Darknet Public Hazard Entity Recognition Based on Deep Learning

PanpanZhang[1,2], Xuebin Wang[1,2], Jing Ya[1], Jiapeng Zhao[1,2], Tingwen Liu[1,2], Jinqiao Shi[3]
[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[3] School of Cyberspace Security, Beijing University of Posts and Telecommunications, China
zhangpanpan@iie.ac.cn, wangxuebin@iie.ac.cn

ABSTRACT

Due to the strong protection of anonymity, Darknet has been exploited by criminals to distribute harmful content and banned items, such as drugs, weapons, and malware, which are regarded as public hazard entities. The task of public hazard entity recognition can help to detect and analyze malicious activities in the Darknet. This paper focuses on the research of Chinese public hazard entity recognition in the field of illegal drugs. To evade detection and surveillance, Chinese public hazard entities in the Darknet usually utilize disguised forms, like homophones and multi-entities in a sentence, which makes it harder to identify them using traditional entity recognition methods. In this paper, we present an effective deep learning-based multi-information fusion model to identify Chinese public hazard entities in the Darknet. Specifically, we introduce the grammatical information by adding Pinyin and lexical features, and strengthen the semantic features by adding the word vectors from one advanced pre-trained language representation model. Then we combine these three parts with a classical sequence annotation architecture used in general named entity recognition to form our ultimate model. At last we construct a real dataset from drugs-related groups in the Darknet and conduct several experiments to evaluate our model. The experimental result verifies that our proposed model gains a good performance on the recognition of Darknet public hazard entities.

KEYWORDS

Public hazard entity, Chinese entity recognition, Darknet, Deep learning, Multi-information fusion

ACM Reference Format:
PanpanZhang[1,2], Xuebin Wang[1,2], Jing Ya[1], Jiapeng Zhao[1,2], Tingwen Liu[1,2], Jinqiao Shi[3], [1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China, [2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China, [3] School of Cyberspace Security, Beijing University of Posts and Telecommunications, China, zhangpanpan@iie.ac.cn, wangxuebin@iie.ac.cn . 2021. Darknet Public Hazard Entity Recognition Based on Deep Learning. In *2021 ACM International Conference on Intelligent Computing and its Emerging Applications (ACM ICEA '21), December 28–29, 2021, Jinan, China.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3491396.3506525

ACM ICEA '21, December 28–29, 2021, Jinan, China

ACM ISBN 978-1-4503-9160-3/21/12.
https://doi.org/10.1145/3491396.3506525

1 INTRODUCTION

Darknet is an overlay network based on anonymous communication technologies, which can only be accessed with specific softwares, configurations, or authorization [1]. Due to the strong protection of anonymity, Darknet has been extensively abused by criminals to implement malicious activities such as selling drugs and distributing terrorism, posing a serious threat to social stability and cyber security [2–5]. Consequently, monitoring malicious activities on the Darknet has been an urgent issue. Public hazard entities (such as drugs, weapons, and malware) are usually mentioned in malicious activities and are an important part of malicious activities. The task of public hazard entity recognition is to identify those public hazard-related entities from unstructured content in the Darknet. It helps to detect and analyze malicious activities and plays a key role in Darknet surveillance. Recognizing public hazard entities can help security departments to make early warnings of dangerous information related to terrorism and gangs, but also help researchers to understand and analyze the specific content of malicious activities on the Darknet. For example, recognition of public hazard entities in a Telegram group of drugs can help to distinguish whether a user is a drug dealer, and also can help to understand the content of messages in drug groups.

In fact, public hazard entities recognition is a domain-specific entity identification task in named entity recognition (NER) [6]. Compared with general NER, which recognizes persons, locations and organizations, public hazard entity recognition is more professional and complex, with smaller recognition range and higher recognition accuracy. It aims to identify corresponding public hazard entities in text data according to a specific field. In this article, we will focus on the recognition of Chinese public hazard entities and take the field of illegal drugs as an example for specific research.

