网络威胁情报实体识别研究综述

王旭仁[1,2], 魏欣欣[1,2], 王媛媛[3], 姜政伟[2,4], 江 钧[2,4],
杨沛安[2,4], 刘润时[1]

[1]首都师范大学信息工程学院 北京 中国 100048
[2]中国科学院信息工程研究所 中国科学院网络测评技术重点实验室 北京 中国 100093
[3]31005 部队 北京 中国 100089
[4]中国科学院大学网络空间安全学院 北京 中国 100049

摘要 由于网络环境愈发复杂, 网络安全形势日渐严峻, 保护网络免受外来攻击成为一项重要的工作。为了使网络空间攻防技术变为主动防御的形式, 网络威胁情报应运而生。通过对网络威胁情报进行分析和检测, 搜集情报证据, 能够预防攻击行为的发生。因此, 通过共享网络威胁情报来抵御网络攻击变得愈发重要。然而, 网络威胁情报通常以非结构化的形式共享, 将其转化为半结构化或结构化数据对后续很多任务来讲尤为重要, 命名实体识别技术能够实现这一点。虽然在通用领域的命名实体识别已经取得了非常不错的成果, 但在网络威胁情报领域却仍然存在很多问题。本文首先介绍威胁情报相关背景, 及其与命名实体识别之间的联系。然后根据命名实体识别技术发展的时间顺序总结基于规则和词典的实体识别技术、基于无监督学习的实体识别技术、基于特征的监督学习实体识别技术、基于深度学习的实体识别技术等, 全面总结威胁情报领域命名实体识别的研究现状和未来的发展方向。最后, 对比研究威胁情报领域命名实体识别所使用的语料库, 使用 SOTA 深度学习方法进行实验, 分析总结出威胁情报领域数据集所存在的问题。提出的 BBC(BERT-BiGRU-CRF) 深度学习实体识别模型具有最好的实验效果, 在 AutoLabel 数据集、DNRTI 数据集、CTIReports 数据集, 以及 APTNER 数据集上分别达到 97.36%、90.40%、82.87%、73.91% 的 F1 值。

关键词 命名实体识别; 网络威胁情报; 深度学习; 网络威胁情报数据集
中图法分类号 TP391.1 **DOI 号** 10.19363/J.cnki.cn10-1380/tn.2024.11.06

A Survey of Cyber Threat Intelligence Entity Recognition Research

WANG Xuren[1,2], WEI Xinxin[1,2], WANG Yuanyuan[3], JIANG Zhengwei[2,4], JIANG Jun[2,4],
YANG Peian[2,4], LIU Runshi[1]

[1]Information Engineering College, Capital Normal University, Beijing 100048, China
[2]Key Laboratory of Network Assessment Technology, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
[3]Unit 31005 of PLA, Beijing 100089, China
[4]School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China

Abstract As the network environment becomes increasingly complex, the security landscape is growing more severe, making the protection of networks from external attacks a crucial task. In order to transform cybersecurity from a reactive defense approach to proactive defense, Cyber Threat Intelligence (CTI) has emerged. By analyzing and detecting CTI, gathering intelligence evidence, potential attacks can be prevented. Therefore, sharing CTI to defend against cyber-attacks has become increasingly important. However, CTI is often shared in an unstructured format, making its conversion to semi-structured or structured data essential for many subsequent tasks. Named Entity Recognition (NER) technology can facilitate this transformation. Although NER has achieved considerable success in general domains, many challenges remain in the field of CTI. This article first introduces the background of threat intelligence and its connection to NER. Then, it summarizes NER technologies in chronological order, covering rule-based and dictionary-based NER, unsupervised learning methods, feature-based supervised learning methods, and deep learning-based NER. It provides a comprehensive overview of the current research status and future directions of NER in the CTI field. Lastly, a comparative study of the corpora used for NER in CTI is conducted, followed by experiments using state-of-the-art (SOTA) deep learning methods. The analysis identifies issues present in CTI datasets. The proposed BBC (BERT-BiGRU-CRF) deep learning entity recognition model achieves the best experimental results, with F1 scores of 97.36%, 90.40%, 82.87%, and

通讯作者: 魏欣欣, 硕士生, Email: 2211002078@cnu.edu.cn。
本课题得到中国科学院青年创新促进会(No. 2020166), 中科院战略先导项目(No. XDC02030200)的资助。
收稿日期: 2023-02-04; 修改日期: 2023-05-15; 定稿日期: 2024-09-05

73.91% on the AutoLabel, DNRTI, CTIReports, and APTNER datasets, respectively.

Key words named entity recognition; Cyber Threat Intelligence (CTI); deep learning; CTI datasets

1 引言

无论是在通用领域, 还是在网络安全领域, 命名实体识别(Named entity recognition, NER)都占据着很重要的地位。在自然语言处理(Natural language processing, NLP)领域中, 命名实体识别是一项非常关键且重要的任务, 是很多任务的基础和前提。因此, 该技术发展非常迅速, 自第六届信息理解会议(Message understanding conference, MUC)提出 NER 的概念后, 二十余年的时间使得 NER 在各个领域的研究逐渐走向热门并取得了极大的发展, 受到了相关人士的密切关注。

在大数据时代, 命名实体识别作为信息抽取中非结构化数据转化为结构化数据的关键步骤, 根据输入数据的结构和语义精确的定位实体, 并追踪在训练数据集中从来没有注意到的实体。从文本中提取出专有名词或特定命名实体[1], 例如通用领域的时间、地点等; 网络威胁情报(Cyber threat intelligence, CTI)中的病毒、漏洞、样本等。

命名实体识别包括两部分内容: 分别是识别实体的类型和检测实体的边界。通常实体的类型是由专业人士根据特定领域环境、特定语料库进行预先定义, 在进行实体识别后, 将每个词分类到某一个类别上。实体通常不是由一个词组成[2], 例如, 人名通常包含姓和名, 至少要两个单词组成该实体, 因此检测实体的范围也很重要。

随着科技的高速发展, 网络使用越来越便捷, 网络安全形势越来越严峻, 网络攻击持续增长。通过共享网络威胁情报, 分析和挖掘情报信息, 能够应对网络攻击, 将网络攻防技术转为主动形式。因此, 网络威胁情报的数量呈现爆发式增长, 但由于网络威胁情报通常以非结构化的形式存在, 这并不利于它的共享和分析。于是, 针对网络威胁情报的命名实体识别技术应运而生。通过使用 NER 技术能够自动化识别并提取出威胁情报中的实体, 从而更好地理解和分析威胁情报中的关系和趋势, 进而为威胁情报技术的分析和处理提供更加准确和全面的基础, 因此 NER 技术与 CTI 技术存在密切的相关性。此外, NER 技术还可以标准化和规范化 CTI 中的实体名称和属性, 使得不同的安全团队和安全工具之间能够更好地交流和协作。

由于 CTI 中存在大量的实体信息, 该信息是情报分析和处理的基础, 如果不基于 NER 对其进行处理, 那么人工处理这些实体信息的方式会十分繁琐和耗时。因此, NER 技术对于 CTI 领域十分重要, 它能够帮助提高情报分析的效率和性能, 更好地保障网络安全。

无论是通用领域还是威胁情报领域都是 NER 技术的不同应用场景, 但二者在数据源、模型训练和调整, 以及使用场景方面都存在明显差异。基于通用领域的 NER 技术主要应用于 NLP 领域, 目的是自动识别文本中涉及的实体。该技术多用于信息抽取、机器翻译等场景, 旨在帮助计算机更好地理解自然语言。基于威胁情报的 NER 技术主要应用于网络安全领域, 关注的是与网络安全相关的特定目标, 目的是自动识别与网络威胁相关的实体, 对模型进行训练时需要结合网络安全领域的特定知识。该技术多用于威胁情报收集、恶意代码分析、网络安全态势感知等场景, 旨在帮助安全分析人员更好地了解网络威胁的来源、手段和目的, 从而更加有效地对网络攻击进行防御和应对。

但是由于网络威胁情报存在自身的特点, 例如边界模糊、一词多义、存在特殊专业词汇等, 这导致它与通用领域的命名实体识别技术有所不同。

相比通用领域的实体识别, 面向网络威胁情报领域的命名实体识别任务是根据网络安全的领域知识识别网络威胁情报中的恶意程序、漏洞等不同类型的安全实体, 对网络安全领域中的专业词汇进行确认和分类。网络安全领域的数据信息可以从很多渠道获得, 例如一些网络安全公司 Kaspersky、FireEye、Twitter 和一些论坛、博客等等。此外, 一些领域相关信息被存储在通用漏洞披露(Common vulnerabilities & exposures, CVE) 和 NVD(National vulnerability database)中, 使用这些数据库资源, 可以帮助研究者及时发现新的威胁和漏洞, 及时采用响应措施。

