融合 Focal Loss 的网络威胁情报实体抽取

郭渊博[1]，李勇飞[1]，陈庆礼[1]，方晨[1]，胡阳阳[2]
（1. 信息工程大学密码工程学院，河南 郑州 450001；2. 加利福尼亚大学河滨分校，河滨 CA92521）

摘 要：网络威胁情报（CTI）蕴含丰富的威胁行为知识，及时分析处理威胁情报能够促进网络攻防由被动防御向主动防御的转变。当前多数威胁情报以自然语言文本的形式存在，包含大量非结构化数据，需要利用实体抽取方法将其转换为结构化数据以便后续处理。然而，由于威胁情报中包含大量漏洞名称、恶意软件、APT 组织等专业词汇，且实体分布极不平衡，导致通用领域的实体抽取方法应用于威胁情报时受到极大限制。为此，提出一种融合 Focal Loss 的实体抽取模型，通过引入平衡因子和调制系数改进交叉熵损失函数，平衡样本分布。此外，针对威胁情报结构复杂且来源广泛，包含大量专业词汇的问题，在模型中增加单词和字符特征，有效改善了威胁情报中的 OOV 问题。实验结果表明，相较于现有主流模型 BiLSTM 和 BiLSTM-CRF，所提模型在 $F1$ 分数上分别提高了 7.07%和 4.79%，验证了引入 Focal Loss 和字符特征的有效性。

关键词：网络安全；威胁情报；实体抽取；样本不平衡
中图分类号：TP391
文献标志码：A
DOI: 10.11959/j.issn.1000−436x.2022132

Fusion of Focal Loss's cyber threat intelligence entity extraction

GUO Yuanbo[1], LI Yongfei[1], CHEN Qingli[1], FANG Chen[1], HU Yangyang[2]
1. Department of Cryptogram Engineering, Information Engineering University, Zhengzhou 450001, China
2. University of California, Riverside, Riverside CA92521, USA

Abstract: Cyber threat intelligence contains a wealth of knowledge of threat behavior. Timely analysis and process of threat intelligence can promote the transformation of defense from passive to active. Nowadays, most threat intelligence that exists in the form of natural language texts contains a large amount of unstructured data, which needs to be converted into structured data for subsequent processing using entity extraction methods. However, since threat intelligence contains numerous terminology such as vulnerability names, malware and APT organizations, and the distribution of entities are extremely unbalanced, the performance of extraction methods in general field are severely limited when applied to threat intelligence. Therefore, an entity extraction model integrated with Focal Loss was proposed, which improved the cross-entropy loss function and balanced sample distribution by introducing balance factor and modulation coefficient. In addition, for the problem that threat intelligence had a complex structure and a wide range of sources, and contained a large number of professional words, token and character features were added to the model, which effectively improved OOV (out of vocabulary) problem in threat intelligence. Experiment results show that compared with existing mainstream model BiLSTM and BiLSTM-CRF, the $F1$ scores of the proposed model is increased by 7.07% and 4.79% respectively, which verifies the effectiveness of introducing Focal Loss and character features.

Keywords: cyber security, threat intelligence, entity extraction, label imbalance

0 引言

一直以来，研究者都在网络安全领域开展了大量科研工作，然而传统的网络被动防御已经难以应对高级持续威胁（APT, advanced persistent threat）[1]等新型攻击，网络威胁情报（CTI, cyber threat intelligence）

收稿日期：2022−03−22；**修回日期**：2022−06−02
通信作者：李勇飞，leekgfly@foxmail.com
基金项目：国家自然科学基金资助项目（No.61501515, No.61601515）
Foundation Item: The National Natural Science Foundation of China (No.61501515, No.61601515)

的出现为态势感知的研究提供了新思路[2]。

2013年，Mcmillan[2]首次提出了关于威胁情报的定义：威胁情报是关于现有或即将出现的对资产有威胁的知识，包括场景、机制、指标、启示和可操作建议等，这些知识可为主体提供针对威胁的应对策略。威胁情报可以帮助各组织改进其网络防御架构，更好地了解威胁状况，协调对未知威胁的响应，以此来减少威胁对组织的影响。

然而，网络安全数据存在海量化、分散化、碎片化以及关系隐蔽化的特点，如何及时、精准地对海量数据进行分析处理，提取关键要素和关联关系，挖掘潜在的有价值信息，是网络安全领域面临的重要问题。威胁情报作为一种网络安全数据，通常以文本形式存在，包含大量非结构化数据。这种多源异构性给安全分析师全面、高效地利用威胁情报带来了巨大挑战[3]。

知识抽取作为文本挖掘的关键技术，能够提取不同来源、不同结构的数据中包含的信息，形成知识（结构化数据）[4]。现阶段知识抽取分为实体抽取、关系抽取、事件抽取三大类。知识抽取面向文本数据，通过自动化/半自动化抽取技术提取相应的知识单元。本文主要研究面向非结构化威胁情报的实体抽取问题。

