

Article

BVTED: A Specialized Bilingual (Chinese–English) Dataset for Vulnerability Triple Extraction Tasks

Kai Liu ¹, Yi Wang ^{1,*}, Zhaoyun Ding ^{1,*}, Aiping Li ² and Weiming Zhang ¹¹ National Key Laboratory of Information Systems Engineering, National University of Defense Technology, Changsha 410073, China; liukai18@nudt.edu.cn (K.L.); zhangweiming@nudt.edu.cn (W.Z.)² School of Computer, National University of Defense Technology, Changsha 410073, China; liaiping@nudt.edu.cn* Correspondence: wyi@nudt.edu.cn (Y.W.); zyding@nudt.edu.cn (Z.D.);
Tel.: +86-18942066739 (Y.W.); +86-17607310865 (Z.D.)

Abstract: Extracting knowledge from cyber threat intelligence is essential for understanding cyber threats and implementing proactive defense measures. However, there is a lack of open datasets in the Chinese cybersecurity field that support both entity and relation extraction tasks. This paper addresses this gap by analyzing vulnerability description texts, which are standardized and knowledge-dense, to create a vulnerability knowledge ontology comprising 13 entities and 15 relations. We annotated 27,311 unique vulnerability description sentences from the China National Vulnerability Database, resulting in a dataset named BVTED for cybersecurity knowledge triple extraction tasks. BVTED contains 97,391 entities and 69,614 relations, with entities expressed in a mix of Chinese and English. To evaluate the dataset's value, we trained five deep learning-based named entity recognition models, two relation extraction models, and two joint entity–relation extraction models on BVTED. Experimental results demonstrate that models trained on this dataset achieve excellent performance in vulnerability knowledge extraction tasks. This work enhances the extraction of cybersecurity knowledge triples from mixed Chinese and English threat intelligence corpora by providing a comprehensive ontology and a new dataset, significantly aiding in the mining, analysis and utilization of the knowledge embedded in cyber threat intelligence.



Citation: Liu, K.; Wang, Y.; Ding, Z.; Li, A.; Zhang, W. BVTED: A Specialized Bilingual (Chinese–English) Dataset for Vulnerability Triple Extraction Tasks. *Appl. Sci.* **2024**, *14*, 7310. <https://doi.org/10.3390/app14167310>

Academic Editor: Christos Bouras

Received: 12 July 2024

Revised: 31 July 2024

Accepted: 12 August 2024

Published: 20 August 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: cyber threat intelligence; information extraction; named entity recognition; relation extraction; Chinese–English hybrid vulnerability descriptions

1. Introduction

Cyber threat intelligence (CTI) is an important source of knowledge for cybersecurity practitioners. Timely acquisition and analysis of CTI information plays a vital role in the process of cyber attack and defense. CTI contains a variety of detailed information about current or upcoming cybersecurity threats, such as strategies, technologies and assets [1]; it can help businesses or organizations implement active cyber defense against cybersecurity threats [2]. However, due to the diversity and concealment of cyber threat knowledge, it is difficult for most security agencies and personnel to efficiently obtain accurate threat knowledge, which leads to the inability to start the defense mechanism accordingly. In addition, with the continuous strengthening of the ability of self-replication, mutation and dissemination of various attack software and malicious programs, their destructiveness is also increasing, which brings great difficulties for data recovery and information system reconstruction at a later stage. Traditional cybersecurity experts need to use CTI to broaden their knowledge boundaries to deal with new cybersecurity issues [3]. Generally speaking, all information related to a cyber threat can be called CTI, such as logs, network traffic, pictures, text, etc. However, most knowledge of threat intelligence is described and published in textual form. Therefore, how to use natural language processing technology to effectively extract threat intelligence knowledge from open source Internet articles or

reports, and transform it into a standardized and structured knowledge organization form, has been of great significance and practical application value for cybersecurity research.

As one of the most active research hotspots in the field of cybersecurity, the task of network threat intelligence knowledge extraction is to identify the security entities in the text by analyzing and learning the context information and classifying the semantic relations contained in it [4]. This task aims to extract knowledge from different sources, including different structural data, and deposit it in a knowledge graph. In the knowledge graph, knowledge is generally composed of the form of the entity relation triples $\langle e_1, r_{1,2}, e_2 \rangle$, where e_1 represents the head entity and e_2 is the tail entity, and $r_{1,2}$ is the relation between the two. Data sources can be structured data (linked data, databases), semi-structured data (tables and lists in the web page), unstructured data (pure text data), etc. This task mainly includes two parts: named entity recognition (NER) and relation extraction (RE). The former is used to identify the boundaries of the entities and classify them into a predefined category, while the latter is used to identify whether there is a certain predefined relation type between the entities in the input text. An important task of CTI knowledge extraction is to identify important entities involved in the threat, such as attackers, network products, vulnerabilities, etc. At present, the more commonly used entity extraction method is based on the sequence prediction model, that is, to obtain the optimal annotated sequence for the input text sequence according to the context semantics. Entity extraction concerns the explicit knowledge of the intelligence text, the implicit semantic relation in the sentence and the interaction between the entities. The goal is to extract possible correlations between the entities, such as an attack organization vulnerability “exploit” relation, an attack organization “Use” relation of malicious software, a software and hardware product version of the “has_version” relation, etc. According to the different datasets used, it can be divided into three categories: template-based relation extraction, relation extraction based on supervised learning and relation extraction based on weak supervised learning. Common implementation methods include the pipeline method and the joint extraction method.

For the field of cybersecurity, in addition to designing an effective information extraction model, it is also necessary to have enough field annotation data as a training corpus to train or fine-tune the model. CTI, usually compiled by experts in the field of cybersecurity, may be a barrier to the general reader because of the expertise needed. In addition, due to individual differences in writing habits, language use and analytical focus, the intelligence content shows significant diversity in consistency and accuracy. Due to the complexity of entities in the field of cybersecurity, the task of extracting cybersecurity knowledge from various channels requires an ontology model that can effectively organize cybersecurity entities and relations as a guide. The ontology can accurately cover the information of the domain entity type, entity relation type and so on.

Extracting information from a corpus presents three main challenges. First, in the field of Chinese cybersecurity, there is no open dataset that satisfies both entity and relation extraction tasks [5]. Existing datasets are either not open or only support general NER tasks, and do not adequately support NER, relation extraction, or joint entity–relation extraction tasks in cybersecurity [6]. Second, the structure of network security text data is complex, with many descriptions mixing Chinese and English [7]. This includes software and vulnerabilities that have both Chinese and English names, and the extensive use of nested structures, abbreviations and obscure terms, which significantly increases the difficulty of entity and relationship extraction. Finally, the Chinese CTI corpus is more complicated than the English CTI corpus, adding another layer of difficulty to the extraction process [8].

To address the aforementioned challenges, we present the Bilingual (Chinese–English) Vulnerability Triple Extraction Dataset (BVTED), the first known dataset capable of supporting Chinese cybersecurity entity–relation triple extraction tasks. We develop an ontology model for describing cybersecurity intelligence knowledge, encompassing all relevant entity and relation types necessary for knowledge extraction. Given the dataset’s char-

acteristics, which include a substantial amount of mixed Chinese and English texts, we trained five deep learning-based named entity recognition models, two standalone relation extraction models, and two joint entity–relation extraction models. The experimental results validate the effectiveness of the BVTED dataset, demonstrating its potential for advancing research in cybersecurity intelligence.

This work provides a significant scientific contribution by addressing the lack of open, comprehensive datasets in Chinese cybersecurity knowledge triple extraction tasks. By offering a robust dataset and demonstrating its utility through extensive experiments, we enhance the ability to mine, analyze and utilize cybersecurity knowledge, thereby advancing the field of cyber threat intelligence. The remainder of this paper is organized as follows:

Section 2 reviews related work, including the existing datasets and information extraction techniques in the cybersecurity domain; Section 3 introduces the construction, annotation and statistical analysis of the BVTED dataset; Section 4 introduces the experimental settings and evaluation metrics in this paper and evaluates the final experimental results; Section 5 summarizes the contributions and limitations of our research; Section 6 provides a detailed outline of our future research directions.

