Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

A Review on Cyber Security Named Entity Recognition

Chen GAO[1], Xuan ZHANG[†‡1, 2, 3], Meng-ting HAN[1], Hui LIU[1]

[1]*School of software, Yunnan University, Kunming 650091, China*

[2]*Key Laboratory of Software Engineering of Yunnan Province, Kunming 650091, China*

[3]*Engineering research center of cyberspace, Kunming 650091, China*

[†]E-mail: zhxuan@ynu.edu.cn

Abstract: With the rapid development of Internet technology and the advent of the era of big data, more and more cyber security texts are provided on the Internet. These texts include not only security concepts, incidents, tools, guidelines, and policies, but also risk management approaches, best practices, assurances, technologies, and more. Through the integration of large-scale, heterogeneous, unstructured cyber security information, and the identification and classification of cyber security entities can effectively help to solve cyber security issues. Due to the complexity and diversity of texts in the cyber security domain, it is difficult to identify security entities in the cyber security domain using the traditional named-entity recognition methods. This paper describes various approaches and techniques for named-entity recognition in the cyber security domain and discusses the problems faced by named-entity recognition research in this domain. Finally, some suggestions for the future direction of named-entity recognition in cyber security are proposed.

Key words: Named-entity recognition; Cyber security; Machine learning; Deep learning

1 Introduction

As a basic component of information extraction tasks, named-entity recognition (NER) (Marrero et al., 2013) plays a very important role in natural language processing tasks, such as knowledge maps, machine translation, automatic text summarization, etc. The NER task is composed of two parts: identifying the entity type and detecting the entity boundary. Entity boundary detection refers to the determination of the scope of an entity. An entity is not usually composed of a word. The complete identification of multiple words constituting the entity needs to determine the boundary of the entity. Entities fall into different categories according to their attributes. In the process of naming entities, we tend to pre-define the categories of entities and give the labels of the corresponding categories to the entities in the text to be recognized.

A named entity refers to a set of concepts that have the same attributes. Specifically, named entities refer to different concepts, depending on the domain. Generically named entities are usually people, places, and organizations. In cyber security, named entities often contain vulnerability names, software names, and security-related terms. These entities are usually composed of numbers, letters, and other characters, and have different lengths, which are non-standard terms. For example, “Anti-Nuclear Worm,” as an entity in the cyber security domain, consists of uppercase and lowercase letters and a dash. The cyber security term “Distributed Denial of Service attack” contains five words. “EternalBlue” and “HeartBleed”

‡ Corresponding author

* Project supported by the National Natural Science Foundation of China under Grant No. 61862063, 61502413, 61262025; the National Social Science Foundation of China under Grant No. 18BJL104; the Natural Science Foundation of Key Laboratory of Software Engineering of Yunnan Province under Grant No. 2020SE301; Yunnan Science and Technology Major Project under Grant No. 202002AE090010, 202002AD080002-5; the Data Driven Software Engineering Innovative Research Team Funding of Yunnan Province under Grant No. 2017HC012;

ORCID: Chen GAO, https://orcid.org/0000-0001-9966-498X_

are specialized terms used in the field of cyber security. With the rapid increase in the frequency of cyber attacks, more and more cyber security data are publicly disclosed in different online resources, such as blogs, forums, and online databases. This information is often collected and stored in semi-structured vulnerability databases, such as the national vulnerability database (NVD) and common vulnerabilities and exposures (CVE). For example, Microsoft Internet Explorer Buffer Error Vulnerability, Juniper Networks Junos OS Security Vulnerability, etc., are stored in these databases. Based on these databases, key concepts are analyzed and extracted automatically, and the knowledge base of cyber security is built. They are helpful for discovering new threats, viruses, or vulnerabilities in time and taking corresponding protective measures.

At present, the NER system has achieved good results in the general field, but there are still many problems in the cyber security domain. In the field of cyber security, there are many entities with complex names, which are constantly updated, such as various Trojan horse viruses and network attack methods, and there are few standard datasets. It is difficult to accurately identify cyber security entities. In short, recognition of named entities in the cyber security domain is still a challenging task.

The rest of this paper is organized as follows: Section 2 discusses some basic NER approaches. Section 3 lists the problems and challenges facing NER in the field of cyber security. Section 4 describes various types of cyber security NER systems. Section 5 summarizes the resources and evaluation paradigms developed to support and standardize research. In Section 6, we conclude this review and list future trends.

2 NER approaches

Different approaches are used to identify named entities from unstructured cyber security data sources. These methods are rule-based NER, dictionary-based NER, and machine learning-based NER. Machine learning methods can be divided into statistical machine learning and deep learning. The classification of NER approaches is shown in Fig. 1.

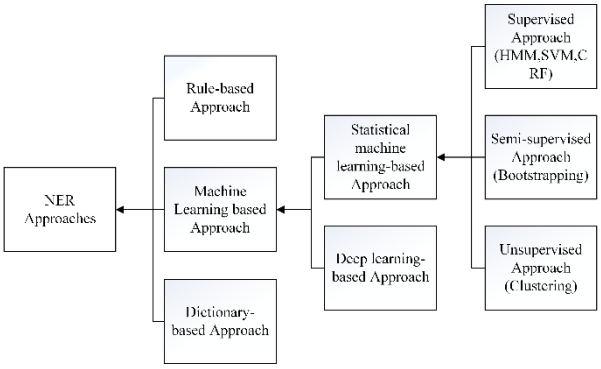

Fig. 1 Classification of NER approaches

2.1 Rule-based approach

In the early stage, the rule-based approach was mostly used in naming entity tasks. The rule-based approach required manual construction of rule templates. In the templates, the selected features generally include statistical information, punctuation symbols, keywords, indicators and positional loci, etc. Pattern and string matching were the main means of identifying the corresponding entities. Most of these systems depend on the establishment of knowledge bases and dictionaries. Because a manually constructed template makes full use of the features of different languages, the rule-based approach is usually accurate and in line with the way people think. However, the rule-based approach has obvious limitations. First, rule template construction mainly relies on manual writing, which is time-consuming and labor-intensive. Second, a large amount of linguistic information is involved in the process of template compilation, which requires a high-performance compiler and professional linguists. Finally, rule templates need to be written for a specific corpus, which lacks reusability and low utilization.

2.2 Dictionary-based approach

A dictionary is a collection of all entity categories. The approach based on a dictionary tries to find the dictionary to match all named entities from the text. This approach settled the problem of entity boundary detection. However, it is difficult to detect noisy entities. At present, the current situation is that there are few training sets for marking in the field of cyber security, so the well-built domain dictionary should be used to pre-mark the text. This greatly reduces the workload of manual knowledge engineering. Riloff et al. (1993) developed a system

called AutoSlog that can be used to automatically extract information from a domain-specific dictionary of concepts from text. They used the system to build a dictionary that can describe terrorist incidents in just five hours. Comparing their dictionary to a hand-made dictionary, AutoSlog is 99.7% as good. But based on the vast array of data sources on the Internet, new information is being added every day. It is impossible to build a dictionary that covers all the different entity categories. Therefore, relying solely on a dictionary-based approach is impractical, takes a lot of work, and is time-sensitive.

2.3 Machine learning-based approach

Due to the limitations of rule-based and dictionary-based approaches, researchers began to gradually adopt machine learning approaches in named-entity recognition. Machine learning is a method that automatically analyzes data and builds models based on data distribution. When using the machine learning approach, named-entity recognition is usually regarded as sequence labeling or a classification task. Machine learning can be divided into two categories: statistical machine learning and deep learning. Compared with the rule-based approach, the machine learning approach does not need to manually define complex rules, and it has better scalability for different forms of text in the same field. Compared with the dictionary-based method, it identifies new entities that do not appear in the dictionary, deals with text ambiguity caused by irregular writing in the text, and reduces dictionary maintenance.

2.3.1 Statistical machine learning-based approach

Statistical machine learning methods require extensive feature engineering, which often requires domain experts to find features before the training process. The feature engineering steps are manual and require a great deal of domain knowledge. Data are important parts of statistical machine learning resources. The quality and size of data determine the quality of a statistical machine learning model. Generally speaking, according to whether the data are marked, the data can be divided into two categories: labeled data and unlabeled data. The labeled data are usually obtained by asking people to make a judgment on the given unmarked data. In statistical machine learning systems, learning methods are divided into three categories: supervised learning, semi-supervised learning, and unsupervised learning (Nadeau et al., 2007).

