Research Article

The Named Entity Recognition of Chinese Cybersecurity Using an Active Learning Strategy

Bo Xie, Guowei Shen, Chun Guo, and Yunhe Cui

1College of Computer Science and Technology, Guizhou University, Guiyang 550025, China
2Guizhou Provincial Key Laboratory of Public Big Data, Guiyang 550025, China

Correspondence should be addressed to Guowei Shen; gwshen@gzu.edu.cn

Received 25 December 2020; Accepted 12 March 2021; Published 21 April 2021

Academic Editor: Zhili Zhou

Copyright © 2021 Bo Xie et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In data-driven big data security analysis, knowledge graph-based multisource heterogeneous threat data organization, association mining, and inference analysis attach increasing interest in the field of cybersecurity. Although the construction of knowledge graph based on deep learning has achieved great success, the construction of a large-scale, high-quality, and domain-specific knowledge graph needs a manual annotation of large corpora, which means it is very difficult. To tackle this problem, we present a straightforward active learning strategy for cybersecurity entity recognition utilizing deep learning technology. BERT pre-trained model and residual dilation convolutional neural networks (RDCNN) are introduced to learn entity context features, and the conditional random field (CRF) layer is employed as a tag decoder. Then, taking advantages of the output results and distribution of cybersecurity entities, we propose an active learning strategy named TPCL that considers the uncertainty, confidence, and diversity. We evaluated TPCL on the general domain datasets and cybersecurity datasets, respectively. The experimental results show that TPCL performs better than the traditional strategies in terms of accuracy and F1. Moreover, compared with the general field, it has better performance in the cybersecurity field and is more suitable for the Chinese entity recognition task in this field.

1. Introduction

In the increasingly complex situation of the cyberspace security situation, threat intelligence-driven cybersecurity defense has become the focus of the industry [1]. From the massive fragmented network data to mine threat intelligence, the use of knowledge graph model on organization, support attack path prediction, attack traceability, etc. can realize the massive data-driven threat intelligence analysis [2].

The cybersecurity entity recognition is a very important basic task in the construction of a threat intelligence knowledge graph [3]; its goal is to extract the semantic classes of cybersecurity entities including attack organizations, enterprises, vulnerabilities, and software from text data in the field of cybersecurity.

Compared with the development of entity recognition in other fields, the development of Chinese entity recognition in the field of cybersecurity is relatively slow. We believe that there are two main reasons. The first main reason is that the entity itself has the following challenges: (1) the identification of network security entities is relatively complex and contains a large number of mixed security entities and (2) inconsistent text. That is, the same entity in an article, paragraph, or long sentence is given different category labels. In the existing work, Qin et al. proposed that it is difficult for the traditional named entity recognition method to identify Chinese cybersecurity entities, and proposed a character-level CNN-BiLSTM-CRF cybersecurity entity recognition model based on feature templates [4]. The advantage of the method proposed by Qin et al. is to prevent the error propagation caused by Chinese word segmentation. The disadvantage of Qin et al.’s method is that due to the limitations of CNN and BiLSTM, only the local features of the entity can be extracted and the long-range dependency relationship between the entities cannot be solved. In addition, the feature template design is complex and not universal. To do ends, we
proposed a method (BERTRDCNN-CRF) in our previous work [5] which is based on a residual dilated convolutional neural network and the BERT model [6].

The second main reason for the slow development of Chinese entity recognition in the field of cybersecurity is that deep learning relies on large-scale labeled data, but there is a lack of large-scale and high-quality cybersecurity entities. This causes a lot of manpower and material resources when labeling the dataset, so the cost of labeling is very high. In addition, the Chinese security entity recognition needs to have a wealth of cybersecurity knowledge background. In the existing work on this problem, some researchers used the CRF model [7–9] to compare the performance of different types of active learning selection strategies in NER tasks. Although these methods have achieved good entity recognition results on the English dataset, they do not perform well in the field of Chinese cybersecurity. Specifically, we will discuss them in detail in the comparative experiment section of this article. In these different types of active learning methods, all are modeled based on textual information. Shen et al. believe that information is determined by the interpretation of uncertainty, representativeness, and diversity [10].

The above methods are all explanations of the uncertainty and representativeness of the information.