2. Related Work

In this section, we provide an overview of the related work in cybersecurity information extraction, structured into three focused subsections. Section 2.1 evaluates existing datasets, emphasizing the need for the BVTED dataset. Section 2.2 discusses the information extraction techniques in cybersecurity, while Section 2.3 explores the construction of the cybersecurity ontology, collectively setting the stage for our dataset’s contribution to the field.

2.1. Datasets for Cybersecurity Information Extraction

For the task of information extraction, substantial efforts have been dedicated to the development of general domain datasets aimed at training and evaluating algorithms. Their extraction goal is to extract entity types that are frequent occurrences of, for example, names of people, places and organizations. Ding et al. [9] released a large-scale manually annotated NER dataset, which contains 8 coarse-grained types and 66 fine-grained types. Xu et al. [10] created a well-defined fine-grained dataset for NER in Chinese, and the entity types are divided into 10 label categories. Similarly in the relationship extraction task, TACRED (TAC Relation Extraction Dataset) [11], which is constructed using newswire and web text, and NYT10 [12], which is constructed using the New York Times as the text and references the Freebase knowledge base for labeling relationships, are both widely used in traditional domains.

Some studies have demonstrated [13] that the performance of models trained by using only the text of the general domain but not the corpus of the professional domain will decline sharply [14]. Therefore, in order to meet the task of cybersecurity intelligence knowledge extraction, it is necessary to build a cybersecurity intelligence field dataset. At present, NER datasets for the field of CTI can basically cover English scenarios. For instance, Wang et al. [15] redefined entity types according to the STIX2.1 standard and published an English NER dataset APTNER for the CTI field. The final dataset contains 10,984 sentences, 260,134 tokens and 39,565 entities. Zhou et al. [5] released a Chinese NER dataset and a Chinese relational extraction dataset, respectively. However, these datasets are based on entity types and relationships designed based on the STIX 2.1 standard. Using such a construction method without a data perspective, it is easy to ignore the important information similar to the affected “product name” and “version” in the text. Moreover, since the datasets for both the NER and relationship extraction tasks are independent, the cross-promoting effect between the entity extraction and relationship extraction tasks is limited. The new malware dataset released by Rastogi et al. [16] is the only dataset available for both the RE and NER tasks. Although they are collected and annotated based on a

variety of data sources, most of them are generated with an English corpus. On the other hand, much security knowledge is published in multiple languages, often interspersed with English, which requires the creation of a multilingual dataset.

2.2. Information Extraction Techniques in Cybersecurity

For information extraction tasks in the field of cybersecurity, NER techniques, RE techniques and joint entity–relation extraction techniques are the main techniques. A comprehensive review of each of these three technologies is provided in the sections to follow.

2.2.1. NER in Cybersecurity

The named entity recognition task consists of two parts: type recognition of the entity and boundary detection of the entity. Since an entity is usually not made up of a single word, the range needs to be determined by boundary detection, so that the entity composed of multiple words can be fully identified [17]. The type identification of entities is used to divide the entities into different categories according to their properties. In this process, we tend to predefine the category of the entity and label the entities in the text of the corresponding category to be identified. Traditional NER techniques can be divided into three mainstream methods: rule-based methods, unsupervised learning models and feature-based supervised learning methods [18].

Cheng et al. [19] have improved the traditional bootstrapping algorithm. Through the process of pattern matching, entity recognition, entity evaluation and other processes, they have continuously improved the rule base, expanded the sample set of entity data, improved the semantic drift problem caused by the increase in the number of iterations in the past, and effectively improved the performance of entity recognition. Sara Mumtaz et al. [20] proposed a scoring function called normalized point mutual information (NPMI) to identify phrases, and the phrases are formed based on the strength of association between words in the phrases. The joint probability of the occurrence of two sets of cybersecurity words i and j and the probability of their independent occurrence are calculated, respectively. If the frequency is higher, the score is positive, proving a strong correlation, and vice versa. A new model, entity extraction with a multi-head attention mechanism, was proposed by Li et al. [21]. The model first obtains the vector representation important to the entity through the multi-head self-attention mechanism, then merges with the feature vectors generated by the recurrent neural network model, and finally inputs the fused results into a linear layer to obtain the sequence labels, thus extracting the entity. Xie [4] proposed a method based on a deep active learning strategy to build the input matrix based on the unique semantic characteristics of the text in the field of cybersecurity, and receive the input matrix through the residual void neural network, so that the model can fully tap the semantic information related to the security entities. In addition, a joint extraction model based on semantic context filtering is proposed, which can better identify different security entities and the relationships between entities. Gao [17] proposed the “BiLSTM-DIC-ATT-CRF” entity recognition model, which combines dictionary features, an attention mechanism and BiLSTM and CRF algorithms to transform the secure entity recognition task in cyberspace into sequence annotation problems, and then labels the cybersecurity entities with semantic features, local context features and dictionary features in the annotated text.

2.2.2. NER in Cybersecurity

Relationship extraction to extract the semantic relationship between entities and the formal description as a relationship triplet form, namely, (entity 1, relationship, entity 2), stored in the knowledge base, utilizes many research-based methods, which overall can be divided into two categories: those based on a knowledge engineering method and those based on machine learning [22]. Since 1998, following the MUC-7 meeting, the entity relationship extraction task, involved relationship extraction mainly using artificial

build knowledge engineering [23], including annotation referring to information, write extraction rules, relationship templates, etc., but the method exhibits low efficiency and has poor portability. In addition, with an increased number of rules, there may also be conflict between different rules, creating later maintenance difficulties. Therefore, machine learning-based relationship extraction techniques have emerged. At present, according to the different machine learning algorithms used, they can be divided into three methods based on feature vectors, kernel functions and neural networks.

Su [24], because of the lack of a vertical annotation corpus and the traditional extraction method relationship overlap problem, proposed a threat intelligence entity relationship joint extraction model based on the BERT network. The method combines the BERT training language model and introduces rich semantic information, enabling entity recognition and relationship extraction using two sub-task parameter sharing. Sun [2] proposed the threat intelligence information extraction method (TIIE) based on a neural network model. On the one hand, the long-short memory neural network (LSTM) is used to perform the named entity recognition task, while the conditional random field (CRF) model is used to realize the constraint between sequence tags. On the other hand, the relationship extraction task is combined with the long-short memory neural network model and the shortest dependence path (SDP) method. Hyeonseong Jo et al. [25] proposed Vulcan (a CTI system) to identify various types of descriptive or static CTI data from unstructured text and to determine the semantic relationship between them. An NER model based on the language model was designed as a threat entity identifier (TEI) to learn the context information of the words, so as to identify the named entities related to ransomware attacks in unstructured text, and through the entity linker to understand the hidden meaning between entities, such as where a single entity may represent different things, or multiple entities with different names represent the same thing. Piyush Ghasiya et al. [26] proposed processing news text in the cybersecurity field by combining information extraction and emotion analysis. This method divides information extraction into two stages: first, keywords are selected using the tf-idf weighting method, and then are used to filter the context of the selected word through the word2vec model. Affective analysis is divided into supervised and unsupervised (vocabulary-based) methods. The former captures the context information of text data through the N-gram feature extraction algorithm, while the latter is suitable for situations where there are no labeled data. Gao [17] proposed the joint extraction method of entity relationship based on relationship decomposition. Firstly, it performs the multi-relationship classification task through the encoder, and then generates triples by identifying the entities under a specific relationship. By capturing the semantic interdependence between entity and relationship extraction steps, this method can effectively solve the triplet overlap problem and reduce the interference of unrelated entity pairs. Cheng [19] and others, based on semantic role annotation relationship extraction, put forward the RE-SRL algorithm. The traditional SRL algorithm is a kind of predicate-centered shallow semantic analysis technology, mainly considering “who to who”, “when and where” and “what” questions. The improved algorithm can be classified on the basis of the relationship between the entity, and effectively solves the field of sample scarcity text complex problem.

