities and Colleges Pearl River Schoalar Funded Scheme(2019) and Guangdong Higher Education Innovation Group 2020KCXTD007 and Guangzhou Higher Education Innovation Group 202032854 and Guangzhou Basic Research Program Jointly Funded by City and University 202201020218.

References

[1] X. Rui, C. Jianfeng, and L. Fang, "Cyber space security threat intelligence and application research," *Communication technology*, vol. 49, no. 6, pp. 758–763, 2016.

[2] Y. Ren, Y. Xiao, Y. Zhou, Z. Zhang, and Z. Tian, "Cskg4apt: A cybersecurity knowledge graph for advanced persistent threat organization attribution," *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[3] W. Chen, X. Helu, C. Jin, M. Zhang, H. Lu, Y. Sun, and Z. Tian, "Advanced persistent threat organization identification based on software gene of malware," *Transactions on Emerging Telecommunications Technologies*, vol. 31, no. 12, p. e3884, 2020.

[4] R. McMillan and K. Pratap, "Market guide for security threat intelligence services," *Gartner report (G00259127)*, 2014.

[5] Y. Zhou, Y. Tang, M. Yi, C. Xi, and H. Lu, "Cti view: Apt threat intelligence analysis system," *Security and Communication Networks*, vol. 2022, 2022.

[6] STIX. (2022, Sep.) Introduction to stix. [Online]. Available: https://oasis-open.github.io/cti-documentation/stix/intro

[7] TAXII. (2022, Sep.) Introduction to taxii. [Online]. Available: https://oasis-open.github.io/cti-documentation/taxii/intro

[8] CybOX. (2022) About cybox. [Online]. Available: http://cyboxproject.github.io/about/

[9] H. E, W. Zhang, S. Xiao, R. Cheng, Y. Hu, X. Zhou, and P. Niu, "Survey of deep learning entity relation extraction," *Software Journal*, vol. 30, no. 6, pp. 1793–1818, 2019.

[10] A. D. Kent, "Comprehensive, multi-source cyber-security events data set," Los Alamos National Lab.(LANL), Los Alamos, NM (United States), Tech. Rep., 2015.

[11] B. Eshete, R. Gjomemo, M. N. Hossain, S. Momeni, R. Sekar, S. Stoller, V. Venkatakrishnan, and J. Wang, "Attack analysis results for adversarial engagement 1 of the darpa transparent computing program," *arXiv preprint arXiv:1610.06936*, 2016.

[12] Z. Liu, C. Zhu, and T. Zhao, "Chinese named entity recognition with a sequence labeling approach: based on characters, or based on words?" in *International Conference on Intelligent Computing*. Springer, 2010, pp. 634–640.

[13] C. Yan, Q. Su, and J. Wang, "Mogcn: Mixture of gated convolutional neural network for named entity recognition of chinese historical texts," *IEEE Access*, vol. 8, pp. 181 629–181 639, 2020.

[14] R. Manikandan, K. Madgula, and S. Saha, "Teamdl at semeval-2018 task 8: cybersecurity text analysis using convolutional neural network and conditional random fields," in *Proceedings of The 12th International Workshop on Semantic Evaluation*, 2018, pp. 868–873.

[15] Y. Keim and A. Mohapatra, "Cyber threat intelligence framework using advanced malware forensics," *International Journal of Information Technology*, pp. 1–10, 2019.

[16] N. Rastogi, S. Dutta, M. J. Zaki, A. Gittens, and C. Aggarwal, "Malont: An ontology for malware threat intelligence," in *International Workshop on Deployable Machine Learning for Security Defense*. Springer, 2020, pp. 28–44.

[17] H. Tan, M. Chandramohan, C. Cifuentes, G. Bai, and R. K. Ko, "Coldpress: An extensible malware analysis platform for threat intelligence," *arXiv preprint arXiv:2103.07012*, 2021.

[18] D. Xu, Z. Tian, R. Lai, X. Kong, Z. Tan, and W. Shi, "Deep learning based emotion analysis of microblog texts," *Information Fusion*, vol. 64, pp. 1–11, 2020.

