基于预训练模型的网络空间安全命名实体识别方法

韩瑶鹏[1,2], 王 璐[1,2], 姜 波[1,2], 卢志刚[1,2], 姜政伟[1,2], 刘玉岭[1,2]

[1]中国科学院信息工程研究所 北京 中国 100093
[2]中国科学院大学 网络空间安全学院 北京 中国 100049

摘要 随着网络空间安全文档数量的快速增长, 网络空间安全领域命名实体识别变的越来越重要。与通用领域命名实体识别任务相比, 网络空间安全领域的命名实体识别面临许多挑战。例如网络空间安全实体类型多样、新词语经常作为新的实体出现并引起超出词表(out-of-vocabulary, OOV)的问题。现有的深度学习识别模型(如循环神经网络、卷积神经网络)的性能不足以应对这些挑战。随着预训练模型的快速发展, 它已被广泛用于许多任务中并获得了最优的表现。但是, 在网络空间安全命名实体识别领域, 很少有关于预训练模型的研究。本文提出了两个基于预训练 pre-training of deep bidirectional transformers(BERT)模型的网络空间安全命名实体识别模型来从网络空间安全文本中提取安全实体, 分别称为“First Subword Replaced(FSR)”和“Masked Cross-Entropy Loss(MCEL)”。FSR 模型和 MCEL 模型还可以解决因 BERT 使用 WordPiece 分词器引起的子词和标签之间的不匹配问题。本文基于真实的网络空间安全文本语料库进行了充分的实验。结果表明, 本文提出基于预训练的模型在网络空间安全数据集上的 F1 值比之前的最优模型高了 1.88%。

关键词 网络空间安全; 命名实体识别; 预训练模型
中图法分类号 TP399 **DOI 号** 10.19363/J.cnki.cn10-1380/tn.2025.01.14

Cybersecurity Named Entity Recognition using the Pre-trained Model

HAN Yaopeng[1,2], WANG Lu[1,2], JIANG Bo[1,2], LU Zhigang[1,2], JIANG Zhengwei[1,2], LIU Yuling[1,2]

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China

Abstract Cybersecurity named entity recognition (NER) is becoming increasingly important as the number of cybersecurity documents rapidly grows. Compared with general domain NER tasks, cybersecurity-domain NER faces many challenges. There are many types of security entities, and new words often appear as entities causing out-of-vocabulary (OOV) problems. Existing deep learning recognition models (RNNs, CNNs) do not perform enough to deal with these challenges. With the rapid development of the pre-trained model, it is widely used in many tasks and achieved state-of-the-art performance. However, in the domain of cybersecurity NER, there are few studies on the pre-trained model. This paper proposes two cybersecurity NER models, named First Subword Replaced (FSR) and Masked Cross-Entropy Loss (MCEL), based on the pre-trained BERT (pre-training of deep bidirectional transformers) model to extract security entities from the cybersecurity dataset. The FSR and MCEL models can also deal with the mismatch between the subwords and labels caused by BERT using WordPiece tokenizer. This paper conducts extensive experiments on a real-world cybersecurity text corpora. The results show that the pre-trained model proposed in this paper outperforms the previous state-of-the-art method by 1.88% F1 score on the cybersecurity dataset.

Key words cybersecurity; named entity recognition; the pre-trained model

1 引言

近年来, 随着网络安全事件的频频发生, 例如应用程序攻击, 恶意软件, 勒索软件和网络钓鱼等, 网络空间安全的重要性日益增加。为提高网络安全意识和防护网络攻击, 大量安全厂商和安全研究人员将安全博客、攻击报告、安全漏洞等网络安全数据发布在各大安全网站上。从大量的网络安全数据

通讯作者: 刘玉岭, 博士, 正高级工程师, Email: liuyuling@iie.ac.cn。

本论文得到国家重点研发计划(No. 2021YFF0307203, No. 2019QY1303, No. 2019QY1302)、中国科学院战略性先导 C 类(No. XDC02040100), 基础加强计划技术领域基金(No. 2021-JCJQ-JJ-0908)、国家自然科学青年基金(No. 61902376)和信息安全等级保护关键技术国家工程实验室(公安部第三研究所)开放课题(No. C21640-3)的资助。这项工作也得到了中国科学院网络评估技术重点实验室和北京市网络安全与保护技术重点实验室的部分支持。
收稿日期: 2020-09-23; 修改日期: 2021-03-26; 定稿日期: 2022-12-08

中提取重要安全信息来构建安全领域的知识图谱[1]可以帮助分析人员更快地检索和发现网络安全威胁情报和安全态势。网络安全知识图谱构建技术包括实体识别、关系抽取等。其中安全命名实体识别可以自动的从大量无标注安全博客等文本数据中提取安全知识。因此安全命名实体识别是构建网络安全知识图谱的基础。

命名实体识别(Named entity recognition, NER)是从文本中提取预定义类型的实体。它不仅是信息抽取的重要组成部分, 而且还在各种自然语言处理应用中发挥重要作用。例如知识图谱、信息检索[2]和问答系统[3]等。命名实体识别在自然语言处理领域中被广泛研究, 方法也从传统的机器学习方法过渡到深度学习模型。传统的机器学习方法主要包括两类: 基于规则的学习方法和基于特征的监督学习方法。基于规则的方法依赖于手工制定的规则。基于特征的监督学习方法依赖于特征工程。近年来, 深度学习模型如卷积神经网络(Convolutional neural network, CNN)、循环神经网络(Recurrent neural network, RNN)逐渐成为命名实体识别任务的主流模型。与传统的机器学习方法相比, 深度学习模型是端到端的, 可以自动挖掘文本中单词的隐藏信息并且可以大大减少人工设计手工特征的工作量。

网络空间安全实体识别是基于特定领域的实体识别任务。基于特定领域的实体识别任务相较于通用领域的实体识别任务文本数据往往集中在特定领域, 特定领域有标签数据相对较少, 且特定领域文本会包含大量领域知识。网络空间安全领域命名实体识别主要任务是在网络空间安全文本中识别安全实体, 其中安全实体的类型主要包括攻击方式、软件名、操作系统以及文件名等。与基于粗粒度的通用领域命名实体识别任务(仅识别人名、地名和组织机构名三种类型实体)相比, 网络空间安全实体抽取更加困难。总的来说, 网络空间安全命名实体识别任务的主要挑战如下:

(1) 网络空间安全领域文本大多是基于特定领域, 安全文本往往含有大量领域知识, 且安全特定领域有标签可用于训练数据相较于通用领域数据往往较少, 训练出安全实体抽取模型往往性能较差。

(2) 网络空间安全实体类型多样, 并且不断出现新的短语作为实体来描述安全事件及过程。例如新的恶意软件、漏洞编号、补丁等实体。

(3) 网络空间安全文本包含很多安全相关术语, 如 DDOS、XSS 等词语, 所以相较于通用领域文本会包含更多的缩写词作为安全实体来描述相关安全术语。同时安全文本也会像通用领域出现词嵌套、一词多义等问题。 例如“Oracle”既可以是软件类型的实体又可以是组织类型实体。