2.2.3. Joint Entity and Relation Extraction in Cybersecurity

The traditional assembly line (pipeline) extraction method involves entity relationship extraction for entity recognition and relationship classification of two independent sub-tasks, namely, on the basis of complete entity recognition and relationship classification. The method is easy to implement, but due to the lack of interaction between the two sub-tasks, cannot capture the internal connection and dependency between tasks, increasing the redundant information and computational complexity. In addition, since the follow-up tasks are completed on the basis of the presequence, the accuracy of the follow-up tasks largely depends on the completion of the presequence, so it will also bring problems such as error accumulation and error propagation. Therefore, it is a new research trend to establish a unified model to complete entity recognition and relationship classification

from unstructured or semi-structured text. At present, the methods of joint extraction of supervised entities mainly include joint extraction based on feature engineering and joint extraction based on neural networks [27].

Wang et al. [28] proposed the joint learning framework TIRECO, which is a model that can complete the extraction and resolution of threat intelligence relations simultaneously. It considers the co-reference relationship as a predefined relationship and completes the graph convolution network (GCN) model. For the problem that document-level text is too long to extract features, the concept of sentence set is proposed, which considers relationship extraction as a multi-classification task at the sentence level, and fully captures the syntax dependency information based on a shortest dependency path pruning strategy (SDP-VP-SET). Li et al. [29], in the implementation of joint extraction into the dynamic attention mechanism, and for the lack of a corpus, based on the idea of active learning, introduced the confrontation learning mechanism. They put forward a kind of active learning framework as a corpus sampling method, marking high quality data using incremental screening and iterative training to improve the performance of the joint extraction model. Liu et al. [30] proposed a joint learning threat intelligence entity recognition extraction method, seeking to address pipeline learning error transmission, information loss and redundancy issues. The method involves parameter sharing, entity recognition and relationship classification, combining two separate tasks into a sequence annotation problem, at the same time identifying the entity and relationship, to avoid the characteristics of complex engineering. Guo et al. [31] put forward a joint entity and relation extraction model with regard to cybersecurity, namely, CyberRel. It generates separate label sequences for different relationships. Joint extraction problem modeling is understood as a multiple sequence tag problem, with each label sequence containing the entities involved in the text and the subject and object information. On the one hand, it effectively solves the common entity overlap problem in the cybersecurity corpus. On the other hand, it also avoids the error propagation associated with the pipeline method. Zuo et al. [32] proposed an end-to-end joint extraction model suitable for the field of cybersecurity intelligence. It first models entity and relationship extraction as a serialized labeling task through a joint labeling strategy, and then constructs an end-to-end sequence labeling model based on BERT-ATT-BiLSTM-CRF. Finally, it involves a new entity and relationship matching rule for the relational overlap problem to extract knowledge triples.

2.3. Cybersecurity Ontology Construction

As per the CNNVD vulnerability content description specification [33] requirement, cybersecurity knowledge, such as product name, version and definition, vendor, vulnerability name, type, reason and consequence are contained in the description. Based on the above official document and the analysis of many vulnerability descriptions, we design an ontology to describe the cybersecurity entities and the relations between them, which will serve as a guide for cybersecurity knowledge triple extraction and annotation processing. This ontology consists of 13 entity types and 26 relationships. As shown in Figure 1, the solid ellipses and dashed segments with arrows are used to draw this ontology. The solid ellipse represents the entity type. The relations between entities are represented by the dashed segments with the relation's type near each corresponding segment. The arrow of each segment represents the direction of the corresponding relation. For example, the relation from "Product" to "Vulnerability" is "has_vul", which means the entity "Product" has a vulnerability named "Vulnerability".

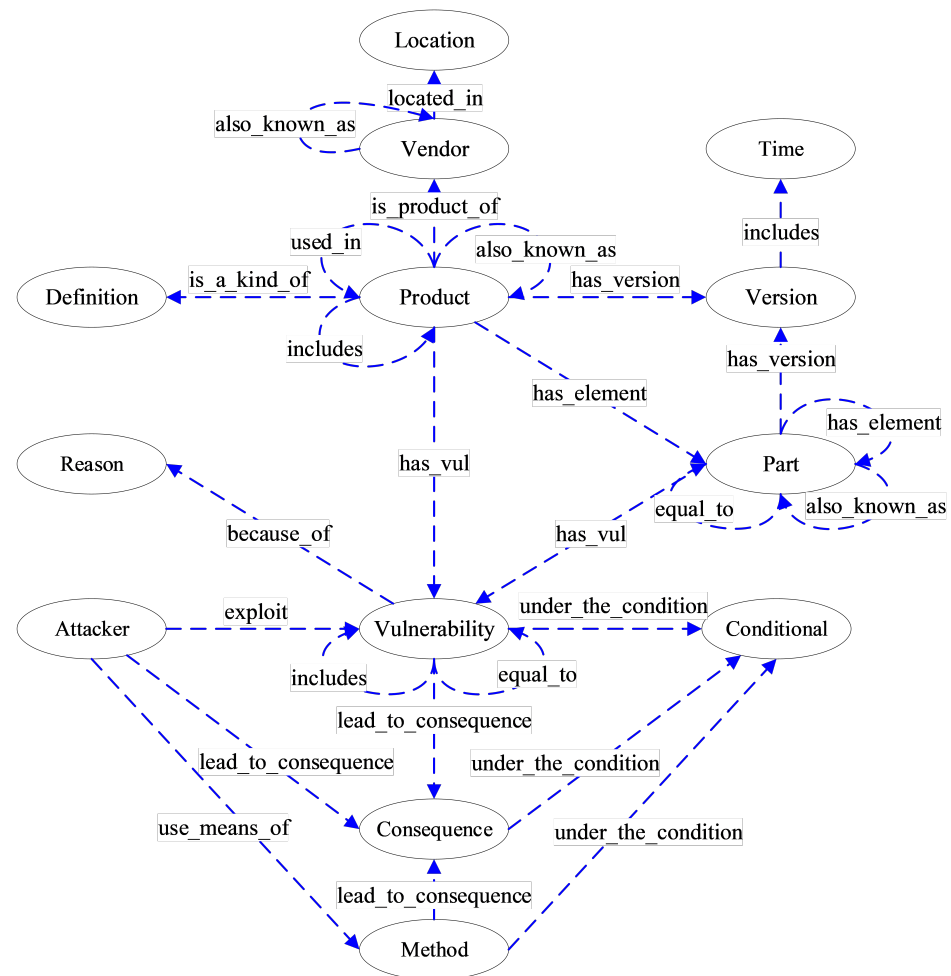


Figure 1. An ontology for vulnerability description in CNNVD.

3. Materials and Methods

To elucidate this dataset, this section presents an overview of the data collection, annotation, statistics and differences from existing datasets.

3.1. Data Collection Process

For supporting the task of cybersecurity triples extraction and cybersecurity knowledge graph construction, a cybersecurity domain dataset with entity and relation annotation is necessary. The corpus data can be sourced from a variety of open sources of heterogeneous threat intelligence data, such as vulnerability databases, cybersecurity news or blogs, cybersecurity industrial technical reports, and hacking forums. Among them, the data of the vulnerability database have a more standardized format, which will facilitate our data processing and annotation. A major contribution of our work is to annotate the first Chinese vulnerability dataset at the sentence level. We build this dataset by crawling 137,625 vulnerability data from the China National Vulnerability Database of Information Security (CNNVD). The data collected covered the period from 1 October 1988 to 3 February 2020. Each vulnerability data record in the CNNVD contains 10 items: vulnerability name, vulnerability type, product, vendor, threat level, CNNVD id, CVE id, recording time, vulnerability textual description, reference URLs, and official patches. Among them, items such as vendor, threat level and vulnerability type are often missing. With the triple extraction technology, these missing entities can be replenished based on the textual description. We acquire the textual description from each vulnerability data element and separate them into individual sentences. Finally, 27,311 non-repeated sentences are randomly selected from a total of 461,199 individual sentences.

3.2. Annotation Tool and Method

Figure 2 shows an annotation example for a vulnerability description sentence. The first line is the tag sequence with the “BIO” tagging method. The second line is a Chinese sentence example from “CNNVD-200212-237”. We connect each character in the Chinese sentence and their label with a dashed arrow. For instance, the first character of the Chinese sentence “C” is connected with its label “B-P” by a red dashed arrow, which means this “C” is the starting part of a “Product” entity. Similarly, the labels of the characters from the second one, “a”, to the 12th one, “n”, are all “I-P”, which means that they are the middle part of the “Product” entity. The 13th character, “is”, does not belong to any entity; hence, its tag is the abbreviation “O” for “Outside”. The English translation for the Chinese example is in the last line. The relations between entities are represented by a line with an arrow under the third line.