将通用领域的 NER 技术进行扩展和优化之后, 能够将其应用到 CTI 领域中。首先是基于规则的方法, 通过添加网络安全领域特定的规则能够应用于 CTI 领域。然后是结合外部知识库的方法, 使用外部知识库来辅助实体识别, 通过结合网络安全领域特有的知识库来适应 CTI 领域的 NER 任务。接下来是基于统计的方法, 使用机器学习算法从大规模文本

数据中学习实体的特征, 通过使用网络安全领域的数据集和特征对模型进行优化, 从而适应 CTI 领域的 NER 任务。最后是基于深度学习的方法, 通过使用网络安全领域的数据集和特征对模型进行训练和优化来完成 CTI 领域的 NER 任务。

本文第二章介绍了威胁情报的国际标准定义和交换标准, 以及威胁情报 NER 的本体定义和下游应用。第三章介绍了通用领域的命名实体识别研究现状, 以实体识别技术出现的时间顺序进行分类, 并总结出每种方法的优缺点, 其中还涉及到了 NLP 领域常见的预训练语言模型。第四章介绍威胁情报领域的命名实体识别研究现状, 重点描述了在网络安全领域上进行实体识别的难点, 以及概括威胁情报领域基于深度学习(Deep learning, DL)技术所使用的基线模型, 同时分门别类总结出该领域 NER 的研究工作。此外, 还总结了基于触发词的 NER 研究现状, 以及网络安全领域最近几年关于 NER 的相关工作。第五章总结了网络安全领域的相关语料库以供研究、学习与使用。第六章介绍 NER 技术的相关评估指标。第七章在开源的威胁情报数据集上进行 NER 实验, 并进行对比分析, 提出一种 BBC(BERT-BiGRU-CRF)威胁情报深度学习实体识别模型。第八章提出网络威胁情报 NER 未来可能的研究方向。最后, 第九章进行全文总结。文章结构如图 1 所示。

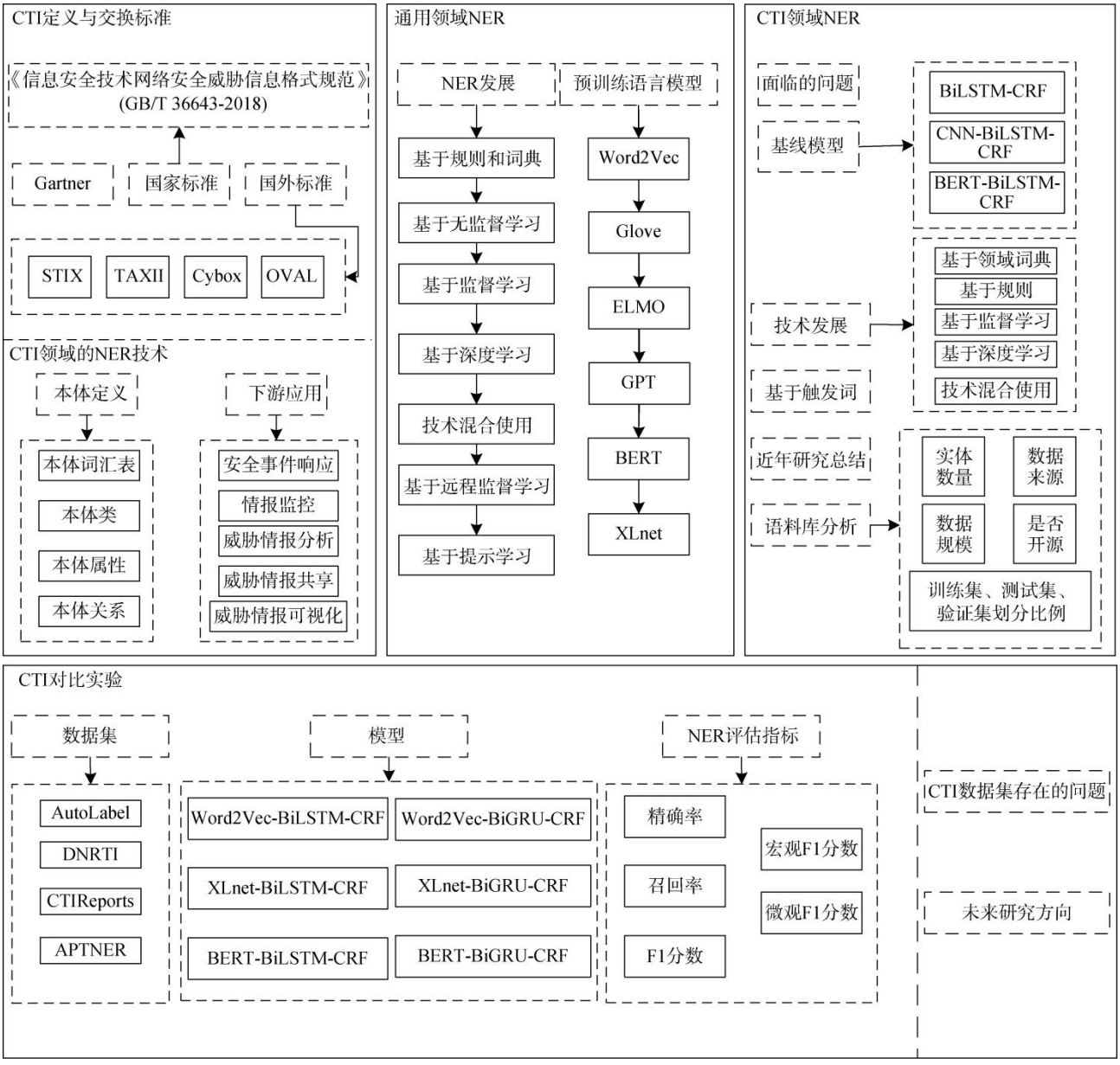

图 1　本文网络威胁情报命名实体识别研究框架

Figure 1　The research framework for named entity recognition in cyber threat intelligence in this paper

2 威胁情报

2.1 威胁情报定义

Gartner 在 2013 年提出的威胁情报定义是: 威胁情报是一种基于证据的知识, 包括上下文、机制、指标、影响、含义和可执行的建议等, 这些知识与资产所面临已有的或潜在的威胁或危害有关, 可为实际的决策提供信息支持[3]。

威胁情报信息通常以非结构化的形式共享, 它们往往来自于一些安全论坛、博客等等, 或者由安全供应商发布。但是, 非结构化的形式不利于研究人员提取其中包含的信息。因此, 使用结构化或者半结构化的形式去共享或者交换威胁情报变得十分重要。命名实体识别技术就是通过将非结构化 CTI 数据转换为结构化或者半结构化形式, 提取其中包含的关键信息, 为网络安全防御做出贡献。

2.2 威胁情报交换标准

为了适应威胁情报本身多源异构的特性, 安全领域制定了一些威胁情报的交换标准, 便于对情报进行表达与共享。

我国在 2018 年 10 月 10 日正式发布《信息安全技术网络安全威胁信息格式规范》(GB/T 36643-2018)[4], 这是我国制定的第一个用于 CTI 共享的国家标准, 该标准在 2019 年 5 月 1 日被正式实施。此标准定义了一个威胁信息模型, 如图 2 所示, 将威胁信息从对象、方法和事件三个维度进行划分, 采用八个威胁信息组件(可观测数据、攻击指标、安全事件、攻击活动、威胁主体、攻击目标、攻击方法、应对措施)来描述网络安全威胁信息, 这 8 个组件又分别被划分到对象域、方法域和事件域。其中, 威胁主体和攻击目标之间的关系属于对象域; 攻击活动、攻击指标、安全事件和可观测数据组成的攻击流程属于事件域; 攻击行动所采取的方法以及防御方所采取的应对攻击的措施属于方法域。