Shen et al. confirmed that in the named entity tasks, the fusion of active learning and deep learning can reduce model cost while enhancing model learning [11]. Therefore, Shen et al. propose an easy active learning method, adopt an easy selection strategy to select a large amount of unlabeled data, and then send the selected data to humans for annotation. This method greatly reduces the amount of data annotation and is able to nearly match state-of-the-art performance with just 30% of the original training data. Contrary to the work of uncertainty or excessive importance emphasized by Claveau and Kijak [8], we improve the three standards of information by integrating Shen et al.’s method. Thus, we propose an active learning strategy which considers uncertainty, confidence, and diversity, called TPCL, which considers not only the output results but also the distribution of cybersecurity entity-based lexicon. The main contributions of this paper are summarized as follows:

1. Integrate active learning strategies and deep learning frameworks to solve the problems of complex structure and inconsistent full text of Chinese cybersecurity entities.

2. A selection strategy is proposed to make the task of recognizing Chinese cybersecurity entities better and reduce the cost of annotation cybersecurity entities.

The remainder of this paper is organized as follows. In the section background, we summarize the related works in cybersecurity entity recognition and active learning. In the section of the proposed method, we introduce our motivation and the active learning strategy framework and describe in detail the active learning selection strategy we propose. Section experiments describe the datasets and experimental setting and discuss the empirical results.

2. Background

2.1. Cybersecurity Entity Recognition. Cybersecurity entity recognition belongs to a specific field of named entity recognition. Georgescu et al. enhanced and detected possible vulnerabilities within an internet of things (IoT) systems by using a named entity recognition solution [12]. Gasmi et al. used a bidirectional long short-term memory network and conditional random fields to identify cybersecurity entities [13]. The method does not rely on any features specific to the entities in the cybersecurity domain and hence does not require expert knowledge to perform feature engineering. However, the method proposed by Gasmi et al. requires a large amount of labeled datasets. Due to the particularity and complexity of cybersecurity named entities, these models ignore the characteristics of cybersecurity data and the correlation between entities. Yi et al. proposed a cybersecurity named entity recognition model with regular expressions, a known entity lexicon and combined with four feature templates [14]. Li et al. presented a cybersecurity named entity recognition neural network model based on self-attention [15]; this model considers that single-word features are not enough to identify entities, and used CNN to extract character features, then connects character features to word features; then, the self-attention mechanism is added on the basis of the BiLSTM-CRF model. These methods all use English datasets. Since English does not require word segmentation, it is easier than Chinese. Since the Chinese will make mistakes in the process of word segmentation, these mistakes will be propagated backward during the model training process, that is, error propagation.

For Chinese cybersecurity entity recognition, Qin et al. proposed that it is difficult for the traditional named entity recognition method to identify network security entities, and introduced a character-level CNN-BiLSTM-CRF network security entity recognition model based on feature templates [4]. Although this method alleviates the problem of error propagation, due to the shortcomings of the deep learning model itself, only the local information of the entity is extracted and the distance feature between the entities is not taken into account.

2.2. Active Learning. The purpose of active learning is to maximize the performance of the model while minimizing the labeled data needed to train the model [16, 17]. Shen et al. defined three broad criteria for determining which data is the most useful in the model after labeling [10]: uncertainty, in which the instances that confuse the model are given priority; diversity, instances that can extend the coverage of the model are given priority; and representativeness, giving priority to the instance closest to the true distribution.

Chen et al. believe that active learning (AL) is a sample selection method that integrates supervised machine learning, which is aimed at minimizing the cost of annotations while maximizing the performance of machine learning-based models [7]. In Chen et al.’s study, the uncertainty of information and the diversity of entity similarities in the selection strategy were considered. Shen et al. proved that when deep learning is integrated with active learning, the amount of...
labeled training data can be greatly reduced [11]. Shen et al. only compare several methods based on information uncertainty. Claveau and Kijak [8] proposed a simple method to correct the deviation of some of the most advanced selection techniques. In addition, this work introduced an original method for selecting examples based on the ratio in the dataset [8]. In these existing works, most researchers only consider the uncertainty and diversity of information about the select strategies. Therefore, in our work, we propose a selection strategy with uncertainty, confidence, and consideration of lexicon.

3. Proposed Method

3.1. Motivation. When constructing a cybersecurity knowledge graph, we often need a large amount of entities and relationships to form a triple. It is convenient to obtain unlabeled text datasets from the Chinese national vulnerability database of information security (CNNVD), blogs, forums, etc. However, in the process of constructing a cybersecurity knowledge graph, Chinese entities consist of many different characters and lack high-quality annotation data. Thus, we propose a straightforward active learning selection strategy to alleviate these problems based on our previous work.