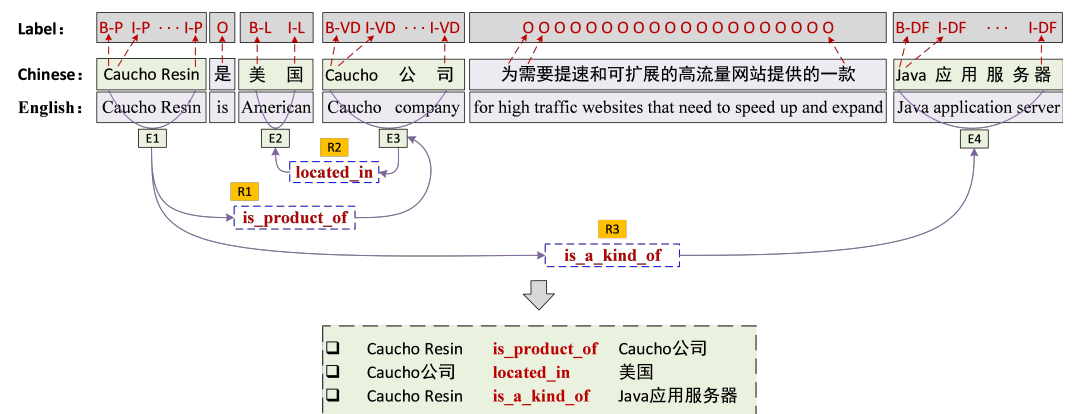


Figure 2. An annotation example for a vulnerability description sentence.

In the dataset, the average length of sentences is 51.39. The length of the longest sentence is 627. The shortest sentence has just 6 characters. Since the open-source automatic information extraction model can not fully identify most vulnerability entities and their relations in the cyber threat intelligent field, we use an open-source annotation tool named “Colabeler” (perhaps an obscure tool) to manually label the entities and relations in CNNVD vulnerability description sentences. Figure 3 shows a labeling example for a CNNVD vulnerability sentence using the mentioned tool “Colabeler” [<http://www.colabeler.com/>]. The entities’ words are labeled with different colors and connected with lines with arrows. The entities’ types are tagged above the entity. The relations’ names are labeled on the line.

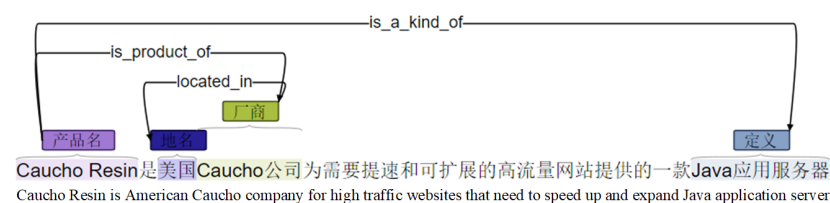


Figure 3. Annotation example using “Colabeler”.

As the input of the target knowledge extraction model, we need to transfer the output of the annotation tool into a standard CoNLL format. Figure 4 shows two kinds of formats of annotation data. Figure 4a represents the data exported from the annotation tool in json format. This part contains entities annotation, relations annotation and target vulnerability textual content. Figure 4b is the labeled data with BIO format transferred from the output of the annotation tool. This part consists of five columns as follows: index, characters in vulnerability sentence, entity tags with BIO format, the relation tags of the corresponding head entity, and the index of tail entities.

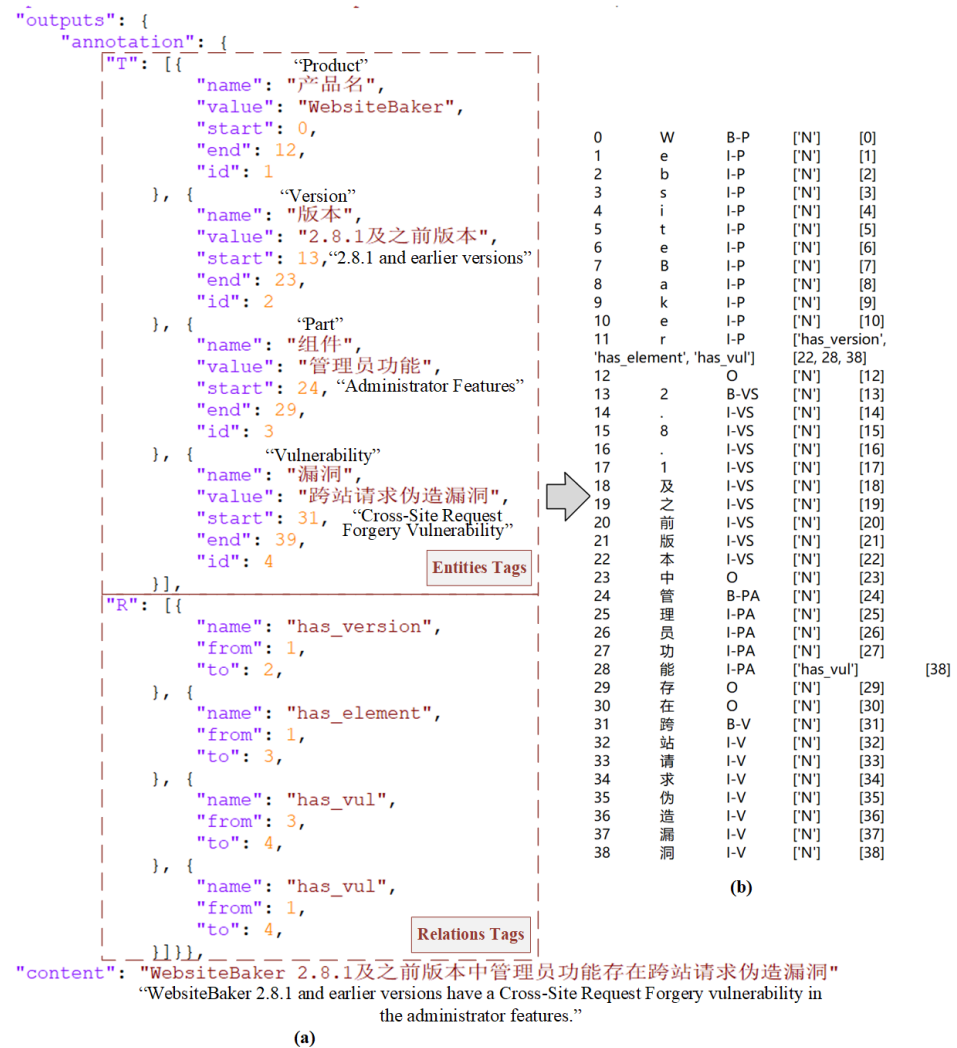


Figure 4. Data transfer for annotation. (a) Json format annotation for Chinese vulnerability sentence. (b) CoNLL format annotation for Chinese vulnerability sentence.

3.3. Statistics of Entities and Relations

As mentioned in the proposed ontology, we use 13 kinds of entity types and 15 kinds of relationships to describe the vulnerability knowledge in CNNVD. Table 1 shows the details of each kind of entity type. The 13 entity types in this dataset are Product, Definition, Vendor, Location, Vulnerability, Version, Part, Consequence, Attacker, Method Reason, Condition and Time. The columns of Table 1 from the left to the right correspond to the following: the Chinese entity type, the English entity type and entity meaning, the abbreviation for the English entity type, the Chinese example for the corresponding entity, the English example for the corresponding entity. We use industrial standard BIO text labeling mode to label the CNNVD vulnerability text. We combine the “B-” with the Chinese entity type or with the abbreviation for the English entity type as the entity start tag, such as “B-Product” or “B-P”. Similarly, the other inside entity tags are combined with “I-” and the Chinese entity type or with the abbreviation for the English entity type, such as “I-Product” or “I-P”.