近年来，基于深度神经网络的方法被广泛应用于自然语言处理（NLP, natural language processing）的各项任务中。典型地，卷积神经网络（CNN, convolutional neural network）、长短期记忆（LSTM, long short-term memory）网络以及双向长短期记忆（BiLSTM, bidirection LSTM）网络均被应用于实体抽取任务。

由于网络安全领域包含大量专业词汇和缩略词，可用数据集较少，缺乏统一、规范的分类标准，且句间结构较复杂，标签分布极不平衡，因此该领域知识抽取难度较大，需要不断改进现有算法，获得更好的性能。

本文在传统模型BiLSTM-CRF（conditional random field）的基础上添加了CNN-BiLSTM网络提取字符特征，在一定程度上改善了OOV（out of vocabulary）问题。针对网络安全领域数据集标签分布极不平衡问题，引入Focal Loss损失函数[5]，平衡样本损失。

本文主要的研究工作如下。

1) 融入Focal Loss，改进损失函数，引入平衡因子与调制系数，减少负类样本和易分类样本的权重，使模型在训练时更关注实体部分及困难样本，提升模型性能。

2) 提取单词的形态信息，不需要考虑语言的语法和语义结构，将字符级特征向量与单词级特征向量拼接，有效改善OOV问题。

3) 针对网络安全领域标注语料缺乏的问题，收集汇总威胁情报并进行人工标注。在此基础上，系统对比本文模型与主流神经网络模型BiLSTM和BiLSTM-CRF在实体抽取任务上的性能。实验结果表明，本文提出的模型优于现有模型。

1 相关工作

实体抽取是自然语言处理的一项基本任务[6]。主要是将非结构化文本中的人名、地名、机构名和具有特定意义的实体抽取出来并加以归类，进而组织成半结构化或结构化的信息，再利用其他技术实现文本分析和理解的目的。

早期实体抽取主要采用基于规则或词典的方法，选用特定特征，包括统计信息、标点符号、指示词、方向词、中心词[7]等，以模式与字符串相匹配为主要手段，这种方法对固定模式的文本较有效，在小规模的数据集上容易实现，但模板规则需要专家构建，且难以将所有可能的模式考虑周全，人工成本过高。另外，该方法难以维护，缺乏稳健性，可移植性较差。

基于统计机器学习的方法在网络安全实体抽取中取得了较好的效果。Mulwad等[8]提出了一个原型系统，通过使用支持向量机（SVM, support vector machine）分类器来提取与漏洞、攻击和威胁相关的概念，但该系统所能识别并提取的概念仅限于两类：攻击手段和攻击结果。Bridges等[9]在不同的安全相关语料库上评估了最大熵模型（MEM, maximum entropy model），以避免在训练模型时出现过拟合，而后他们开发了3个网络安全实体抽取器。Jones等[10]描述了一种半监督机器学习方法，并结合主动学习机制，提取网络安全实体和关系。由于其使用简单的Bootstrapping算法，导致提取的结果中含有大量的噪声。除此之外，基于统计机器学习的方法严重依赖特征工程，在面对样本数较少的数据集时往往难以保证模型的学习效果。

随着深度学习在多领域的兴起，Huang等[11]首次将BiLSTM-CRF模型应用于自然语言处理基准序列数据集。Sarhan等[12]引入注意力机制进行安全命名实

体识别，并借助词嵌入技术融合构建知识图谱。Zhao 等[13]提出基于 CNN 的领域识别算法，利用 BiLSTM-CRF 识别妥协指标，将其集成并生成带有领域标签的 CTI。Gasmi 等[14]利用 BiLSTM-CRF 模型提取网络安全概念和实体，比较 3 种基于 LSTM 的模型，这些模型结合依赖特征来提取语义关系。王伟平等[15]将 CNN 与 LSTM 进行组合，构建了一个用于识别包含妥协指标语句的分类器。Wu 等[16]利用依存解析的方法提取电子商务威胁情报中的战术、技术、过程实体，发现了新的攻击模式。

针对已发布的数据集 SemEval-2018 Task 8，Manikandan 等[17]利用 CNN-CRF 模型来识别恶意软件相关的条目，以便进行进一步的网络安全文本分析。与此同时，基于文献[18]工作，Fu 等[19]针对网络安全报告，开发了一个端到端的序列标注系统，其采用 BiLSTM-CNN-CRF 模型在没有特征工程的情况下处理网络安全文本，并在语义评测中取得了较好的排名。

目前，实体抽取最新进展主要集中在基于大量人工标注的数据集训练神经网络模型，从而产生较好的结果。然而人工标注昂贵耗时，尤其是在威胁情报领域，数据标注需要掌握一定的网络安全知识。除此之外，威胁情报中实体出现的频率较低，标签分布极不平衡，且包含大量漏洞名称 APT 组织等专业词汇，为实体抽取带来了巨大挑战。针对上述威胁情报领域的相关问题，本文提出了一种新的实体抽取模型，并通过实验证明了该模型在威胁情报实体抽取方面的有效性。