As a basic task of natural language processing (NLP), named entity recognition has a lot of related research. They can be classified into three kinds: rule-based approaches, statistical learning-based approaches, and deep learning-based approaches. The rule-based approaches usually use manually defined rules to identify entities and do not need annotated data, such as the LTC system [7] which formulates a series of grammatical rules for named entity recognition and the ProMiner [8] that uses pre-processed dictionaries to identify "protein" and "gene" entities in the field of medicine. The statistical learning-based approaches usually use sequence labeling methods based on various language model and machine learning

to recognize entities, such as Maximum Entropy (ME) [9], Hidden Markov Models (HMM) [10], Support Vector Machine (SVM) [11], and Conditional Random Fields (CRF) [12]. The deep learning-based approaches [13, 14] utilize deep neural networks and deep learning methods to automatically extract related features in the text and finish the entity recognition in an end-to-end manner. For example, Lample *et al.*[15] presents a neural framework named BiLSTM-CRF, which uses a BiLSTM layer (short for Bi-directional Long Short-Term Memory) and a CRF layer to identify named entities in four languages. Compared with the former two, the deep learning-based approaches have better generalization ability and automatic feature learning ability, and do not require much domain knowledge and engineering skills. Therefore they are the mainstream researches of NER today. In our study, we follow the deep learning-based approaches and utilize the abovementioned BiLSTM-CRF model as the baseline and the base network structure.

Due to the particularity of Darknet, it is tough to identify public hazard entities in the Darknet accurately by applying ordinary NER deep learning approaches directly. First of all, there is a lack of large-scale labeled data of public hazard entities in the Darknet. Such a situation makes general NER models can not be well trained to identify public hazard entities effectively. For the domain-specific task, we have to construct a specific dataset to train and evaluate models for public hazard entity recognition. Moreover, Chinese public hazard entities in the Darknet usually utilize disguised forms to evade detection, which makes it harder than traditional entity recognition tasks. The disguised forms are always reflected in the textual expressions, such as homophones and abbreviations. Using these tricks, users can better protect themselves in the network environment and avoid supervision. However, it brings challenges to the Chinese public hazard entity recognition task, which makes it more difficult for simple deep learning-based models to locate public hazard entities. Overall, compared with traditional NER tasks, it is more challenging for Chinese public hazard entity recognition in the Darknet.

To tackle the above problems, we propose a effective deep learning-based model with multi-information fusion to identify Chinese public hazard entities in the Darknet. The model firstly introduces the semantic and grammatical information together to get the comprehensive representation of the input sentence. To be specific, we apply Pinyin and Part of Speech (POS) tags of words as auxiliary information to enhance semantic understanding and entity location judgment. These two measures can effectively alleviate the recognition difficulties caused by homophones and multiple entities. Meanwhile, we employ the advanced pre-trained language representation model named BERT (short for Bidirectional Encoder Representation form Transformers) to obtain the entire sentence embedding vectors. At last, the model combines these three parts with the BiLSTM-CRF sequence annotation model to mine potential public hazard entities. Because the BiLSTM-CRF model has been proved to be the mainstream model and achieves a good performance for NER tasks [16–18].

To verify the effectiveness of our model, we construct a real dataset from drugs-related groups in the Darknet and conduct comparative experiments. The experimental results indicate that our model is better than the baseline models in identifying Chinese public hazard entities.

The contributions of this paper are summarized as follows:

- We present a effective deep learning-based model with multiple information fusion for Chinese public hazard entity recognition in the Darknet. The model can not only effectively exploit the semantic and grammatical information, but also capture the long dependencies of the entire sentence.
- We introduce Pinyin and POS tags of the words as auxiliary information to enhance semantic understanding and entity location judgment, for addressing the problems of homophones and multiple entities in a long sentence.
- We construct a real dataset and conduct extensive experiments. And the results show that our proposed model can identify Darknet Chinese public hazard entities significantly.

In the rest of paper, we describe our approach detailedly in Section 2. And then the implementation details and experimental results are present in Section 3. Finally, we summarize the related work and conclude our work in Section 4 and Section 5.

2 CHINESE PUBLIC HAZARD ENTITY RECOGNITION MODEL

2.1 Problem Statement

In our work, the task of Chinese public hazard entity recognition in the Darknet is to identify potential illegal drug entities from Chinese text data generated by Darknet users. Referring to the universal NER method, we also turn this task into a sequence labeling problem. And we apply the BIO tagging scheme, where "B" indicates the beginning of the entity, "I" indicates the middle or end of the entity, and "O" indicates other words without target information. For eradicating the influence of Chinese word segmentation error propagation, we use Chinese character-level BiLSTM-CRF as our fundamental structure.

We take sentences as inputs to our model. Formally, let $S = (w_1, w_2, w_3, ..., w_n)$ denotes an input sentence, where w_i denotes the i-th word of the input, and n denotes the sentence length. Every word has a corresponding BIO tag in the input, which is labeled "B-D", "T-D" or "O".

2.2 Model

The overview architecture of our proposed model is shown in Fig. 1. The whole model includes 4 layers, namely the input layer, the embedding layer, the BiLSTM layer and the CRF layer.