国外的网络威胁情报交换标准主要由 MITRE 发布, MITRE 是由美国政府支持赞助的一个非盈利组织。

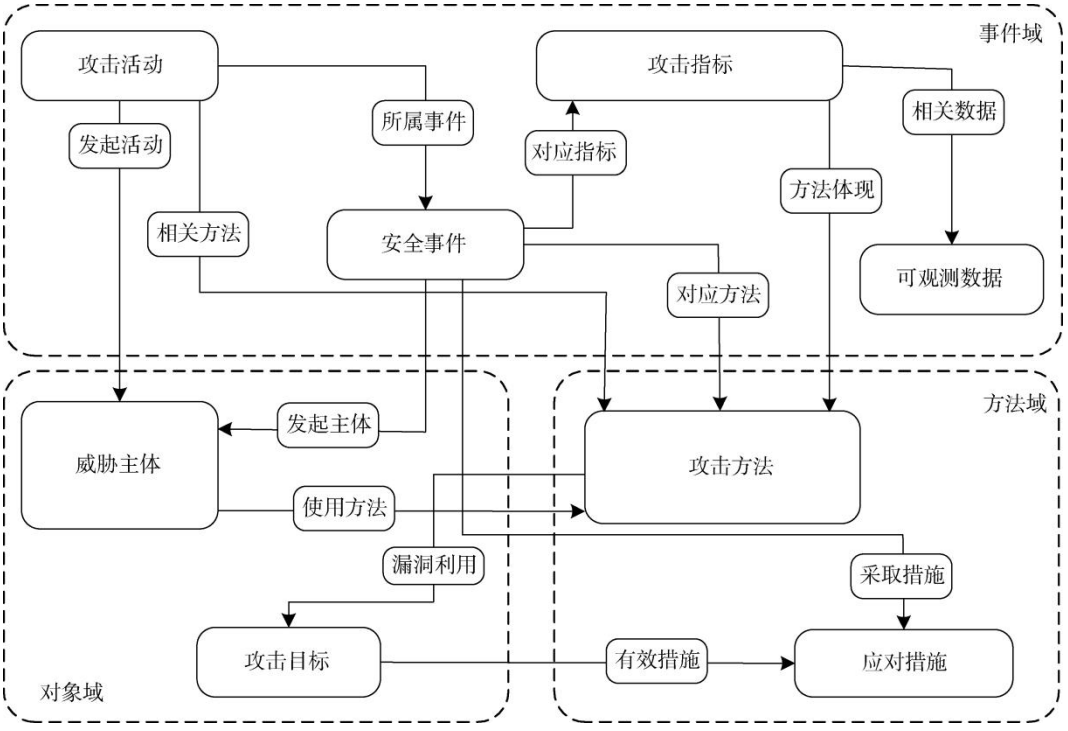

图 2 威胁信息表达模型[4]

Figure 2 Threat information expression model [4]

(1) STIX

结构化威胁信息表达式(Structured threat information eXpression, STIX)作为一种表示 CTI 数据的标准化通信语言, 它规定了威胁情报的内容。通过使用 STIX 规范, 可以通过对象和描述关系表示威胁情报中的威胁活动、威胁因素等多方面特征[5]。

STIX1.0 版本是使用 XML 语言进行表示的, 它定义了 8 种威胁情报域对象。而 STIX2.x 版本是基于 JSON 语言表示的, 最新的版本是 STIX2.1, 它定义了 18 种威胁情报域对象(如表 1 所示)和 2 种关系对象。此外, STIX 是开源且免费的, 方便学者研究与使用。

表 1 STIX2.1 数据对象

Table 1 STIX2.1 data objects

类别名称	描述
攻击方式 Attack Pattern	描述攻击方破坏目标的方式, 能够帮助对攻击进行分类
攻击活动 Campaign	描述在一段时间内对特定目标发起的一组恶意活动或攻击
攻击应对措施 Course of Action	为预防攻击或对攻击做出反应而采取的措施
分组 Grouping	分析和调查时产生的数据
身份 Identity	代表个人、组织或团体以及个人、组织或团体的类别
威胁指标 Indicator	包含可用于检测可疑或恶意网络活动的模式
基础设施 Infrastructure	描述系统、服务等物理或虚拟资源
入侵特征集 Intrusion Set	一组具有共同属性的敌对行为和资源
位置 Location	具体地点
恶意软件 Malware	插入系统中用于破坏数据或系统机密性、完整性或可用性的程序
恶意软件分析 Malware Analysis	恶意软件或家族分析过程中的结果
注意 Note	其他对象中不存在的额外信息
可观察数据 Observed Data	在系统和网络上观测到的信息(例如网络连接、IP 地址)
意见 Opinion	对 STIX 对象中信息正确性的评估
威胁报告 Report	针对一个或多个主题的威胁情报集合(例如恶意软件或攻击技术的描述, 包含上下文信息)
威胁行为体 Threat Actor	带有恶意意图的个人、组织或团体(攻击行动的发起者)
工具 Tool	威胁行为体发起攻击所利用的合法的软件
漏洞 Vulnerability	黑客能够直接利用的一种软件错误, 用其访问系统或网络

(2) TAXII

情报信息的可信自动化交换(Trusted automated exchange of intelligence information, TAXII)定义了威胁情报的传递方式, 是基于 HTTPS 并交换情报信息的应用层协议, 可以将 STIX 规范化后的信息通过 TAXII 进行交换, TAXII 在提供安全传输的同时不用考虑拓扑结构、授权管理等问题。由于 STIX 和 TAXII 是机器可读的, 因此很容易被自动化。虽然说使用 TAXII 可以传输 STIX 数据, 并且 TAXII 也是为了传输 STIX 格式的数据而专门设计的, 但是它们之间是相互独立的。因此, 使用 TAXII 也能够传输非 STIX 数据。通常的做法是使用 TAXII 传输数据, 使用 STIX 描述威胁情报。TAXII 包含一个服务信息交换集合和一个客户端服务器需求集合[5], 支持三种方式的威胁情报共享模型, 分别是点对点、订阅型以及辐射型, 如图 3 所示。

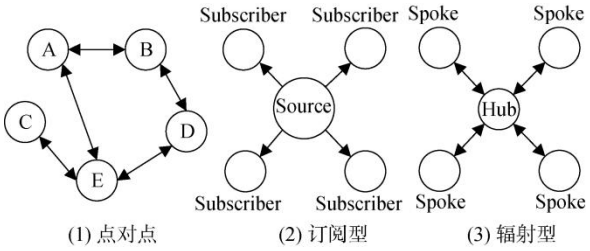

图 3 TAXII 支持的三种共享模型

Figure 3 Three sharing models supported by TAXII

(3) Cybox

网络可观察表达式(Cyber observable eXpression, Cybox)是一种能够表示计算机中所有可观察对象(文件、证书、IP 地址等)的标准化语言, 这些对象能够作为威胁判断的评估指标, Cybox 已经被集成到 STIX2.0 版本中, Cybox 数据框架结构如图 4 所示。

图 5 是来自 github 上展示域对象的基本示例, 展示了如何在实例中使用它来捕获域名。

(4) OVAL

开放式漏洞与评估语言(Open vulnerability and assessment language, OVAL)是一种描述语言, 用来在漏洞分析中定义检查项、脆弱点等技术细节[5]。OVAL 内容使用标准 XML 格式, 具有很强的灵活性, 能够分析多种操作系统的安全状态、漏洞、补丁等情况, 还能够描述系统配置信息和测试报告, 具有很好的可机读性。其中 OVAL 包含三种 XML 格式, 分别是 OVAL 定义格式、OVAL 系统特性格式和 OVAL 结果格式, 它们分别对应评估过程的三个步骤: 表示特定计算机状态的 OVAL 定义架构、表示系统信息的 OVAL 系统特征架构以及表示报告评估结果的 OVAL 结果架构。其中最重要的是 OVAL 定义格式, 它通过机器可读的方式提供了一种对系统进行安全评估的操作指南, 用于描述系统的配置信息、分析系统的安全状态以及报告评估结果等[5]。定义格式的 XML 结构如图 6 所示, 其中主要将定义(Definition)、测试(Test)、对象(Object)、状态(State)和变量(Variable)进行枚举。此外, OVAL 系统特性格式用来描述系统信息快照, OVAL 结果格式用于描述评估结果。但是由于其应用场景的局限性, 不能满足威胁情报共享需求, 因此已经逐渐被 STIX 做替代。

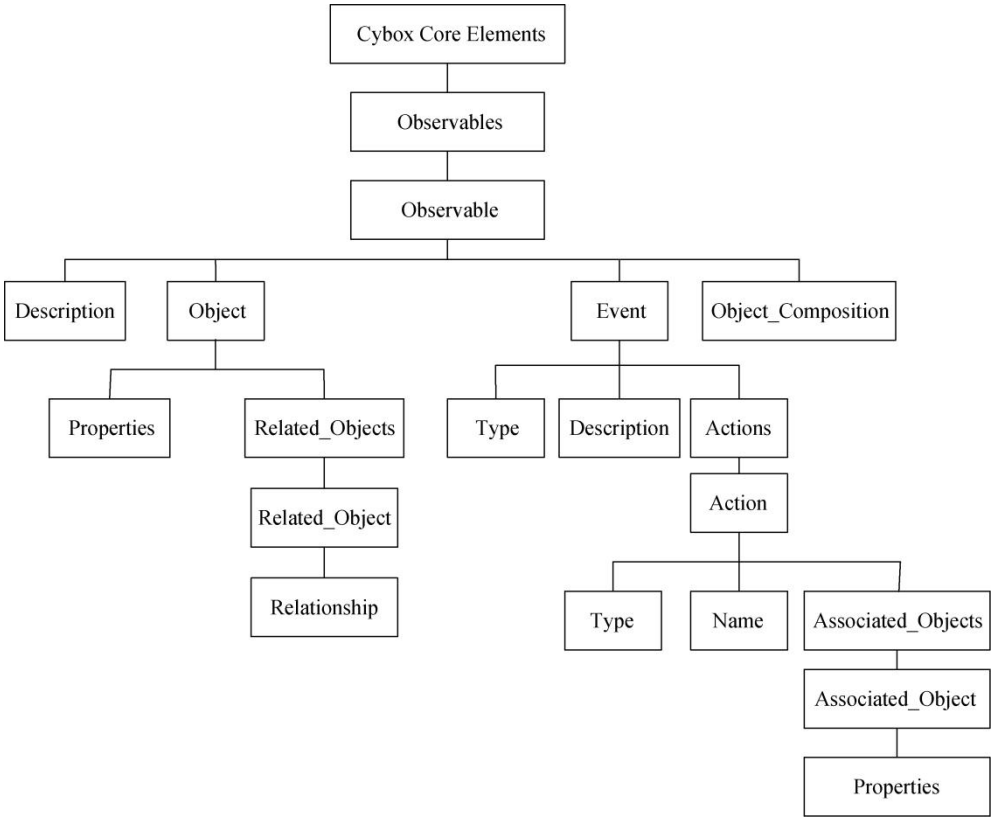

图 4　Cybox 数据框架结构图[5]