1. Supervised learning approach

The supervised learning approach requires a large amount of labeled data to train the model as a training set. After the model runs, the named entity in the identified text is classified and detected. In NER, labels refer to categories of different entities in text, usually annotated by domain-specific experts. Supervised learning, by far the most common type of machine learning, learns to map input data to known targets given a set of tag samples. The main models for NER research using supervised learning methods include the Hidden Markov Model (HMM) (Eddy et al., 1996), support vector machine (SVM) (Hearst et al., 1998), and Conditional Random Fields (CRF) (Lafferty et al., 2001). The processing steps of these approaches are mainly divided as follows: text preprocessing, feature selection, model training, model optimization, entity recognition, and evaluation. Fig. 2. Shows the basic flow of named-entity recognition for the supervised learning approach.

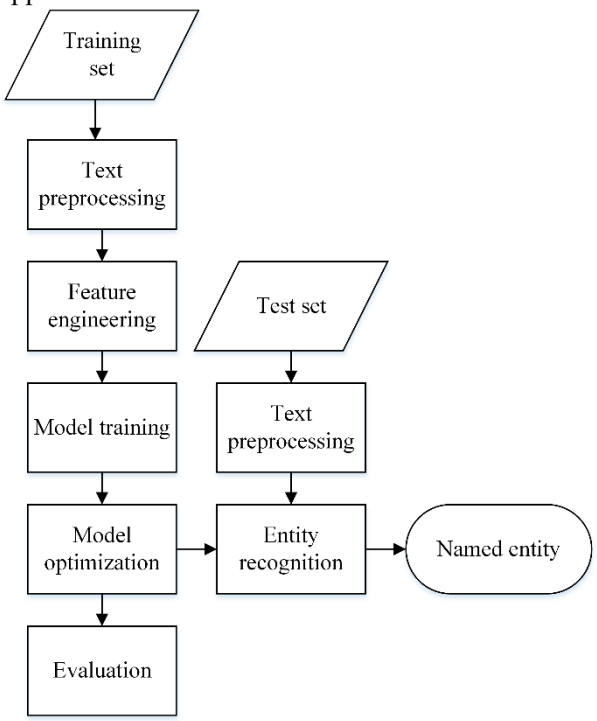

Fig. 2 NER process based on the supervised learning approach.

2. Semi-supervised approach

Semi-supervised learning is a special learning method that is somewhere between supervised learning and unsupervised learning. The statistical machine learning method, based on supervised learning, requires a large amount of labeled data as a training set for learning. In many fields where machine learning is applied, there is a large amount of unlabeled data. Marking the data requires experts in different fields to spend a lot of time and energy, and the efficiency is low. This results in a very small number of samples with class tags and a surplus of samples without class tags. Therefore, people try to add a large number of classless label examples to a limited class of labeled samples and train them together to learn, hoping to improve the learning performance. This is why semi-supervised learning is presented.

As a representative semi-supervised learning method, Bootstrapping uses a small number of training samples called "seeds" to learn and supplement from a large number of unlabeled texts to generate labeled samples. The results are then used to retrain the system to generate more tag examples to increase the size of the training dataset. Learning decisions are improved by repeating the process.

3. Unsupervised approach

Semi-supervised learning is essentially the same as supervised learning, and requires a large number of features to be specified from the dataset through feature engineering. Learning a model often requires good feature sets and large labeled corpora. However, these labeled training sets are expensive to obtain and have domain limitations. Unsupervised learning makes full use of a large amount of unlabeled data and mines some potential structures of the data. It mainly includes two algorithms: reducing dimensionality and clustering. An unsupervised learning algorithm based on clustering plays an important role in NER. The NER system that is based on clustering extracts named entities from the cluster based on context similarity and infers the names of named entities by calculating lexical resources, lexical patterns, and statistical information.

2.3.2 Deep learning-based approach

Deep learning involves training the model by mining the character features through the neural network. In recent years, with the continuous improvement in computing power and the arrival of the era of big data, neural networks have developed rapidly, especially in the field of deep learning. Deep learning (LeCun et al., 2015) is a branch of machine learning that learns data representations with multiple levels of abstraction through computer models that have multiple processing layers. It is a feature learning method that transforms raw data through simple but nonlinear models into higher-level and more abstract representations. With enough combinations of transformations, very complex functions can also be learned. A process of neural network training generally includes two steps: forward propagation and back propagation. The process of forward propagation involves a weighted sum of all inputs and passing the result to an activation function. The purpose of back propagation is to recalculate the gradient of the loss function in the network through the chain derivative rule of the compound function. This process adjusts each weight value to minimize the loss function and obtain the optimal weight parameter, which is usually used in combination with the optimization method (such as the gradient descent method (Ruder, 2016)). The loss function is the reflection of the model to the degree of data fitting. It is the difference between the predicted value and the actual value of the sample.

The NER model based on deep learning has become the leading approach and provides the best experimental result. Compared with the traditional feature-based method, deep learning automatically excavates the hidden features by constructing neural layer and has good feature-learning ability. The deep learning model is widely used in NER for the following reasons. First, the neural network generates a nonlinear mapping from input to output. Compared with the traditional linear model HMM or linear chain CRF, the deep learning model learns complex features from data through the nonlinear activation function. Second, the statistics-based method needs to find the appropriate feature design model and design the appropriate combination of features for different tasks. Only through extensive feature engineering, can good experimental results can be achieved. In deep learning, the features of the data do not need to be designed by hand, but are learned from the data using a general

learning process. It automatically learns useful representations of data, and model training is a data-driven process, independent from feature engineering.

The process of using a deep learning model to carry out NER is mainly divided into the following steps. The first step is preprocessing of input data, which is the extraction of character features and the distributed representation of input data by the word vector method. Second, the context dependence of input data is captured through the deep neural network model. Generally, network structures are selected, such as the Recurrent Neural network (RNN), Convolutional Neural Network (CNN), or Long Short-Term Memory (LSTM). Finally, the tag decoder is used to predict the token tags in the input sequence. Two commonly used tag decoders are CRF and softmax. The input sequence tag scheme has a variety of different options, but they are generally similar and usually marking the beginning, ending, and non-entity of the entity.

3 Problems with cyber security NER

Compared with the traditional domain, NER in the cyber security domain is more challenging for the following three reasons. First, the cyber security domain has many technical terms and complex naming conventions. Some examples are as follows:

- Conjunction and disjunction: two or more cyber security entity names share a common header noun by using conjunction or disjunction. For example, “mail and USB stick worm” contains two entity names: “mail worm” and “USB stick worm,” and “worm” itself can be used as a cyber security entity name.
- Non-standardized naming convention: entity names in the cyber security domain do not follow a regular naming convention and there are many special terms. These terms include capital letters, numbers, or hyphens, e.g., “EternalBlue” and “HeartBleed” are domain nouns that are specific to the cyber security domain. In addition, the same cyber security entity name often has different spelling forms, such as “Zero-day” and “0-day” all refer to the same entity.
- Abbreviation: text in cyber security often uses abbreviations, which lead to a lot of ambiguity. For example, “CSRF” means “cross-site request forgery” and “XSS” means “cross-site scripting.” These abbreviated entities are often highly ambiguous. It is not possible to categorize them based solely on existing dictionaries. “JS” may represent “JavaScript” or “JScript” in different text, which are two different entities.
- Massive nesting: another major challenge in the cyber security domain is that one entity name may often be embedded in another entity name.

Second, in the traditional NER domain, there are many standard, exact public datasets that can be used. In the cyber security domain, there are few available datasets. There is also a lack of unified and standardized classification standards.

Finally, the cyber security entity categories are unevenly distributed. For categories containing an abundant number of entities, the training set can be well identified by the model. However, there are still entity classes that contain only a few entities that are sparsely distributed in the massive network space security text. It is difficult to identify these entities based on the traditional NER method, thus affecting the performance of the whole classifier. Next, in response to these cyber security NER issues, we summarize and analyze the literature on solving related issues. We use machine learning methods to classify and describe these works, mainly analyze their methodologies and shortcomings, and compare the technologies they used and their F1 values.