The input layer contains three kinds of feature sequences: textual word sequence, Pinyin sequence, and POS sequence, respectively. The embedding layer obtains the corresponding embedding vectors by the pre-trained BERT model or transformation matrices. The BiLSTM layer uses the representation containing the above three features for global feature extraction. Finally, the output of the BiLSTM layer that combines semantic and grammatical information is inputted into the CRF layer for the global optimal label annotation. Next, we will detailedly introduce each layer of our model.

2.3 Input Layer

As a sequence labeling task, NER needs to recognize each segment in the sentence. In general, the solution for NER is firstly locating

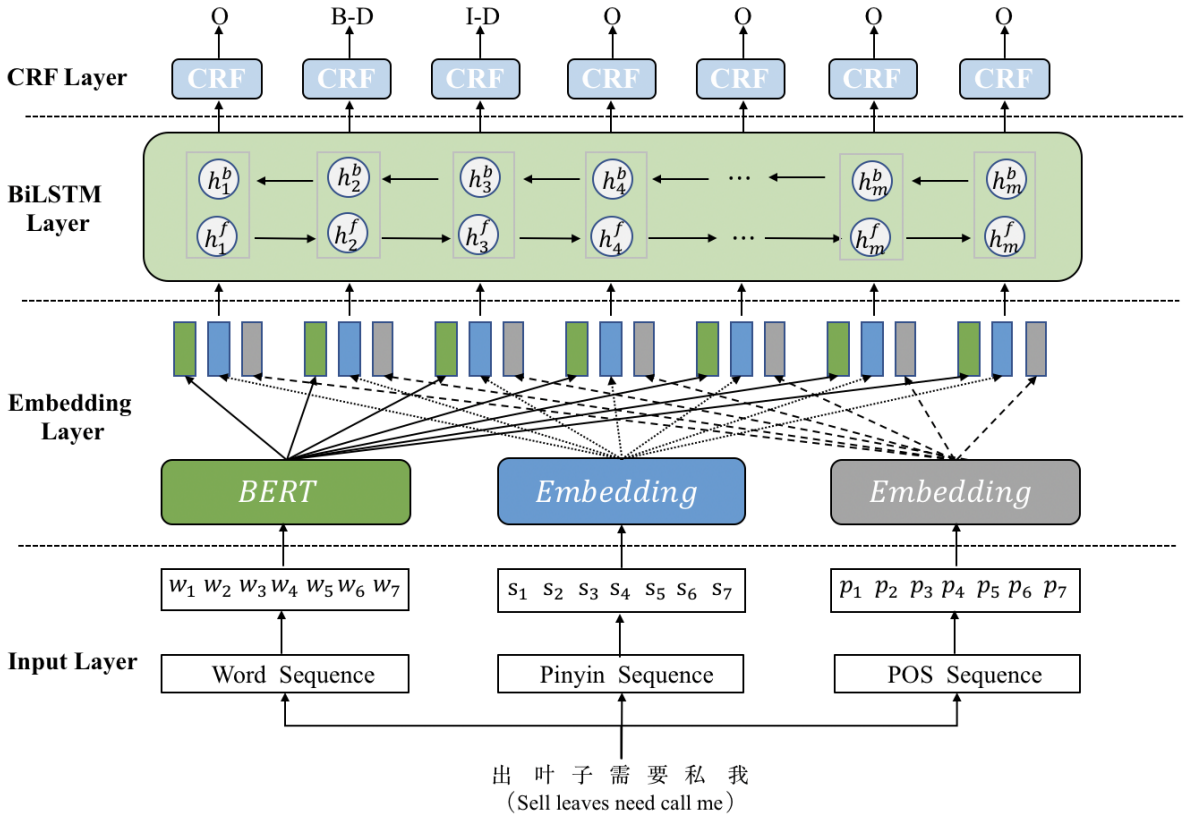

Figure 1: The overview architecture of our model

potential entity mentions, by means of the sentence's syntax information. And then according to the semantic information of the context, it is easy to judge whether the categories of the potential entities are related to the target category.

In the process of Darknet monitoring, we find that in some forums or chat rooms on public hazard topics, the user's language expression is usually informal and irregular. There are two specific expressions, named homophonic expression and multi-entities expression. As shown in Table 1, "库什" and "库士" all mean "Kush", although they are different words in Chinese. And the third sentence S3 has multiple entities simultaneously because it is a common advertisement. These special measures can protect users to hide as much as possible, and avoid being supervised. However, they have brought inconvenience to the public hazard entity recognition task, making it more difficult to locate the location of the entity and to identify the target entity. Moreover, these challenges also affect the generalization ability of the NER models.