The problem of the cybersecurity entity recognition task is defined as follows. In this paper, we take the characters in the Chinese texts of cybersecurity as the basic unit, given a dataset $D = X, Y$, where $X = \{s_1, s_2, \ldots, s_N\}$ and $Y$ is the corresponding label. For any of the sentences, $s_j = [x_1^N] = \{x_1, x_2, \ldots, x_N\}$, $x_i$ is the $i$th character in sentence $s_j$. Under the guidance of the BIO tagging method [4], identifying the security entity in a sentence $s_j$ is equivalent to giving a tagging sequence.

3.2. Design of Architecture. The proposed active learning framework for the cybersecurity entity recognition is shown in Figure 1. The framework mainly includes three parts: the distributed spider, BERT-RDCNN-CRF model (described briefly in this section), and selection strategies (described in detail in section of the active learning strategies). Through multiple iterations of model training and selection strategies, we can get ideal results at a relatively low annotation cost (discussed in Results and Discussion).

Specifically, the distributed spider uses the Scrapy-redis framework (https://github.com/rmax/scrapy-redis) to crawl multisource heterogeneous threat data on the network space. The active learning method includes the input layer, pretrained language model (BERT), residual dilation convolutional neural network (RDCNN), conditional random field layer (CRF), and selection strategy. At the input layer, we first perform some preprocessing on the original data, such as removing some special punctuation. Then, the sentence is processed by word segmentation, while the cybersecurity datasets are divided into training datasets and test datasets. On the training datasets, we further divide them into labeled datasets and unlabeled datasets and use the labeled datasets as the initial training datasets of the model. The CRF layer models the label dependencies in the output sentences. Finally, we select a part of the data from the unlabeled datasets and submit it to the crowdsourcing platform for annotation through our selection strategy. In addition, we use the test datasets to evaluate the cybersecurity entity recognition performance, which is a sign that indicates our active learning framework has stopped running. It is worth noting that in our previous work, we found that the performance of BERT-CRF in the cybersecurity entity recognition is very poor, and the performance of BERT-RDCNN-CRF is better than that of BERT-CRF.

3.3. Active Learning Strategies. As a pretraining model, BERT has been successfully applied in many natural language processing tasks. In this work, BERT is used as a character-level encoder. Given a sentence $s_j = [x]^N = \{x_1, \ldots, x_i, \ldots, x_N\}$, the BERT model converts each character in sentence $s_j$ into a fixed vector. In particular, we use linear-chain CRF as the tag decoder, and we use the Viterbi algorithm for decoding. The parameterized brief form of the posterior probability
formula of the label $y$ corresponding to the linear chain CRF model given the sentence $s_i$ is

$$p(y_j|s_i; A) \propto \exp \left( h^1(y_j; s_i) + \sum_{k=1}^{n-1} h^{k+1}(y_{k+1}; s_i) + A_{y_k, y_{k+1}}\right),$$

(1)

where $Z(s_i)$ is the normalization factor of all possible $s_i$ tags. $h^1(y_j; s_i)$ represents the probability of $y_j$ at $k$, which is the output of the softmax layer of BERT-RDCNN. $A$ is the transition matrix. In this paper, it is automatically learned during the model training process. $A_{y_k, y_{k+1}}$ represents the probability of transition from label state $y_k$ to label state $y_{k+1}$; thus, it is a square matrix. We use $\hat{y}$ to denote the most likely sequence of labels for $s_i$:

$$\hat{y} = \arg \max_y p(y | s_i; A).$$

(2)

Culotta and McCallum adopted a simple uncertainty-based strategy for the sequence model [18], called least confidence (LC), which is assigned by the model to the most likely sequence of tags, sorts each sentence in an ascending order, and then selects a part data to the labeling expert:

$$LC(X) = 1 - p(\hat{y} | s_i; A).$$

(3)

This uncertainty can be calculated using the posterior probability given by Equation (1). Preliminary analysis found that LC strategy is more inclined to choose longer sentences:

$$p(\hat{y} | s_i; A) \propto \exp \left( h^1(y_j; s_i) + \sum_{k=1}^{n-1} h^{k+1}(y_{k+1}; s_i) + A_{y_k, y_{k+1}}\right).$$

(4)

Because Equation (4) sums all the tags, the LC method naturally tends to long sentences, but our cybersecurity Chinese text is mostly short sentences. Although the LC method is very simple and has some shortcomings, many works have proved the effectiveness of the method in sequence labeling tasks.

