A Self-Attention-Based Approach for Named Entity Recognition in Cybersecurity

Tao Li, Yuanbo Guo, Ankang Ju
Zhengzhou Institute of Information Science and Technology
Zhengzhou, China
wade.heat@163.com

***Abstract*—With cybersecurity situation more and more complex, data-driven security has become indispensable. Numerous cybersecurity data exists in textual sources and data analysis is difficult for both security analyst and the machine. To convert the textual information into structured data for further automatic analysis, we extract cybersecurity-related entities and propose a self-attention-based neural network model for the named entity recognition in cybersecurity. Considering the single word feature not enough for identifying the entity, we introduce CNN to extract character feature which is then concatenated into the word feature. Then we add the self-attention mechanism based on the existing BiLSTM-CRF model. Finally, we evaluate the proposed model on the labelled dataset and obtain a better performance than the previous entity extraction model.**

***Keywords*-cybersecurity; entity recognition; BiLSTM; Self-Attention mechanism; CRF**

I. INTRODUCTION

Since the continuous cyber-attacks makes cybersecurity situation more complicated, there has been increasingly challenging in cybersecurity protection. To better improve cybersecurity, it is of great significance and indispensability to keep abreast of cyber threats in time by focusing on the emerging vulnerabilities, exploitation, cyber-attack modes, etc. [1]. However, such massive volumes of cybersecurity information exist in the textual sources, like cybersecurity whitepapers, blogs, vendors' bulletins and hacker forums [2]. Traditional analysis of threat-related information from textual sources relies on the numerous manual efforts, which is inefficient and time-consuming. As a result, the security analysts cannot make full use of the confirmed cybersecurity information to respond to cyber threats timely and exactly. Once converting the threat information into structured and machine-readable format, it can enhance the efficiency of utilizing the cyber threat intelligence and assisting the analysts to respond to cyber threats. In such procession, it is vitally important to identify threat related entities and relationships. In the paper, we discuss the application of the Named Entity Recognition (NER) [3], a major focus on information extraction in NLP techniques, for entities identifying in cybersecurity.

The earlier NER methods mainly rely on the pattern recognition, which needs experts' knowledge to design identifying rules and is poor portability. With the successful applications of statistical machine learning in the recognition of image and speech, the scholars also use machine learning to implement named entities recognition. In this phase, several machine learning models were proposed, including Maximum Entropy Models (MEM) [4], Support Vector Machines, SVM [5], Hidden Markov Models (HMM) [6] and Conditional Random Fields (CRF) [7], and have achieved pretty good results. However, such machine learning methods need great manual efforts to obtain domain features, which may decline robustness and generalization of the model, especially in specific domain.

Compared with statistical machine learning, since the neural networks can automatically extract features and be of better generalization, the utilization of deep learning has made greater achievements in image recognition, speech recognition, natural language processing, etc. In the NER task, there have been many neural networks-based methods. Typically, the deep learning model CNN or BiLSTM combining with the CRF model has been applied [8, 9], and the performance has been greatly improved with both precision and f value beyond 90% on the general evaluation dataset. Moreover, some scholars integrated the complementary character feature extracted by CNN or BiLSTM model into the input of the deep neural network model, resulting in the further improvement of model performance [10, 11].

Different from the NER in general domain, entities identified in cybersecurity mainly refer to as software and hardware names, attack organizations, attack methods, malicious files, vulnerability names, etc. Nowadays, the entity-recognition result has been continuously improved with the novel models proposed successively. In this paper, considering the poor performance of utilization of the off-the-shelf NER tools in cybersecurity, we propose a self-attention-based approach to identify the entities in cybersecurity. Finally, we evaluate the proposed method on the dataset and obtain the better result exactly.

The rest of the paper is organized as follows. In section 2, we describe the proposed model and the whole process. In section 3, we conduct the experiment on the given dataset and evaluate the performance of our model with the

978-1-7281-6092-4/19/$31.00 ©2019 IEEE
DOI 10.1109/CIS.2019.00039

comparison by some typical previous models. Finally, the summary is given in section 4.

II. SELF-ATTENTION-BASED BILSTM-CRF MODEL

In the proposed model, we first obtain the word embedding using the Word2Vec [12] on plenty of unlabeled cybersecurity data. In addition, we introduce the CNN model to extract character feature of each word. Then the concatenation of the word embedding and the character embedding is inputted to the proposed model. It is notable that we add the self-attention mechanism [13] between BiLSTM layer and CRF layer, which can better obtain context representation of the current word and obtain more information on the current word. The architecture of the proposed model is depicted in Figure 1.