因此, 网络空间安全实体识别吸引了许多来自不同角度的研究工作, 文献[4]提出了两种自动提取漏洞相关信息的方法, 一种是 Conditional random fields(CRF)[5]方法, 另一种是基于特征的标记方法。文献[6]提出了一种使用正则表达式、安全术语和相关语法知识的 iACE 框架来从非结构化文本中自动提取 Indicators of compromise (IOC)指标。IOC 指标是入侵的取证产物, 例如病毒签名、僵尸网络、IP 、域名和攻击文件等, IOC 且积极的在机构组织之间收集和交换, 并可以直接用于安全系统来为组织机构提供即时保护。但是, IOC 指标并不是那么直观, 无法帮助人们更好的了解正在发生的入侵。网络安全文本中包含的安全实体对帮助把握网络安全威胁会更有帮助。但是这些传统模型不能很好的解决网络空间安全实体识别中存在的实体类型多样、多义词等问题。

最近, 预训练模型在自然语言处理的许多任务中取得了巨大的进展。语言模型的迁移学习证明了基于大量文本的无监督预训练是许多语言理解任务的重要组成部分。文献[7]采用了 Transformer[8]网络结构来解决命名实体识别任务, 证明了 Transformer 网络提取文本特征的能力强于卷积神经网络和循环神经网络, 并且可以得到更加精确的文本语义表示。采用 Transformer 网络结构的预训练模型 Bidirectional encoder representations from transformers(BERT)[9]可以先从大量未标注的文本中学习到高质量的单词语义表示, 进而在对下游任务微调时可以根据文本的上下文信息获得动态的词语义表示, 因此可以较好的解决一词多义的问题。

BERT 在众多预训练模型中表现突出, 同时 BERT 也是其他许多预训练模型[10-11]的基础。因此, 本文采用 BERT 作为本文提出模型的主要网络架构。此外, BERT 使用 WordPiece[12]作为英文文本的单词分词器, 会将每个英文单词分成若干子词, 这样可以大大减少词表的词量, 而且还可以使拥有相同前后缀的英文单词语义更加相近。另外, 由于词表之外的单词(Out-of-vocabulary, OOV)也会被 WordPiece 切分成若干子词, 并且每个子词都可以在预训练词表中找到, 所以这种分词方法可以较好解决 OOV 的问题。然而, 使用 WordPiece 分词容易导致单词与标签无法一一对应的问题。当使用 BERT 在对序列标注下游任务微调时, WordPiece 首先会将输入文本的所

有单词切分成若干子词, 这些子词不能与单词的标签一一对应, 将导致不能直接计算模型损失。例如单词“exploit”会被 WordPiece 切分成“ex”, “##p”, “##lo”, “##it”。单词“exploit”的标签是“I-MEANS”, 所以被切分的这 4 个子词不能与标签对应。而一般交叉熵公式为:

$$CE(p,q) = -\sum_{i=1}^{C} p_i \log(q_i) \tag{1}$$

其中 p_i 代表每个词的真实标签, q_i 代表每个词的预测标签。所以在计算交叉熵的公式时需要每个词的真实标签与自己的预测标签进行一一对应, 但是经过 WordPiece 分词后会导致一个单词的多个部分对应这一个单词的标签, 导致不能直接计算该单词的交叉熵损失。

针对上述问题, 本文提出了两个网络空间安全实体识别的模型 First subword replaced(FSR) 和 Masked cross-entropy loss(MCEL)来实现对安全实体的准确识别。本文的主要贡献如下:

(1) 本文提出了一种基于预训练模型 BERT 新颖的安全实体识别模型, 可以有效解决网络空间安全文本中存在的实体类型多样、缩写词以及 OOV 等问题。本文提出的模型结合 CRF 和 Highway[13]网络来进一步提升安全实体识别模型的性能。

(2) 本文提出的模型还可以进一步解决因预训练模型 BERT 使用 WordPiece 分词给英文文本带来的子词和标签不能一一对应的通用问题。本文提出的模型还可以应用在其他的领域和其他序列标注任务上, 证明了模型的通用性。

(3) 本文将提出的模型应用在真实的网络空间安全文本数据集上, 并进行了详细的实验对比, 本文提出的模型达到了目前最优的效果, 相较于之前的最优模型 F1 值提升了 1.88%。

2 相关工作

近年来, 随着在命名实体识别任务上深度学习模型的性能超过传统的机器学习方法, 许多深度学习研究大多都集中在长短期记忆网络(Long short-term memory, LSTM) 和 CRF 上。文献[14]使用了基于词级别的双向 LSTM(Bidirectional LSTM, BiLSTM)和 CRF 模型, 该论文展示了在解码层加上 CRF 可以在 CoNLL2003 数据集上提高 F1 值。文献[15-17]不仅将词级别的向量表示作为输入, 还用了字符级别的向量表示并且获得了更好的结果。在提取字符级别的向量表示有两种广泛使用的结构, 一种是基于卷积神经网络, 另一种是基于循环神经网络。文献[15]证明了基于卷积神经网络的字符级别提取模型在 CoNLL 数据集上比循环神经网络可以获得更好的效果。文献[16]提出了一种基于注意力机制来确定分别使用多少的词级别信息和字符级别信息。文献[17]利用卷积神经网络提取单词的字符级别向量表示, 并将其与词级别向量表示拼接起来输入给 BiLSTM-CRF 模型中, 该模型在很多命名实体识别任务上都取得了良好的性能。

在基于网络空间安全特定领域的命名实体识别任务上, 文献[18]提供了一个包含数种网络空间安全实体类别的英文数据集, 并使用 CRF 模型来解决。文献[19]提出了使用 XBiLSTM-CRF 来提取英文安全文本中的实体, 相较于 BiLSTM-CRF 模型可以提高实体识别的精准率和召回率, 但是使用的是基于字符级别的评价指标。不如基于实体级别的评价指标可以更为直观准确的判断实体抽取的效果。文献[20]提出了结合正则表达式、先验词典和 CRF 模型来抽取安全文本中的安全实体。文献[21]提出了使用基于分布式计算框架的与规则结合的 CRF 算法对中文安全实体的高效识别。文献[22]提出了基于特征模板的深度学习模型 FT-CNN-BiLSTM-CRF 来提升安全实体识别模型的性能。可以看到, 之前基于网络空间安全特定领域的实体识别模型的研究大多采用基于规则或者循环神经网络的方法。然而, 基于传统的规则方法需要手工寻找特征, 大大增加了安全分析人员的时间精力。基于循环神经网络的深度学习模型虽然可以减少人工寻找特征的时间, 但是循环神经网络模型使用的都是类似 Word2vec[23]、GloVe[24]等静态词向量, 静态词向量对于相同词语在不同文本中使用的是相同的向量表示, 因此不能很好的解决安全文本中多义词的问题。