Liu et al. believe that the sequence selected by CRF is valuable [19]. Liu et al. look for the most likely sequence assignment and hope that each label in the sequence has a high probability:

$$LTP(s_i) = 1 - \min_{\hat{y}, \tilde{y}} h^1(\hat{y}_j | s_i; A).$$

(5)

In addition, although Liu et al.‘s method has achieved relatively good results, we need to consider the confidence of the label of the sentence in the decoding process in the Chinese
cybersecurity entity recognition. Our confidence calculation formula is as follows:

\[
\text{Confidence}(s_i) = \text{Viterbi}(k, t),
\]

where Viterbi\((k, t)\) is the Viterbi algorithm and represents the probability of the most likely path to state \(k\) at time \(t\).

In addition, because there may not be a cybersecurity entity in some sentences of the corpus in the Chinese security field and the model will incorrectly sent to the labeling experts, then we also need to consider the differences of the samples. Given a sentence \(s = [x_1, \ldots, x_N]\) and a lexicon \(w_1, \ldots, w_M\), use the entities in the lexicon to match the sentence, so the matching frequency of the sentence is

\[
\text{Freq}(s_i) = \frac{1}{\sigma} \sum_{i=1}^{M} I(w_i \subseteq s_i),
\]

where \(M\) is the number of entities in the lexicon, \(\sigma > 1\) is the normalization factor, and \(I(\bullet)\) is the indicator function; that is, \(w_i \subseteq s_i = 1\); otherwise, it is 0. We evaluate whether a sentence is worthy of annotation to the labeling expert based on a comprehensive consideration of the uncertainty of this sentence, the confidence level, and the matching frequency in the lexicon. The final score calculation formula of the sentence is as follows:

\[
\text{Score}(s_i) = \min_{y_i} h(y_i \mid s_i; A) + \text{Viterbi}(k, t) + \frac{1}{\sigma} \sum_{i=1}^{M} I(w_i \subseteq s_i).
\]

Finally, use our approach to calculate the score of Equation (8) for each sentence in the unlabeled data, and then, sort the sentences in the unlabeled data in descending order, select the top \(k\) sentences, and send them to the labeling expert for annotation. In the experiment of this paper, we select the top 500 sentences after each iteration of model training for annotation.

### Table 1: Statistic of datasets.

| Entity type | Train | Dev | Test | Total |
|-------------|-------|-----|------|-------|
| PER         | 8144  | 884 | 1864 | 12563 |
| LOC         | 16571 | 1951| 3658 | 23716 |
| ORG         | 9277  | 984 | 2185 | 18134 |

### Table 2: Statistic of datasets of cybersecurity.

| Entity type | Train | Dev | Test | Total |
|-------------|-------|-----|------|-------|
| PER         | 8737  | 1256| 2570 | 12563 |
| LOC         | 16452 | 2372| 4890 | 23716 |
| ORG         | 12755 | 1813| 3566 | 18134 |
| SW          | 4649  | 673 | 1407 | 6729  |
| RT          | 56190 | 8045| 16212| 80447 |
| VUL_ID      | 4649  | 673 | 1407 | 6729  |
| Total       | 103434| 14832|30052|148318 |

### Table 3: Statistic of initial pool.

| Entity type | Initially labeled | Initially unlabeled |
|-------------|-------------------|--------------------|
| PER         | 492               | 8245               |
| LOC         | 864               | 15590              |
| ORG         | 782               | 11973              |
| SW          | 241               | 4408               |
| RT          | 3279              | 52911              |
| VUL_ID      | 16                | 4633               |
| Total       | 5674              | 97760              |
same time. The crowdsourcing annotation system is shown in Figure 2.

3.5. Model Training Process. This paper implements the cybersecurity entity recognition method proposed in this paper based on TensorFlow (https://www.tensorflow.org/).

The specific training process is shown in Algorithm 1. The active learning framework proposed in this paper first uses the initially labeled datasets to train the BERTRDCNN-CRF model and then uses this trained model to calculate the score of each sentence through Equation (8). Subsequently, these sentences are sorted in

![Figure 4: Comparison of F1 of different models on the general datasets.](image)
a descending order. A part of the data are selected to be annotated by the crowdsourcing platform. Finally, the labeled datasets are added.