A. Embedding Layer

Embedding layer means that the input is the vector representation learning from the data corpus. In our proposed model, the input embedding is composed of two parts by word embedding and char embedding. Firstly, we aggregate and preprocess plenty of corpora crawled from the cybersecurity blogs and reports. Then, using the Word2Vec tool to train the unlabeled data, we obtain the word embedding. And looking up the trained word embedding, each word in cybersecurity data can be mapped into a low-dimensional and dense word vector. In addition, since the character feature of each word also influence the entity recognition, we use the CNN model to extract character feature, named as char embedding. Finally, the concatenation of the word embedding and char embedding is input to the neural networks as the input.

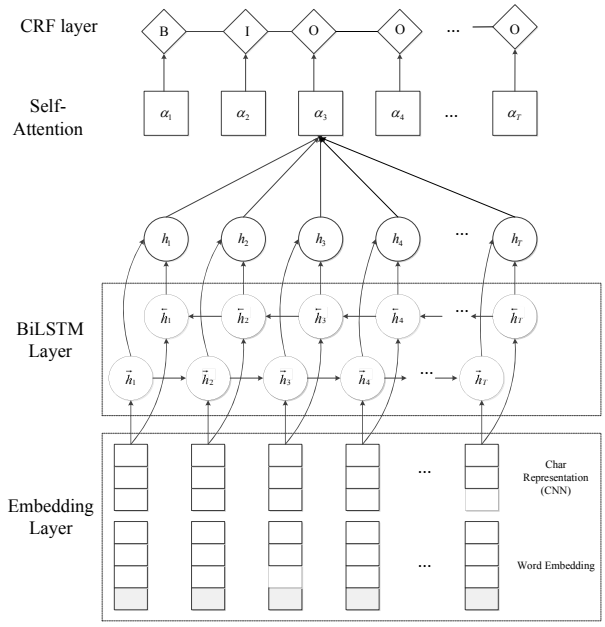

Figure 1. The architecture of the Self-Attention-based NER model

B. Self-Attention-based BiLSTM Layer

(1) BiLSTM Layer

As a special recurrent neural network, the LSTM model has been widely used to the NLP tasks due to its excellent performance in solving the long-distance dependence problem. Especially for sequence labeling, the current output is not only relied on the current input, but also influenced by the previous output. Since NER is a typical sequence labeling task in NLP, it is suitable to use LSTM model to obtain the context information of a word. And the implementation of LSTM is denoted as follows:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \quad (1)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \quad (2)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \quad (3)$$

$$c_t = i_t \cdot \tanh(W_{xc}x_t + W_{hc}h_{t-1} + W_{cc}c_{t-1} + b_c) + f_t \cdot c_{t-1} \quad (4)$$

$$h_t = o_t \cdot \tanh(c_t) \quad (5)$$

In the above formulate, i_t along with the weight matrix W_{xi}, W_{hi}, W_{ci}, b_i indicates the input gate of LSTM; f_t along with the weight matrix W_{xf}, W_{hf}, W_{cf}, b_f indicates the forget gate; o_t along with the weight matrix W_{xo}, W_{ho}, W_{co}, b_o indicates the output gate; c_t along with the weight matrix W_{xc}, W_{hc}, W_{cc}, b_c and previous cell state c_{t-1} indicates the current cell state. Finally, the calculation of h_t forms the whole hidden state of the LSTM model. However, in the manipulation of sequence labeling, we use the BiLSTM to get the past and future information related to the current word simultaneously. Thus, we denote $h_i = [\overrightarrow{h_i}; \overleftarrow{h_i}]$ as the hidden state of the i_{th} word.

(2) Self-Attention Layer

In the paper, we add the Self-Attention mechanism after BiLSTM layer to extract more context information related to the current word within a sentence. In self-attention process, as the hidden state of a sentence output by BiLSTM layer is denoted as $H = \{h_1, h_2, \cdots, h_T\}$, we map the vector H respectively with the trained weight matrices W_i^Q, W_i^K, W_i^V into new vectors Q_i, K_i, V_i. Then the function $\text{Attention}(Q_i, K_i, V_i) = \text{soft m ax}\left(Q_i K_i^T / \sqrt{d_k}\right) V_i$ is activated, where $1/\sqrt{d_k}$ indicates a calculating factor to avoid the gradient vanishing. Repeating the above implementation for h times, we concatenate the h results calculated by the

Attention function. Finally, we also conduct a linear mapping for the above concatenation and obtain the output vector $A = \{\alpha_1, \alpha_2, \cdots, \alpha_T\}$.

C. CRF Layer

For sequence labeling task, CRF model is necessary. With the calculation of the transition probability between labels and the probability of the whole labeling sequence, the global optimal labeling sequence can be obtained. Specifically, the probability that the sentence sequence X is labeled as $Y = \{y_1, y_2, \cdots, y_n\}$ denotes as the follows:

$$P(Y \mid X) = \frac{e^{Score(x,y)}}{\sum e^{Score(x,y)}} \tag{6}$$

$$Score(X,Y) = \sum_{I=1}^{n} (P_{i,y_i} + T_{y_i, y_{i+1}}) \tag{7}$$

where P_{i,y_i} corresponds to the probability that the i_{th} word is labeled as y_i , and $T_{y_i, y_{i+1}}$ means the label transition probability from y_i to y_{i+1} .