[19] M. Shafiq, Z. Tian, A. K. Bashir, X. Du, and M. Guizani, "Corrauc: a malicious bot-iot traffic detection method in iot network using machine-learning techniques," *IEEE Internet of Things Journal*, vol. 8, no. 5, pp. 3242–3254, 2020.

[20] X. Liao, K. Yuan, X. Wang, Z. Li, L. Xing, and R. Beyah, "Acing the ioc game: Toward automatic discovery and analysis of open-source cyber threat intelligence," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 2016, pp. 755–766.

[21] Z. Long, L. Tan, S. Zhou, C. He, and X. Liu, "Collecting indicators of compromise from unstructured text of cybersecurity articles using neural-based sequence labelling," in *2019 international joint conference on neural networks (IJCNN)*. IEEE, 2019, pp. 1–8.

[22] J. Zhao, Q. Yan, J. Li, M. Shao, Z. He, and B. Li, "Timiner: Automatically extracting and analyzing categorized cyber threat intelligence from social data," *Computers & Security*, vol. 95, p. 101867, 2020.

[23] Y. Gao, L. Xiaoyong, P. Hao, B. Fang, and P. Yu, "Hincti: A cyber threat intelligence modeling and identification system based on heterogeneous information network," *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[24] A. Panwar, "igen: Toward automatic generation and analysis of indicators of compromise (iocs) using convolutional neural network," Ph.D. dissertation, Arizona State University, 2017.

[25] G. Husari, E. Al-Shaer, M. Ahmed, B. Chu, and X. 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*, 2017, pp. 103–115.

[26] J. Yan, Z. Du, J. Li, S. Yang, J. Li, and J. Li, "A threat intelligence analysis method based on feature weighting and bert-bigru for industrial internet of things," *Security and Communication Networks*, vol. 2022, 2022.

[27] L. Zongxun, L. Yujun, Z. Haojie, and L. Juan, "Construction of ttps from apt reports using bert," in *2021 18th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*. IEEE, 2021, pp. 260–263.

[28] N. Dionísio, F. Alves, P. M. Ferreira, and A. Bessani, "Cyberthreat detection from twitter using deep neural networks," in *2019 international joint conference on neural networks (IJCNN)*. IEEE, 2019, pp. 1–8.

[29] H. Gasmi, J. Laval, and A. Bouras, "Information extraction of cybersecurity concepts: an lstm approach," *Applied Sciences*, vol. 9, no. 19, p. 3945, 2019.

[30] T. Satyapanich, F. Ferraro, and T. Finin, "Casie: Extracting cybersecurity event information from text," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 8749–8757.

[31] T. Li, Y. Guo, and A. Ju, "A self-attention-based approach for named entity recognition in cybersecurity," in *2019 15th International Conference on Computational Intelligence and Security (CIS)*. IEEE, 2019, pp. 147–150.

[32] K. Satvat, R. Gjomemo, and V. Venkatakrishnan, "Extractor: Extracting attack behavior from threat reports," in *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2021, pp. 598–615.

[33] Matrix. (2022, Sep.) Enterprise matrix. [Online]. Available: https://attack.mitre.org/matrices/enterprise/#

[34] J. Yang, Y. Zhang, L. Li, and X. Li, "Yedda: A lightweight collaborative text span annotation tool," *arXiv preprint arXiv:1711.03759*, 2017.

[35] D. E. Knuth, J. H. Morris, Jr, and V. R. Pratt, "Fast pattern matching in strings," *SIAM journal on computing*, vol. 6, no. 2, pp. 323–350, 1977.

[36] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[38] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural networks*, vol. 18, no. 5-6, pp. 602–610, 2005.

[39] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[40] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[41] L. D. Li DM, Zhang Yang and L. DanQiong, "Survey of entity relation extraction methods," *Journal of Computer Research and Development*, vol. 57, no. 7, pp. 1424–1448, 2020.

[42] G. Q. E. Liu Hui, Jiang QianJun, "Survey on the research progress of entity relation extraction technology," *Application Research of Computers*, vol. 37, no. S2, pp. 1–5, 2020.

[43] X. Han, T. Gao, Y. Yao, D. Ye, Z. Liu, and M. Sun, "Opennre: An open and extensible toolkit for neural relation extraction," *arXiv preprint arXiv:1909.13078*, 2019.