如今, 预训练模型已经在自然语言处理的众多任务中广泛应用并且大多获得了最优的结果。文献[9]提出了一种新的语言模型称为 BERT, 它是基于 Transformer 网络的双向编码器使用 Masked 语言模型(Masked language model, MLM)来实现预训练的深度双向表示。文献[25]使用 BERT 模型结合全局的上下文信息解决通用领域命名实体识别任务, 但是他们仅将 BERT 的输出向量表示作为特征进行使用。此外, 在不同的特定领域也有一些关于预训练模型的最新研究。文献[26]提出了一种针对科学特定领域使用 BERT 预训练来解决缺乏高质量和大规模有标注数据的科学领域问题。文献[27]提出了一种基于大规模医疗语料库的 BERT 架构预训练医疗领域特定语言表示模型。

预训练模型已在其他特定领域, 例如科学和医疗等领域得到广泛应用, 但在网络空间安全领域中, 仍然采用传统的基于规则和基于循环神经网络的方法, 这些方法无法满足当今网络安全文本存在的实体类型多样、多义词、OOV 等问题。因此基于网络空间安全领域研究预训练模型具有重要意义。

3 模型

本章首先介绍本文使用的预训练模型架构 BERT, 然后介绍本文提出的网络空间安全命名实体识别模型 FSR 和 MCEL 模型具体的实现以及核心创新部分。

3.1 BERT 模型

BERT 模型采用了多层双向 Transformer 网络结构, 可以很好地捕获单词的上下文表示。BERT 在预训练阶段提出了两个新颖的预训练目标, 一个是 MLM, 另一个是下句预测(Next sentence prediction, NSP)目标。相较于传统的从左到右的语言建模目标, MLM 预测文本中被随机 mask 掉的词, 这样可以很好的应用单词的双向上下文表示。NSP 任务旨在预测两个文本是否连续, 可以更好地学习文本间的关系, 更适用于自然语言推理任务。BERT 首先根据两个预训练目标 MLM 和 NSP 从大量的无标签数据中学习子词的向量表示, 然后根据不同的任务目标在下游任务中进行少量的结构修改, 最后可以在大多数具体的下游任务中获得最优的性能。

使用 BERT 针对英文文本进行命名实体识别下游任务微调时, 输入为英文文本, BERT 会将文本进行切词预处理, 之后将切好的子词对应的嵌入向量、位置嵌入向量和段嵌入向量拼接起来作为 BERT 模型的输入。模型上 BERT 采用多层 Transformer 进行编码, Transformer 通过多头注意力机制(Multi-head attention mechanism)来捕获每个子词与文本中其他子词的相互关系, 可以帮助模型捕获来自不同位置的不同表示子空间的信息。多头注意力机制的基础是自注意力, 自注意力机制的计算公式如下:

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = softmax(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\boldsymbol{V}) \tag{2}$$

$$\text{head}_i = \text{Attention}(\boldsymbol{Q}\boldsymbol{W}_i^Q, \boldsymbol{K}\boldsymbol{W}_i^K, \boldsymbol{V}\boldsymbol{W}_i^V) \tag{3}$$

$$\text{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{Concat}(\text{head}_1,...,\text{head}_h)\boldsymbol{W}^O \tag{4}$$

其中 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 分别是自注意力机制中的查询向量、键向量和值向量。d_k 是自注意力机制中的输出向量维度。$\boldsymbol{W}_i^Q$、$\boldsymbol{W}_i^K$、$\boldsymbol{W}_i^V$、$\boldsymbol{W}^O$ 是权重矩阵模型学习参数。

最后经过两个线性变换层, 中间用激活函数 ReLU。最终经过 BERT 模型得到文本的输出嵌入向量表示:

$$\text{BERT}(x) = \max(0, x\boldsymbol{W}^1 + b^1)\boldsymbol{W}^2 + b^2 \tag{5}$$

其中 $\boldsymbol{W}^1$、$\boldsymbol{W}^2$ 为权重矩阵, b^1、b^2 为偏置。

在处理英文文本的时候, BERT 会使用 WordPiece 作为单词分词器, WordPiece 与字节对编码类似, 依赖于子词的词汇表。词汇表的构建使其包含最常用的词或者子词。BERT 分词器用“##”表示分割词。使用 WordPiece, 任何 OOV 词都可以切分成若干子词, 这可以有效减轻 OOV 以及解决多义词的问题。例如文本“Expert warn of Zero-Day exploit”在使用 WordPiece 后被切分为: “Ex、##pert、##s、war、##n、of、Zero、-、Day、ex、##p、##lo、##it”。文本中各单词对应的标签: “<O>、<O>、<O>、<B-MEANS>、<I-MEANS>”。然而, 在 WordPiece 对文本中的单词进行切分后, 每个单词子词的数量都会大于等于该单词对应的标签, 导致无法直接计算交叉熵损失。针对此问题, 本文的模型 FSR 和 MCEL 可以有效进行解决。

3.2 FSR 模型

针对网络空间安全命名实体任务, 本文提出在 BERT 预训练模型基础上, 结合 Highway 和 CRF 的安全实体识别模型 FSR-CRF, 模型架构图如图 1 所示。

形式上, 本文使用 $X = \{x_1, \cdots, x_n\}$ 来代表输入的英文文本, $Y = \{y_1, \cdots, y_n\}$ 代表文本 X 的标签序列。n 是文本中的单词的数量。在经过 BERT 分词器 WordPiece 后, 文本中的单词被切分成若干子词 $X' = \{x_1, \cdots, x_m\}$。$m$ 是文本中经过 WordPiece 分词后子词的数量, $m \geqslant n$。

如图 1 输入的文本首先经过 BERT 的分词器 WordPiece, 会将文本中单词切分成若干子词。之后经过 First Subword Replaced 模块使用每个单词被切分后的第一个子词表示来代表文本中的每个单词表示。在经过此模块后, 序列子词的长度等于标签序列的长度 $\hat{X} = \{x_1, \cdots, x_n\}$。进一步, 将 $\hat{X}$ 输入给图 1 中的 BERT 模型, BERT 模型使用序列 $\hat{X}$ 的嵌入向量、位置嵌入向量和段嵌入向量并将其相加输入给 L 层 Transformer 架构, 这样最终可以得到 $\hat{X}$ 中每个子词对应的上下文输出表示 $h^L = \{h_1,...,h_n\} = \text{BERT}(\hat{X})$。基本模型 BERT 层数有 12 层, 而较大的