4. Experiments

4.1. Datasets. This paper first uses a general dataset (https://github.com/zjy-ucas/ChinesNER) to verify the effectiveness of this method on the named entity recognition task in the general domain. The specific statistical information of the general datasets is shown in Table 1.

Secondly, the cybersecurity entity recognition datasets constructed by our laboratory are used to verify the effectiveness of this method on the cybersecurity entity recognition task. This experiment data mainly comes from the public data of mainstream cloud security platforms such as Wuyun Vulnerability Database, FreeBuf website, and the Chinese national vulnerability database of information security. There are six main types of cybersecurity entities, viz., person (PER), location (LOC), organization (ORG), software (SW), the relevant term (RT), and vulnerability ID (VUL_ID). Cybersecurity entity data adopts BIO named entity labeling strategy. The specific statistical information on the datasets is shown in Table 2. In the experiment, the labeled datasets are divided into the training set and test set, accounting for 80% and 20% of the total datasets, respectively. Among them, we divide the training set into the initially labeled pool and initially unlabeled pool, accounting for 5% and 95% of the training datasets, respectively. The statistics of the data are shown in Table 3. In the model training process, we first use the initially labeled pool to train the model, and then, we propose an active learning selection strategy to select data from the initially unlabeled pool to the crowdsourcing platform for labeling. It is worth noting that this paper uses the train_test_split() method in the sklearn machine learning package (https://scikit-learn.org/stable/) to split the datasets.

4.2. Experimental Setting. In this paper, precision, recall, F1, and microaveraging are used as performance metrics:

\[
\text{Micro} - P = \frac{\sum_{i=1}^{n} TP_i}{(\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FP_i)},
\]

\[
\text{Micro} - R = \frac{\sum_{i=1}^{n} TP_i}{(\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FN_i)},
\]

\[
\text{Micro} - F = \frac{2 \times \text{micro} - P \times \text{micro} - R}{\text{micro} - P + \text{micro} - R}.
\]

In order to verify the effectiveness of the active learning selection strategy proposed in this paper, the MTP method of Shen et al. [11], the LTC method of Calotta and McCallum [18], and the randomly selected method were used as baseline methods for comparative experiments. During the experiment, the hyperparameters of the BERT-based model were the same as those of our previous work. The hyperparameters of the BiLSTM model are shown in Table 4. In addition, we use the Tensorflow deep learning framework to implement the experiments in the experiment and use the Adam optimization algorithm [20].

5. Results and Discussion

In order to verify the performance of the proposed method on the general datasets, as shown in Figure 3, we conducted experiments on the general datasets. From the learning curves of Figure 3, it is clear that our approach TPCL compares with the baseline model in the F1 value and is almost the same as the performance of the baseline model. The main reason is that the structure of the entity in the general datasets itself is relatively simple, and secondly, a large number of entities in the general datasets are duplicated, which does not take advantage of the TPCL.

Further, we compare the recognition performance of different methods on different entities on general datasets, as shown in Figure 4. Figure 4 shows the result of the LOC.
(a) F1 comparison of different methods on LOC

(b) F1 comparison of different methods on ORG

(c) F1 comparison of different methods on PER

(d) F1 comparison of different methods on RT

Figure 6: Continued.
and PER entities; our approach TPCL has the same recognition performance as the baseline model, but on the ORG entity, the TPCL does not have the recognition performance of the two baseline methods, viz., LTC and MTP well, and the TPCL has the same performance as the random selection strategy method.

In order to verify the effectiveness of the TPCL in the Chinese entity recognition task in the field of cybersecurity, we conducted a comparative experiment on BERT-RDCNN and a typical BiLSTM-CRF entity recognition model. This paper conducted experiments on the cybersecurity entity datasets and analyzed the performance of the TPCL and the baseline method. Table 5 shows the named entity recognition accuracy and recall in the cybersecurity datasets.

It can be seen from the experimental results in Table 5 that the overall TPCL achieves better accuracy and recall on the cybersecurity datasets than the baseline method. In addition, due to the differences between the deep learning models, the BiLSTM-CRF model’s cybersecurity entity recognition performance is worse than the BERT-RDCNN-CRF recognition performance. However, Table 5 shows that whether we use the BERT model or the BiLSTM model, the TPCL is better than the baseline selection strategy method on accuracy and recall, which verifies our approach is effective.