In this dataset, we count and rank the frequency of each entity type and show them in Figure 5. In our dataset, there are a total of 97,391 entities, with an average of 3.5 entities in every sentence. As shown in Figure 5, the three most numerous entity types in our dataset are “Product”, “Definition” and “Vendor” (32,328, 19,887 and 13,279, respectively), indicating that, on average, each sentence contains 1.2 “Product” entities. “Reason”, “Condition” and “Time” are the minimum three entity types, and the frequency of these

is 318, 156 and 82, respectively. It is worth mentioning that there is a data imbalance where the number of the last six kinds of entities is less than 1200. As a result, some data augmentation technologies need to be applied to alleviate this data imbalance problem during the model training stage.

Table 1. The predefinition entity types and examples.

Entity Type	Abbr.	Examples
Product	P	MoinMoin, Netware
Definition	D	Email Service Program
Vendor	VD	Cisco, Huawei
Location	L	China, Germany
Vulnerability	V	SQL Injection Vulnerability
Version	VS	Versions prior to 1.2.9
Part	PA	index.php, uum program
Consequence	C	Bypass authentication
Attacker	A	Remote attacker
Method	M	Through unknown vectors
Reason	R	Improper memory management
Condition	CD	When handling URL requests
Time	T	10 January 2007
Other	O	Other words

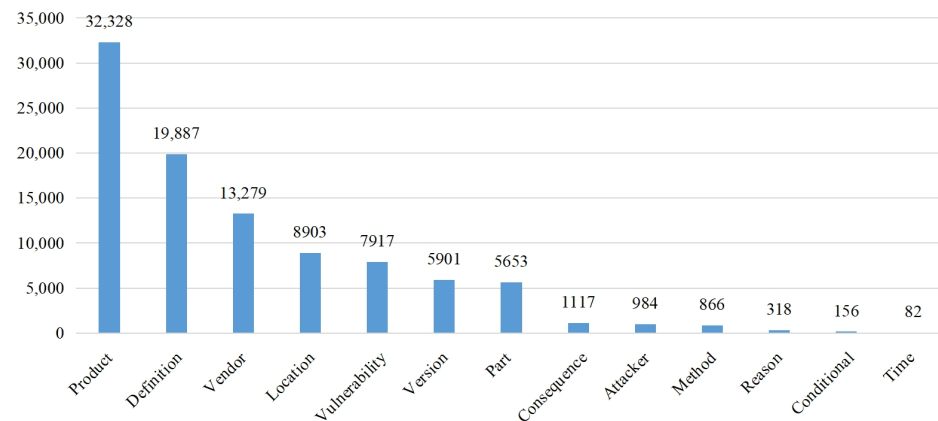


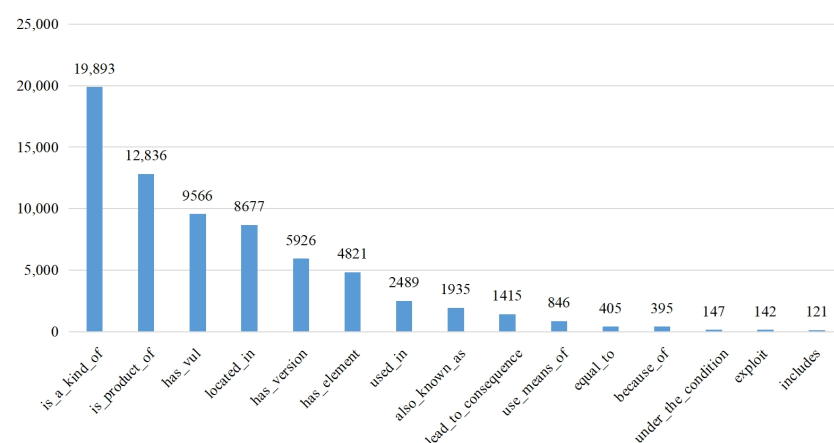
Figure 5. The frequency of each entity type.

In this section, 15 kinds of entity relationships will be introduced to describe the organization of the vulnerability entities. Although the ontology in Figure 1 shows 26 entity relationships, there are 15 non-repetitive entity relationship types since the head entity and tail entity of the same relationship can be different. They are chosen because these 15 relationships are those mostly described in the CNNVD vulnerability textual description, and other relationships are rarely involved. In order to facilitate the understanding of the relations, Table 2 presents the details of each relationship.

Figure 6 is a frequency bar chart of the relationships in the dataset based on frequency ordering. We counted 69,614 relationships from 27,311 non-repeated sentences in our dataset, which means there are 2.55 triples in one sentence on average. As shown in this figure, the number of each of the last six relationships is less than 1000, which may become an important factor affecting the training and extraction effects of these relations. Nonetheless, these relationships are still non-negligible components in describing the relationship between the vulnerability-related entities. As mentioned in the entity annotation work, data augmentation techniques are also needed in the training phase for these types of relationships.

Table 2. Details of the predefinition relationship types and their description.

Relation Types	1st Entity	2nd Entity	Description
is_a_kind_of	Product	Definition	Information such as software, solutions, systems, tools or function.
is_product_of	Product	Vendor	One product usually provided by a person, company or organization.
has_vul	Product	Vulnerability	Vulnerabilities are usually found in a product or components that make it up.
located_in	Vendor	Location	The address information of vendor.
has_version	Product	Version	Vulnerabilities exist in specific versions of software or components.
has_element	Product	Part	A product or component could consist of multiple parts.
used_in	Product	Product	A product may be used in another product such as a bigger system.
also_known_as	Part	Part	The same product, part, and vendor may be described with different names.
lead_to_consequence	Attacker	Consequence	Attackers, vulnerabilities and exploitation methods can have consequences for one product.
use_means_of	Attacker	Method	A perpetrator usually uses some means of attack against the target product.
equal_to	Vulnerability	Vulnerability	A vulnerability or part in a sentence may be referred to by multiple pronouns.
because_of	Vulnerability	Reason	A vulnerability exists for reasons.
under_the_condition	Method	Condition	The existence of vulnerabilities, the use of attack methods and the consequences are usually based on specific conditions.
exploit	Attacker	Vulnerability	Attackers usually exploit vulnerabilities for attack purposes.
includes	Product	Vulnerability	Some products are sets with several sub-products. One product may have a complex vulnerability that contains multiple vulnerabilities.

**Figure 6.** Frequency of each relation type.

3.4. Comparison with Existing Datasets

Unlike previous information extraction datasets, the dataset we constructed is a mixed Chinese and English dataset, which means that almost every sentence contains both Chinese and English. Among all the 27,311 non-repeated sentences, there are 27,279 mixed Chinese and English sentences, accounting for 99.89%. It is because of this sentence feature that

the target entities we want to extract are also likely to be composed of mixed Chinese and English entities. In the dataset we constructed, there are 32,721 full Chinese entities, 35,334 English entities, and 29,336 mixed Chinese and English entities. Among them, the number of mixed Chinese and English entities accounts for 30.12% of all 97,391 entities.

4. Experiments and Evaluation—Results

In this part, we use and validate multiple baselines on the BVTED dataset. For the named entity recognition task, we select five baselines, including a traditional deep learning model and a pre-training model, and, in particular, use a multi-language pre-training model to reduce the out-of-vocabulary (OOV) problems due to the mixture of Chinese and English in the dataset. As baselines, the BiLSTM-based algorithm model and the CNN-based algorithm model are applied to the relationship classification task. In addition to the pipeline's method, we also select two state-of-the-art baselines to test the effect of the joint extraction model of entity relations on the BVTED dataset.

In order to make the distribution of each entity in the training set, validation set and test set roughly the same, we separately divide the entity types with a small annotated number in the dataset. From Figure 5, the samples of "Reason", "Condition" and "Time" are small, with 318,156 and 82 frequencies, respectively. Therefore, we separately divide each of the three entity types into the training set, validation set and test set by an 8:1:1 random classification. Finally, these three entity types are guaranteed in all three parts of the dataset, and the data distribution of the three entity types is roughly equivalent. Similar to the processing of the entity types, we separately divide nine relationship types with less than 3000 frequencies ('used_in', 'also_know_as', 'lead_to_consequence', 'use_means_of', 'equal_to', 'because_of', 'under_the_condition', 'exploit', 'includes'), dividing the remaining data at a unified ratio of 8:1:1. On the one hand, this division ensures that each part will contain each relationship type, and at the same time, each relationship type has the same distribution in each part of the dataset.