Table 1: Special expressions

Examples	Types	Details
S1: 出og 库什, 叶子 S1: Sell og Kush, Leaves	Homophone	库什,库士 Kush
S2:最好的库士, 需要私我 S2: Best Kush, need call me		
S3: 健康菜单:CBD油, CBD糖浆, CBD pen, 纯cbd电子烟, cbd 烟丝, 货物当天下午三点前付款,顺丰发售 S3: Health menu: CBD oil, CBD syrup, CBD pen, pure CBD e-cigarettes and CBD cut, tobacco will be paid before 3 p.m. and sold by SF	Multi-entities	CBD油, CBD糖浆, CBD pen, 纯cbd电子烟, cbd 烟丝 CBD oil, CBD syrup, CBD pen, CBD e-cigarettes, CBD cut

To solve the above problems, we propose three strategies for improving the ability to model the syntactic and semantic information about the input sentence. For the homophonic expression problem, we add Pinyin of Chinese words as auxiliary information, to enhance the semantic understanding of words. For the multi-entities expression problem, we utilize the POS tagging tool to tag words and obtain the syntactic feature, which helps to mine locations of potential entities. For the robustness of our model, we introduce BERT, which is a classic pre-trained language model to get the comprehensive representation and improve the ability to model the whole sentence (details in the embedding layer).

In practice, we use the public Python library *pypinyin* to get the Pinyin expression of Chinese words. Take the sentences of "最好的库什" and "最好的库士", meaning "Best Kush" in English as

examples, we can get their corresponding Pinyin expressions are "zui hao de ku shen" and "zui hao de ku shi". Observation shows that although their Pinyin of "ku shen" and "ku shi" are somewhat different, they have the same initial part of " k sh". Therefore, we select the initials of Chinese words as the feature of the pronunciation and use them in our model.

To obtain the POS features of the input sentence, we employ the public tool *Jieba*. It is a well-known Chinese word segmentation tool, that not only has the function of word segmentation, but also can tag words as part-of-speech. The POS features can be used as auxiliary information for entity location judgment. For example, nouns and prepositions often appear at the boundary of entities. When getting each token's POS tags, we assign them to each word in the token.

2.4 Embedding Layer

Our embedding layer contains two kinds of embedding: syntactic information embedding and semantic information embedding.

2.4.1 Syntactic Embedding. Given the word sequence of the input sentence $S = (w_1, w_2, w_3, ..., w_n)$, we can get the initial sequence of Pinyin $I = (s_1, s_2, s_3, ..., s_n)$ and the POS tags sequence $P = (p_1, p_2, p_3, ..., p_n)$, as described in the above layer. They are both auxiliary information that helps for the entity location judgment. Then, we use the embedding part to obtain a distributed feature representation of the discrete characters. The vector representations are generated as follows:

$$e_{s_i} = Embedd(I) = W^I \cdot I + b_I \tag{1}$$

$$e_{p_i} = Embedd(P) = W^P \cdot P + b_P \tag{2}$$

where s_i and p_i are the i-th character in the sequences I and P, W^I and W^P denotes the corresponding embedding matrix weights, and b_I and b_P are the biases.

2.4.2 Semantic Embedding. In order to strengthen the understanding of the overall semantics of the input sentence, we use the BERT model to acquire the encoding of the word sequence. BERT is a pre-trained representation model based on fine-tuning, which combines the strengths of GPT (short for Generative Pre-Training) [19] and ELMo (short for Embeddings from Language Models) [20]. It generates the word embedding by using transformer [21] and attention mechanism [22] and considering the contextual semantic information. Take the word sequence $S = (w_1, w_2, w_3, ..., w_n)$ as the input, the output representation of BERT is as follows:

$$e_{w_i} = BERT(S) \tag{3}$$

where w_i is the i-th word in the sequences S.

Through syntax analysis and semantic representation, we obtain three different meanings of word vectors. To input into the BiLSTM layer for further fusion representation, these three vectors are concatenated together as E_{w_i}, described as follows:

$$E_{w_i} = [e_{s_i} \oplus e_{p_i} \oplus e_{w_i}] \tag{4}$$

where the $\oplus$ is the connection operator.