2　方法

2.1　模型架构

模型架构如图 1 所示，主要包括嵌入层、编码层和带有 Focal Loss 的解码层。

首先，利用随机嵌入的方式获取词嵌入，相较于文献[11,13]中的主流模型，本文在嵌入层利用 BiLSTM 和 CNN 提取字符特征，连接向量获得低维单词表示，有效改善威胁情报包含大量专业词汇的 OOV 问题；然后，利用 BiLSTM 作为编码层，对单词表示进行编码；最后，利用 CRF 对文本向量进行解码，摒弃传统的交叉熵函数，融合 Focal Loss 平衡样本分布，使模型更加关注文本中的实体标签。

2.2　嵌入层

自然语言文本不能被神经网络直接编码，在本文模型中，对每个单词提取其词嵌入特征及字符嵌入特征，并进行连接。嵌入层的输入为单词序列 $\boldsymbol{S}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_i,\cdots,\boldsymbol{x}_m\}$，其中 $\boldsymbol{S}$ 表示输入句子，m 表示句子的长度，$\boldsymbol{x}_i$ 表示句子的第 i 个单词。每个单词可表示为 $\boldsymbol{x}=\{c_1,c_2,\cdots,c_i,\cdots,c_p\}$，其中 c_i 表示单词的第 i 个字符，p 表示单词的长度。

2.2.1　词嵌入层

收集训练集中所有单词，基于随机嵌入的方式将其映射到维度为 d 的单词嵌入矩阵 $\boldsymbol{W}_n^d$，n 表示词汇表的大小（训练语料库中单词的数目）。将输入序列 $\boldsymbol{S}$ 表示为词嵌入序列 $\boldsymbol{\omega}=\{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2,\cdots,\boldsymbol{\omega}_i,\cdots,\boldsymbol{\omega}_m\}$，$\boldsymbol{\omega}_i$ 表示第 i 个词的嵌入向量。

2.2.2　字符嵌入层

文献[20-21]已经证明，CNN 能够通过单词中的字符有效提取形态信息（如单词的前缀或后缀），并将其编码为神经网络表示。图 2 展示了 CNN 的网络结构。利用该网络获得单词中每个字符的特征向量，将其与 BiLSTM 层获得的字符特征向量、词嵌入层获得的单词特征向量拼接，获得原始输入序列低层特征的向量表示。