III. Experimental Analysis

A. Dataset construction

Given the scarce dataset of information extraction in cybersecurity, to better adapt to our entity identification, we manually labelled a certain amount of corpora which are the aggregation of the existing APT reports and alienvault' blogs with the mode BIO, where B indicates the beginning of an entity, I indicates the inside and O indicates the outside meaning the word is not an stated entity. In the implementation, we preliminarily labelled four types of entities to train the model, including software, attack organization, attack method and malicious file, which is shown in Table I.

TABLE I. Entity labelled mode

Entity type	BIO Mode
software	B-SW / I-SW
attack organization	B-ORG / I-ORG
attack method	B-MTH / I-MTH
malicious file	B-FIL / I-FIL
non entity	O

According to experience, we select 70% and 10% of the constructed corpora as the training set and testing test respectively. Then the rest 20% is applied to test model. The information about the labelled corpora is shown in Table II.

TABLE II. Information about labelled corpora

Corpora	Training set	Evaluating set	Testing set
Sentences	16190	2313	4625
software	5824	703	1637
attck organization	1016	297	583
attack method	2862	737	1286
malicious file	4728	752	1874

B. Hyper-Parameters and evaluating result

In the training of the proposed model, we set the hyper-parameters as follows:

TABLE III. Hyper-Parameters setting

parameters	value	parameters	value
word_embedding	200	batch_size	100
char_embedding	100	epoch	100
BiLSTM_dim	100	dropout	0.5
Self-attention_h	5	learning rate	0.001

We compare our model with the previous typical NER model which are also train and test on our cybersecurity corpora. And we also adopt the general evaluation indexes, including the Precision, Recall and F value. The experiment result of the comparison among several models is depicted in Table III.

TABLE IV. Comparison of different models

Models	P	R	F
LSTM-CRF	80.79	76.74	78.71
BiLSTM-CRF	82.36	79.33	80.82
CNN-BiLSTM-CRF	83.91	78.85	81.30
Our Model	86.24	83.75	84.98

As shown in Table 1, with the identical CRF model, the performance of BiLSTM is better than a single LSTM, due to the former's full extraction of context information. In addition, introducing the CNN model to extract the character feature really has an enhance in both precision and F value, which means the character feature is indispensable. Finally,

our proposed Self-Att-CNN-BiLSTM-CRF model obtains the best performance compared with others, illustrating the Self-Attention mechanism has significant improvement in entity identifying.

IV. CONCLUSION

Aimed at the generation of the structured data in cybersecurity, the paper proposed a Self-Attention-based model to identify the entities. We added the self-attention mechanism to capture more related feature information for the current word from the sentence itself. Through the evaluation on the labelled dataset that we constructed with the sequence labeling rules, we obtained a better result than the existing model, which fully illustrate the performance improvement because of the self-attention mechanism.

V. REFERENCES

[1] Narayanan S N, Ganesan A, Joshi K, et al. Early Detection of Cybersecurity Threats Using Collaborative Cognition[C]. 2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC). IEEE, 2018.

[2] Liao X. Towards automatically evaluating security risks and providing cyber intelligence[D]. Georgia Institute of Technology, 2017.

[3] Joshi A, Lal R, Finin T, et al. Extracting cybersecurity related linked data from text[C]. 2013 IEEE Seventh International Conference on Semantic Computing. IEEE, 2013: 252-259.

[4] Koeling R. Chunking with maximum entropy models[C]//Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop. 2000.

[5] Talha M, Boulaknadel S, Aboutajdine D. Performance Evaluation of SVM-Based Amazighe Named Entity Recognition[C]//International Conference on Advanced Machine Learning Technologies and Applications. Springer, Cham, 2018: 232-241.

[6] Alam T M, Awan M J. Domain Analysis of Information Extraction Techniques[J]. INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY SCIENCES AND ENGINEERING, 2018, 9: 1-9.

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

[8] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch[J]. Journal of machine learning research, 2011, 12(Aug): 2493-2537.

[9] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv:1508.01991, 2015.

[10] Ma X, Hovy E. End-to-end sequence labeling via bi-directional lstm-cnns-crf[J]. arXiv preprint arXiv:1603.01354, 2016.

[11] Chiu J P C, Nichols E. Named entity recognition with bidirectional LSTM-CNNs[J]. Transactions of the Association for Computational Linguistics, 2016, 4: 357-370.

[12] T. Mikolov, I. Sutskever, K. Chen, G. S Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, pages 3111–3119.

[13] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.