2.5 BiLSTM Layer

Long short-term memory (LSTM) is a classical Recurrent Neural Network (RNN), and it addresses the problems of the vanishing and exploding of gradients in traditional RNNs [23]. It uses the adaptive gating mechanism and the memory cell to better capture long-distance dependencies. Figure 2 shows the structure of LSTM. A basic LSTM unit contains an input gate i, a forget gate f, an output gate o and a memory cell c, to control the selection and forgetting of information, defined as follows:

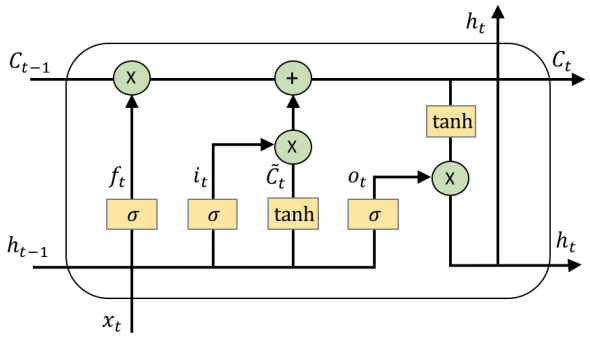

Figure 2: The structure of LSTM

$$\mathbf{i}_t = \delta\left(\mathbf{W}_i\mathbf{x}_t + \mathbf{U}_i\mathbf{h}_{t-1} + \mathbf{b}_i\right) \tag{5}$$

$$\mathbf{f}_t = \delta\left(\mathbf{W}_f\mathbf{x}_t + \mathbf{U}_f\mathbf{h}_{t-1} + \mathbf{b}_f\right) \tag{6}$$

$$\tilde{\mathbf{c}_t} = \tanh(\mathbf{W}_c\mathbf{x}_t + \mathbf{U}_c\mathbf{h}_{t-1} + \mathbf{b}_c) \tag{7}$$

$$\mathbf{c}_t = \mathbf{f}_t * \mathbf{c}_{t-1} + \mathbf{i}_t * \tilde{\mathbf{c}_t} \tag{8}$$

$$\mathbf{o}_t = \delta\left(\mathbf{W}_o\mathbf{x}_t + \mathbf{U}_o\mathbf{h}_{t-1} + \mathbf{b}_o\right) \tag{9}$$

$$\mathbf{h}_t = \mathbf{o}_t * \tanh\left(\mathbf{c}_t\right) \tag{10}$$

where δ and $*$ are the element-wise sigmoid function and product, $\mathbf{W}_{(\cdot)}$ and $\mathbf{U}_{(\cdot)}$ are the weights matrices of the input and recurrent connections, and $\mathbf{b}_{(\cdot)}$ are the biases.

The unidirectional LSTM can only retain the related information from the past sequence, but in the task of sequence labeling, the output at the current time is not only related to the previous sequence, but may also be related to the future sequence. Therefore, we use a Bidirectional LSTM (BiLSTM) to get the contextual features of the sentence [24]. In the BiLSTM layer, a forward LSTM and another backward LSTM are used at the same time, and the comprehensive representation of the sequence is calculated from two opposite directions. Consequently, we represent the BiLSTM hidden state $\mathbf{h}_t$ as below:

$$\mathbf{h}_t = \left[\overrightarrow{\mathbf{h}_t}, \overleftarrow{\mathbf{h}_t}\right] \tag{11}$$

2.6 CRF Layer

The traditional NER methods use the output layer of neural networks to make labeling decisions independently. In this way, the tags of words are regarded as independent. In fact, the tags of adjacent words have a strong dependency. For instance, the beginning word of a sentence can only be marked as "O" or "B-D", but not as "I-D". And the word marked "I-D" cannot appear alone, and must follow the word marked "B-D". Therefore, in the CRF layer, we use the CRF model to jointly infer the entity tag output of the sequence. Note that CRF is an undirected graphical model, and it emphasizes the sentence level rather than individual position of a word.

Given a sentence $X = w_1, w_2, ..., w_n$, its most possible tag sequence $y = y_1, y_2, ..., y_n$ can be predicted using the CRF layer. We

consider the output of the BiLSTM layer is the matrix of scores **P**. The matrix element $P_{i,j}$ is the score of the j-th tag of the i-th word in the sentence. So we can calculate the score of the tag sequence as follows:

$$s(X, y) = \sum_{i=0}^{n} A_{y_i, y_i+1} + \sum_{i=1}^{n} P_{i, y_i} \tag{12}$$

where **A** is the transition score matrix and $A_{i,j}$ denotes the transition score from tag i to tag j. Note that y_0 and y_n denote the start and stop tags of the sentence.

Then we use the softmax function to normalize the transition score of all likely tag sequences, and generate the conditional probability of the sequence y :

$$p(y|X) = \frac{e^{s(X,y)}}{\sum_{\tilde{y} \in Y_x} e^{s(X,\tilde{y})}} \tag{13}$$

where $\tilde{y}$ is an optional tag sequence, and Y_x is the collection of all likely output tag sequences for the input X.