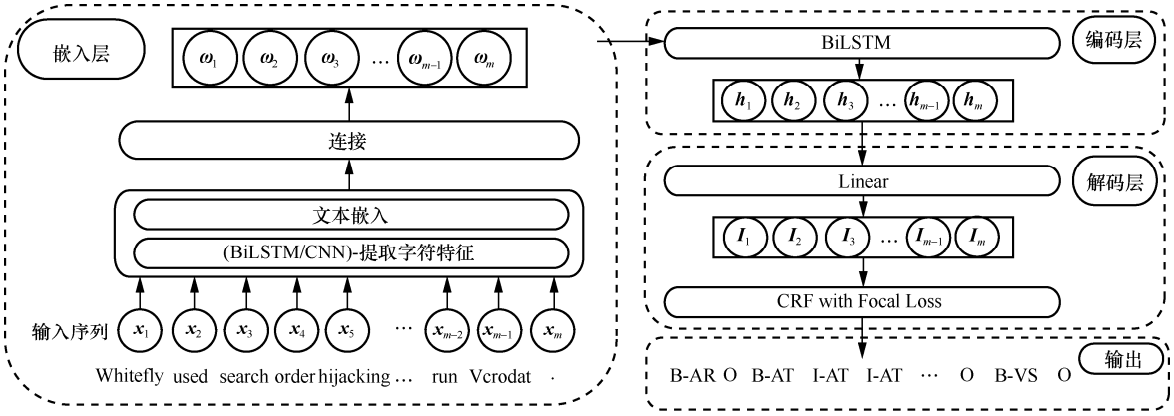


图 1　模型架构

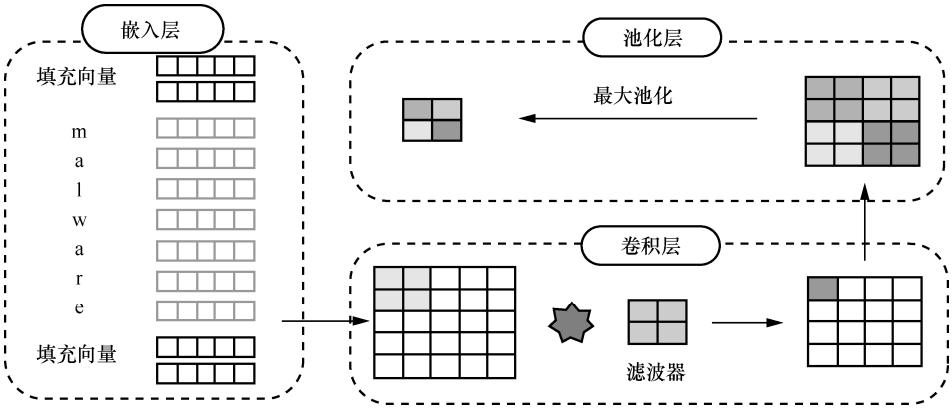


图 2　CNN 的网络结构

2.3　BiLSTM 编码层

在序列标注任务中，利用当前单词之前的文本信息（历史信息）和当前单词之后的文本信息（未来信息）能够有效提升模型的性能，但 LSTM 隐藏层仅能获取历史信息。在文献[11]中，BiLSTM 编码层展示了其捕获每个单词语义信息的有效性。BiLSTM 包括前向 LSTM 层、后向 LSTM 层和连接层。每个 LSTM 包含一组循环连接子网络，称为存储模块。每个时间步都是一个 LSTM 存储模块，基于前一个时刻隐藏向量 $\boldsymbol{h}_{t-1}$、前一个时刻存储单元向量 $\boldsymbol{c}_{t-1}$ 和当前输入单词嵌入向量 $\boldsymbol{\omega}_t$ 运算获得。

2.4　CRF 解码层

对于序列标注任务，考虑相邻标签之间的相关性，联合解码出标签序列能够有效提升模型性能。例如，在实体抽取中，“I-ORG”不可能在“I-VIRUS”之后。因此，本文引入了 CRF，损失函数包括 2 种分数：发射分数和转移分数。发射分数是 BiLSTM 编码层的输出矩阵 $\boldsymbol{P}_{i,y_i}$，表示单词 i 对应的标签为 y_i 的概率；转移分数是 CRF 层的矩阵 $\boldsymbol{A}_{y_i,y_{i+1}}$，表示标签之间的转移关系。分数定义为

$$\text{Score}(x,y)=\sum \boldsymbol{P}_{i,y_i}+\sum \boldsymbol{A}_{y_i,y_{i+1}} \tag{1}$$

损失函数定义为

$$\text{Loss}_{\text{CRF}}=\ln\left(\sum_{\tilde{y}\in Y_x}\text{e}^{\text{Score}(x,\tilde{y})}\right)-\text{Score}(x,y) \tag{2}$$

其中，$\tilde{y}$ 表示 y 的预测值，Y_x 表示该句所有可能的标签序列。

2.5　Focal Loss 损失函数

网络威胁情报通常为长文本，数据中包含大量的非实体词（标签为 O），二元交叉熵损失函数迭代缓慢且可能偏离正确的优化方向，无法调整至最优。另外，在网络安全领域实体抽取任务中，存在严重的标签分布不平衡问题，因此本文引入 Focal Loss 函数对模型进行优化。Focal Loss 函数是二元交叉熵损失函数的一个变种，通过改变正负样本的相对频率和降低简单样本的贡献权重来解决二元交叉熵损失函数中的类不平衡问题，使学习更难的样本成为可能，Focal Loss 定义为

$$\text{Loss}_{\text{Focal}}=-\alpha\left(1-P(y|x)\right)^{\gamma}\ln\left(P(y|x)\right) \tag{3}$$

其中，$\alpha\in[0,1]$ 是平衡因子，用于平衡正负样本的数量；$\gamma\geqslant 0$ 是调制系数，用于减少非实体样本（简单样本）的损失，使模型更加关注于实体标签（困难样本）；$P(y|x)$ 是单词 x 的标签为 y 的概率。例如，当 γ=2 时，对于置信度为 0.9 的简单样本与置信度为 0.6 的困难样本，其权重比例由交叉熵中的 1:4 变为 1:16，有效增强了困难样本的影响。Focal Loss 能够削弱简单样本对梯度更新方向的主导作用，避免网络学习到大量无用的信息。同时能够避免模型向样本多的类别偏移，缓解类别不平衡问题。

3　实验

3.1　数据集

由于网络安全领域数据集较少，本文从 Help Net Security、The Hacker News 等网站预先爬取 80 篇威胁情报并进行人工标注。为验证模型在数据样本较少时的训练效果，从中选出 25 篇作为训练样本，55 篇作为测试样本。训练集共包括 1 199 个实体，其中各实体样本分布如表 1 所示。实体标签由实体边界和实体类别组成，采用“BIO”模型来识别单词在实体中的具体位置，B（Begin）表示实