4.1. Evaluation Metrics

The three commonly used measures in the named entity recognition and relationship extraction tasks are precision, recall and the F1-score. These metrics were chosen for their effectiveness in evaluating the performance of the NER and RE models, particularly in imbalanced datasets commonly encountered in cybersecurity applications.

Precision is the ratio of correctly predicted positive categories to all samples predicted to be positive. It helps minimize false positives. Recall is the ratio of the correctly predicted positive category to all the positive samples in the original sample. It helps minimize false positives. From the single performance indices of precision and recall, the harmonic average of the two indexes is calculated to obtain the F1-score. This is particularly useful for imbalanced datasets. The formulas are as follows:

$$Precision = \frac{TP}{TP + FP}, \quad (1)$$

$$Recall = \frac{TP}{TP + FN}, \quad (2)$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN}, \quad (3)$$

TP represents the number of positive cases that are correctly predicted; TN represents the number of negative cases that are correctly predicted; FP represents the number of negative cases that are incorrectly predicted; and FN represents the number of positive cases that are incorrectly predicted.

4.2. Experimental Settings

The software and software configurations used in the experiments in this paper are shown in Table 3. All the experiments in this paper are expanded on the BVTED dataset, and the parameter settings of the model in the experiments are shown in Tables 4 and 5. To meet the memory requirements, the gradient accumulation algorithm is set in the two models based on BERT to reduce the number of model operations and calculate a loss value for every four steps. The training process uses the Adam optimization algorithm [34], which adjusts the learning rate and momentum parameters to optimize the weight of the model. With a low memory footprint, it can converge faster and adapt to different parameter characteristics.

Table 3. Experimental environment configuration.

Hardware	Configuration	Software	Configuration
Graphics Card	NVIDIA Tesla V100	Algorithm	Based on Ubuntu system
RAM	16 GB	Program Language	Python 3.7
Processor	Intel(R) Xeon(R) Platinum 8163 CPU @ 2.50 GHz	Deep Learning Framework	Pytorch 1.8.0

Table 4. BiLSTM + CRF model parameter settings.

Parameter Type	Explanation	Value
Max_length	The maximum truncation length of the text	512
Batch_size	The size of each GPU/CPU training batch	32
Learning_rate	The initial learning rate of the algorithm	0.001

Table 5. BERT + CRF, BERT + BiLSTM + CRF and ERNIE + BiLSTM + CRF model parameter settings.

Parameter Type	Explanation	Value
Max_length	The maximum truncation length of the text	512
Batch_size	The size of each GPU/CPU training batch	16
Learning_rate	Iterative updating of neural network weights based on training data	3×10^{-5}
Optimization Algorithm	The optimization algorithm used for training	Adam

4.3. Experimental Result

In this section, we introduce the experimental results of applying existing NER models, RE models and joint entity relation extraction models on the BVTED dataset to demonstrate the effectiveness of the newly established dataset.

4.3.1. Baseline Models of Cybersecurity Named Entity Recognition

In this section, as shown in Table 6, we use baselines of five named entity recognition tasks for training and testing on the BVTED dataset and use F1 values obtained on the test set to evaluate the performance of individual models. In the first model (BERT + CRF) [35], we add a softmax layer and a CRF layer as a decoding module after encoding the BERT output sequence. To assess the impact of various embedding methods on the BiLSTM + CRF model [36], we individually employ three distinct representation models (Word2Vec, BERT and ERNIE) in the embedding layer to capture the sequence information. The output of the embedding layer undergoes encoding by BiLSTM, extracting contextual semantics. Similar to the first baseline model, both softmax and CRF layers are incorporated for decoding the encoded information. The amalgamation of these three representation models with the BiLSTM-CRF encoder–decoder model constitutes the other baselines.

From Table 6, we can see that the overall F1 value of the five algorithm models has reached more than 91% when the entity extraction completes. Comparing the F1 value of the BERT + CRF model and the BERT + BiLSTM + CRF model, the F1 value of the latter is generally smaller than the F1 value of the former, probably because the BiLSTM layer increases the complexity of the model and introduces the redundancy of information. In the same upper code model, the lower word embedding mode plays an important role in the information representation in the text. The model based on the Word2Vec embedding has a higher F1 value than the Chinese BERT model. Therefore, the static Word2Vec embedding based on the text training in the field of CTI will be more suitable for the specific words and terms in this field. The pre-trained knowledge-based knowledge enhancement mechanism language model ERNIE [37] and Multilingual BERT [38] both achieve better F1 values than using the Chinese BERT model as the vector representation layer. This shows that the knowledge-enhanced pre-training model can improve the ability to represent professional words and terms in threat intelligence texts. The Multilingual BERT-based model achieves the best F1 value for most entity types and the overall extraction effect (96.67%). This shows that the Multilingual BERT model can more effectively represent the text with a mixture of Chinese and English.

For entity types such as “definition”, “place name” and “attacker” that can be expressed in a limited dictionary, or can be extracted based on rules, the extraction effect of the F1 value on the five training models used reaches more than 99%. This shows that these entity types are easier to extract with a relatively single expression. The small number of annotated samples limited the performance of the Chinese BERT models; for example, the “Reason”, “Condition” and “Time” samples are less than 350, so the Chinese BERT models obtain relatively small F1 values. Especially for the “Time” entity type with the number of annotated data less than 100, the model extraction effect is poor because the parameters are not effectively adjusted, all falling below 62%.

Table 6. Comparison of F1 values of our cybersecurity entity extraction model and other baseline methods, i.e., CRF and BiLSTM-CRF, with multiple embedding models based on BVTED.

Entity Types	CRF		BiLSTM + CRF		
	BERT	Word2Vec	BERT	ERNIE	Multilingual BERT
Product	0.9154	0.9504	0.8997	0.9497	0.9537
Definition	0.9906	0.9911	0.9911	0.9920	0.9962
Vendor	0.8867	0.9283	0.8690	0.9323	0.9410
Location	0.9978	0.9961	0.9978	0.9934	0.9978
Vulnerability	0.9639	0.9756	0.9546	0.9842	0.9874
Version	0.8909	0.9286	0.8660	0.9430	0.9462
Part	0.8740	0.9185	0.8309	0.9155	0.9196
Consequence	0.9127	0.9577	0.8550	0.9922	0.9844
Attacker	0.9901	0.9975	0.9901	0.9999	0.9988
Method	0.8494	0.9280	0.7732	0.9777	0.9779
Reason	0.8619	0.9402	0.7368	0.9699	0.9787
Condition	0.6753	0.9483	0.7153	0.9464	0.9913
Time	0.6190	0.5641	0.3673	0.7692	0.8000
Whole F1	0.9290	0.9575	0.9113	0.9628	0.9667

Bold values indicate the best performance value for each entity type.

4.3.2. Baseline Models of Cybersecurity Relation Extraction

To train and test the BVTED, the experiments use two groups of common relation extraction models. The first group are classification-based relation extraction approaches, which include classification methods based on BiLSTM and classification methods based on CNN. In the second group, CasRel and OneRel are the joint relation extraction state-of-the-art baselines.

ERNIE + BiLSTM: Zhang et al. [39] proposed bidirectional long short-term memory networks (BiLSTM) to model the sentence and two entities with complete, sequential information about all words and classified relations using a softmax classifier. We utilize the knowledge-enhanced pre-training model ERNIE as an embedding layer to provide rich semantic context for the following feature extraction step. These two modules are combined to form the first baseline model ERNIE + BiLSTM.

PCNN: Zeng et al. [40] proposed a novel model dubbed the Piecewise Convolutional Neural Network (PCNN), with multi-instance learning for a distant supervised relation extraction task. This baseline model solved the wrong label problem and error propagation or accumulation problem, avoiding feature engineering and capturing the structural information between two entities using a CNN architecture with a piecewise max pooling method.