4 Cyber security NER systems

Cyber security NER systems use three different learning approaches: supervised learning, semi-supervised learning, and unsupervised learning. Table 1 lists the works related to cyber security NER in the past ten years. As can be seen from Table 1, most of the literature was published in the last three years.

Table 1 List of literatures on cyber security NER systems (ordered by the number of citations)

No.	Work	Method Title	Country	Citations
1	Joshi et al. (2013)	Extracting cyber security-related linked data from text	USA	70

Table 1 (Continued)

No.	Work	Method Title	Country	Citations
2	Mulwad et al. (2011)	Extracting information about security vulnerabilities from web text	USA	63
3	Bridges et al. (2013)	Automatic labeling for entity extraction in cyber security	USA	25
4	McNeil et al. (2013)	PACE: Pattern accurate computationally efficient bootstrapping for timely discovery of cyber-security concepts	USA	22
5	Lal. (2013)	Information extraction of security related entities and concepts from unstructured text	USA	15
6	Dionísio et al. (2019)	Cyberthreat Detection from Twitter using Deep Neural Networks.	Portugal	6
7	Weerawardhana et al. (2014)	Automated Extraction of Vulnerability Information for Home Computer Security	USA	9
8	Gasmi et al. (2018)	LSTM Recurrent Neural Networks for cyber security Named-Entity Recognition	USA	4
9	Zhou et al. (2018)	Automatic Identification of Indicators of Compromise using Neural-Based Sequence Labelling	China	3
10	Xiao (2017)	Towards a Two-phase Unsupervised System for cyber security Concepts Extraction	USA	1
11	Shang et al. (2017)	A Framework to Construct Knowledge Base for Cyber Security	China	1
12	Mazharov and Dobrov. (2018)	Named-Entity Recognition for Information Security Domain	Russia	1
13	Tikhomirov et al. (2020)	Using BERT and Augmentation in Named-Entity Recognition for Cybersecurity Domain	Russia	1
14	Qin et al. (2019)	A network security entity recognition method based on feature template and CNN-BiLSTM-CRF	China	0

Table 1 (Continued)

No.	Work	Method Title	Country	Citations
15	Pingchuan et al. (2019)	Cyber security Named-Entity Recognition Using Bidirectional Long Short-Term Memory with Conditional Random Fields	China	0
16	Zhang et al. (2019)	Multifeature Named-Entity Recognition in Information Security Based on Adversarial Learning	China	0
17	Long et al. (2019)	Collecting Indicators of Compromise from Unstructured Text of cyber security Articles using Neural-Based Sequence Labelling	China	0
18	Gu et al. (2019)	Tweet malware Name Recognition based on enhanced BiLSTM-CRF model	China	0
19	Georgescu et al. (2019)	Named-Entity-Recognition-Based Automated System for Diagnosing Cybersecurity Situations in IoT Networks	Romania	0
20	Li et al. (2019)	A Self-Attention-Based Approach for Named-Entity Recognition in Cybersecurity	China	0
21	Liu. (2020)	Network Security Entity Recognition Methods Based on the Deep Neural Network	China	0
22	Wu et al. (2020)	An Effective Approach of Named-Entity Recognition for Cyber Threat Intelligence	China	0
23	Simran et al. (2020)	Deep Learning Approach for Intelligent Named-Entity Recognition of Cyber Security	USA	0
24	Kim et al. (2020)	Automatic extraction of named entities of cyber threats using a deep Bi-LSTM-CRF network	Korea	0
25	Wang et al. (2020)	NER in Threat Intelligence Domain with TSFL	China	0

Fig. 3 shows how the F1 values of the cyber security methods change over time. As can be seen from the figure, the research methods gradually change from the statistical machine learning method to the deep learning method, and the F1 value continuously increases. Before 2018, most methods were based on statistical learning, which represented about 20% of the total. The F1 values of these methods were very low, below 80%. With the continuous development of deep learning, since 2018, deep learning-based methods have become the mainstream method of network security NER. More than 80% of the work is based on the deep learning method, most of which have F1 values close to 90%.

Among them, the F1 values of two jobs are lower than 80% (Tikhomirov et al., 2020; Kim et al., 2020). This is caused by the dataset they used. Their datasets are small and the distributions of category data are uneven.

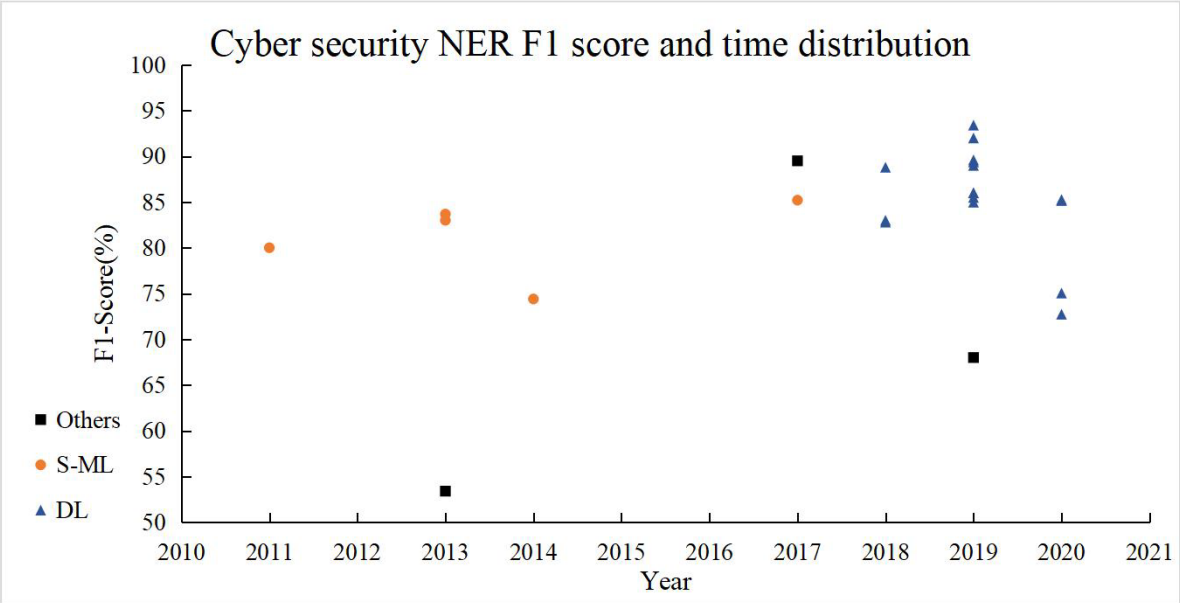

Fig. 3 Cyber security NER F1 score and time distribution (DL: deep learning; S-ML: Statistical machine learning; Others: Semi-supervised and unsupervised).

Next, supervised, semi-supervised, and unsupervised cyber security NER systems will be introduced in detail.

4.1 Supervised cyber security NER systems

The supervised learning method is the main learning method in machine learning systems. Early supervised learning models based on statistical machine learning mainly include HMM, SVM, and CRF. With the continuous development of neural networks, deep learning technology has gained a leading position in Natural language processing (NLP). Compared with the statistical machine learning method, deep learning has achieved better learning results. More importantly, the deep learning model can be combined with the statistical machine learning method, which is the mainstream method in cyber security NER at present. These methods and related studies are described below.

4.1.1 Statistical machine learning-based systems

Because works using HMM in the field of cyber security is very rare, we mainly introduce cyber security NER systems that are based on SVM and CRF.

1. SVM-based systems

SVM is a supervised and discriminant learning model. It is a binary classifier that looks for the optimal linear hyperplane between positive and negative samples. As an improbability linear classifier, it handles a large number of features with high accuracy without falling into overfitting. SVM is mainly used to deal with classification and regression problems. However, in the cyber security NER task, the data category distribution is severely unbalanced. The number of uncorrelated entities far exceeds the number of cyber security entities, resulting in sparse data issues, which have a great impact on the training of the SVM classifier. Therefore, SVM has only been used by Mulwad et al. (2011).