表 1　　数据集实体样本分布

实体类型	释义	数量/个	样例
TIME	时间	146	Apparently, the breach started on **September 6, 2021.**
VIRUS	恶意脚本	232	Buckeye uses a variant of **DoublePulsar.**
SOFTWARE	合法软件	46	The group has succeed in comprising **Microsoft Exchange.**
TYPE	攻击类型	27	Conducting **DdoS attacks** on key historical dates is not new.
LOCATION	国家地区	185	The top 10 regions targeted by BEC scammers are led by the **U.S.** .
ORG	组织厂商	181	**Symantec** will continue to monitor their activities and respond in kind.
ATTACKER	攻击者	195	**Lazarus** was subsequently implicated in the WannaCry ransomware attacks.
VERSION	版本	5	Both Windows **7** and Windows Server **2008** ceased receiving support.
OS	操作系统	30	Researchers implemented ESP in Simple Gallery, a popular app on **Android.**
EVENT	攻击事件	39	Symantec believes that the attackers behind **the Anthem breach** are part of a highly resourceful cyberespionage group called Black Vine.
ATTACK	攻击方式	113	Leafminer attempts to infiltrate target networks through various means of intrusion: **watering hole websites**, **vulnerability scans** of network services

体的开始位置，I（Inside）表示实体的内部或结尾，O（Outside）表示该单词为非实体词，在其之后连接实体类型。

3.2　实验设置

本文采用实体抽取任务常用的指标对模型性能进行评价，即精确率（P, precision）、召回率（R, recall）和 $F1$ 分数（F1-score），并选取 $F1$ 作为综合性指标。

统计训练集句子的长度分布如图 3 所示，句子长度由单词数量衡量。模型超参数设置如下：输入句子的最大长度为 50，对于长度不足 50 的句子使用“<pad>”进行填充；词嵌入维度 $d^{w}=50$，基于 BiLSTM 和 CNN 提取的字符特征 $d^{c}=25$，编码层神经元数量 $d^{e}=100$，CRF 解码层维度 $d^{d}=100$；批处理大小 batch size=64，训练轮次 epoch=100，学习率 lr=0.025，平衡因子 $\alpha=0.96$、γ=2。实验采用 AdamW 算法更新模型参数，并对模型进行优化。

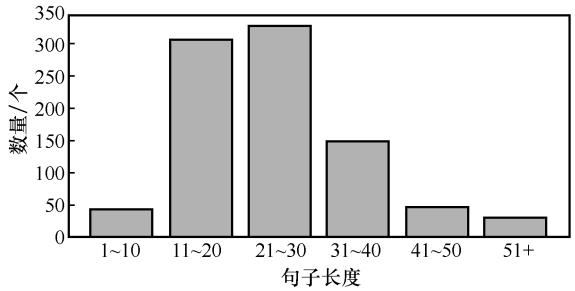


图 3　统计训练集句子的长度分布

针对具有特定形式的网络安全实体 CVE 漏洞编码、Email 地址等，本文模型采用基于正则匹配的方法进行实体识别，如表 2 所示。

表 2　　部分实体的正则表达式

实体类型	正则表达式
CVE 漏洞编码	CVE\-[0-9]{4}\-[0,9]{4,6}
Email 地址	[a-zA-Z0-9_-]+@[a-zA-Z0-9_-]+(\.[a-zA-Z0-9_-]+)+
SHA1 码	[a-f0-9]{40}|[A-F0-9]{40}
MD5 码	[a-f0-9]{32}|[A-F0-9]{32}
IP 地址	\d{1,3}\.\d{1,3}\.\d{1,3}\.\d{1,3}

3.3　实验结果

为进一步说明本文模型的优势，复现 BiLSTM 和 BiLSTM-CRF 模型，并将其作为基线模型，与本文模型进行比较，结果如表 3 所示。由表 3 可以看出，融合单词的字符特征能够有效解决网络安全领域的 OOV 问题。在 CRF 层中融入 Focal Loss 能够平衡样本分布，使模型更加关注实体标签，减少非实体样本给模型带来的影响。本文提出的融合 Focal Loss 的(CNN-BiLSTM)-BiLSTM-CRF 模型在实体抽取任务上表现出最优的性能，相较于主流模型 BiLSTM 和 BiLSTM-CRF 在 $F1$ 分数上分别提升了 7.07%和 4.79%。

表 3　　模型性能比较

模型	P	R	$F1$
BiLSTM	**76.32%**	31.50%	44.59%
BiLSTM-CRF	71.77%	34.80%	46.87%
(CNN-BiLSTM)-BiLSTM-CRF$_{FL}$	72.06%	**40.26%**	**51.66%**

4 结果分析

4.1 消融实验

相较于现有工作，本文主要增加了基于 CNN 模型和 BiLSTM 模型的字符特征，以及在 CRF 解码层中增加了 Focal Loss 损失函数。为验证其必要性，本节进行消融实验，观察其对模型性能的改进效果。

表 4 展示了不同损失函数对模型性能的影响。由于标签分布不平衡，移除 Focal Loss 后，使用交叉熵损失函数使模型 *F*1 分数下降了 0.56%，证明了其对模型性能提升的效果。