In the process of training, we choose to maximize the log-probability of the correct tag sequence as the objective of the model. In the process pf decoding, we use the Viterbi algorithm [25] to predict the best tag sequence that will obtain the largest score, as shown below:

$$\bar{y} = argmax_{\tilde{y} \in Y_x} s(X, \tilde{y}) \tag{14}$$

3 EXPERIMENTS

Next, we first introduce our experimental setup, including the constructed dataset, validation metrics and training parameters. Then we evaluate the effectiveness of our model on Chinese public hazard entity recognition in the Darknet and make a comparison with the baseline approaches.

3.1 Experimental Setup

1) Dataset: Since there is currently no available Chinese Darknet public hazard entities corpus, we manually construct a real dataset. The data is crawled from drug-related groups in the Darknet. Based on the collected conversation data, we manually label public hazard entities with the BIO tagging scheme, where each word in a sentence is marked as one of "B-D", "I-D", and "O". In order to ensure the accuracy of annotation, three annotators annotate the data at the same time, and inconsistent sentences will be annotated again. Table 2 displays some examples from the annotated dataset. It should be noted that the blue words in each sentence are the parts of specific drugs, and the black words are non-entity parts.

Table 2: Annotated examples

实力出, 支持各种验证, 菠萝, 北极光 , 白寡妇 Power out, support various verifications, pineapple, northern lights, white widow.
谁知道蘑菇现在多少钱? Who knows how much are mushrooms now ?
出草,lsd , 猪肉 , 需要私聊, 支持担保 Sell leaves, lsd and pork, if need call me, support verifications.

In our study, we finally annotated a dataset that contains 12175 sentences. According to experience, we select 75% of the dataset as the training set and the remaining 25% as the test set to evaluate the performance of our model. For further evaluating the effectiveness of our model in solving the problems of the homophonic expression and multi-entities expression, we also use two specific corresponding subsets. Specifically, we select three typical examples of the homonymic entities, that are "库士/库什(Kush)", "海洛因/海洛英(Heroin)", and "bong/泵(Pump)" respectively. For the multi-entities judgment, we define a criterion that whether the number of entities contained in a sentence is greater than or equal to 5. Table 3 describes the details of our dataset.

Table 3: Datasets statistics

Types	Training set	Test set
Sentences	9238	2397
Entities	6818	4391
Homonymic entities	#	168
Multi-entities sentences	#	228
Multi-entities entities	#	1952

2) Metrics: We use the three metrics: Precision (P), Recall (R), and F1 score (F1) to evaluate the effectiveness of our model. Precision and recall are important evaluating indicators, but they are contradictory. The F1 score is the harmonic mean of the two. The closer the F1 score is to 1, the better the performance of the model.

3) Others: The pre-trained model BERT we used is Google's open-source model: BERT-Base, Chinese. It has 12 encoder layers, 768 hidden units and 12-head mode with a total of 110M parameters.

During the training, we select the hidden size and number of layers of BiLSTM to 300 dims and 1 layer, and the batch size of training and testing to 32 and 8. And we exploit Adam as the model optimization with an initial learning rate of 1e-5. To avoid overfitting, we use the dropout rate of 0.5.

3.2 Experiment results

In order to verify the performance of our model, we use the classical BiLSTM-CRF as the baseline model. The second comparable model is BERT-BiLSTM-CRF, which firstly uses the BERT model to get semantic embedding, and then adds this output to the baseline model. Our proposed model is Syntax-BERT-BiLSTM-CRF, which adds semantic features based on the BERT model as well as two kinds of syntax features based on Pinyin and POS information.

The final results of all models are listed in Table 4, and the best performance is in bold. It is obvious that our proposed model Syntax-BERT-BiLSTM-CRF achieves the best results, with the precision rate, recall rate and F1 score are 89%, 79% and 84%, respectively. And our model significantly outperforms the baseline model of BiLSTM-CRF, with improvements of 1%, 6% and 4% in the three metrics. This result not only confirms the effectiveness of our Chinese public hazard entities recognition model, but also confirms the rationality and feasibility of adding semantic and syntactic information to public hazard entity recognition.

Table 4: Experimental results of different models

Models	P	R	F1
BiLSTM-CRF	0.88	0.73	0.80
BERT-BiLSTM-CRF	0.88	0.78	0.83
Syntax-BERT-BiLSTM-CRF	**0.89**	**0.79**	**0.84**

Compared with the BiLSTM-CRF model, the BERT-BiLSTM-CRF model has better results in the recall rate and F1 score, and the specific increases are 5% and 3%. We analyze the results and find that the reason is the BERT model can improve the efficiency of entity recognition by capturing more semantic information. Because the BERT model has a large number of pre-training corpus and a complex network structure, it provides a strong ability to capture semantic features.