Mulwad et al. (2011) proposed a prototype framework based on the SVM classifier. They detected and extracted information about vulnerabilities and attacks from web page text. Their framework consists of three parts. The first part is to train an SVM classifier to identify vulnerability and threat text. The second part is an information extraction system. This system extracted relevant concepts, relationships, and events from the vulnerability and threat text according to the Wikitology (Syed et al., 2010) knowledge base and computer security vulnerability ontology. Then, using Web Ontology Language (OWL) coding, machine-

readable assertions were generated from these data. Finally, the prototype system was evaluated based on the Vulnerability text description set of the NVD.

2. CRF-based systems

CRF is a probabilistic graph model. It is widely used in sequential marking tasks such as NER, speech tagging, and speech recognition. CRF combines the characteristics of the maximum entropy model and HMM. It relaxes the conditional independence assumption required by HMM and uses more global features. Moreover, it is based on the same exponential form as the maximum entropy model, which carries out complete and non-greedy finite state derivation and training effectively, and trains the model with less training data.

Joshi et al. (2013) developed an information extraction framework. It extracts cyber security-related entities, terms, and concepts from the NVD Database and unstructured text. Their framework is divided into three main components. The first is a CRF classifier, which is used to identify terms and concepts related to cyber security from text. Next is an ontology-based RDF triplet generator that generates triples based on the extracted information provided by the classifier. Finally, there is a link generator that uses DBpedia Spotlight (Mendes et al., 2011) to link the extracted entities and concepts to the DBpedia repository. The system used 50% cross validation to evaluate the CRF classifier and obtained Precision of 0.83 and F1 of 0.80.

Based on Joshi's work, Weerawardhana et al. (2014) used two different NER methods to identify critical Personalized Attack Graph (PAG) parameters embedded in vulnerability description text. Their first method is based on machine learning and is divided into two steps: feature selection and model training. Another method uses the Part of Speech (POS) to mark the lexical pattern of text, which mainly defines a series of rules and dictionaries to match text to get entities according to the grammatical structure of text description. According to the data comparison, the POS method is generally better, especially when identifying a small number of implicit entity classes in the vulnerability description. In addition, POS solutions outperform machine learning solutions in identifying explicit entities such as file names. However, using POS relies too much on the grammatical structure in the sentence, and the entity type recognized is too narrow. When faced with a large number of heterogeneous data sources, there are a large number of entities of different language classes, so machine learning methods are more efficient.

Lal (2013) used NLP and text mining methods to implement a system that automatically extracted terms from cyber security blogs and security announcements. The model in this system is the Stanford NER model based on CRF. This system is divided into three parts: training module, Stanford named-entity identifier, and feature set project. Nine features were selected to form the text's feature set, including word order, string length, the relationship between the preceding and the following words, and part of speech. To evaluate the system, their experiment adopted the method of quadruple cross validation and trained four NER recognizers with different data blocks; the results were quite consistent.

Shang et al. (2017) proposed a framework to integrate an existing cyber security knowledge base and extract security-related information from text. In their framework, a vulnerability-centric ontology is built to assist in information extraction, and a Stanford named-entity identifier is trained to extract security-related entities from text. Secure data sources in cyberspace are mainly divided into structured data and unstructured data. For structured data, the D2R mapping tool was used to transform the data in the relational database into RDF data. For unstructured data, the method based on rules and the Stanford NER recognizer based on the linear chain random field sequence model were used, respectively. In their experiment, the 10-fold cross-validation method was used to evaluate the model.

Table 2 lists the work related to machine learning network security NER.

Table 2 Summary of related work on statistical machine learning cyber security NER

Algorithm	Method	Dataset	F-Score	Pros	Cons
Mulwad et al. (2011)	SVM, Wikitology	Internet Cybersecurity Text	80.0%	Effective way to link entities to knowledge base	Fully dependent on Wikiology, Feature Engineering

Table 3 (Continued)

Algorithm	Method	Dataset	F-Score	Pros	Cons
Werawardhana et al(2014)	CRF, Rule template	Bridges et al. (2013)	74.4%, 67.8%	Compare different methods	Feature Engineering, poor mobility
Joshi et al. (2013)	CRF, DBpedia	Internet Cybersecurity Text	83.0%	Effective way to link entities to knowledge base	Fully dependent on DBpedia; the method is difficult to reproduce
Lal. (2013)	CRF	Internet Cybersecurity Text	83.68%	Compare different data sets	Feature Engineering Size of the dataset is small
Shang et al. (2017)	CRF, Rule template	Lal. (2013)	85.2%	Combining rule and machine learning	Feature Engineering Rely on experts to define rules.

4.1.2 Deep learning-based systems

LSTM is the dominant model for performing NER in deep learning. As a recurrent neural network structure, it is an improvement on the RNN and learns arbitrary long-term dependencies.

Lample et al. (2016) combined LSTM with the traditional CRF model and achieved good NER results. In the LSTM-CRF model, LSTM automatically extracts character features in the sequence, and replaces feature engineering in traditional machine learning. The introduction of the CRF model after the LSTM model enables the model to effectively use the dependencies while combining context information to generate the optimal tag sequence.

Gasmi et al. (2018) proposed a domain-independent method to extract entities in the cyber security domain. Their method is LSTM-CRF, which does not rely on specific knowledge in the field of cyber security and performs feature engineering without expert knowledge. Their experiments were carried out on the dataset published by Bridges et al. (2013). The experimental results showed that the LSTM-CRF method was generally better than the CRF method in all categories. Mazharov and Dobrov (2018) studied NER in Russian text and proposed two artificial neural network-based methods for information extraction. They are fully connected neural network models and LSTM models. The inputs to the models include character embedding and word embedding to better capture the morphological features of the text, while the external information security domain dictionaries add external features to words. Experimental results show that the LSTM network structure captures text sequence information more efficiently. It is superior to fully connected neural networks, and the identification of relevant entities can be significantly improved by introducing an external dictionary. Wu et al. (2020) proposed a method based on an LSTM-CRF model and domain dictionary matching. Their method uses Bi-directional Long Short-Term Memory (BiLSTM) to automatically capture context features. CRF is used to learn label constraint rules, and an ontology-based domain dictionary is used for matching correction.

Pingchuan et al. (2019) proposed a NER model for cyber security: XBILSTM-CRF. The model consists of a word embedding layer, a BiLSTM layer, and a CRF layer. Compared with the traditional BiLSTM-CRF model, it combines the output of the BiLSTM layer and the word-embedded vector as the input of the CRF layer, which improves the information contained in the feature vector. In the experiment, the open-source secure unstructured text dataset (Lal, 2013) is used, and the experimental results show that the optimal precision rate is 90.54%, the recall rate is 88.26%, and F1 is 89.38%. The XBILSTM-CRF model obtains the best result.

The above methods are all the applications of deep learning methods in cyber security NER. They avoided the complex feature engineering in traditional machine learning methods and achieved better results. However, due to the complexity of secure text in cyberspace, it is difficult to accurately identify secure entities in cyberspace solely based on LSTM. The ability of LSTM to extract features is limited when considering complex long text, and a lot of useful

information may be lost. One possible solution is to combine multiple neural networks for feature extraction. Simran et al. (2019) evaluated several deep learning architectures and adopted the Gated Recurrent Unit (GRU) as the basic model. The model consists of three parts: the GRU layer refines the vectors and feeds them to the CNN layer, the CNN network generates more optimal features, and the CNN network feeds them to the CRF layer to enhance learning. The GRU also has a memory mechanism as a variant of LSTM, but there are fewer GRU parameters and they converge more quickly.

Qin et al. (2019) combined deep learning methods and feature templates to propose the FT-CNN-BILSTM-CRF model to identify hybrid cyber security entities. Their method requires manual formulation of a small number of features, the formation of a feature template, the extraction of local contextual features, and the combination of local contextual features with the global features of the text extracted by the neural network to form feature vectors for entity recognition. Based on Qin et al. (2019), Liu (2020) made further improvements. Liu improved the concatenation of the word vector and character vector to dynamically use the word vector and character vector information. In addition, the pre-training language model (Peters et al., 2017) is introduced to improve NER on small data sets.