表 5 展示了字符特征对模型性能的影响。由于威胁情报包含大量专业词汇，分别移除 BiLSTM 和 CNN 字符特征，模型 *F*1 分数下降了 2.07%和 0.62%。同时可以分析得到，CNN 的优势在于能够提取一定的结构特性，而威胁情报中单词通常较长，没有固定的命名标准，因此，捕获长距离依赖信息的 BiLSTM 对模型性能影响更大。

表 6 详细展示了单类别细粒度性能。从表 6 可以直观地发现，引入字符特征能够显著提升 ORG、ATTACKER 等类别的分类能力。这一现象表明，传统方法受限于这些类别中存在的 OOV 问题，而本文模型能够通过学习字符上下文信息对该问题进行改善。

4.2 平衡因子与调制系数对模型性能的影响

为进一步研究平衡因子与调制系数对模型性能的影响，本文进行了平衡因子与调制系数不同组合的对比实验，性能变化曲线如图 4 所示。

表 4 不同损失函数对模型性能的影响

模型	损失函数	*P*	*R*	*F*1
(CNN-BiLSTM)-BiLSTM-CRF$_{FL}$	Focal Loss	72.06%	**40.26%**	**51.66%**
(CNN-BiLSTM)-BiLSTM-CRF	BCE	**73.30%**	39.22%	51.10%

表 5 字符特征对模型性能的影响

模型	字符特征	*P*	*R*	*F*1
(CNN-BiLSTM)-BiLSTM-CRF$_{FL}$	CNNChar+BiLSTMChar+Word	72.06%	**40.26%**	**51.66%**
(BiLSTM)-BiLSTM-CRF$_{FL}$	BiLSTMChar+Word	**72.28%**	39.45%	51.04%
(CNN)-BiLSTM-CRF$_{FL}$	CNNChar +Word	68.46%	38.87%	49.59%
BiLSTM-CRF$_{FL}$	Word2Vec	71.45%	35.96%	47.84%

表 6 单类别细粒度性能

实体类型	BiLSTM-CRF$_{FL}$			(BiLSTM)-BiLSTM-CRF$_{FL}$			(CNN)-BiLSTM-CRF$_{FL}$			(CNN-BiLSTM)-BiLSTM-CRF$_{FL}$		
	P	*R*	*F*1	*P*	*R*	*F*1	*P*	*R*	*F*1	*P*	*R*	*F*1
TIME	83.79%	75.96%	79.68%	88.42%	59.59%	60.00%	86.40%	77.71%	81.82%	89.11%	79.14%	**83.83%**
VIRUS	51.11%	19.26%	27.98%	59.59%	29.12%	**39.12%**	50.06%	22.83%	31.36%	54.47%	27.75%	36.77%
SOFTWARE	35.81%	19.26%	25.05%	34.41%	21.53%	26.49%	32.64%	20.59%	25.25%	32.56%	22.36%	**26.51%**
TYPE	54.47%	38.09%	44.83%	53.67%	38.84%	**45.07%**	53.65%	35.62%	42.81%	55.81%	36.20%	43.92%
LOCATION	86.35%	83.13%	84.71%	87.03%	83.77%	85.37%	85.94%	83.51%	84.71%	86.45%	84.82%	**85.63%**
ORG	61.37%	42.11%	49.95%	63.78%	44.42%	52.37%	62.52%	44.66%	52.10%	65.80%	44.63%	**53.19%**
ATTACKER	63.87%	17.49%	27.46%	60.00%	17.33%	26.89%	58.57%	17.50%	26.95%	60.94%	20.05%	**30.17%**
VERSION	0.25%	0.10%	0.14%	0.25%	0.10%	0.14%	0.50%	0.08%	0.14%	—	—	—
OS	50.98%	45.78%	48.24%	48.20%	44.47%	46.26%	47.83%	43.80%	45.73%	51.06%	46.30%	**48.56%**
EVENT	35.47%	8.04%	13.11%	36.08%	9.46%	**14.99%**	34.34%	9.37%	14.72%	30.34%	9.33%	14.27%
ATTACK	60.50%	17.83%	27.54%	59.89%	18.86%	28.69%	61.80%	20.11%	**30.35%**	59.29%	18.58%	28.29%

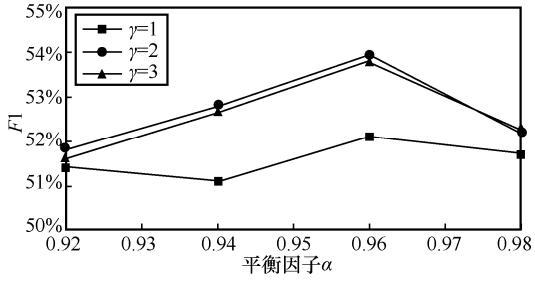


图 4 平衡因子与调制系数对系统性能的影响

由图 4 可以看出，当$\alpha = 0.96$、γ=2 时，系统能够获得最好的性能。

4.3 实例分析

3.3 节中对比了几种实体抽取方法，本节利用实例进一步比较模型的抽取效果。模型抽取实例如表 7 所示，“[]”表示能够正确识别的实体，“[]”表示未被正确识别的实体，实体下标表示实体类型。