Compared with the BERT-BiLSTM-CRF model, our proposed model adds additional syntax information, such as Pinyin features and part of speech features. These two kinds of features are useful to solve informal and irregular expressions in the Darknet. As expected, the performance of our model is better than that of the BERT-BiLSTM-CRF model. This also reflects adding appropriate external features is conducive to improve the recognition result.

In addition to directly comparing the overall performance of the models, we also analyze the identification results of these models in the two special expressions mentioned above. As shown in Table 5, they are the entity recognition rates of different models on the homophonic entities and multiple entities test sets.

Table 5: Results on two expressions

Models	Homophone	Multi-entities
BiLSTM-CRF	0.53	0.60
BERT-BiLSTM-CRF	0.65	0.67
Syntax-BERT-BiLSTM-CRF	**0.85**	**0.71**

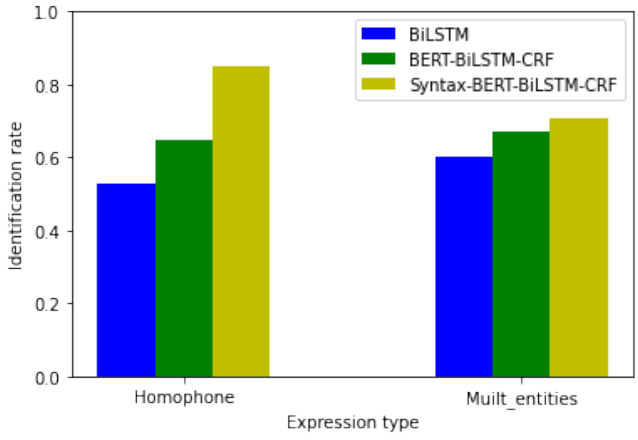

Figure 3: Comparison on two special expressions

The intuitive comparison is shown in Fig. 3 below It is obvious that our model has the highest recognition rate on homophonic entities. This is because we observe the homophonic entities have the same initial consonant, and introduce the Pinyin information to our model. Moreover, our model also has the best performance on multiply entities identification, which also benefits from the addition of the syntactic information.

4 RELATED WORK

In this section, we will briefly introduce the related work of named entity recognition based on deep learning and the basic principles of the BERT model.

In spired by the advancement of neural networks, there have been many deep learning-based NER approaches on multiple domains. Dong *et al.*[26] proposed a character-based LSTM-CRF model with radical-level features to identify Chinese named entities. Zeng *et al.*[27] proposed a system based on LSTM-CRF model to recognize biomedical named entities. Zhang *et al.*[28] proposed a lattice-structured LSTM model to identify candidate lexicon words from the sentence. Luo *et al.*[29] proposed an attention-based BiLSTM-CRF approach to recognize chemical named entities. Gao *et al.*[30] proposed a data and knowledge-driven NER approach for cyber security.

BERT is an advanced pre-trained language model proposed by Google [31], which successfully achieves impressive results in 11 NLP tasks that year. The overall framework of BERT includes two stages, namely the pre-training stage and fine-tuning stage. In the first stage, the model is trained on the unlabeled labeled data and in the second stage, the model is trained all initialized parameters from the pre-train model on the downstream labeled data. The BERT model uses a bidirectional transformer neural network [21] as the encoder, which can further enhance the generalization ability of the pre-training model and fully describe the relationship features between characters, words, and sentences. Jawagar *et al.*[32] conducted further research on the internal mechanism of the BERT model and found out that the features learned in each layer of the BERT model are different.

5 CONCLUSION

In this paper, we present the task of public hazard entity recognition in the Darknet, which can help to discover malicious activities and play an important role in Darknet monitoring. And we present a simple yet effective model based on deep learning, which combines syntactic and semantic information to identify Chinese public hazard entities in the field of illegal drugs. In oder to overcome the problems of the homophonic and multi-entities expressions, we specially introduce Pinyin and POS information into our model. And we employ the BERT model to capture more semantic information. Finally, we utilize the classical BiLSTM-CRF model as the base network structure to recognize public hazard entities. To evaluate the effectiveness of our model, we construct a real dataset from drug-related groups and conduct several experiments. The results validate that our model has the best performance than the baseline models. Further analysis also shows that the syntactic and semantic information we proposed arr very effective.