To effectively identify an entity list of low-frequency indicators of compromise (IOCs) in cyber security reports, Zhou et al. (2018) used an attention mechanism (Vaswani et al., 2017) to help LSTM accurately encode the input sequence. At the same time, some spelling features are introduced to define IOCs to improve the performance of the model on a small amount of training data. It was shown that introducing the attention mechanism can effectively encode important information in sentences and improve the accuracy of entity recognition. However, manually defining features not only requires a great deal of knowledge, but is also inefficient and takes effect for specific data sets. Based on the work of Zhou et al. (2018), Long et al. (2019) reduced false extractions caused by spelling features and introduced multiple self-attention mechanisms and contextual information from unstructured cyber security text for IOC recognition. The multi-headed self-attention mechanism was proposed by Google's machine translation team to extract more features of the text itself from multiple perspectives and levels of perspective. Li et al. (2019) considers that a single word feature is not sufficient to identify entities. So, CNN is introduced to extract character features to splice into word features, and a self-attention mechanism is added to the existing BiLSTM-CRF model.

Twitter includes a large number of cyber security tweets, most of which are short and informal. Recognizing cyber security entities from Twitter is a challenge. Dionísio et al. (2019) developed a feature engineering independent neural network-based end-to-end threat intelligence identification and detection tool. They use a deep neural network to process data streams, identify security-related information, and extract relevant entities. The system is based on a binary classifier that uses CNN architecture to identify the security-related tweet text. Named entities were extracted from these tweets using a BiLSTM model.

To solve the problems of short and informal text, the single entity category, and entity ambiguity in malware tweets, Gu et al. (2019) proposed an entity recognition method based on BERT-BILSTM-Attention-CRF, which was the automatic recognition of malware names in tweets. Based on the BiLSTM-CRF model, the Bidirectional Encoder Representation from Transformers (BERT) model (Devlin et al., 2018) was used to encode the word context information. This improves the quality of the context semantics that are embedded in the word and enhances the semantic disambiguation capability of the original model. At the same time, the self-attention mechanism is used to learn the relationship between words and sentence structure, to alleviate the imbalance between entity classes and improve the recognition effect of malware name entities.

To better encode the context, Tikhomirov et al. (2020) proposed a new model called RuBERT, which uses BERT coding as the basic structure, pre-trained a large number of Russian-related corpora, and significantly improved the performance of the three NLP tasks in Russian. In addition, they discussed a special data augmentation method for NER. In their method, annotated data were obtained by inserting named entities in appropriate sentences and contexts. The specific augmentation methods are divided into two subtypes: internal expansion and external expansion. Internal expansion refers to replacing descriptors with specific names in sentences containing related descriptors in the training set. External expansion is searching for sentences with related descriptors in a set of unannotated cyber security texts.

However, in the face of uneven tag distribution, the above methods have a greater impact on the overall recognition effect. In response to this problem,

Wang et al. (2020) proposed a new loss function. It is a triple loss function based on metric learning and classification, which is used to solve the problem of unbalanced data label distribution. In addition, they combined word2vec and BERT to alleviate the out-of-vocabulary (OOV) problem, and introduced an attention mechanism to better encode long sentences. Experiments show that F1 values have been improved in public domain data sets and threat intelligence.

Zhang et al. (2019) introduced an adversarial learning mechanism (Lowd et al., 2005) in cyber security NER and proposed the BILSTM-Attention-CRF-crowd model. To integrate the best single consensus annotation, a generative adversarial network (GAN) is used to generate data. An attention layer is added to the network structure to deal with long sentences in the text. Kim et al. (2020) presented a neural network model for obtaining data from secure documents in cyberspace. They constructed a gold-standard corpus from unstructured PDF text. In addition, using RNN, CNN, and bag-of-character (BOC) methods, the characters were combined with word units to form dynamic inputs and applied to BILSTM-CRF models to predict entities in the cyber security domain. The results show that the proposed BOC character-level feature representation has faster speed and higher performance than the character-level feature representation based on CNN or RNN. Table 3 summarizes the cyber security NER research based on deep learning.

Table 4 Summary of deep learning cyber security NER research

Algorithm	Method	Dataset	F-Score	Pros	Cons
Gasmi et al. (2018)	LSTM, CRF	Bridges et al. (2013)	82.8%	No feature engineering	Difficult to identify complex entities
Mazharov and Dobrov (2018)	LSTM, CRF	Internet Cybersecurity Text	83.0%	Improved performance	Small examples provide small improvement
Wu et al. (2020)	LSTM, CRF, Dictionary	Internet Cybersecurity Text	85.27%	Improved performance	Need to build a dictionary manually
Pingchuan et al. (2019)	LSTM, CRF	Lal. (2013)	89.38%	Improve network structure	More features are needed
Simran et al. (2019)	GRU, RNN, CRF	Bridges et al. (2013)	93.4%	Improved performance and efficiency	Need a lot of label data
Qin et al. (2019)	CNN, LSTM, CRF, Feature template	Internet Cybersecurity Text	86%	Identify mixed security entities in Chinese and English	Manually define feature templates
Liu. (2020)	CNN, LSTM, CRF, Unified Language Model Pre-training	Internet Cybersecurity Text	86%	Improved NER effect on small data sets	Manually define feature templates
Zhou et al. (2018)	LSTM, CRF, Attention, token spelling features	Internet Cybersecurity Text	88.8%	Improved low-frequency entity recognition	It is difficult to distinguish tokens similar to IOCs but not malicious
Long et al. (2019)	LSTM, CRF, Multi-Attention, token spelling features	Internet Cybersecurity Text	89.6%	Better contextual expression learning	No pre-embedded language model is used

Table 3 (Continued)

Algorithm	Method	Dataset	F-Score	Pros	Cons
Li et al. (2019)	LSTM, CRF, Mutli-Attention	Internet Cybersecurity Text	84.98%	Improved performance	More features are needed
Dionísio et al. (2019)	CNN, LSTM, CRF	Internet Cybersecurity Text	92%	End-to-end processing of short text data	Fewer entity categories
Gu et al. (2019)	BERT, LSTM, CRF, Multi-attention	Internet Cybersecurity Text	85.5%	Improve the recognition effect of a single complex entity category	Need improvement in efficiency
Wang et al. (2020)	BERT, word2vec, new loss function, attention	Internet Cybersecurity Text	85.16%	solve the problem of unbalanced data label distribution.	Time complexity
Tikhomirov et al. (2020)	BERT, dataset augmentation	Internet Cybersecurity Text	72.74%	Propose a new method of dataset augmentation for NER tasks	Unbalanced data categories lead to poor recognition
Zhang et al. (2019)	GAN, LSTM, CRF	Internet Cybersecurity Text	89%	Improve the quality of crowdsourced annotations in information security	Time complexity
Kim et al. (2020)	LSTM, CRF, BOC (Bag-of-Character)	Internet Cybersecurity Text	75.05%	BOC character embedding effective than RNN and CNN	Need improvement in performance and accuracy

4.2 Semi-Supervised cyber security NER systems

McNeil et al. (2013) proposed a new semi-supervised learning algorithm. This method uses the bootstrapping algorithm learning heuristic to identify and classify cyber security entities in unstructured text sources, and to identify other entities through an iterative loop. They improved on the traditional bootstrapping method and adopted a compromise method of time memory to extract entities. It avoids the tedious corpus search and strengthens the model nomination, which is beneficial in improving the accuracy. The algorithm achieved a precision value of 0.9 and recall value of 0.12. When corpora containing fewer seeds were removed, their recall value was 0.38.

Georgescu et al. (2019) designed an automated diagnostic system for cyber security situations in IoT networks. The architecture of the system consists of four components: data input, data analysis, data storage, and data output. In the data analysis module, a domain ontology is built, on which the NER model for semantic text analysis is developed to identify specific vulnerabilities in the IoT from a large number of heterogeneous IoTs data sources. Their NER used the Watson Knowledge Studio tool to train the model, which was a semi-supervised learning approach that combines manually annotated data with context-based knowledge extraction. Their experimental data were mainly obtained from CVE. The data processing was divided into three steps. First, a large number of relevant instances were identified by ontology-based model; then, text was automatically annotated by dictionary and unrecognized parts were manually annotated to obtain the final training data set.