[44] N. Zhang, X. Xu, L. Tao, H. Yu, H. Ye, X. Xie, X. Chen, Z. Li, L. Li, X. Liang *et al.*, "Deepke: A deep learning based knowledge extraction toolkit for knowledge base population," *arXiv preprint arXiv:2201.03335*, 2022.

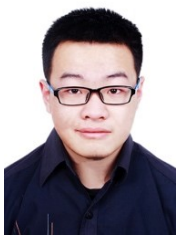
[45] Q. Xipeng, *Neural Networks and Deep Learning*. China Machine Press, 2019.

[46] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[47] XuYibin, "Research on document level relation extraction based on graph convolution network," Master's thesis, Dalian University of Technology, 2021.

[48] KnowSec. Patchwork herbminister's operational arsenal revealed. (2013, Jun 14). [Online]. Available: https://mp.weixin.qq.com/s/XMrWLx6KVeoDQ7WzvOcwqA

[49] NSFOCUS. Apt organization lorec53 (lori bear) recent large-scale cyber attack activities against ukraine. (2013, Jun 14). [Online]. Available: http://blog.nsfocus.net/apt-lorec53-20220216/



Yinghai Zhou is currently pursuing the Ph.D. degree with the Cyberspace Institute of Advanced Technology (CIAT), Minis-try of Education, Guangzhou University, Guangzhou, China. His current research interest is Cyberspace Threat Intelligence and intelligent detection and traceability of APT threat in Cyberspace.

Yitong Ren is currently pursuing the Ph.D. degree with the Cyberspace Institute of Advanced Technology (CIAT), Minis-try of Education, Guangzhou University, Guangzhou, China. She has published some papers in journals and conference proceedings. Her current research interests include APT, Knowledge Graph, Graph Neural Network and Causal Inference.

Ming Yi, M.S. candidate, Institute of Computer Application, China Academy of Engineering Physics. His current research interest is Adversarial Attacks in Machine Learning.

Yanjun Xiao is currently working on PINGXING Lab, Nsfocus Technology Group Company.

Zhiyuan Tan received his Ph.D. degree from the University of Technology Sydney, Australia, in 2014. His research focus on Cybersecurity, Machine Learning, Cognitive Computing. He was a Postdoctoral Researcher with the University of Twente between 2014 and 2016. He is currently an Associate Professor in the School of Computing, Edinburgh Napier University, UK.

Nour Moustafa is Postgraduate Discipline Coordinator (Cyber) and Senior Lecturer in Cyber Security & Computing at the School of Engineering and Information Technology (SEIT), University of New South Wales (UNSW)'s UNSW Canberra, Australia. He was a Postdoctoral Fellow in Cyber Security at UNSW Canberra from June 2017 till February 2019. He received his PhD degree in the field of Cyber Security from UNSW in 2017. He obtained his Bachelor's and master's degrees in Information Systems in 2009 and 2014, respectively, from the Faculty of Computer and Information, Helwan University, Egypt. His areas of interest include Cyber Security, in particular, Network Security, host- and network- intrusion detection systems, statistics, deep learning and machine learning techniques. He is interested in designing and developing threat detection and forensic mechanisms to the Industry 4.0 technology for identifying malicious activities from cloud computing, fog computing, IoT and industrial control systems over virtual machines and physical systems.

Zhihong Tian is currently a Professor and Dean, with the Cyberspace Institute of Advanced Technology, Guangzhou University, Guangdong Province, China. He received his B.S., M.S. and Ph.D. degree in Computer Science and Technology from Harbin Institute of Technology, Harbin, China in 2001, 2003 and 2006 respectively. He is honored as Pearl River Scholar in Guangdong Province. He is also a part-time Professor at Carlton University, Ottawa, Canada. Previously, he served in different academic and administrative positions at the Harbin Institute of Technology. He has authored over 200 journal and conference papers. His research interests include computer networks and cyberspace security. His research has been supported in part by the National Natural Science Foundation of China, National Key research and Development Plan of China, National High-tech R&D Program of China (863 Program). He also served as a member, Chair and General Chair of a number of international conferences. He is a Distinguished Member of the China Computer Federation, and Senior Member of IEEE.