REFERENCES

[1] Luo Junzhou, Yang Ming, Ling Zhen, Wu Wenjia, and Gu Xiaodan. Anonymous communication and darknet: A survey. *Journal of Computer Research and Development*, 56(1):103, 2019.

[2] Joe Van Buskirk, Sundresan Naicker, Raimondo Bruno, Lucy Burns, C Breen, and Amanda Roxburgh. Drugs and the internet. *Drugs and the Internet.[7 ed.]*, pages 1–14, 2016.

[3] Roderic Broadhurst, Jack Foye, Chuxian Jiang, and Matthew Ball. Illicit firearms and weapons on darknet markets. *Broadhurst, R, J. Foye, J, Jiang and M. Ball, Illicit Firearms and Weapons on Darknet Markets, Trends and Issues in Criminal Justice, AIC Cannberra, Forthcoming*, 2020.

[4] George Hurlburt. Shining light on the dark web. *IEEE Computer Architecture Letters*, 50(04):100–105, 2017.

[5] Matthias Schäfer, Markus Fuchs, Martin Strohmeier, Markus Engel, Marc Liechti, and Vincent Lenders. Blackwidow: Monitoring the dark web for cyber security information. In *2019 11th International Conference on Cyber Conflict (CyCon)*, volume 900, pages 1–21. IEEE, 2019.

[6] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[7] Andrei Mikheev, Marc Moens, and Claire Grover. Named entity recognition without gazetteers. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–8, 1999.

[8] Daniel Hanisch, Katrin Fundel, Heinz-Theodor Mevissen, Ralf Zimmer, and Juliane Fluck. Prominer: rule-based protein and gene entity recognition. *BMC bioinformatics*, 6(1):1–9, 2005.

[9] Jagat Narain Kapur. *Maximum-entropy models in science and engineering*. John Wiley & Sons, 1989.

[10] Sean R Eddy. Hidden markov models. *Current opinion in structural biology*, 6(3):361–365, 1996.

[11] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998.

[12] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[13] Vikas Yadav and Steven Bethard. A survey on recent advances in named entity recognition from deep learning models. *arXiv preprint arXiv:1910.11470*, 2019.

[14] Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[15] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016.

[16] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.

[17] Yanliang Jin, Jinfei Xie, Weisi Guo, Can Luo, Dijia Wu, and Rui Wang. Lstm-crf neural network with gated self attention for chinese ner. *IEEE Access*, 7:136694–136703, 2019.

[18] Wenchao Gao, Xiaohui Zheng, and Shanshan Zhao. Named entity recognition method of chinese emr based on bert-bilstm-crf. In *Journal of Physics: Conference Series*, volume 1848, page 012083. IOP Publishing, 2021.

[19] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

[20] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

[21] Vaswani Ashish, Shazeer Noam, Parmar Niki, Uszkoreit Jakob, Jones Llion, N. Gomez Aidan, and Polosukhin Illia. Attention Is All You Need. In *Proc. of NIPS*, pages 5998–6008, 2017.

[22] Chongyang Tao, Shen Gao, Mingyue Shang, Wei Wu, Dongyan Zhao, and Rui Yan. Get the point of my utterance! learning towards effective responses with multi-head attention mechanism. In *IJCAI*, pages 4418–4424, 2018.

[23] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[24] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610, 2005.

[25] G David Forney. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278, 1973.

[26] Chuanhai Dong, Jiajun Zhang, Chengqing Zong, Masanori Hattori, and Hui Di. Character-based lstm-crf with radical-level features for chinese named entity recognition. In *Natural Language Understanding and Intelligent Applications*, pages 239–250. Springer, 2016.

[27] Donghuo Zeng, Chengjie Sun, Lei Lin, and Bingquan Liu. Lstm-crf for drug-named entity recognition. *Entropy*, 19(6):283, 2017.

[28] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International conference on machine learning*, pages 933–941. PMLR, 2017.

[29] Ling Luo, Zhihao Yang, Pei Yang, Yin Zhang, Lei Wang, Hongfei Lin, and Jian Wang. An attention-based bilstm-crf approach to document-level chemical named entity recognition. *Bioinformatics*, 34(8):1381–1388, 2018.

[30] Chen Gao, Xuan Zhang, and Hui Liu. Data and knowledge-driven named entity recognition for cyber security. *Cybersecurity*, 4(1):1–13, 2021.

[31] Devlin Jacob, Chang Ming-Wei, Lee Kenton, and Toutanova Kristina. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. of NAACL-HLT*, pages 4171–4186, 2019.

[32] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*, 2019.