In order to balance the accuracy and recall metrics, we also calculated the F1 value of different methods, as shown in Figure 5. In Figure 5, whether it is the BERT model or the BiLSTM model, our approach TPCL is better than the baseline method, which further verifies the effectiveness of the TPCL. Among them, for the BERT-based model, the F1 value of the TPCL is 88.54%, compared to 87.6%, 85.2%, and 87.1% of the LTC, MTP, and RANDOM methods, which are improved by 0.94%, 3.34%, and 1.44%, respectively. For the BiLSTM model, the F1 value of the TPCL is 80.17%, compared to 78.81%, 77%, and 77.24% of LTC, MTP, and RANDOM methods, which are improved by 1.36%, 3.17%, and 2.93%, respectively.

In order to further compare the recognition effect of the TPCL and baseline method on different cybersecurity entities, we calculate the PER, LOC, ORG, SW, RT, and VUL_ID F1 values of these six cybersecurity entities, as shown in Figure 6.

Figure 6 shows that the result of F1 value on the cybersecurity entity SW, whether it is BiLSTM or BERT model, is quite different. Among them, for the BERT-based model, the F1 value of the TPCL on the entity type SW is 59.01%, compared to 54.5%, 43.3%, and 46.17% of the LTC, MTP, and RANDOM methods, which are increased by 4.51%, 15.7%, and 12.84%, respectively. For the BiLSTM model, the F1 value of the TPCL on the entity type SW is 60%, compared to 61%, 53%, and 55% of the LTC, MTP, and RANDOM methods, which are improved by -1%, 7%, and 5%, respectively. It can be seen that using the BiLSTM model...
on our approach TPCL and LTC can achieve good results, among which LTC-BiLSTM has achieved the best entity recognition performance.

In Figure 6, it can be seen that the identification of the entity VUL_ID has achieved the best recognition performance on the BERT-based model, with an F1 value of 97.98%. However, it can also be seen in Figure 6 that there is little difference in entity recognition performance between BiLSTM-based and BERT-based different selection strategies. This is because the structure of the entity of VUL_ID is fixed, and both BiLSTM and BERT have extracted similar structural features. In addition, in the recognition result of the entity SW, the TPCL-BERTRDCNN-CRF has a slightly higher security entity recognition performance than the LTC-BERT-RDCNN-CRF model. However, the LTC-BiLSTM-CRF performance is lower than the TPCL-BiLSTM-CRF model and even higher than the TPLC-BERT-RDCNN-CRF model. From the perspective of the model, it shows that the BERTRDCNN and BiLSTM models are not good at identifying the cybersecurity entity SW. This is because, on the one hand, the entity has a small number of datasets in the cybersecurity entity. On the other hand, this kind of entity is usually composed of numbers, letters, and Chinese characters. The composition is very complex, and the features are not easy to extract. In short, for the two entities SW and VUL_ID, it is not very important which model to use, but from the perspective of active learning, the selection strategy proposed in this paper is much better in terms of the performance of cybersecurity entity recognition.

From Figure 6, the TPCL is better on the BERT model or BiLSTM model, and the entity recognition effect on the cybersecurity entities LOC, ORG, RT, and PER is better than the baseline method, indicating our approach can select sentences containing LOC, ORG, RT, and PER entities to the crowdsourcing platform to improve the cybersecurity entity recognition performance, and the selected sentences have a higher annotation value. Compared with the baseline selection strategy, it further shows that the sentence selected by the TPCL has more value than the sentence selected by the baseline selection strategy. In other words, the sentence contains rich information. The above result discussion indicates that the selection strategy proposed in this paper is robust and effective in various security entity recognition tasks.

Through the above analysis of the experimental results, the entity recognition performance of the TPCL on the general datasets is not much different from the baseline method. At the same time, the experimental results show that the TPCL is more suitable for the Chinese entity recognition task in the field of cybersecurity.

6. Conclusions

Aiming at the problem of large data annotation cost in the open network text data, we propose a new active learning selection strategy. This strategy is based on uncertainty, confidence, and lexicon considerations, it not only considers the output but also the distribution of cybersecurity entities. The experimental results on the cybersecurity entity recognition datasets show that the active learning selection strategy proposed in this paper is superior to the baseline method in accuracy, recall rate, and F1 value. In addition, the experimental results show that the strategy proposed in this paper is more suitable for entity recognition tasks in the field of cybersecurity.