针对实例 1，BiLSTM 模型未能很好地考虑标签之间的相互关系，错误地将“them”识别为“I-ATTACK”；而另外 2 种模型则准确识别出相应实体。

针对实例 2，2 种基线模型都未能识别出恶意软件“Android.Doublehidden”，此结果可能是由训练集中缺乏相关词汇，且类似正样本较少导致。而本文模型融入了字符特征，利用 Focal Loss 平衡了正负样本分布，因此可正确识别出该未知单词并分类。

针对实例 3，3 种模型均未能准确识别出“VERSION”类型的实体“7”“2003”，且实体“Windows Server”的边界识别有误，由此说明本文模型在少样本实体抽取中还有待改进和优化。

5 结束语

为实现面向网络威胁情报的实体抽取，解决该领域存在的样本分布不平衡问题，本文提出了一种融合 Focal Loss 的实体抽取模型，并通过实验验证了模型的有效性。所提模型在主流模型的基础上增加了字符特征，有效解决了威胁情报中存在的 OOV 问题。在未来工作中，笔者将进一步探索威胁情报中的关系抽取问题，构建知识图谱，实现网络安全实体的关联分析。

参考文献：

[1] DALY M K. Advanced persistent threat[C]//Proceedings of 23rd Large Installation System Administration Conference. Berkeley: USENIX Association, 2009: 1-6.

[2] MCMILLAN R. Definition: threat intelligence[R]. Garter Research, 2013.

[3] 李涛, 郭渊博, 琚安康. 融合对抗主动学习的网络安全知识三元组抽取[J]. 通信学报, 2020, 41(10): 80-91.
LI T, GUO Y B, JU A K. Knowledge triple extraction in cybersecurity with adversarial active learning[J]. Journal on Communications, 2020, 41(10): 80-91.

[4] HOHENECKER P, MTUMBUKA F, KOCIJAN V, et al. Systematic

表 7　　模型抽取实例

模型	实体抽取结果
实例 1	We reported these apps to **[Microsoft]$_{ORG}$** and they subsequently removed them from their **[store]$_{SOFTWARE}$**.
BiLSTM	We reported these apps to **[Microsoft]$_{ORG}$** and they subsequently removed [them] from their [store].
BiLSTM-CRF	We reported these apps to **[Microsoft]$_{ORG}$** and they subsequently removed them from their **[store]$_{SOFTWARE}$**.
(CNN-BiLSTM)-BiLSTM-CRF$_{FL}$	We reported these apps to **[Microsoft]$_{ORG}$** and they subsequently removed them from their **[store]$_{SOFTWARE}$**.
实例 2	The malware (**[Android.Doublehidden]$_{VIRUS}$**) is localized in the Persian language, which aims to compromise Android apps.
BiLSTM	The malware ([Android.Doublehidden]) is localized in the Persian language, which aims to compromise **[Android]$_{OS}$** [apps].
BiLSTM-CRF	The malware ([Android.Doublehidden]) is localized in the Persian language, which aims to compromise **[Android]$_{OS}$** [apps].
(CNN-BiLSTM)-BiLSTM-CRF$_{FL}$	The malware (**[Android.Doublehidden]$_{VIRUS}$**) is localized in the Persian language, which aims to compromise **[Android]$_{OS}$** apps.
实例 3	This command was then seen searching for all the computer objects in the Active Directory database with filter condition like *server* or *2003* or *7* (returning all **[Windows Server]$_{OS}$**, **[Windows Server]$_{OS}$ [2003]$_{VERSION}$**, or **[Windows]$_{OS}$ [7]$_{VERSION}$** instances).
BiLSTM	This command was then seen searching for all the computer objects in the Active Directory database with filter condition like *server* or *2003* or *7* (returning all **[Windows]$_{OS}$** [Server], **[Windows]$_{OS}$** [Server] [2003], or **[Windows]$_{OS}$** [7] instances).
BiLSTM-CRF	This command was then seen searching for all the computer objects in the Active Directory database with filter condition like *server* or *2003* or *7* (returning all **[Windows]$_{OS}$** [Server], [Windows Server] [2003], or **[Windows]$_{OS}$** [7] instances).
(CNN-BiLSTM)-BiLSTM-CRF$_{FL}$	This command was then seen searching for all the computer objects in the Active Directory database with filter condition like *server* or *2003* or *7* (returning all **[Windows]$_{OS}$** [Server], **[Windows]$_{OS}$** [Server] [2003], or **[Windows]$_{OS}$** [7] instances).

comparison of neural architectures and training approaches for open information extraction[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg: Association for Computational Linguistics, 2020: 8554-8565.