Figure 4　Cybox data frame structure [5]

```
<?xml version="1.0" encoding="UTF-8"?>
<cybox:Observables xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xmlns:cybox="http://docs.oasis-open.org/cti/ns/cybox/core-2"
    xmlns:cyboxCommon="http://docs.oasis-open.org/cti/ns/cybox/common-2"
    xmlns:URIObject="http://docs.oasis-open.org/cti/ns/cybox/objects/uri-2"
    xmlns:example="http://example.com/"
    xsi:schemaLocation="
    http://docs.oasis-open.org/cti/ns/cybox/core-2 ../core.xsd
    http://docs.oasis-open.org/cti/ns/cybox/objects/uri-2 ../objects/URI_Object.xsd"
    cybox_major_version="2" cybox_minor_version="1" cybox_update_version="1">
    <cybox:Observable id="example:Observable-0b9af310-0d5a-4c44-bdd7-aea3d99f13b6">
        <cybox:Object id="example:Object-15be6630-b2df-4bf9-8750-3f45ca9e19cf">
            <cybox:Properties xsi:type="URIObject:URIObjectType" type="Domain Name">
                <URIObject:Value>example.com</URIObject:Value>
            </cybox:Properties>
        </cybox:Object>
    </cybox:Observable>
</cybox:Observables>
```

图 5　Cybox 示例

Figure 5　Cybox example

2.3　威胁情报 NER 本体定义

对于威胁情报领域而言, NER 的本体定义具有十分重要的作用和意义, 因为它能够帮助安全团队高效便捷地理解和分析威胁情报。NER 的本体定义可以理解为一种表示实体、属性、关系和上下文的结构化模型。

NER 本体定义通常由本体词汇表、本体类、本体属性和本体关系组成。本体词汇表是指用来描述威胁情报领域中实体的词汇表。它的设计和维护需要考虑到 CTI 领域的变化, 例如新出现的攻击工具或者攻击组织等等。此外, 对本体词汇表进行标准化能够帮助不同的安全团队之间更好地进行交流, 还能够帮助不同的安全工具之间更好地协作。本体类是指具有共同特征的一组实体, 例如攻击者等, 它的定义需要考虑到实体之间的关联性和交互性。本体属性是指实体的特征, 例如 IP 地址、域名等, 它的定义需要考虑到实体的特征是否能够进行标准化和共享。本体关系是指实体之间的关系, 例如攻击者使用恶意软件, 攻击者攻击受害者等等, 它的定义需要考虑到实体之间的复杂关系和多样性。

NER 本体定义能够帮助安全团队更好地了解 CTI 中实体的属性和关系, 从而更好地进行分析和共享。本体定义的设计和维护需要充分考虑到威胁情报领域的变化及其发展趋势, 以及标准化和共享威胁情报的需要。

2.4　威胁情报 NER 下游应用

在安全领域的多个下游应用中都有威胁情报 NER 的贡献:

(1) 安全事件响应: 由于 NER 技术能够自动化识别事件中出现的实体, 并对其进行分类。因此安全分析人员能够快速确定事件的严重程度和来源, 进而采取合适的响应措施。

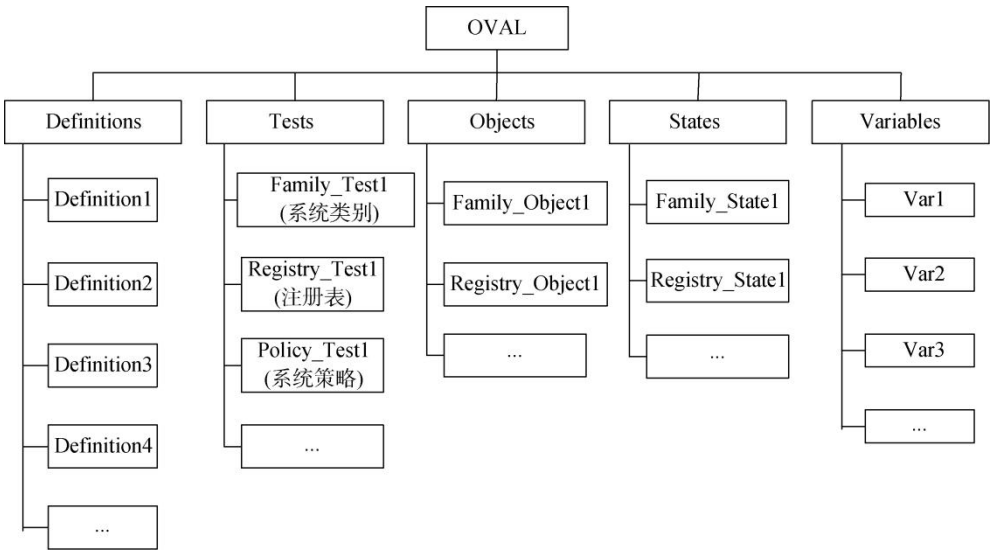

图 6　OVAL 架构[5]

Figure 6　OVAL schema [5]

(2) 情报监控: 由于 NER 技术能够监控网络和社交媒体上的实体, 因此安全分析人员能够及时了解威胁情报, 从而采取有力的措施。

(3) 威胁情报分析: 由于 NER 技术能够从网络等各种途径提取出实体信息, 因此安全分析人员能够对威胁情报进行分析。

(4) 威胁情报共享: 使用 NER 技术能够自动化标记威胁情报中的实体, 这有利于标准化共享威胁情报的格式, 便于了解和共享威胁情报。

(5) 威胁情报可视化: 使用 NER 技术能够创建威胁情报可视化, 例如散点图等, 这便于安全分析人员更好地理解威胁情报, 并与其他团队分享情报信息。

3　通用领域命名实体识别研究现状

由于命名实体的复杂性、多变性等特点, 使得实体标注代价过大, 因此导致高质量数据集过少。为克服以上困难, 实体识别使用的方法一直在不断改进, 如图 7 所示。主流的方法有[6]: 1) 基于规则和词典的实体识别技术; 2) 无监督学习的实体识别技术; 3) 基于特征的监督学习实体识别技术; 4) 基于深度学习的实体识别技术和 5) 以上各种技术的混合使用。

直到现在非常受欢迎的注意力机制、图神经网络等方法也开始被用在 NER 上, 它们可以更好地处理文本数据的结构特征, 精确度越来越好, 效果也越来越优秀。Kruengkrai 等人[7]提出了一个支持多类分类的联合模型, 并引入了一个简单的自注意变体, 允许模型学习缩放因子。Chen 等人[8]通过图神经网络结合了实体提及关系, 提高了来自不同领域的两个数据集的 NER 性能。然而, 命名实体识别还是有一些局限性, 比如说实体的多样性和歧义性、实体的复杂性以及超出词典范围(Out of vocabulary, OOV)等问题。

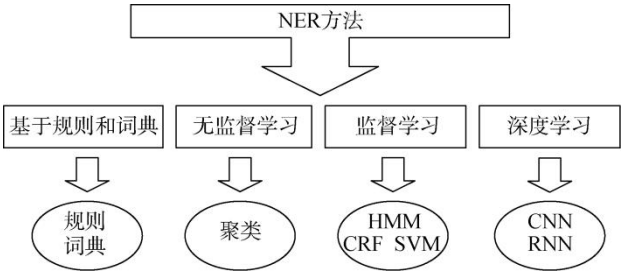

图 7　NER 方法分类

Figure 7　NER method classification

本章介绍了几种关于命名实体识别的技术, 总结了它们的优缺点以及学者们的相关工作, 其中还包含了对 NLP 领域的预训练语言模型及其特点的介绍。

3.1　基于规则和词典的方法

基于规则匹配的方法需要人工来手动构建规则模板, 使用模式匹配或者字符串匹配的方法进行命名实体识别。基于词典的方法首先要根据全部实体得到类别构建词典, 然后将文本中的实体与词典进行匹配, 当实体存在噪声时, 这种方法很难识别到实体[2]。

由于此类方法主要依赖于特定专业领域内的专

家针对特定语料库标注词典和构建规则, 所以只能应用于特定的领域或者场景中, 无法很好地适应新出现的实体类型和新兴领域。而且该方法需要专家进行维护, 需要大量的人力和时间成本。此外, 由于该方法只考虑了已知的规则和词典, 因此无法处理同实体的不同变体和多义性。