CasRel: Wei et al. [41] proposed a novel cascade binary tagging framework (CasRel) that models relationships as mapping subjects and objects, instead of classifying discrete relation labels of entity pairs. This architecture addresses the overlapping triple problem using three components. First, a BERT encoder module is used to extract the feature information of the input sentence. Second, a subject tagger of a cascade decoder is designed to detect all possible subjects in the input sentence. Third, a relation-specific object tagger of a cascade encoder, a high level tagging module, is used to identify objects and the involved relations with respect to the subjects obtained from a lower level.

OneRel: Shang et al. [42] proposed a novel model named OneRel, designed for joint entity and relation extraction. This model frames the joint extraction task as a fine-grained triple classification problem. Specifically, OneRel incorporates a classifier to assess whether a token pair and a relation form part of a factual triple. Additionally, it features a relation-specific horn tagging strategy, ensuring a straightforward yet effective decoding process.

Table 7 shows the experimental results of four baselines on the dataset BVTED. The experiments use precision, recall and the F1-score.

to evaluate the performance of each model on the dataset BVTED. From the experimental results, the three evaluation indexes of the PCNN model are better than the other models. The BiLSTM relationship extraction model based on the ERNIE pre-trained model performs the worst. Therefore, the most important thing in the relationship extraction task for this dataset is to capture the structural information between two entities, rather than the context semantics of the upper distance. The experimental results show that CasRel and ERNIE + BiLSTM have advantages in handling complex relationships and long-range dependencies, but fail to fully exploit their strengths in the datasets and tasks of this experiment. The reasons may include the following two aspects: (1) The text of the dataset describing threat intelligence information is roughly similar and has a clear sharding structure; (2) The relationships in this dataset are mainly dependent on local context rather than long-distance dependence. The reason why the OneRel model performs less well than PCNN may be that its integrated architecture is less efficient than directly extracting local features in this specific scenario.

Table 7. Main results for the relational judgement model and other baseline methods, i.e., ERNIE-BiLSTM, PCNN, CasRel, and OneRel based on BVTED.

Model	Precision	Recall	F1-Score
ERNIE_BiLSTM [39]	56.95%	72.92%	63.58%
PCNN [40]	95.35%	95.12%	95.19%
CasRel [41]	77.19%	65.73%	71.00%
OneRel [42]	93.08%	90.73%	91.89%

Bold values indicate the best performance value of each metric.

4.4. Limitations of Proposed Methodology

Despite achieving promising results on real-world datasets, the proposed pre-trained models for NER and RE exhibit several limitations in enhancing the efficiency and gen-

eralization performance of cybersecurity knowledge triplets. From a dataset perspective, the BVTED dataset is solely sourced from CNNVD vulnerability descriptions, which constrains its generalizability. A single data source may not capture the diverse spectrum of cybersecurity threats and vulnerability narratives, potentially diminishing the model's performance on unseen or stylistically different data. Additionally, the dataset's annotation process, which combines rule-based and manual methods, incurs high labor and time costs, limiting the scalability and efficiency of dataset construction. Furthermore, the imbalance in data distribution, where certain entities or relations, such as "Product" entities and "is_a_kind_of" relations, are overrepresented compared to others like "Condition" entities and "exploit" relations, can adversely affect the performance of knowledge triplet extraction models.

From the perspective of the NER and RE models, there is a significant dependency on large amounts of training data. As demonstrated in Table 6, the performance of these models declines markedly when the training data are insufficient, indicating a challenge in maintaining high performance under data-constrained conditions. Additionally, the models may not achieve the desired effectiveness when applied to security datasets with different linguistic structures and styles. This suggests a limitation in the cross-linguistic and cross-domain generalization capabilities of the trained models, necessitating further fine-tuning and adaptation to suit various languages and data formats. These limitations highlight the need for future research to incorporate more diverse data sources, develop more efficient annotation methods, address data imbalance, and enhance model adaptability to ensure broader applicability and robustness of the proposed methodology.

5. Conclusions

In the realm of cyber threat intelligence, cybersecurity vulnerability intelligence plays a crucial role. We developed an ontology for vulnerability knowledge, encompassing 13 entity types and 15 unique relation types, which captures the common knowledge found in vulnerability description texts. This ontology serves as a foundational framework for extracting and structuring information, enabling more effective analysis and utilization of cybersecurity vulnerability data. Leveraging this ontology, we construct a comprehensive dataset for the cybersecurity domain, named BVTED, which supports the task of entity–relation triple extraction. The dataset comprises 27,311 unique vulnerability description sentences with mixed Chinese and English expressions. It contains 97,391 entities and 69,614 relations, including 32,721 Chinese entities, 35,334 English entities and 29,336 entities with mixed Chinese and English.

To evaluate the dataset's effectiveness, we conduct performance tests using state-of-the-art baselines for named entity recognition and relation extraction tasks. The experimental results confirm the validity and utility of the BVTED dataset. Specifically, our cybersecurity entity extraction model with BiLSTM+CRF and Multilingual BERT achieved an overall F1-score of 0.9667. For relational judgment, the PCNN model attained a precision of 95.35%, a recall of 95.12% and an F1-score of 95.19%. These results substantiate the feasibility and efficacy of utilizing the BVTED dataset for cybersecurity applications.

This research contributes significantly to the field of CTI analysis by demonstrating the potential of advanced NER and RE techniques in enhancing the accuracy and efficiency of knowledge extraction processes. Despite these achievements, our methodology faces certain limitations. The BVTED dataset's reliance on CNNVD data limits its generalizability, and the high labor and time costs associated with our combined rule-based and manual annotation methods restrict the scalability of our dataset construction. Additionally, the imbalance in data distribution poses challenges for knowledge triplet extraction models. Addressing these issues is crucial for advancing the practical applications of our proposed approach.

6. Future Work

To overcome these limitations and further advance the field, we plan to undertake several key initiatives in our future work: (1) Diversification of Data Sources: We will expand the BVTED dataset to include data from a broader range of CTI sources. This diversification will help ensure that our models can generalize across different types of cybersecurity threats and descriptions, enhancing their robustness and applicability in real-world scenarios. (2) Automated Annotation Methods: To improve the efficiency and scalability of our dataset annotation process, we aim to incorporate more automated annotation methods. Leveraging techniques such as prompt-based learning with large models and semi-supervised learning, we hope to reduce the dependency on manual labor while maintaining high annotation accuracy. (3) Addressing Data Imbalance: We will implement strategies to balance the distribution of entity and relation types within the dataset. Techniques such as data augmentation and synthetic data generation will be explored to ensure that our models perform well across all categories of data.

By addressing these future directions, we aim to significantly enhance the practical impact and generalizability of our work, contributing to more effective mining, analysis and utilization of cybersecurity threat intelligence knowledge.