In the future, we will build a complete automated web tool for cybersecurity entity cognition and relationship extraction based on this work to serve as a crowdsourcing data annotation platform. In addition, we will perform knowledge reasoning with the entities and relationships obtained on this web crowdsourcing platform in order to build a complete cybersecurity knowledge graph.

Data Availability

This paper’s dataset can be downloaded through the following link: https://github.com/xiebo123/NER/tree/master/Data.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant no. 61802081) and the Science and Technology Project of Guizhou Province (Grant no. [2018]3001).

References

[1] X. Shu, F. Araujo, D. L. Schales et al., “Threat intelligence computing,” in Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, pp. 1883–1898, Toronto, Canada, 2018.
[2] G. Shen, W. Wang, M. Qilin, P. Yanhong, Y. Qin, and Y. Miao, “Data-driven cybersecurity knowledge graph construction for industrial control system security,” Wireless Communications and Mobile Computing, vol. 2020, 2020.
[3] T. Wang, A. Zhongliang, and Z. Xiangguo, “Knowledge graph construction of threat intelligence based on deep learning,” Computer and Modernization, no. 12, pp. 21–26, 2018.
[4] Y. Qin, G. Shen, W. Zhao, Y. Chen, M. Yu, and X. Jin, “A network security entity recognition method based on feature template and CNN-BiLSTM-CRF,” Frontiers of Information Technology & Electronic Engineering, vol. 20, no. 6, pp. 872–884, 2019.
[5] B. Xie, G. Shen, C. Guo, Y. Zhou, and Y. Miao, “Cyber security entity recognition method based on residual dilation convolution neural network,” Chinese Journal of Network and Information Security, vol. 6, no. 5, pp. 126–138, 2020.
[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” 2018, https://arxiv.org/abs/1810.04805.
[7] Y. Chen, T. A. Lasko, Q. Mei, J. C. Denny, and H. Xu, “A study of active learning methods for named entity recognition in clinical text,” Journal of biomedical informatics, vol. 58, pp. 11–18, 2015.
[8] V. Claveau and E. Kijak, “Strategies to select examples for active learning with conditional random fields,” in
D. Marcheggiani and T. Artieres, "An experimental comparison of active learning strategies for partially labeled sequences," in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 898–906, Doha, Qatar, 2014.

D. Shen, J. Zhang, S. Jian, G. Zhou, and C. L. Tan, "Multi-criteria-based active learning for named entity recognition," in Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04), pp. 589–596, Barcelona, Spain, 2004.

Y. Shen, H. Yun, Z. C. Lipton, Y. Kronrod, and A. Anandkumar, "Deep active learning for named entity recognition," 2017, https://arxiv.org/abs/1707.05928.

T.-M. Georgescu, B. Iancu, and M. Zurini, "Named-entity-recognition-based automated system for diagnosing cybersecurity situations in IoT networks," Sensors, vol. 19, no. 15, p. 3380, 2019.

H. Gasmi, A. Bouras, and J. Laval, "LSTM recurrent neural networks for cybersecurity named entity recognition," ICSEA, vol. 2018, p. 11, 2018.

F. Yi, B. Jiang, L. Wang, and J. Wu, "Cybersecurity named entity recognition using multi-modal ensemble learning." IEEE Access, vol. 8, pp. 63214–63224, 2020.

T. Li, Y. Guo, and J. Ankang, "A self-attention-based approach for named entity recognition in cybersecurity," in 2019 15th International Conference on Computational Intelligence and Security (CIS), pp. 147–150, Macao, China, December 2019.

D. Yoo and I. S. Kweon, "Learning loss for active learning," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 93–102, Long Beach, USA, 2019.

S. Sinha, S. Ebrahimi, and T. Darrell, "Variational adversarial active learning," in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5972–5981, Seoul, Korea (South), November 2019.

A. Culotta and A. McCallum, "Reducing labeling effort for structured prediction tasks," AAAI, vol. 5, pp. 746–751, 2005.

M. Liu, T. Zhiying, Z. Wang, and X. Xiaofei, "LTP: a new active learning strategy for BERT-CRF based named entity recognition," 2020, https://arxiv.org/abs/2001.02524.

D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," 2014, https://arxiv.org/abs/1412.6980.