虽然有局限性, 比如只能对普通的实体进行识别, 但是对于固定格式的实体识别效率可以达到百分百, 因此, 它具有准确、可靠、高效的优点。同时由于该方法中使用的规则和词典是人工制定的, 因此具有高可解释性。此外, 通过修改和更新规则和词典能够提高实体识别的性能, 具有易于调整和优化的优点。

1998 年, Black 等人[9]提出的 FACILE 系统就是一个支持上下文分析规则的系统, 封闭测试结果查全率和查准率都在 90%以上。这虽然取得了较好的识别效果, 但人工指定词典和规则的成本过高, 对于大规模, 多领域的实体识别任务而言代价过大, 可移植性较弱。如 FACILE 系统的开放测试结果查全率较封闭测试结果下降了 14%。虽然有相当一部分的研究者已经在尝试使用机器生成规则以减少人工成本, 如 Collins 等人[10]1999 年提出的 DL-Co Train 方法, 但生成的规则可移植性依然不强。

2000 年, Kim 等人[6,11]提出一个基于规则的命名实体识别系统, 根据 Brill 的词性标记器自动生成规则。 2016 年, Quimbaya 等人[6,12]提出了一种基于词典的方法, 用于提高电子健康记录中命名实体识别的召回率。

2021 年, Fang 等人[13]提出了一种新颖的字典扩展方法, 该方法通过类型扩展模型提取新实体。还设计了一个多粒度边界感知网络, 从本地和全局角度检测实体边界。

2022 年, Salah 等人[14]使用基于规则的命名实体识别方法, 该方法依赖于阿拉伯语实体信息生成的触发词、模式、地名词典、规则和黑名单, 这不仅提高了实体识别的性能还能够自动更新规则。

3.2 无监督学习的方法

无监督学习使用未标记数据, 主要通过降维和聚类[15]进行推断。基于聚类的无监督学习方法通过将数据进行分组, 相似度高的数据之间的距离就越近, 反之越远。

与之前的基于规则和词典的方法相比, 无需人工定制规则, 增强了模型的扩展性, 且识别的实体不再限制于字典, 减少了对字典的维护。此外, 该方法无需标注数据, 节省了大量的人力和时间成本。并且由于该方法不依赖于特定领域的标注数据, 所以能够适应于各种领域和语言。

也正是因为该方法不依赖标注数据, 所有准确率相对较低。同时由于该方法需要大量的计算资源来进行训练, 需要消耗的时间和算力也相对较大。

Zhang 等人[16]在生物医学领域文本中应用无监督方法提取命名实体, 他们使用术语、浅层句法知识以及语料库统计, 在两个生物医学数据集上证明了无监督方法的有效性。

Brooke 等人[17]在没有标注数据的情况下对特定领域进行实体识别, 文中没有使用手工标注的数据或地名词典, 而是利用预先切分好的语料做 Brown 聚类, 将聚类结果分为三个类别并训练模型, 得到了很好的实验效果。

3.3 监督学习方法

监督学习方法通过序列标注表示特征, 使用特征工程提取特征, 这需要大量人工标注的训练数据集, 人力和时间成本较高, 同时还会受到标注数据质量的影响。由于该方法依赖于标注数据的分布情况, 因此在新兴领域或新的语言上, 需要重新标注数据并从头训练模型。此外, 由于该方法手动定义实体类型, 并且只能识别已经定义过的实体类型, 因此无法自动地发现新的实体类型。

但该方法也有它的优势, 使用标注数据训练模型使其具有较好的准确性。同时, 该方法还具有可解释性。此外, 通过调整模型的参数和特征选择等方式能够对模型进行优化, 从而提高实体识别的准确性。

通常使用输入文本序列 X, 预测输出序列 Y 的概率 $P(Y/X;\theta)$, 对于 NER 任务, Y 通常是标签序列。为了学习模型的参数 θ, 往往需要使用带有输入输出对的数据集 $D=\{X,Y\}$ 来训练模型。其中有监督的机器学习常用的算法主要有隐马尔可夫模型[18](Hidden markov mode, HMM) 、 最 大 熵 模 型[19](Maximum entropy markov model, MEMM)、条件随机场[20](Conditional random fields, CRF) 和支持向量机[21](Support vector machine, SVM)等等。

与基于规则的方法相比, 这些方法在很大程度上提高了算法的安全性和可移植性。当然, 相对于大规模的训练数据, 由于数据特征的增多, 模型的训练难度也相对加大。监督学习有分类和回归两种主要任务, 命名实体识别则属于分类任务。图 8 为监督学习方法基本流程。

1998 年, Bikel 等人[6,22]提出了第一个名为 IdentiFinder 的基于 HMM 的 NER 系统。但 HMM 模型得到的实验结果往往会忽略大量的上下文相关信

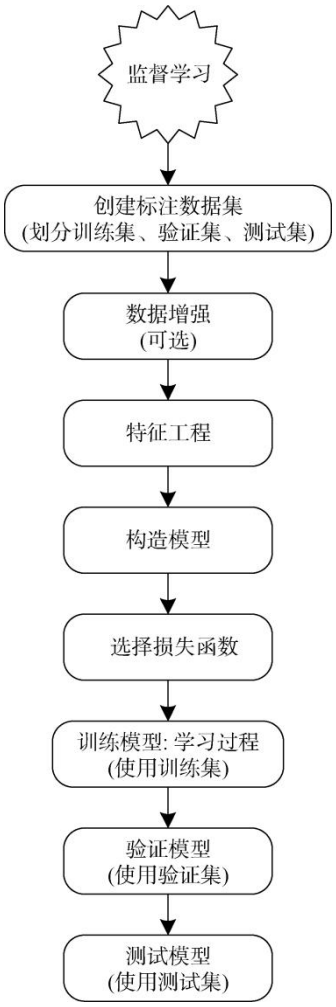

图 8 监督学习算法基本流程

Figure 8 The basic process of supervised learning algorithm

息, 实用性和迁移性较差。2003 年, Curran 等人[23]使用最大熵模型对信息进行编码, MEMM 相对于 HMM 可以定义更复杂的特征, 结果表明, 命名实体识别产生非常高的精度。McCallum 等人[6,24]使用 CRF 进行命名实体识别, 数据集使用 CoNLL03, 得到的 F1 分数为 84.04%。CRF 结合了 HMM 和 MEM 的特点, 通常使用线性链结构, 能够很好地解决 MEMM 的缺点, 但是训练时间相对较长。Ju 等人[25]在生物医学领域使用 SVM 进行命名实体识别, 得到了 84.24%的准确率和 80.76%的召回率。SVM 比较适合在小样本、高维度的语料库中进行学习。

相对于使用上述单个模型, 采用混合模型对数据进行训练往往能够得到更好的效果。Tang 等人[26]提出结构支持向量机(Structural support vector machines, SSVMs), 该算法结合了 CRF 和 SVM 的优点, 在临床文本的实体识别中取得了不错的效果。

3.4 深度学习方法

随着大数据时代的到来和深度学习技术的发展, 深度学习技术在计算机视觉和自然语言处理上都取得了极大的发展, 因此使用基于深度学习的实体识别技术[27]也逐渐成为主流, 命名实体识别通常会被视为序列标注任务或是分类任务。与传统的方法相比, 此类技术几乎独立于领域, 几乎不需要特征工程和领域知识, 通过神经网络逐层挖掘文本中的隐藏特征来训练数据, 发现数据中隐藏的复杂特征, 完成多层神经网络的训练和预测任务。

由于是基于深度神经网络模型进行训练和实体识别, 因此具有较好的准确性。同时该方法不需要手动提取特征, 因此具有更好的适应性和泛化能力。此外, 基于深度学习的 NER 技术可以支持多种语言, 不需要重新训练模型。

深度学习模型的缺点也显而易见, 由于需要大量的标注数据进行训练, 因此对人力和时间的需求很大。同时训练模型需要的时间较长, 并且需要大量的算力, 因此训练速度较慢。此外, 深度学习模型内部具有高度复杂性, 可解释性很差。

深度学习中命名实体识别模型大致可以分为三个步骤, 首先对输入数据进行预训练, 对原始输入进行嵌入。常用的预处理模型有 Glove[28]、Word2Vec[29]、ELMo[30]、GPT[31]、BERT[32]等。

这是因为计算机不能识别语言, 因此要将其转换为机器可读形式。词嵌入就是将词转换为词向量的过程, 有动态词嵌入和静态词嵌入两种形式。其中, 静态词嵌入得到的词向量是固定不变的, 而动态词嵌入得到的词向量通常是可变的, 往往包含一定的上下文语义信息。下面重点介绍上述几种预训练模型。