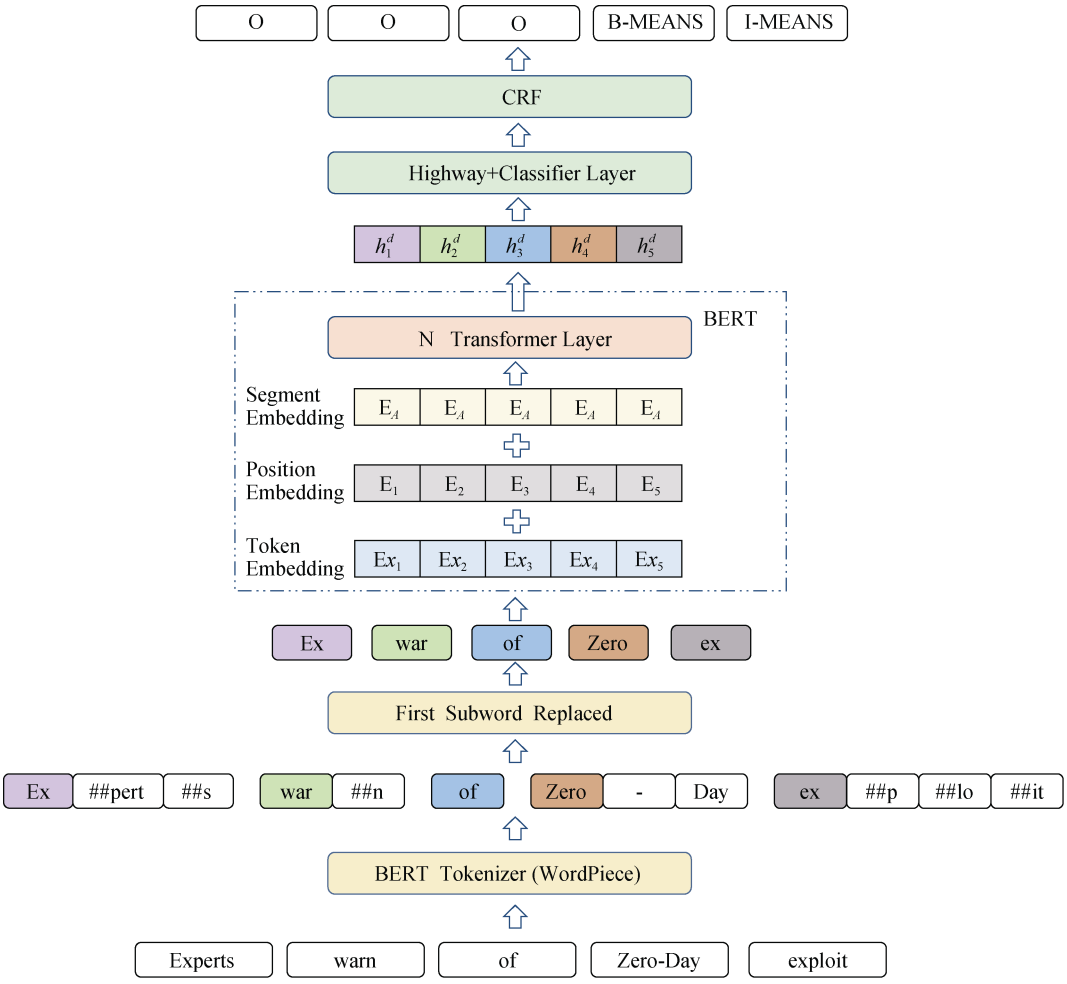

图 1 FSR-CRF 整体模型架构图

Figure 1 Overview of the FSR-CRF model

BERT 模型有 24 层, 接下来本文采用了 Highway 网络, Highway 网络采用了门机制可以帮助模型较好的训练深层网络:

$$H_n = Highway\left(h^L\right) = G\left(h^L, W_G\right)\cdot h^L + \left(1 - G\left(h^L, W_G\right)\right)\cdot T\left(h^L, W_T\right) \tag{6}$$

其中 $h^L \in R^{n\times d}$, $H_n \in R^{n\times d}$, d 是 BERT 中隐状态的维度。$T(\cdot)$ 和 $G(\cdot)$ 是非线性转换函数。W_G 和 W_T 是模型参数。图 1 中的分类器(Classifier)是一个线性层用于将 H_n 的维度映射到与标签类别数量相等的维度 k。然后将其通过 *softmax* 激活函数来计算子词级别的预测:

$$p(y \mid X) = softmax\left(W_0 H_n + b_0\right) \tag{7}$$

其中 $W_0 \in R^{d\times k}$ 和 b_0 是分类器可学习的参数。最终使用交叉熵做为损失函数:

$$L_{FSR} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{n}\tilde{y}\log\left(p\left(y_t \mid X\right)\right) \tag{8}$$

其中 $\tilde{y}$ 是每个子词的真实标签。N 是样本的总数量。此外 FSR 模型还可以在解码层使用 CRF 模型。

CRF: 对于序列标注任务例如命名实体识别, 考虑给定输入文本的相邻标签的相关性是有用的。例如在使用“BIO”标注格式的命名实体识别任务中, “I-MEANS”不能接在“I-ATTACK”之后。因此模型使用 CRF 可以对标签序列联合解码而不是独立的解码。CRF 是命名实体识别任务上最常见的标签解码器选择。假如在 FSR 模型的解码层采用 CRF, 则对于任一输入文本 $X = \{x_1, \cdots, x_n\}$, H 是 FSR 模型经过 Highway 网络得到的输出得分矩阵, $H_{i,j}$ 表示第 i 个词的第 j 个标签的分数, 引入转移得分矩阵 $\boldsymbol{A}$, 为了使转移分数矩阵 $\boldsymbol{A}$ 更具鲁棒性, 在句子开始和结尾加上 START 和 END 两类标签, START 代表一个句子的开始, END 代表一个句子的结束。对预测序列 $Y = \{y_1, \cdots, y_n\}$ 而言, 得到该标签序列的总得分为:

$$S(X, Y) = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} H_{i, y_i} \tag{9}$$

其中转移分数矩阵 $\boldsymbol{A}$ 是模型学习参数，y_0 和 y_{n+1} 是句中的起始标签和终止标签，k 是标签个数，A 的大小为 $k+2$，$A_{y_i,y_{i+1}}$ 代表从标签 y_i 到 y_{i+1} 的分数。

对所有可能的序列路径进行归一化，计算关于输出序列 Y 的概率分布:

$$p(Y\mid X)=\frac{e^{(S(X,Y))}}{\sum_{Y'\in\tilde{Y}}e^{(S(X,Y'))}} \tag{10}$$

$\tilde{Y}$ 表示文本 X 的所有可能标签序列的集合。对于训练，该模型会最大化正确序列的对数概率:

$$\log\left(p(Y\mid X)\right)=S(X,Y)-\log\left(\sum_{Y'\in\tilde{Y}}e^{(S(X,Y'))}\right) \tag{11}$$

通常，使用维特比算法[28]查找所有标签序列上得分最高的标记序列:

$$Y^{*}=\arg\max_{Y'\in\tilde{Y}}S(X,Y') \tag{12}$$

FSR-CRF 模型仅仅使用每个单词被切分后的第一个子词的信息来解决子词数量与标签数量不匹配的问题，但是忽略了其他子词的信息。

3.3 MCEL 模型

本文在 BERT 预训练模型的基础上，为了在训练过程中充分使用英文文本中单词所有子词的信息，结合 Highway 网络提出了安全实体识别模型 MCEL，整体架构图如图 2 所示。

MCEL 模型首先保持所有的被 BERT 分词器 WordPiece 切分后的子词 $X'=\{x_1,\cdots,x_m\}$ 输入至 BERT 模型，和 FSR 模型一样经过 BERT 模型后可以得到 X' 中所有子词的输出向量表示 $\boldsymbol{h}_T^L=\{h_1,\ldots,h_m\}=$ BERT(X')。与 FSR 模型相比，MCEL 模型使用了文本中单词的所有子词信息经过了 BERT 模型。然后将 h_T^L 输送至与 FSR 模型中公式(6)一样的 Highway 网络和分类器:

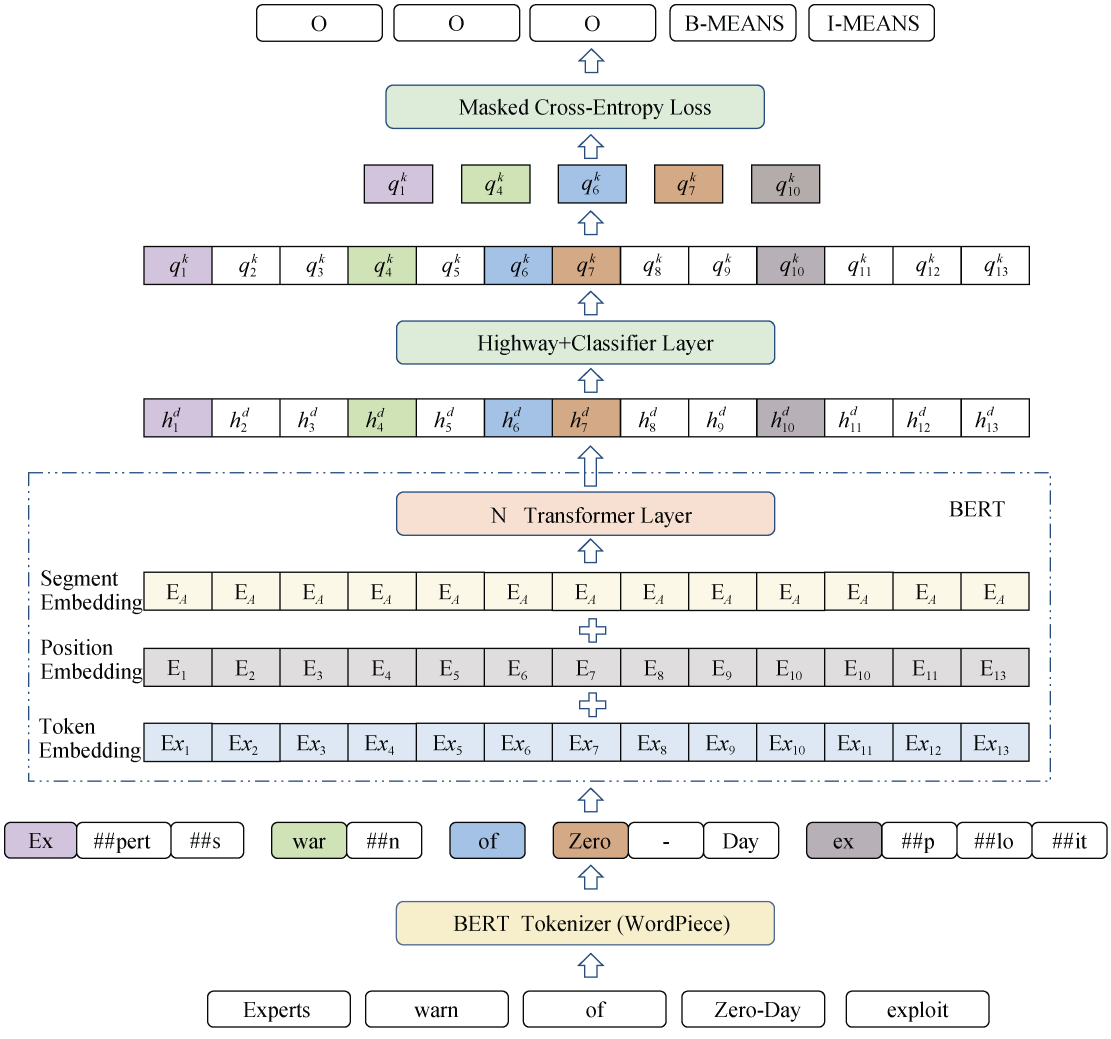

图 2 MCEL 整体模型架构图

Figure 2 Overview of the MCEL model

$$H_m = W_1 \cdot Highway\left(h_T^L\right) + b_1 \tag{13}$$

其中 $W_1 \in R^{d\times k}$, $h_T^L \in R^{m\times d}$, $H_m \in R^{m\times k}$。MCEL 模型同样使用交叉熵作为损失函数。在计算交叉熵

损失的时候, 因为需要每个子词表示和标签进行一对一对应, MCEL 模型使用文本中每个单词的第一个子词表示 H_n^T, 单词的其他子词表示将会被 mask。因为 BERT 模型是采用了多层 Transformer 网络, 它能很好的融合信息, 其他子词的信息会被融合到第一个子词向量表示中, 所以相较于 FSR 模型文本信息丢失就会大大减少。最后使用图 2 中的 Masked Cross-Entropy Loss 模块来计算模型损失:

$$H_n^T = \gamma\left(H_m\right)$$

$$p\left(y \mid X\right) = softmax\left(H_n^T\right) \tag{14}$$

$$L_{MCEL} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{n}\tilde{y}\log\left(p\left(y_t \mid X\right)\right)$$

其中 $H_n^T \in R^{n\times k}$, $\gamma(\cdot)$ 是采用每个单词中第一个子词信息表示的函数。MCEL 模型不仅能使子词与标签一一对应, 而且还可以充分地使用到文本中每个单词中所有子词的信息。

当在测试集上预测时, 只预测文本每个单词中第一个子词的标签, 最后将这些标签组合起来就是测试集对应文本的标签序列。

4 实验

为了评估所提出两个模型的性能, 本章依次介绍一下使用的数据集、用于对比的基线模型、评估指标以及实施细节。

4.1 数据集

本文使用开源的安全领域的非结构化文本数据[18]。该数据集由人工从网络空间安全领域收集的漏洞报告、微软安全公告和各种安全博客文章并进行手动标注的数据。该网络空间安全数据集包含超过 45000 单词和 5000 个安全实体。这个安全领域数据的实体类别代表识别和描述攻击的关键方面, 如下所示:

(1) Software (软件名, 例如: Microsoft .NET Framework 3.5)

(2) Operating_System (操作系统名, 例如: Linux Ubuntu 10.4)

(3) Network_Terms (网络术语, 例如: SSL, IP Address, HTTP)

(4) Attack

a) Means (攻击方法, 例如: Buffer overflow)

b) Consequence (最终攻击结果, 例如: Denial of Service)

(5) File_Name (文件名, 例如: index.php)

(6) Hardware (硬件, 例如: IBM Mainframe B152)

(7) NER_Modifier (一般在 Software 或 Operating_System 类前或者之后, 可以帮助识别软件产品或操作系统的版本信息)