Table 4 summarizes cyber security NER research based on semi-supervised learning.

Table 4 Summary of related work on semi-supervised learning cyber security NER

Algorithm	Method	Dataset	F-Score	Pros	Cons
McNeil et al. (2013)	Bootstrapping Pattern	Internet Cybersecurity Text	53.4%	Selection of informative samples from the unlabeled data	The data is small and performance is very low
Georgescu et al. (2019)	Ontology Watson Knowledge Studio tool	Internet Cybersecurity Text	68.0%	Combining manually annotated data with context-based knowledge extraction	Needs improvement in performance and accuracy

4.3 Unsupervised cyber security NER systems

Xiao (2018) proposed an information extraction system based on unsupervised learning for locating and classifying cyber security concepts in unstructured texts. Their system is mainly composed of two parts. The first part is an unsupervised NER system that does not require manual annotation of data. The second part of their system represents domain knowledge in a domain model and a domain ontology. For the unstructured text input, first, the named entity in the text is located and identified according to their NER system. Then the entity is classified by using the previously constructed domain knowledge. The experimental results have achieved high accuracy for all categories, but there are still some problems. Their system cannot handle name disambiguation, cannot accurately identify the same security entity with different representations, and has limited domain knowledge coverage. Their system cannot satisfactorily handle named entity instances that never appear in the domain model and domain ontology.

Table 5 summarizes the NER cyber security research based on unsupervised learning. According to the above statistics, it can be concluded that cyber security NER research is mainly focused on supervised learning, while there is less research on semi-supervised and unsupervised learning.

Table 5 Summary of related work on unsupervised learning cyber security

Algorithm	Method	Dataset	F-Score	Pros	Cons
Xiao. (2018)	word2vec ontology	Internet Cybersecurity Text	89.50%	Exploits a large amount of unlabeled data	Entity ambiguity and difficulty in identifying new entities

4.4 Technology development

Fig. 4 shows how methods have changed over the years. In Fig. 4, FT refers to Feature template, WP refers to Word embedding paradigm, and PP refers to Pre-training paradigm. As can be seen, before 2017, cyber security NER adopted a feature template-based approach. Starting in 2017, Word embedding paradigm methods took the lead. The Pre-training paradigm method has been gradually applied to cyber security NER since 2019.

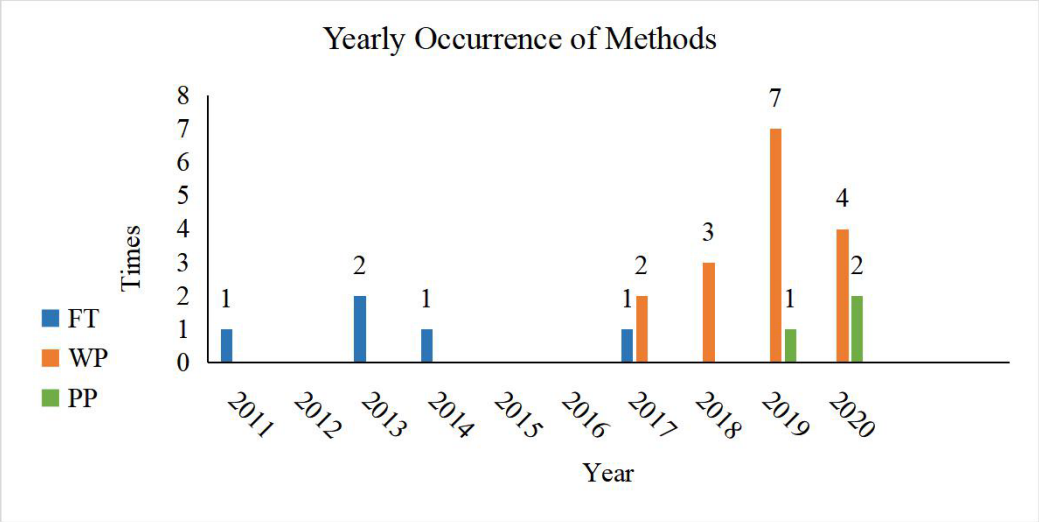

Fig. 4 Yearly occurrence of methods.

4.4.1 Feature template

Early cyber security NER mainly adopted methods based on statistical learning, such as SVM and CRF. Such methods usually require complicated feature engineering to construct the feature template required by the model to form a feature vector representation. Feature vector representation is an abstraction over text where a word is represented by one or many Boolean, numeric, or nominal values. Word-level features (such as capitalization, location, lexical, and part-of-speech tags) (Weerawardhana et al., 2014; Lal, 2013; Shang et al., 2017), list lookup features (such as Wikipedia and DBpedia) (Mulwad et al., 2011; Joshi et al., 2013), and document and corpus features (such as local grammar and multiple appearances) have been widely used in various NER systems based on statistical machine learning.

4.4.2 Word embedding paradigm

With the rise of deep learning technology, more and more cyber security NER methods have begun to adopt deep learning. Using neural networks for NER can automatically discover hidden text features. First, the input sentence needs to be vectorized, usually in the form of distributed representation. The distributed representation is automatically learned from the text. It captures the semantic and syntactic attributes of words, which are not explicitly displayed in the NER input. The word embedding representation used in the NER model mainly includes the following three types: word-level, character-level, and hybrid representations.

1. Word-level Representation

Common word embedding only refers to Google Word2Vec. Many researchers employ word-level representation (Gasmi et al., 2018; Mazharov and Dobrov., 2018; Pingchuan et al., 2019; Simran et al., 2019; Dionísio et al., 2019; Xiao, 2018), which is also the most extensively used method. However, the traditional word-level representation is based on a large-scale open-source corpus for training. The vocabulary contained in the dictionary is mainly common words. When used in a specific field, it will cause OOV problems, so it is difficult to obtain a specific field sentence vector representation. Roy et al. (2017) trained one million word embedding vector vocabulary representations in the cyber security domain, which can effectively improve the NER vector representation.

2. Character-level representations

Character-level representation has been found useful for exploiting explicit sub-word-level information such as prefix and suffix. Another advantage of character-level representation is that it can naturally handle OOV. Therefore, character-based models infer the representation of invisible words and share morpheme-level regular information. Kim et al. (2020) found that using the BOC character embedding method to model sentences is better than traditional RNN and CNN methods. Li et al. (2019) considered that single word features were not enough to recognize entities, and CNN was introduced to extract character features and then connect them to word

features.

3. Hybrid Representation

In addition to word-level and character-level representations, some studies also incorporate other information in the final representation of the word, and then send it to the context encoding layer; for example, spelling features (Zhou et al., 2018; Long et al., 2019), dictionary features (Wu et al., 2020; Zhang et al., 2019), and context feature templates (Qin et al., 2019). DL-based representation and feature-based methods are combined in a hybrid manner. Adding additional information may lead to improved NER performance, but at the cost of impairing the versatility of these systems.

4.4.3 Pre-training paradigm

The main idea of word2vec is the distributed hypothesis of word meaning, and mapping each word to a unique dense vector. The resulting word representation is static and does not consider context, so it is difficult to solve the polysemy problem. BERT uses Transformer as a feature extractor to fully learn contextual features and can better model polysemous words. Some researchers use BERT for cyber security NER (Gu et al., 2019; Tikhomirov et al., 2020; Wang et al., 2020). Compared with the traditional LSTM extraction structure, the experimental results have been significantly improved, but the network structure of BERT is more complicated and has more layers, so the time and space cost of the model is high.

5 Resources available for cyber security NER

With the development of NER research on cyber security, more and more related resources have been created and utilized by researchers to facilitate the development, evaluation, and comparison of such systems. Common cyber security resources include corpus, security ontology, and security NER evaluation. This section focuses on these three resources.