Author Contributions: Methodology, K.L.; validation, K.L.; investigation, Y.W.; writing—original draft, K.L. and Y.W.; writing—review and editing, K.L. and Y.W.; supervision, Z.D., A.L. and W.Z.; project administration, Z.D., A.L. and W.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: All data included in this study are available upon request by contact with the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Definition: Threat Intelligence. Available online: <https://www.gartner.com/en/documents/2487216> (accessed on 21 June 2020).
2. Sun T. Research on Threat Intelligence Knowledge Extraction Based on Deep Learning. Master's Thesis, Sichuan University, Chengdu, China, 2021.
3. Auer M. Lack of Experts in Cyber Security. 2020. Available online: <https://www.threatq.com/lack-of-experts-in-cyber-security/> (accessed on 3 April 2022).
4. Xie, B. Research on Chinese Network Threat Intelligence Knowledge Extraction Technology Based on Deep Learning. Master's Thesis, Guizhou University, Guizhou, China, 2022.
5. Zhou, Y.; Ren, Y.; Yi, M.; Xiao, Y.; Tan, Z.; Moustafa, N.; Tian, Z. Cdtier: A Chinese dataset of threat intelligence entity relationships. *IEEE Trans. Sustain. Comput.* **2023**, *8*, 627–638. [\[CrossRef\]](#)
6. Liu, K.; Wang, F.; Ding, Z.; Liang, S.; Yu, Z.; Zhou, Y. Recent progress of using knowledge graph for cybersecurity. *Electronics* **2022**, *11*, 2287. [\[CrossRef\]](#)
7. Wei, X.; Qin, Y.; Chen, Y. A Network Security Named Entity Recognition Method Based on Component CNN. *Comput. Digit. Eng.* **2020**, *48*, 106–111.
8. Yan, C.; Su, Q.; Wang, J. Mogcn: Mixture of gated convolutional neural network for named entity recognition of chinese historical texts. *IEEE Access* **2020**, *8*, 181629–181639. [\[CrossRef\]](#)
9. Ding, N.; Xu, G.; Chen, Y.; Wang, X.; Han, X.; Xie, P.; Zheng, H.T.; Liu, Z. Few-NERD: A Few-shot Named Entity Recognition Dataset. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Online, 1–6 August 2021; pp. 3198–3213.
10. Xu, L.; Dong, Q.; Liao, Y.; Yu, C.; Tian, Y.; Liu, W.; Li, L.; Liu, C.; Zhang, X. CLUENER2020: Fine-grained named entity recognition dataset and benchmark for Chinese. *arXiv* **2020**, arXiv:2001.04351.
11. Zhang, Y.; Zhong, V.; Chen, D.; Angeli, G.; Manning, C.D. Position-aware Attention and Supervised Data Improve Slot Filling. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; pp. 35–45.
12. Christou, D.; Tsoumakas, G. Improving distantly-supervised relation extraction through bert-based label and instance embeddings. *IEEE Access* **2021**, *9*, 62574–62582. [\[CrossRef\]](#)
13. Bridges, R.A.; Hufer, K.M.T.; Jones, C.L.; Iannaccone, M.D.; Goodall, J.R. Cybersecurity automated information extraction techniques: Drawbacks of current methods, and enhanced extractors. In Proceedings of the 16th IEEE International Conference on Machine Learning and Applications, ICMLA 2017, Cancun, Mexico, 18–21 December 2017; pp. 437–442.

14. Lim, S.K.; Muis, A.O.; Lu, W.; Hui, O.C. MalwareTextDB: A database for annotated malware articles. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, BC, Canada, 30 July–4 August 2017; Volume 1: Long Papers, pp. 1557–1567.
15. Wang, X.; He, S.; Xiong, Z.; Wei, X.; Jiang, Z.; Chen, S.; Jiang, J. Aptner: A specific dataset for ner missions in cyber threat intelligence field. In Proceedings of the 2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD), Hangzhou, China, 4–6 May 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1233–1238.
16. Rastogi, N.; Dutta, S.; Christian, R.; Gridley, J.; Zaki, M.; Gittens, A.; Aggarwal, C. Predicting malware threat intelligence using KGs. *arXiv* **2021**, arXiv:2102.05571.
17. Gao, C. Research on Entity Relationship Extraction of Cyberspace Security Knowledge. Master's Thesis, Yunnan University, Kunming, China, 2023.
18. Li, J.; Sun, A.; Han, J.; Li, C. A survey on deep learning for named entity recognition. *IEEE Trans. Knowl. Data Eng.* **2020**, *34*, 50–70. [\[CrossRef\]](#)
19. Cheng, S.; Li, Z.; Wei, T. A Threat Intelligence Entity Relationship Extraction Method Combining Bootstrap and Semantic Role Annotation. *Comput. Appl.* **2023**, *43*, 1445.
20. Mumtaz, S.; Rodriguez, C.; Benatallah, B.; Al-Banna, M.; Zamanirad, S. Learning word representation for the cyber security vulnerability domain. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–8.
21. Li, Y.; Guo, Y.; Fang, C.; Liu, Y.; Chen, Q. A Novel Threat Intelligence Information Extraction System Combining Multiple Models. *Secur. Commun. Netw.* **2022**, *2022*, 8477260. [\[CrossRef\]](#)
22. Huang, X.; You, H.; Yu, Y. A review of research on relationship extraction techniques. *Mod. Libr. Inf. Technol.* **2013**, *29*, 30–39. [\[CrossRef\]](#)
23. Li, D.M.; Zhang, Y.; Li, D.Y. A review of research on entity relationship extraction methods. *Comput. Res. Dev.* **2020**, *57*, 1424–1448.
24. Su, C. Research on Network Threat Intelligence Extraction Technology for Unstructured Text. Master's Thesis, University of Chinese Academy of Sciences, Beijing, China, 2020.
25. Jo, H.; Lee, Y.; Shin, S. Vulcan: Automatic extraction and analysis of cyber threat intelligence from unstructured text. *Comput. Secur.* **2022**, *120*, 102763. [\[CrossRef\]](#)
26. Ghasiya, P.; Okamura, K. A hybrid approach to analyze cybersecurity news articles by utilizing information extraction & sentiment analysis methods. *Int. J. Semant. Comput.* **2022**, *16*, 135–160.
27. Zhang, S.; Wang, X.; Chen, Z.; Wang, L.; Xu, D.; Jia, Y. A review of research on supervised entity relationship joint extraction methods. *Comput. Sci. Explor.* **2022**, *16*, 713.
28. Wang, X.; Xiong, M.; Luo, Y.; Li, N.; Jiang, Z.; Xiong, Z. Joint learning for document-level threat intelligence relation extraction and coreference resolution based on gcn. In Proceedings of the 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), Guangzhou, China, 29 December 2020–1 January 2021; IEEE: Piscataway, NJ, USA, 2020; pp. 584–591.
29. Li, T.; Guo, Y.; Ju, A. Triple Extraction of Network Security Knowledge Based on Fusion Adversarial Active Learning. *J. Commun./Tongxin Xuebao* **2020**, *41*, 80–91.
30. Liu, Q.; Zhu, P. A method for constructing an end-to-end threat intelligence knowledge graph based on federated learning. *Mod. Comput.* **2021**, *16*, 1–6.
31. Guo, Y.; Liu, Z.; Huang, C.; Liu, J.; Jing, W.; Wang, Z.; Wang, Y. CyberRel: Joint entity and relation extraction for cybersecurity concepts. In Proceedings of the Information and Communications Security: 23rd International Conference, ICICS 2021, Chongqing, China, 19–21 November 2021; Proceedings, Part I 23; Springer International Publishing: Cham, Switzerland, 2021; pp. 447–463.
32. Zuo, J.; Gao, Y.; Li, X.; Yuan, J. An End-to-end Entity and Relation Joint Extraction Model for Cyber Threat Intelligence. In Proceedings of the 2022 7th International Conference on Big Data Analytics (ICBDA), Guangzhou, China, 4–6 March 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 204–209.
33. CNNVD Vulnerability Content Description Specification. Available online: <http://123.124.177.30/web/wz/bzxqById.tag?id=5&mkid=5> (accessed on 1 March 2020).
34. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
35. Hu, S.; Zhang, H.; Hu, X.; Du, J. Chinese Named Entity Recognition based on BERT-CRF Model. In Proceedings of the 2022 IEEE/ACIS 22nd International Conference on Computer and Information Science (ICIS), Zhuhai, China, 26–28 June 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 105–108.
36. Wang, X.; Liu, R.; Yang, J.; Chen, R.; Ling, Z.; Yang, P.; Zhang, K. Cyber threat intelligence entity extraction based on deep learning and field knowledge engineering. In Proceedings of the 2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD), Hangzhou, China, 4–6 May 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 406–413.
37. Zhang, Z.; Han, X.; Liu, Z.; Jiang, X.; Sun, M.; Liu, Q. ERNIE: Enhanced Language Representation with Informative Entities. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 1441–1451.
38. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.

39. Zhang, S.; Zheng, D.; Hu, X.; Yang, M. Bidirectional long short-term memory networks for relation classification. In Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation, Shanghai, China, 30 October–1 November 2015; pp. 73–78.
40. Zeng, D.; Liu, K.; Chen, Y.; Zhao, J. Distant supervision for relation extraction via piecewise convolutional neural networks. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 1753–1762.
41. Wei, Z.; Su, J.; Wang, Y.; Tian, Y.; Chang, Y. A novel cascade binary tagging framework for relational triple extraction. *arXiv* **2019**, arXiv:1909.03227.
42. Shang, Y.M.; Huang, H.; Mao, X. Onerel: Joint entity and relation extraction with one module in one step. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 22 February–1 March 2022; Volume 36, pp. 11285–11293.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.