该数据共有 7 大类, 9 小类实体类型。其中“Network_Terms”被认为是重要的一类, 因为如今大多数攻击都会用到网络术语。因此, 提取文本中的“Network_Terms”类尤为重要。“Attack”类可以进一步分类为帮助识别攻击方法的“Means”小类或描述攻击最终结果的“Consequence”小类。例如, “缓冲区溢出(Buffer overflow) ”被认为是“Means”的一个实例, 因为它不是攻击者的最终目标, 而仅仅是实现期望结果的一步。例如“拒绝服务(denial of service)”就是“Consequence”类别。在给定的文本中, 是否总是将短语视为“Means”或“Consequence”的实例不总是很清楚。当很难在一个短语之间决定它们时, 定为“Attack”类别。“NER_Modifier”类可以确认文本中软件产品或操作系统的版本信息。例如“This vulnerability is present in Adobe Acrobat X and earlier versions...”, 短语“and earlier version”表示版本 X 之前的所有 Adobe Acrobat 版本也容易受到攻击威胁, 可以帮助识别文本中未记录但是仍容易受到威胁的产品版本。

本文采用五折交叉验证进行评估模型性能, 其中四个数据块作为模型的训练集, 一个数据块作为测试集。训练集平均包含 3800 个实体和超过 37000 个英文单词, 测试集平均包含超过 1200 个实体和 8000 个英文单词。

4.2 基线模型

(1) BiLSTM-Softmax: 文献[14]采用了 BiLSTM 作为文本编码层, 输出层使用 Softmax 进行多分类来解决命名实体识别任务。

(2) BiLSTM-CRF: 文献[14]提出了使用 BiLSTM 作为编码层, CRF 作为解码层来进行命名实体识别模型。

(3) BiLSTM-CharCNN-CRF: 文献[17]提出使用 CNN 来提取单词的字符级别信息, 并将其和词级别信息拼接在一起作为 BiLSTM-CRF 模型的输入进行实体识别。此模型被广泛使用在各个领域的命名实体识别任务上并且表现很好。通过将本文提出的模

型与该模型进行比较, 可以更好的解释使用预训练模型对网络空间安全领域命名实体识别任务改进的效果。

4.3 实现细节

本文使用基于实体的精准率召回率以及F1值作为评估指标。只识别实体中某一个单词或者几个单词都不算正确识别, 只有将整个安全实体识别出来才算正确识别。首先计算每折测试集的精准率和召回率, 之后求平均值并使用公式(17)来计算得到最终的F1值。

$$Precision = \frac{TP}{TP + FP} \times 100\% \tag{15}$$

$$Recall = \frac{TP}{TP + FN} \times 100\% \tag{16}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \times 100\% \tag{17}$$

本文采用PyTorch框架实现所有的模型。详细的参数设置如下:

(1) 基线模型

超参数是基于验证数据集上的初步实验进行设置的。训练使用批大小为10的Adam优化器。初始学习率为0.015。训练epoch数为100。LSTM中的隐向量维度设置为100, 因为使用的是双向LSTM, 所以隐状态向量维度是200。词向量使用基于Glove模型对维基百科和网络文本进行预训练得到的300维词向量[23]进行初始化。为了防止过拟合, 在词向量embedding层后加了25%的dropout层。经过分析, 有99%的文本长度小于128, 因此将文本最大长度设置为128。在解码层使用“BIO”的标注格式。此外, 还使用学习率衰减机制:

$$LR_{new} = LR_{old} \times \frac{1}{1 + 0.05 \times epoch} \tag{18}$$

(2) FSR 和 MCEL 模型

基于预训练模型对比了区分大小写和不区分大小写分别在BASE和LARGE不同模型大小上的结果。此外, 还添加了一个基于整词mask(Whole word masking, WWM)的预训练BERT模型实验对比。WWM在预训练过程中不单mask单词的某些子词而mask每个单词的所有子词部分, 可以更好的在预训练阶段保留整个单词的所有子词信息连贯性。

(1) $\text{BERT}_{\text{BASE}}$: 12层, 隐层768维度和12个注意力头, 总参数量1.1亿。

(2) $\text{BERT}_{\text{LARGE}}$: 24层, 隐层1024维度和16个注意力头, 总参数量3.4亿。

(3) BERT_{WWM}: 24层, 隐层1024维度和16个注意力头, 总参数量3.4亿。

FSR和MCEL模型添加“[CLS]”符号在每个文本的开头, 并且使用“[SEP]”符号用作文本的结尾。解码层标注采用“BIO”格式。并且在BERT层之后采用了两层Highway网络。所有这些经过预训练的模型都可以在Github[29]上找到, 此外本文使用的是PyTorch版本[30]。在下游网络空间安全命名实体识别任务微调阶段, FSR和MCEL模型的绝大多数超参和BERT论文里设置基本一致。最大文本长度设置为128, Dropout设置为10%, 学习率设置为5e-5, adam_epsilon设置为1e-8, warmup_proportion设置为0.4, batch_size设置为10。

5 结果分析

5.1 性能对比

表1基于网络空间安全命名实体识别数据上对比了本文提出的模型和基线模型的效果。其中预训练模型FSR和MCEL采用的是基于区分大小写单词(cased)的BERT-BASE模型权重即$\text{BERT}_{\text{BASE-cased}}$。从表中可以看出, 在BiLSTM模型上添加字符级别信息和在解码层采用CRF模型都可以提升F1值。这证明了添加字符级别信息可以更加充分利用文本的语义信息从而带来效果的提升。对FSR模型加上CRF层也可以提升F1值, 这证明了在解码层使用CRF模型可以在命名实体识别任务上获得性能提升。本文的模型FSR-CRF和MCEL在网络空间安全数据集上都超过了基线方法中最优模型BiLSTM-CharCNN-CRF, 证明了基于预训练的模型在网络安全特定领域的命名实体识别任务上是有帮助的。此外, 基于$\text{BERT}_{\text{BASE-cased}}$模型大小的MCEL模型F1值可以达到87.13%, 优于FSR-CRF模型。因为MCEL充分利用了单词的所有子词的信息, FSR模型仅仅使用了每个单词的第一个子词信息。

图3展示了不同模型对于提取每种安全实体类型的实验结果。可以看出基于预训练模型的FSR和MCEL的结果均优于LSTM模型。此外基于预训练的模型在提取“Network_Terms”、“File_Name”和“Hardware”三种安全实体上获取了较大的提升。“Network_Terms”多为网络术语例如IP, HTTP, 大多都是词表中没有覆盖到的OOV词, 进而可以证明基于预训练的模型可以很好的解决安全文本中OOV的问题, 也可以很好的识别网络空间安全数据集中重要的实体“Network_Terms”。