3.4.1 预训练语言模型

(1) Word2Vec

最早出现的词嵌入模型是 2013 年谷歌提出的 WordVec[29], 它基于神经网络训练语料库。该模型根据上下文直接的出现关系训练词向量, 缺少整体的统计信息, 因此这是一种静态词向量模型(即对于任意一个词, 它的向量表示是唯一不变的, 不随着上下文的变化而变化), 也就导致它不能适应一词多义的情况(例如 windows 在通用领域多指窗口, 而在计算机领域多指系统名称)。

Word2Vec 包含两种词向量学习模型, 分别是 skip-gram 模型和 CBOW(Continuous bag of words)模型, 两种模型都包含输入层、隐藏层和输出层。其中 skip-gram 通过中心词预测上下文一定窗口内的单词出现的概率, 例如“Bill Gates was born in Washington.”, 使用“born”预测“was”和“in”这两个单词; CBOW 正好相反, 它是通过上下文预测中心词出现的概率, 使用“was”和“in”预测中心词“born”。最后使用模型的一

部分参数作为词向量。由于无法处理一词多义的情况, 因此要使用融合上下文表示的方法。

(2) Glove

Glove[28]是一个轻量级的非深度学习语言模型, 全称为 Global Vectors for Word Representation, 它基于全局词频统计, 有效的利用了语料库的全局统计信息, 把单词表示为向量, 这些向量之间存在着语义相似性。Glove 模型是基于计数的, 首先根据语料库构造一个共现矩阵, 然后基于共现矩阵和Glove模型学习词向量。

(3) ELMO

2018 年 Peters[30]推出了一种新的基于深度学习框架的词向量表征模型 ELMO, 它由多层 BiLSTM 神经网络训练而来。这种模型不仅能够表征词汇的语法和语义层面的特征, 也能够随着上下文语境的变换而改变, 是一种深层双向的动态语言模型, 也成为了后面预训练思想的萌芽。实验证明, ELMO 模型能够很轻松的与 NER 现有主流模型(Bi-LSTM-CRF)相结合, 且结果有显著提升。Akbik 等人[33]利用训练后的字符语言模型的内部状态来产生一种新型的单词嵌入(上下文字符串嵌入), 对以前的嵌入进行了比较评估, 发现本文的嵌入对下游任务非常有用。

(4) GPT

GPT[31]是一种生成式预训练模型, 核心思想是先利于大规模无监督语料库进行预训练, 然后使用小规模有监督语料库进行微调, 从而解决许多 NLP 领域中的下游任务。GPT-1 使用基于 Transformer 的解码器作为模型的结构, 但是由于 GPT 是单向的语言模型, 只能获取句子从左到右的语义信息, 无法获取完整的上下文信息。GPT-1 的成功为 GPT-2[34]和 GPT-3[35]的出现打下了很好的基础。

GPT-2 的目标是训练泛化能力更强的词嵌入模型, 相比 GPT-1 而言, 它使用了更多的网络参数和更大规模的语料库。然后随着模型容量和数据量的增大, 它的潜能还有进一步的提升空间。

GPT-3 在 GPT-2 的结构上, 很大程度到地提升了网络容量, 它采用 96 层的多头 Transformer, 词向量长度设置为 12888, 上下文滑动的窗口大小设置为 2048 个 token。GPT-3 所拥有的强大能力, 使得其在零样本和小样本上也有出色的表现。

(5) BERT

2019 年 Devlin 等人[32]提出了一种名为 BERT 的新的语言表示模型(基于自注意力机制的 Transformers 双向编码器表示), 预先训练的 BERT 模型只需一个额外的输出层即可进行微调。BERT 模型有两个学习目标, 分别是掩码语言模型(Masked language model, MLM), 它根据周围的上下文预测掩码文本片段, 以及下一句预测(Next sentence prediction, NSP)。BERT 可以适用于多种 NLP 任务, 例如文本分类、序列标注、阅读理解等。

BERT 预训练模型能够产生词向量, 主要应用 Transformer 的编码器结构, 编码器由输入、多头注意力机制和前馈神经网络组成, 多个编码器堆叠在一起组成 BERT 模型。由于 BERT 具有很强的语义表征特性, 因此很多模型使用 BERT 作为词嵌入模型以提高 NER 任务的性能。BERT 的输入由三部分组成(如图 9 所示), 公式如下:

$$\text{Input=Token embedding+}\\ \text{Segment embedding+Position embedding} \tag{1}$$

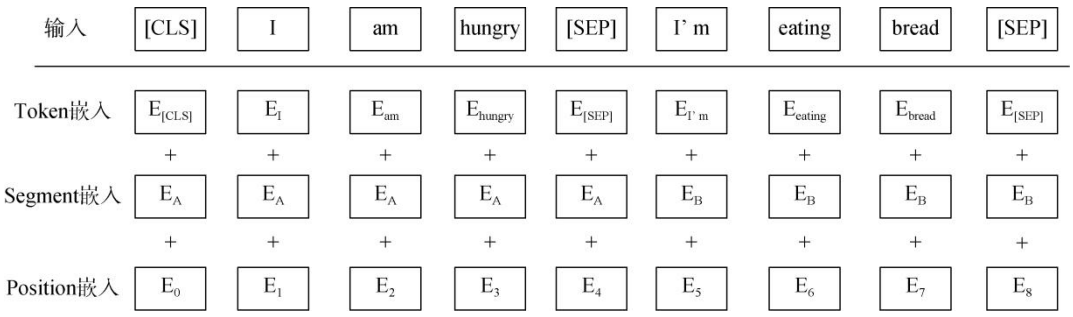

图 9 BERT 预训练模型的输入

Figure 9 Input of BERT pretrained model

(6) XLnet

虽然 BERT 模型被广泛使用, 并且实验效果非常优秀, 但是不可否认的是它仍然存在一些限制[36], 由于 BERT 使用掩码语言策略, 导致训练数据与测试数据的分布不一致。此外, BERT 忽略了被 mask 词语之间的上下文依赖关系。因此, 为了解决上述问题, Yang 等人[37]提出了基于自回归方法的 XLnet 预训练模型。XLnet 使用自回归代替 BERT 所使用的自编码语言模型, 从而避免掩码带来的负面效果。XLnet 还使用双流注意力机制, 并且引入

Transformer-XL 语言模型[38], 通过使用该模型可以获得长期依赖关系, 还可以解决上下文碎片问题。实验证明, Transformer-XL 比 Transformer 模型快 1800 倍, 并且在长短序列上性能都很好。

基于深度学习进行 NER 的第二步是使用神经网络层进行编码, 例如卷积神经网络(Convolutional neural networks, CNN)、循环神经网络(Recurrent neural network, RNN)、长短期记忆网络(Long short-term memory, LSTM)、门控循环单元(Gate recurrent unit, GRU), 从而获取语义、语法等特征。下面总结 RNN、LSTM 和 GRU 之间的区别和联系。

3.4.2 神经网络层

(1) RNN

RNN 能够记住之前出现的特征, 采用线性结构对序列数据编码, 按时间顺序对神经元进行计算, 是一种记忆模型, 它是 NER 任务中较早提出的方法之一。RNN 的机制就是将神经网络的输出保存在一个记忆单元里, 再将该记忆单元与下一次输入一起放入网络进行训练。RNN 的输入数据通常是序列, 通过网络内部的结构捕获序列之间的关系, 通常 RNN 的输出也为序列。由于 RNN 善于处理序列数据, 因此经常被用于 NLP 任务中。但随着序列长度的增加, 在进行反向传播时, RNN 会出现梯度爆炸和梯度消失问题, 梯度消失会造成网络权重无法被更新, 最终模型训练失败。梯度爆炸会使网络大幅度更新网络参数, 造成结果溢出现象。因此 RNN 适合短时依赖问题。

(2) LSTM

LSTM 是对传统 RNN 的一种改进模型, 它能够学习长期依赖关系, 善于处理序列数据, 能够捕获长序列之间的语义关系。由于 LSTM 结构包含遗忘门、输入门、输出门和细胞状态。因此可以对数据进行有选择的记忆, 从而缓解了 RNN 训练时出现的梯度消失或梯度爆炸问题。但由于 LTSM 的内部结构更加复杂, 模型的训练时间也会增加。

由于 LSTM 只能捕获单方向的信息, 无法获得完整的上下文信息, 因此 Lample 等人[39]提出使用双向长短期记忆网络 (Bidirectional long short-term memory, BiLSTM)作为 LSTM 的改进, 弥补 LSTM 的缺点。BiLSTM 的结构就是一个前向 LSTM 和一个后向 LSTM 的结合, 最终输出为两个单项 LSTM 的拼接。结果表明, 双向 LSTM 在某些场景表现更好, 使用双向的 LSTM 能够学到更多的语义信息。但是由于 BiLSTM 不能进行并行计算, 而且当序列太长时, 它的长距离信息也会被弱化, 因此效果也会下降。