[5] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//Proceedings of 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 2999-3007.

[6] LEI J B, TANG B Z, LU X Q, et al. A comprehensive study of named entity recognition in Chinese clinical text[J]. Journal of the American Medical Informatics Association, 2013, 21(5): 808-814.

[7] 刘显敏, 李建中. 基于键规则的 XML 实体抽取方法[J]. 计算机研究与发展, 2014, 51(1): 64-75.
LIU X M, LI J Z. Key-based method for extracting entities from XML data[J]. Journal of Computer Research and Development, 2014, 51(1): 64-75.

[8] MULWAD V, LI W J, JOSHI A, et al. Extracting information about security vulnerabilities from web text[C]//Proceedings of 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology. Piscataway: IEEE Press, 2011: 257-260.

[9] BRIDGES R A, HUFFER K M T, JONES C L, et al. Cybersecurity automated information extraction techniques: drawbacks of current methods, and enhanced extractors[C]//Proceedings of 2017 16th IEEE International Conference on Machine Learning and Applications. Piscataway: IEEE Press, 2017: 437-442.

[10] JONES C L, BRIDGES R A, HUFFER K M T, et al. Towards a relation extraction framework for cyber-security concepts[C]//Proceedings of the 10th Annual Cyber and Information Security Research Conference. [S.l.:s.n.], 2015: 1-4.

[11] HUANG Z, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv Preprint, arXiv:150801991, 2015.

[12] SARHAN I, SPRUIT M. Open-CyKG: an open cyber threat intelligence knowledge graph[J]. Knowledge-Based Systems, 2021, 233: 107524.

[13] ZHAO J, YAN Q B, LI J X, et al. TIMiner: automatically extracting and analyzing categorized cyber threat intelligence from social data[J]. Computers & Security, 2020, 95: 101867.

[14] GASMI H, LAVAL J, BOURAS A. Information extraction of cybersecurity concepts: an LSTM approach[J]. Applied Sciences, 2019, 9(19): 3945.

[15] 王伟平, 宁翔凯, 宋虹, 等. iAES: 面向网络安全博客的 IOC 自动抽取方法[J]. 计算机学报, 2021, 44(5): 882-896.
WANG W P, NING X K, SONG H, et al. An indicator of compromise extraction method based on deep learning[J]. Chinese Journal of Computers, 2021, 44(5): 882-896.

[16] WU Y M, LIU Q J, LIAO X J, et al. Price TAG: towards semi-automatically discovery tactics, techniques and procedures of E-commerce cyber threat intelligence[J]. IEEE Transactions on Dependable and Secure Computing, 2021, PP(99): 1.

[17] MANIKANDAN R, MADGULA K, SAHA S. TeamDL at SemEval-2018 task 8: cybersecurity text analysis using convolutional neural network and conditional random fields[C]//Proceedings of the 12th International Workshop on Semantic Evaluation. Stroudsburg: Association for Computational Linguistics, 2018: 868-873 .

[18] MA X Z, HOVY E. End-to-end sequence labeling via bi-directional LSTM-CNNS-CRF[J]. arXiv Preprint, arXiv: 160301354, 2016.

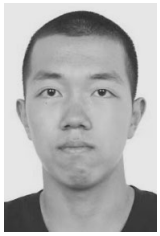
[19] FU M M, ZHAO X M, YAN Y H. HCCL at SemEval-2018 task 8: an end-to-end system for sequence labeling from cybersecurity reports[C]//Proceedings of the 12th International Workshop on Semantic Evaluation. Stroudsburg: Association for Computational Linguistics, 2018: 874-877.

[20] CHIU J P C, NICHOLS E. Named entity recognition with bidirectional LSTM-CNNs[J]. arXiv Preprint, arXiv: 1511.08308, 2015.

[21] SANTOS C N D, ZADROZNY B. Learning character-level representations for part-of-speech tagging[C]//Proceedings of International Conference on Machine Learning. New York: ACM Press, 2014: 1818-1826.

[作者简介]

郭渊博（1975− ），男，陕西周至人，博士，信息工程大学教授、博士生导师，主要研究方向为网络防御、数据挖掘、机器学习和人工智能安全等。



李勇飞（1998− ），男，河南开封人，信息工程大学硕士生，主要研究方向为威胁情报实体抽取及关系抽取等。

陈庆礼（1998− ），男，河南新乡人，信息工程大学硕士生，主要研究方向为人工智能安全。

方晨（1993− ），男，安徽宿松人，博士，信息工程大学讲师，主要研究方向为机器学习、隐私安全。

胡阳阳（1990− ），男，江苏南京人，加利福尼亚大学河滨分校博士生，主要研究方向为机器学习。