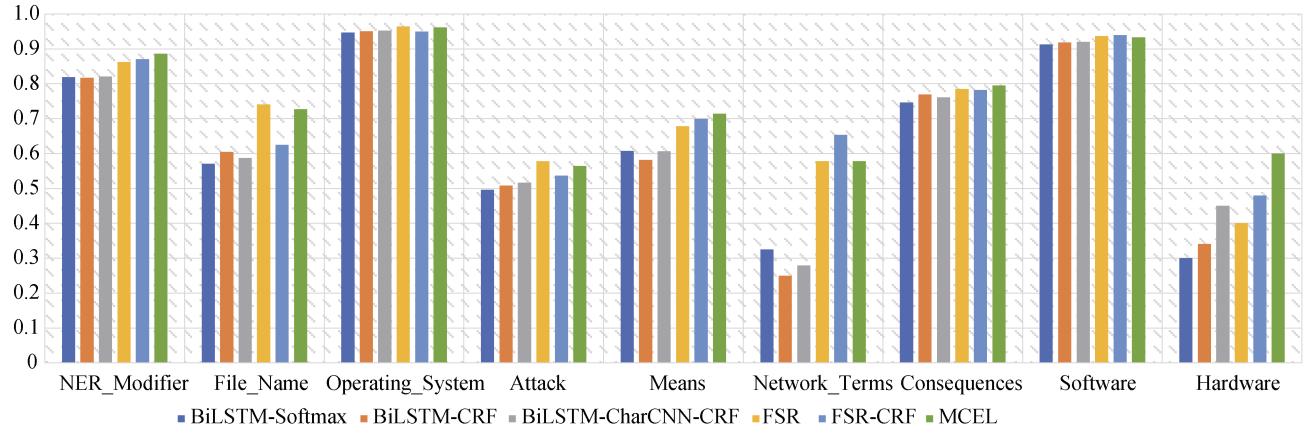

图 3　不同模型在网络空间安全数据集上每个实体类别的 F1 值。其中预训练模型 FSR、FSR-CRF 和 MCEL 采用 $BERT_{BASE\text{-}cased}$ 的模型大小

Figure 3　The F1 scores of different models for each entity category on the cybersecurity dataset. The pre-trained FSR, FSR-CRF and MCEL models are based on $BERT_{BASE\text{-}cased}$ model

表 1　网络空间安全命名实体识别数据集上结果对比

Table 1　Results on the cybersecurity NER dataset

模型	Precision	Recall	F1
BiLSTM-Softmax	85.69	84.41	85.04
BiLSTM-CRF	86.12	84.79	85.45
BiLSTM-CharCNN-CRF	86.67	85.06	85.85
FSR-$BERT_{BASE\text{-}cased}$	85.75	85.92	85.84
FSR-CRF-$BERT_{BASE\text{-}cased}$	85.01	**86.98**	85.98
MCEL-$BERT_{BASE\text{-}cased}$	**87.43**	86.83	**87.13**

如表 2 所示, 本文还基于 MCEL 模型探索了不同模型大小和是否区分单词大小写对实验结果的影响并进行了详细的实验比较。其中包括 $BERT_{BASE\text{-}cased}$、$BERT_{BASE\text{-}uncased}$、$BERT_{LARGE\text{-}cased}$、$BERT_{LARGE\text{-}uncased}$、$BERT_{WWM\text{-}cased}$ 和 $BERT_{WWM\text{-}uncased}$。通过实验结果可以观察到基于单词区分大小写的模型效果优于不区分单词大小写, 基于 WWM 的模型可以获取最高的 F1 值, WWM 可以减轻在 BERT 预训练中只 mask 部分子词的缺点。基于 $BERT_{WWM\text{-}cased}$ 的 MCEL 模型获得了 87.73%的 F1 值, 相较于 BiLSTM-CharCNN-CRF 提升了 1.88%。

5.2　消融研究

上述实验结果已经证明了基于预训练模型 FSR 和 MCEL 的有效性, 本文还希望了解模型中组件的具体贡献。为此本文探索“[SEP]”符号、在解码层使用不同标注格式“BIEOS”、“BIO”以及 Highway 网络层数的贡献。相关实验基于 MCEL 模型, 模型权重采用 $BERT_{BASE\text{-}cased}$。

表 3 展示了不同组件对模型的影响结果。可以观察到使用“BIEOS”的标注格式会大大降低模型对安全实体识别的精准率, 网络空间安全命名实体识别数据一共 9 类实体类型, 如果使用“BIEOS”标注格式, 标签的总类别会增多, 精度低于“BIO”的标注格式。同时在文本句尾不添加“[SEP]”符号也会影响模型的精准率。“[SEP]”符号放在文本的末尾,

表 2　MCEL 基于不同模型大小在网络空间安全命名实体识别数据集上的结果对比。U 代表 uncased, C 代表 cased

Table 2　Results on the cybersecurity NER dataset with different model Sizes based on MCEL model. U means uncased, C means cased

模型	Precision	Recall	F1
BiLSTM-CharCNN-CRF	86.67	85.06	85.85
U-MCEL-$BERT_{BASE}$	87.44	86.40	86.92
U-MCEL-$BERT_{LARGE}$	87.23	87.44	87.33
U-MCEL-$BERT_{WWM}$	87.56	87.32	87.44
C-MCEL-$BERT_{BASE}$	87.73	86.83	87.13
C-MCEL-$BERT_{LARGE}$	86.90	**88.32**	87.61
C-MCEL-$BERT_{WWM}$	**87.96**	87.50	**87.73**

表 3　基于网络空间安全命名实体识别数据集的消融研究

Table 3　Ablation study on the cybersecurity NER dataset

模型	Precision	Recall	F1
BiLSTM-CharCNN-CRF	86.67	85.06	85.85
$MCEL_{BIO\text{-}SEP\text{-}Highway\text{-}2}$	87.43	86.83	**87.13**
w/ BIEOS	85.31	87.08	86.19
w/o SEP	86.46	**87.49**	86.97
w/o Highway	86.53	87.48	87.01
w/ Highway-1	**87.72**	86.21	86.96
w/ Highwway-3	86.99	87.08	87.04

可以更好的确定测试集中每条文本的真实长度, 从而可以获取更高的精度。Highway 网络使用门机制有效解决网络层数过深不容易训练的问题, 可以提升模型的泛化性。通过表 3 观察到使用 Highway 网络可以有效提高安全实体识别精准率, 但是模型性能并不会随着 Highway 层数的增加而增加, 通过实验发现使用两层 Highway 网络可以获得较高的 F1 值。

6 结论和未来展望

针对网络空间安全特定领域的命名实体识别存在的实体类型多样、多义词、OOV 等问题, 本文提出了两个基于预训练的新模型 FSR 和 MCEL, 用来提取网络空间安全文本中的实体。此外, 本文提出的模型很好的解决了因 BERT 模型中使用 WordPiece 分词器带来的子词和标签不匹配的问题, 证明了本文提出模型的通用性。

本文基于真实网络空间安全文本语料库进行了广泛的实验。实验结果表明, 本文提出的模型优于之前最优模型, 尤其 MCEL 模型在网络空间安全数据集上的 F1 值达到了 87.73%, 提升了 1.88%。

未来, 我们希望提出基于大规模网络空间安全文本上预训练的特定安全领域的语言表示模型。

致 谢 感谢中国科学院网络测评技术重点实验室的各位老师和同学提出的有益建议。感谢审稿专家和编辑部老师对本文提出的有益建议及指导。

参考文献

[1] Liu Q, Li Y, Duan H, et al. Knowledge Graph Construction Techniques[J]. *Journal of Computer Research and Development*, 2016, 53(3): 582-600.
(刘峤, 李杨, 段宏, 等. 知识图谱构建技术综述[J]. *计算机研究与发展*, 2016, 53(3): 582-600.)