5.1 Corpus

A corpus is a set of text documents containing one or more entity-type annotations, and is often used to train machine learning models to identify other similar entities in the relevant text. A high-quality annotated corpus is crucial to model learning and evaluation. Corpus data mainly comes from cyber security domain Security blogs, CVE, NVD, Security Bulletins, and so on. Data annotation is usually divided into manual annotation, dictionary annotation, and algorithm annotation. Manual labeling has a high accuracy rate, but it requires a lot of manpower and time to perform inefficient work. When facing different domains, the data are not universal and the availability is low. For cyber security, relevant data are updated very quickly and new entity names emerge in an endless stream. It is difficult to build and maintain a cyber security dictionary. Therefore, it is difficult to form a large-scale and accurate annotation corpus through dictionary annotation. At present, there are two relatively complete cyber security corpus types. One is data annotation based on an algorithm. Bridges et al. (2013) automatically marked text from multiple data sources by using a large amount of structured data available in the cyber security domain. Another is the data set formed by manual annotation by Lal (2013), Mazharov and Dobrov (2018), and Kim et al. (2020). Table 6 summarizes the publicly available datasets.

Table 6 Corpus for cyber security

Corpus	Main Entity	Type Size	Language
Lal. (2013)	Software, modifier, operating system, consequences, attack, means, file name, network, hardware	More than 45,000 tokens, 5000 tagged entities	English
Bridges et al. (2013)	Vendor, application, version, edition, OS, hardware file	853,560 tokens, 73,964 tags	English
Mazharov and Dobrov (2018)	Hacker, hacker_group, virus, device. tech, program	377,364 tokens, 13,076 tags	Russian
Kim et al. (2020)	Hash, malware, IP, URL	498,000 tokens, 15,720 tags	English

5.2 Cyber security ontology

Cyber security ontology refers to the fusion and reasoning of a cyber security knowledge base using the ontology method to obtain a clear and standardized formal explanation. The construction of cyber security ontology is helpful in correctly expressing the concept characteristics of the cyber security domain and provides standard concept classification for marking cyber security data. In particular, these ontologies contribute to the construction of cyber security dictionaries, which are at the heart of developing dictionary-based NER technologies and are often used to improve the performance of various machine learning and rule-based approaches. Syed et al. (2016) developed an ontology for cyber security in the OWL language. The ontology incorporates and integrates heterogeneous data and knowledge schemas from different cyber security systems and most commonly used cyber security standards for information sharing and exchange.

5.3 Evaluation criteria

The evaluation of system performance is a very important step for the NER task. By evaluating the system, we can analyze the problems existing in the system in the process of identifying named entities to improve the system. The performance of a NER system is evaluated by comparing the output with manual annotations. Three common metrics are Precision, Recall, and F-score.

If there is only one entity category in the sample, the above metrics can be used for evaluation. However, due to the diversity of named entity categories in different domains, most NER systems often involve multiple entity types. Therefore, comprehensive evaluation of multiple dichotomous confusion matrices is required to evaluate the performance of all entity categories. For this purpose, there are two commonly used evaluation metrics: macro average F-score and micro average F-score. The calculation of the macro average F-score requires statistical values for each class and arithmetic mean values for all classes. However, the micro-average F-score calculation method is designed to establish a global obfuscation matrix for every instance in the data set, regardless of category, and then calculate the corresponding metrics.

6 Conclusions and future trends

In this paper, we summarize the research on NER in the cyber security domain, and introduce the common NER models, methods, and related resources. Compared with traditional NER systems, the research on NER in cyber security is more complicated. At present, by using rich feature sets and supervised machine learning, relatively high accuracy has been achieved, but the research focuses on the method of supervised learning, whereas application of semi-supervised learning and unsupervised learning is scarce. In this section, we will discuss some of the trends and issues on which future research may focus.

Supervised learning makes the NER system of cyber security much more efficient than the traditional rule-based or dictionary-based approach. However, most cyber security datasets come from the Internet, with fast update speed and large data volume. Therefore, it is difficult to get an accurate and practical training set, which greatly limits the performance of the supervised learning method. In the future, the application of unsupervised or semi-supervised technology should be explored. Such techniques leverage context patterns and have been quite successful in open named-entity extraction tasks.

Developing a more comprehensive cyber security ontology is another trend. At present, there is no unified standard for entity classification in the field of cyber security. For supervised machine learning methods, switching to a different ontology means changing the entity categories, which is not feasible because it requires reannotating the corpus. The same entity may be divided into different categories under different annotators. This is not only detrimental to the effective use of resources, but also means that system evaluation is not based on common standard datasets, which makes it difficult to compare different works.

A more comprehensive deep learning model is also needed. Deep learning methods have become one of the main methods in NER and have achieved the best results. Compared with traditional machine learning methods, deep learning automatically mines the features in the text. In the cyber security domain, when facing more complex entity names and more diverse entity categories, it is necessary to continuously improve the existing deep learning model. The ongoing work is as follows: neural attention and deep adversarial learning have been applied in cyber security NER. By applying an attention mechanism, a NER model could capture the most informative elements in the inputs, and adversarial learning is used to make the model more robust to attack or to reduce its test error on clean inputs. More deep learning technologies will be used to improve the effect of the model and its recognition ability; for example, Multi-Task Learning (Caruana et

al., 1997), Active Learning (Shen et al., 2017), Transfer Learning (Lee et al., 2018), and Reinforcement Learning (Kaelbling et al., 1996).

Contributors

Chen GAO designed and organized the research. Xuan ZHANG guided the research and revised the manuscript. Hui LIU collected the relevant materials. Meng-ting HAN helped organize the manuscript.

Compliance with ethics guidelines

Chen GAO, Xuan ZHANG, Hui LIU and Meng-ting HAN declare that they have no conflict of interest.

References

Bridges RA, Jones CL, Iannacone MD, et al., 2013. Automatic labeling for entity extraction in cyber security. https://arxiv.org/abs/1308.4941

Caruana R, 1997. Multitask learning. Machine learning, 28(1): 41–75. https://doi.org/10.1023/A:1007379606734

Devlin J, Chang M, Lee K, 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. https://arxiv.org/abs/1810.04805

Dionísio N, Alves F, Ferreira PM, et al., 2019. Cyberthreat Detection from Twitter using Deep Neural Networks. 2019 International Joint Conference on Neural Networks, p.1-8. https://doi.org/10.1109/ijcnn.2019.8852475

Eddy SR, 1996. Hidden markov models. Current opinion in structural biology, 6(3): 361-365. https://doi.org/10.1016/s0959-440x(96)80056-x

Gasmi H, Bouras A, Laval J, 2018. LSTM Recurrent Neural Networks for cyber security Named Entity Recognition. Thirteenth International Conference on Software Engineering Advances.

Gu X, Liu J, Cheng F, et al., 2019. Tweet malware name recognition based on enhanced BILSTM-CRF model. Computer science, 47(2): 245-250 (in Chinese).

Georgescu T-M, Iancu B, Zurini M, 2019. Named-Entity-Recognition-Based Automated System for Diagnosing Cybersecurity Situations in IoT Networks, Sensors, 19(15):3380. https://doi.org/10.3390/s19153380

Hearst MA, Dumais ST, Osuna E, et al., 1998. Support vector machines. IEEE Intelligent Systems and their applications, 13(4): 18–28. https://doi.org/10.1109/5254.708428

Joshi A, Lal R, Finin T, 2013. Extracting cyber security related linked data from text. Seventh International Conference on Semantic Computing, p.252-259. https://doi.org/10.1109/icsc.2013.50

Kaelbling LP, Littman ML, Moore AW, 1996. Reinforcement learning: A survey. Journal of artificial intelligence research, 4:237–285.

Kim G, Lee C, Jo J, et al., 2020. Automatic extraction of named entities of cyber threats using a deep Bi-LSTM-CRF network. International Journal of Machine Learning and Cybernetics, 11(10):2341-2355. https://doi.org/10.1007/s13042-020-01122-6

Lafferty J, McCallum A, Pereira F, 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th International Conference on Machine Learning, p.282–289. https://dl.acm.org/doi/10.5555/645530.655813

Lowd D, Meek C, 2005. Adversarial learning. in Proc. SIGKDD, p.641–647. https://dl.acm.org/doi/abs/10.1145/1081870.1081950

Lal R, 2013. Information Extraction of Security related entities and concepts from unstructured text. MS Thesis, University of Maryland Baltimore County, Baltimore.