(3) GRU

传统 RNN 的另一种变体就是 GRU, GRU 和 LSTM 都能捕获长期依赖关系, 但是 GRU 结构相比 LSTM 优化了参数量, 结构和计算更加简单。GRU 的核心结构分别是更新门和重置门, 使用更新门和重置门控制数据的记忆程度, 决定有多少信息被保留带入下一时刻, 同时 GRU 模型的训练时间也相对减少。

基于深度学习进行 NER 的最后一步是解码, 预测输入序列的标签序列。通常使用 CRF 和 softmax 进行标签解码。下面总结一些基于深度学习进行命名实体识别的工作。

3.4.3 基于深度学习的命名实体识别工作

2019 年 Jana 等人[40]结合预训练模型, 提出了两种用于嵌套命名实体识别的神经网络架构, 使用线性方式对嵌套标签进行编码, 在 CoNLL2003 上 F1 值达到 93.38%, 再次提升了实验结果。

2020 年 Simran 等人[41]使用门控循环单元作为基本模型评估深度学习架构。Wang 等人[42]提出了一种新的损失函数 TSFL, 用于解决数据标签分布不平衡的问题。Tikhomirov 等人[43]提出了一个 RuBERT 模型, 以 BERT 编码为基本结构, 显着提升了一些 NLP 任务的性能。

2021 年 Li 等人[44]提出了一种新颖的模块化交互网络模型, 利用片段级信息和词级依赖关系, 并结合了一种交互机制来支持边界检测和类型预测之间的信息共享, 以提高 NER 任务的性能。Ushio[45]介绍了名为 T-NER(基于 Transformer 的命名实体识别)的用于 NER 语言模型微调的 Python 库, 对 NLP 下游任务进行持续改进。Amin[46]提出一个基于 Transformers 的迁移学习框架, 用于在 PyTorch 中创建的命名实体识别 (T2NER)。Tu 等人[47]提出了 Tough Mentions Recall (TMR) 指标来补充传统的 NER 评估指标。TMR 指标能够区分其他相似的评分系统, 并识别性能模式, 这些模式不会从整体精度、召回率和 F1 中被注意到。

2022 年, Zhang 等人[48]提出使用扩张卷积神经网络和双向长短期记忆网络, 解决 CNN 在医学方面命名实体识别任务的劣势。

3.5 各种技术混合使用

为了提高实体识别的效率, 现在基本选择对上述方法进行组合使用, 从而实现最好的实验效果。

Liu 等人[49]利用词典加入单词的最终表征中, 提高神经网络模型的性能。Gao 等人[50]使用 BERT-BiLSTM-CRF 命名实体识别模型, 在 BiLSTM 层和

CRF 层之间加了一个注意力网络, 提高实体识别准确率。Liou 等人[51]在 2022 年提出基于 BERT-BiLSTM-ATT-CRF 的中文实体识别模型, 首先使用 BERT 根据单词的上下文语义信息获取向量, 然后将向量输入带有注意力机制的双向长短期记忆网络中(BiLSTM-ATT), 最后, 使用 CRF 获取句子标签序列。

除了上述五种方法之外, 本文还总结了远程监督方法和提示学习方法。

3.6 远程监督方法

在实际应用深度学习时, 纯人工标注的确能达到很好的实验效果, 同时模型训练的好坏受到数据规模的影响, 对于大型细粒度实体分类模型[52]来说, 这项工作是巨大的。因此, 使用深度学习来解决问题的最大阻碍就是缺少丰富的标记数据。此外, 在很多低资源场景中, 标注数据的缺失也成为阻碍 NLP 发展的重要因素, 远程监督技术便能够很好的解决这个问题。

Hedderich 等人[53]调查了使用远程监督对低资源数据进行命名实体识别的相关工作, 通常能够使用自动化技术在外部信息源中为原始未标记文本生成标签。例如, 通过使用外部词典获得关于地名的实体列表, 当原始数据中的标记与列表中的实体匹配时, 则该标记会被自动标注为地点。所需要的外部信息通常可以从知识库、外部词典等途径获得。

虽然远程监督节省了大量的人工标注工作, 降低了训练成本, 并且还带来了更多的标注数据, 但该方法会带来噪声标签以及不完整标注的问题。如果直接对这些数据进行训练, 很大程度上对降低算法的性能。为了解决上述两个问题, 不断有学者在探索解决方案。

文献[53]中总结有两种方法可以对噪声进行处理, 从而减轻对性能的负面影响, 分别是噪声过滤和噪声建模。Ni 等人[54]为了解决噪声问题, 提出使用启发式规则, 进而筛选出高质量标注数据。Chen 等人[55]提出使用远距离监督进行细粒度实体分类, 通过压缩隐空间簇(Compact latent space clustering, CLSC)[56]的方法高效利用噪声数据。

Shang 等人[57]为了解决标注不完整的问题, 提出了新的神经模型, 使用 CRF 层处理未标记数据的标签。

Yang 等人[58]为了解决不完整标注问题, 提出使用部分标注数据, 从而减少未知字符标签的影响; 此外, 他们还提出使用强化学习对自动标注进行筛选, 过滤掉噪声标签。

在本文的调查中, 总结了一些在命名实体识别上应用远程监督的工作:

(1) 知识库

Liang 等人[59]通过使用外部知识库, 提出 BOND 框架, 第一阶段利用远程标签使预训练语言模型适用 NER 任务, 从本质上转移预训练模型的语义, 为所有数据得到一组预测标签; 第二阶段使用预测标签替换远程标签, 通过自训练来提高模型的性能。

(2) 外部词典

远程监督技术通常依赖于像实体列表这样的辅助数据。当使用预定义的外部实体词典时, 若某个字符串出现在词典中, 则该字符串可能是一个实体。Fries 等人[60]针对生物医学领域提出使用远程监督进行命名实体识别, 他们使用词典和启发式算法构建 NER 模型。Peng 等人[61]使用未标记数据和命名词典进行命名实体识别工作, 提出新的 PU(Positive-Unlabeled)学习算法, 降低对词典的要求, 使该方法对简单词典具有通用性。Hedderich 等人[62]提出了 ANEA, 一个基于实体列表自动注释文本中命名实体的工具。

综上所述, 虽然远程监督方法通过使用未标注语料库进行训练, 能够在较短时间内训练好模型, 提高训练效率, 但该方法容易忽略一些上下文信息, 因此会导致识别精度有限的问题。同时, 由于误差传递和精度有限的问题, 该方法只适用于相对简单和实体类型和领域, 对于一些复杂的应用场景可能效果不佳。此外, 据本文了解, 目前的远程监督工作大多集中在特定领域, 例如生物医学领域[60], 在威胁情报领域的研究少之又少, 因此, 这也为该领域的 NER 发展提供了新思路。

3.7 提示学习方法

NLP 经历了两次巨变[63], 第一次是 2017 年到 2019 年, 从完全监督范式(Fully supervised learning)转向预训练、微调范式(Pre-train, fine-tune)。如今 NLP 正处于第二次巨变中, 从预训练、微调范式转向预训练、提示、预测范式(Pre-train, prompt, and predict)。在这种新范式中, 不像之前一样使预训练模型适应下游任务, 而是重新制定下游任务去适应语言模型。这里主要有两种模式, 一种是基于掩码语言模型的提示学习, 另一种是基于生成式语言模型的提示学习。

基于掩码的语言模型在预训练期间已经获得了丰富的知识, 因此, 模型可以计算掩码位置的概率分布情况。对于 NER 任务, 首先设计一个提示模板, 例如[E] is a [MASK] entity, 其中被掩码的位置代表实体类型, E 代表句子中的实体。然后将其拼接到原始输入序列 X 的后面, 再将其一起输入到语言模型

中。这种完形填空的形式将原本的分类问题转化为了预测[MASK]的问题, 这也是掩码语言模型所擅长的[64]。添加提示能够充分利用模型在预训练阶段学习到的信息, 降低对大量标注数据的需要, 利用适合的提示能够减少预训练和微调之间的差异, 使得模型在小样本上也能取得不错的成绩[65]。

基于生成式语言模型的方法主要是利用预训练模型良好的生成能力, 文献[66]中手动创建实体模板, 例如<Entity Span> is a person entity, 手动创建非实体模板, 例如<Entity Span> is not a named entity。模型的编码器端输入原始序列 X, 得到句子的隐藏表示 H^{enc}, 如公式(2)所示。然后在解码器的第 i 步, 将提示 T 的第 1 到 $i-1$ 个标记和 H^{enc} 一起输入 BART 解码器, 如公式(3)所示。

$$H^{enc} = \text{encoder}(X) \tag{2}$$

$$H_i^{dec} = \text{decoder}(H^{enc}, T_{1:i-1}) \tag{3}$$

文献[66]所使用的方法需要枚举所有的跨度, 存在效率低下的问题。文献[67]也将序列标注问题建模为生成问题。它在 Transformer 的注意力机制中融合提示信息, 利用提示去指导注意力的分配。这种方法通过仅优化与提示有关的参数, 能够让模型更加轻量。