[2] Chen Y B, Xu L H, Liu K, et al. Event Extraction *via* Dynamic Multi-Pooling Convolutional Neural Networks[C]. *The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (*Volume 1*: *Long Papers*), 2015: 167-176.

[3] Diefenbach D, Lopez V, Singh K, et al. Core Techniques of Question Answering Systems over Knowledge Bases: A Survey[J]. *Knowledge and Information Systems*, 2018, 55(3): 529-569.

[4] Weerawardhana S, Mukherjee S, Ray I, et al. Automated Extraction of Vulnerability Information for Home Computer Security[M]. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2015: 356-366.

[5] John L, Andrew M, Pereira Fernando C N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[J]. *ICML '01*: *Proceedings of the Eighteenth International Conference on Machine Learning*, 2001: 282-289.

[6] Liao X J, Yuan K, Wang X F, et al. Acing the IOC Game: Toward Automatic Discovery and Analysis of Open-Source Cyber Threat Intelligence[C]. *The 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016: 755-766.

[7] Yan H, Deng B C, Li X N, et al. TENER: Adapting Transformer Encoder for Named Entity Recognition[EB/OL]. 2019: 1911.04474. https://arxiv.org/abs/1911.04474v3.

[8] Vaswani A, Shazeer N, Parmar N, et al. Attention is all You Need[C]. *The 31st International Conference on Neural Information Processing Systems*, 2017: 6000-6010.

[9] Devlin J, Chang M W , Lee K , et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. (Volume 1: Long and Short Papers). 2019: 4171-4186.

[10] Lan Z Z, Chen M D, Goodman S, et al. ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations[EB/OL]. 2019: 1909.11942. https://arxiv.org/abs/1909.11942v6.

[11] Sanh V, Debut L, Chaumond J, et al. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter[EB/OL]. 2019: arXiv preprint arXiv:1910.01108.

[12] Wu Y H, Schuster M, Chen Z F, et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation[EB/OL]. 2016: arXiv: 1609.08144. https://arxiv.org/abs/1609.08144.

[13] Srivastava R K, Greff K, Schmidhuber J, et al. Highway Networks[EB/OL]. 2015: 1505.00387. https://arxiv.org/abs/1505.00387v2.

[14] Huang Z H, Xu W, Yu K. Bidirectional LSTM-CRF Models for Sequence Tagging[EB/OL]. 2015: 1508.01991. https://arxiv.org/abs/1508.01991v1.

[15] Yang J, Liang S L, Zhang Y. Design Challenges and Misconceptions in Neural Sequence Labeling[EB/OL]. 2018: 1806.04470. https://arxiv.org/abs/1806.04470v2.

[16] Rei M, Crichton G K O, Pyysalo S. Attending to Characters in Neural Sequence Labeling Models[EB/OL]. 2016: 1611.04361. https://arxiv.org/abs/1611.04361v1.

[17] Ma X Z, Hovy E. End-to-End Sequence Labeling *via* Bi-Directional LSTM-CNNS-CRF[EB/OL]. 2016: 1603.01354. https://arxiv.org/abs/1603.01354v5.

[18] Joshi A, Lal R, Finin T, et al. Extracting Cybersecurity Related Linked Data from Text[C]. *2013 IEEE Seventh International Conference on Semantic Computing*, 2013: 252-259.

[19] Ma P C, Jiang B, Lu Z G, et al. Cybersecurity Named Entity Recognition Using Bidirectional Long Short-Term Memory with Conditional Random Fields[J]. *Tsinghua Science and Technology*, 2021, 26(3): 259-265.

[20] Yi F, Jiang B, Wang L, et al. Cybersecurity Named Entity Recognition Using Multi-Modal Ensemble Learning[J]. *IEEE Access*, 2020, 8: 63214-63224.

[21] Qin Y, Shen G W, Yu H X. Large-Scale Network Security Entity Recognition Method Based on Hadoop[J]. *CAAI Transactions on Intelligent Systems*, 2019, 14(5): 1017-1025.

(秦娅, 申国伟, 余红星. 基于Hadoop的大规模网络安全实体识别方法[J]. *智能系统学报*, 2019, 14(5): 1017-1025.)

[22] Qin Y, Shen G W, Zhao W B, et al. A Network Security Entity Recognition Method Based on Feature Template and CNN-BiLSTM-CRF[J]. *Frontiers of Information Technology & Electronic Engineering*, 2019, 20(6): 872-884.

[23] Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and Their Compositionality[EB/OL]. 2013: 1310.4546. https://arxiv.org/abs/1310.4546v1.

[24] Pennington J, Socher R, Manning C. Glove: Global Vectors for Word Representation[C]. *The 2014 Conference on Empirical Methods in Natural Language Processing*, 2014: 1532-1543.

[25] Liu Y J, Meng F D, Zhang J C, et al. GCDT: A Global Context Enhanced Deep Transition Architecture for Sequence Labeling[C]. *The 57th Annual Meeting of the Association for Computational Linguistics*, 2019: 2431-2441.

[26] Beltagy I, Lo K, Cohan A. SciBERT: A Pretrained Language Model for Scientific Text[EB/OL]. 2019: 1903.10676. https://arxiv.org/abs/1903.10676v3.

[27] Lee J, Yoon W, Kim S, et al. BioBERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining[J]. *Bioinformatics*, 2020, 36(4): 1234-1240.

[28] Forney G D. The Viterbi Algorithm[J]. *Proceedings of the IEEE*, 1973, 61(3): 268-278.

[29] https://github.com/google-research/bert.

[30] https://github.com/huggingface/transformer.

韩瑶鹏 于 2017 年在中国矿业大学计算机科学与技术专业获得学士学位。现在中国科学院信息工程研究所第六研究室攻读硕士学位。研究领域为网络安全态势感知、知识图谱等。Email: hanyaopeng@iie.ac.cn

王璐 于 2018 年在河北工业大学软件工程专业获得学士学位。现在中国科学院大学信息工程研究所第六研究室攻读硕士学位。研究领域为网络安全态势感知、知识图谱。Email: wanglu@iie.ac.cn

姜波 于 2016 年在中国科学院大学计算机系统结构专业获得博士学位。现任中国科学院 信息工程研究所副研究员。研究领域为网络 安全态势感知、知识图谱、数据挖掘等。Email: jiangbo@iie.ac.cn

卢志刚 于 2010 年在中国科学院研究生院获得博士学位。现任中国科学院信息工程研究所高级工程师, 中国科学院网络空间安全学院副教授。研究领域为网络安全态势感知、网络攻击检测、移动终端安全等。Email: luzhigang@iie.ac.cn

姜政伟 于 2014 年在中国科学院大学获得博士学位。现任中国科学院信息工程研究所高级工程师, 中国科学院网络空间安全学院副教授。研究领域为威胁情报、态势感知、网络威胁发现。Email: jiangzhengwei@iie.ac.cn

刘玉岭 于 2013 年在中国科学院软件研究所获得博士学位。现任中国科学院信息工程研究所高级工程师。研究领域为网络安全态势感知、网安大数据分析、安全测评认证等。Email: liuyuling@iie.ac.cn