LeCun Y, Bengio Y, Hinton G, 2015. Deep learning. Nature, 521(7553):436-444. https://doi.org/10.1038/nature14539

Lample G, Ballesteros M, Subramanian S, et al., 2016. Neural architectures for name entity recognition. https://arxiv.org/abs/1603.01360

Lee JY, Dernoncourt F, Szolovits P, 2018. Transfer learning for named-entity recognition with neural networks. Proc. LERC 2018, P.4471-4473. https://doi.org/10.1093/bioinformatics/bty449

Long Z, Tan L, Zhou S, et al., 2019. Collecting Indicators of Compromise from Unstructured Text of Cybersecurity Articles using Neural-Based Sequence Labelling. 2019 International Joint Conference on Neural Networks, p.1-8.

Li T, Guo Y, Ju A, 2019. A Self-Attention-Based Approach for Named Entity Recognition in Cybersecurity. 2019 15th International Conference on Computational Intelligence and Security (CIS), p.147-150. https://doi.org/10.1109/cis.2019.00039

Liu W, 2020. Network Security Entity Recognition Methods Based on the Deep Neural Network. In Data Processing Techniques and Applications for Cyber-Physical Systems, p.1687-1692. https://doi.org/10.1007/978-981-15-1468-5_201

Mulwad V, Li W, Joshi A, T, et al., 2011. Extracting

information about security vulnerabilities from web text. IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 3:257-260. https://doi.org/10.1109/wi-iat.2011.26

Mendes PN, Jakob M, García-Silva A, et al., 2011. Dbpedia spotlight: shedding light on the web of documents. Proceedings of the 7th Inter-national Conference on Semantic Systems, p.1-8. https://doi.org/10.1145/2063518.2063519

Marrero M, Urbano J, Sánchez-Cuadrado S, et al., 2013. Named entity recognition: fallacies, challenges and opportunities. Computer Standards & Interfaces, 35(5): 482–489. https://doi.org/10.1016/j.csi.2012.09.004

McNeil N, Bridges RA, Iannacone MD, et al., 2013. Pace: Pattern accurate computationally efficient bootstrapping for timely discovery of cyber-security concepts. Machine Learning and Applications (ICMLA) 2013 12th International Conference on, 2:60-65. https://doi.org/10.1109/icmla.2013.106

Mazharov I, Dobrov BV, 2018. Named Entity Recognition for Information Security Domain. Data Analytics and Management in Data Intensive Domains, P.200-207.

Nadeau D, Sekine S, 2007. A survey of named entity recognition and classification. Lingvisticae Investigationes, 30(1):3–26. https://doi.org/10.1075/bct.19.03nad

Peters ME, Ammar W, Bhagavatula C, et al., 2017. Semi-supervised sequence tagging with bidirectional language models. https://arxiv.org/abs/1705.00108

Pingchuan M, Bo J, Zhigang L, et al., 2019. Cyber security Named Entity Recognition Using Bidirectional Long Short-Term Memory with Conditional Random Fields. Tsinghua Science and Technology, p.1-7.

Qin Y, Shen G, Zhao W, et al., 2019. A network security entity recognition method based on feature template and CNN-BiLSTM-CRF. Frontiers of Information Technology & Electronic Engineering, 20(6):872-884. https://doi.org/10.1631/fitee.1800520

Riloff E, 1993. Automatically constructing a dictionary for information extraction tasks. Proceedings of the Eleventh National Conference on Artificial Intelligence, p.811–816. https://dl.acm.org/doi/10.5555/1867270.1867391

Ruder S, 2016. An overview of gradient descent optimization algorithms. http://arxiv.org/abs/1609.04747

Roy A, Park Y, Pan S H, 2017. Learning domain-specific word embeddings from sparse cybersecurity texts.

Syed Z, 2010. Wikitology: A novel hybrid knowledge base derived from wikipedia.

Syed Z, Padia A, Mathews ML, et al., 2016. UCO: A Unified Cybersecurity Ontology. Proceedings of the AAAI Workshop on Artificial Intelligence for Cyber Security, p.14-21.

Shang H, Jiang R, Li A, et al., 2017. A framework to construct knowledge base for cyber security. 2017 IEEE Second International Conference on Data Science in Cyberspace, p.242-248. https://doi.org/10.1109/dsc.2017.55

Shen Y, Yun H, Lipton Z, et al., 2017. Deep active learning for named entity recognition. Proceedings of the 2nd Workshop on Representation Learning for NLP. https://arxiv.org/abs/1707.05928

Simran K, Sriram S, Vinayakumar R, et al., 2019. Deep Learning Approach for Intelligent Named Entity Recognition of Cyber Security. In International Symposium on Signal Processing and Intelligent Recognition Systems, p.163-172. https://arxiv.org/abs/2004.00502

Tikhomirov M, Loukachevitch N, Sirotina A, et al., 2020, Using BERT and Augmentation in Named Entity Recognition for Cybersecurity Domain[C]//International Conference on Applications of Natural Language to Information Systems, 12089:16-24. https://doi.org/10.1007/978-3-030-51310-8_2

Vaswani A, Shazeer N, Parmar N, et al., 2017. Atten tion is all you need, Advances in neural information processing systems, p.5998-6008. https://arxiv.org/abs/1706.03762

Weerawardhana S, Mukherjee S, Ray I, et al., 2014. Automated Extraction of Vulnerability Information for Home Computer Security. International Symposium on Foundations and Practice of Security, p.356-366. https://doi.org/10.1007/978-3-319-17040-4_24

Wu H, Li X, Gao Y, 2020. An Effective Approach of Named Entity Recognition for Cyber Threat Intelligence, 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), p.1370-1374. https://doi.org/10.1109/itnec48623.2020.9085102

Wang X, Xiong Z, Du X, et al., 2020. NER in Threat Intelligence Domain with TSFL, CCF

International Conference on Natural Language Processing and Chinese Computing, p. 157-169. https://doi.org/10.1007/978-3-030-60450-9_13

Xiao Z, 2018. Towards a two-phase unsupervised system for cybersecurity concepts extraction. 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery, p.2161-2168. https://doi.org/10.1109/fskd.2017.8393106

Zhou S, Long Z, Tan L, et al., 2018. Automatic identification of indicators of compromise using neural-based sequence labelling. Proc. PACLIC 2018 https://arxiv.org/abs/1810.10156

Zhang H, Guo Y, Li T, 2019. Multifeature Named Entity Recognition in Information Security Based on Adversarial Learning. Security and Communication Networks, 2:1-9. https://doi.org/10.1155/2019/6417407

主　题:	ZUSC: Your manuscript entitled A Review on Cyber Security Named Entity Recognition - [EMID:800027b38a658fca]
发件人:	"Yizhou Miao" <em@editorialmanager.com> 2021-2-1 9:23:33
收件人:	"Xuan Zhang" <zhxuan@ynu.edu.cn>

CC: fitee@zju.edu.cn, jzus_zzy@zju.edu.cn

Ref.: Ms. No. ZUSC-D-20-00286R2
A Review on Cyber Security Named Entity Recognition
Frontiers of Information Technology & Electronic Engineering

Dear Prof. Zhang,

I am pleased to tell you that your work has now been accepted for publication in Frontiers of Information Technology & Electronic Engineering.

Thank you for submitting your work to this journal.

With kind regards

Yizhou Miao
Editor-in-Chief
Frontiers of Information Technology & Electronic Engineering

—

Our flexible approach during the COVID-19 pandemic

If you need more time at any stage of the peer-review process, please do let us know. While our systems will continue to remind you of the original timelines, we aim to be as flexible as possible during the current pandemic.

This letter contains confidential information, is for your own use, and should not be forwarded to third parties.

Recipients of this email are registered users within the Editorial Manager database for this journal. We will keep your information on file to use in the process of submitting, evaluating and publishing a manuscript. For more information on how we use your personal details please see our privacy policy at https://www.springernature.com/production-privacy-policy. If you no longer wish to receive messages from this journal or you have questions regarding database management, please contact the Publication Office at the link below.

In compliance with data protection regulations, you may request that we remove your personal registration details at any time. (Use the following URL: https://www.editorialmanager.com/zusc/login.asp?a=r). Please contact the publication office if you have any questions.