文献[68]提出自描述网络 SDNet, 并使用大型语料库对其进行预训练, SDNet 连续执行两个生成任务, 第一个任务是生成实体提及的概念描述, 第二个任务是自适应生成实体。不同的任务具有不同的提示, 概念描述生成任务使用[MD]作为任务描述符, 实体生成任务使用[EG]作为任务描述符, 使用不同的提示来引导模型生成不同的输出。

提示学习(Prompt learning)如今也已经成为自然语言处理中的第四范式, 将其用在命名实体识别上产生了很好的效果。 本文总结了使用提示学习进行实体识别的相关场景, 见表 2。能够发现, 提示学习不光对资源丰富的场景有效, 还能够应用于小样本甚至零样本场景, 适应标签数据很少或没有的情况。

表 2 提示学习所用场景

Table 2 Prompt learning scenarios

参考文献	预训练语言模型	资源丰富	小样本	零样本
[64]	BERT-BASE	√	√	√
[65]	BERT-BASE		√	
[69]	BERT-BASE		√	
[66]	BART	√	√	
[67]	BART	√	√	√
[68]	T5	√	√	√

综上所述, 基于提示学习的 NER 技术有效结合了有监督学习和自监督学习技术, 通过将已知的实体类型和上下文信息作为提示, 提高实体识别的准确性和泛化能力。该方法通过利用命名实体类型和上下文信息, 能够在不需要大量标注数据的情况下提高实体识别的准确性, 但也正因为如此, 该方法对标注数据的质量有较高要求。

该方法通过学习通用的文本表示, 能够提高对新兴领域和新实体类型的泛化能力。并且模型的结构相对简单, 具有可解释性。但由于该方法需要同时进行有监督和自监督学习, 因此具有相对复杂的训练过程。此外, 提示学习的性能受限于已知的命名实体类型和上下文信息的质量, 如果提示信息不准确或者不具代表性, 那么可能会对实体识别的准确性造成不好的影响。

4 威胁情报领域命名实体识别研究现状

目前命名实体识别技术在通用领域取得了很好的成果, 但在网络安全领域仍然面临着许多的问题。

(1) 实体类型很多, 分类工作相对繁琐复杂;

(2) 实体名称复杂, 由各种字符和数字组成, 往往会存在单词很长的情况;

(3) 实体有别名, 如 APT 组织往往有多个别名;

(4) 实体通常不是仅有一个单词组成, 实体边界模糊, 非安全领域专业人士难以区分辨别;

(5) 实体更新频繁, 随着网络攻击和防御的不断升级, 新的攻击手段、攻击工具等不断出现;

(6) 网络安全领域术语类别多, 且有些单词与通用领域含义不同;

(7) 语料库往往不够全面, 实体类型少, 不涉及足够的领域知识;

(8) 标注完整、语料丰富的数据集相对很少;

(9) 语料库中存在大量 other 类型。

由于以上种种困难, 导致网络安全领域的命名实体识别仍然是具有挑战性的, 对于网络空间安全领域的新兴词汇、术语、攻击指标等抽取仍有很长的路, 等待研究人员突破创新。

在许多网络安全应用程序中, 命名实体识别已被用于识别感兴趣的实体, 例如易受攻击的软件的名称和版本、易受攻击的组件以及易受攻击的软件所依赖的底层软件系统的名称和版本[70-71]。目前, 网络威胁情报领域已经应用了命名实体识别的各种相关技术对其实体和关系进行提取。

最开始的基于规则匹配、词典的命名实体识别

方法针对固定格式的威胁情报实体依然能适用。与之后的基于特征的方法相比, 基于统计学习的方法效率更高, 减少了对领域知识的依赖, 提升了模型的健全性、可移植性和面对大规模复杂语料的应对能力。不过, 基于统计学习的方法, 对语料的要求更强, 需要大量的标注数据, 这会消耗大量的人力物力。

现如今的发展趋势无疑是基于深度学习的方法, 融合深度学习模型, 开展威胁情报领域实体识别的研究, 这可以避免繁琐的特征工程。深度学习神经网络一般采用端到端的方法, 并尝试直接从大规模的标注数据中学习文本中的隐藏特征。

深度学习相关的许多应用已经在网络安全领域发挥作用[72]。但是, 由于网络安全威胁情报领域的特殊性, 网络安全威胁情报中文本的复杂性, 仅仅依靠单独的深度学习神经网络很难准确的对该领域进行实体识别和特征提取。针对网络安全威胁情报领域进行模型调整, 包括和早期的两类方法进行结合, 更好的提高实体识别准确率。

如今威胁情报领域出现越来越多的学者开始研究命名实体识别, 他们将 CTI 数据进行规范化处理, 从而进行模型训练, 近年来出现了许多经典的模型。同时, 关于该领域的 NER 研究还在持续发展。本章重点总结经典的方法和工作。

4.1 基线模型

4.1.1 BiLSTM-CRF 模型

为了能够得到序列的历史信息和未来信息, 模型通常使用基于 BiLSTM 的上下文编码层。

CRF 是一个概率图模型, 经常作为 NER 模型的最后一层, 即标签解码层, 是近年来使用很广泛并且效果最好的。CRF 模型根据给定的输入序列 X, 得到输出序列 Y 的概率分布。用数学表达为 $P(Y|X)$, 其中 X 和 Y 为随机变量, 则概率 $P(Y|X)$为给定 X 的条件下 Y 的概率分布。若此时 Y 构成的是马尔可夫随机场, 那么 $P(Y|X)$即为条件随机场。

目前, 在命名实体识别任务中加入 CRF 取到了不错的成果。CRF 与深度学习的结合, 例如 2016 年 Lample 等人[39]提出了 BiLSTM-CRF 模型, Ma 等人[73]提出的 CNN-BiLSTM-CRF 模型, 都取得了不错的实验效果。由于 LSTM 和 CRF 的组合特点, 使得它们经常成为后期模型的基础结构。

4.1.2 CNN-BiLSTM-CRF 模型

Ma 等人[73]在 2016 年提出了使用 CNN-BiLSTM-CRF 模型进行命名实体识别任务, 他们使用 CNN 获取字符级特征, 同时使用静态词嵌入进行拼接, 作为 BiLSTM-CRF 架构的输入。文中所使用的字符级分布式表达, 可以有效的利用单词的前缀、中缀以及后缀信息, 还能够捕获单词的大小写等特征信息。

CNN 是一种深层的前馈神经网络[74], 卷积模型的并行计算效率较高, 多用于处理类似网格结构数据。图 10 展示了使用 CNN 提取单词的字符级表示的模型结构。该模型首先使用查找表将输入的字符转化为向量, 即字符嵌入。然后将该嵌入输入 CNN 模型中进行字符级特征提取。最后通过池化层得到最终的字符级表示。

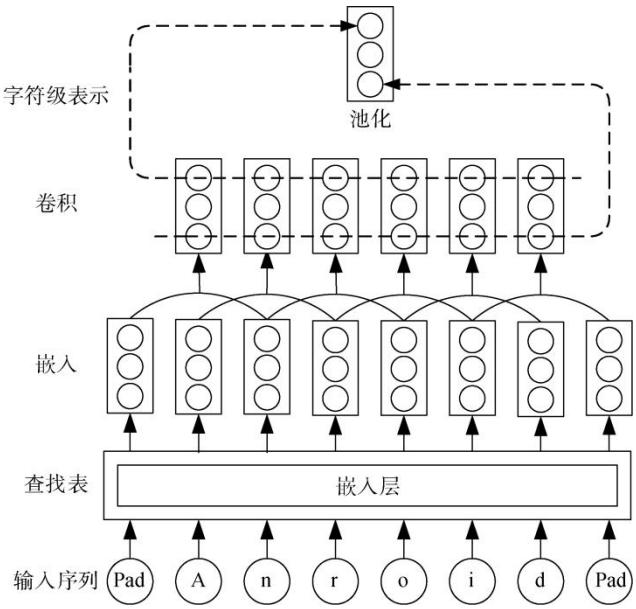

图 10 基于卷积神经网的字符级表示[6]

Figure 10 Character-level representation based on CNN [6]

4.1.3 BERT-BiLSTM-CRF 模型

首先, 将标注语料库输入 BERT 模型自动提取出相应的语义特征, 然后将得到的词向量输入 BiLSTM, 最后将输出结果放入 CRF 层进行标签解码, 最终得到输入序列的标签预测结果。

文献[75]提出基于 BERT-BiLSTM-CRF 的命名实体识别模型, 如图 11 所示, 取得了很好的效果。

与该模型结构类似的还有 XLnet- BiLSTM-CRF 模型, 2021 年郑等人[36]在 BiLSTM-CRF 的基础上加入 XLnet 预训练模型, 将其应用在中文命名实体识别任务上, 得到了很好的实验结果。

4.2 网络安全领域 NER 的发展

4.2.1 基于领域词典

Jones 等人[76]采用半监督的方式, 通过少量的输入数据, 结合主动学习, 对少数示例进行标注。通过使用人工设计的地名词典实现命名实体识别。

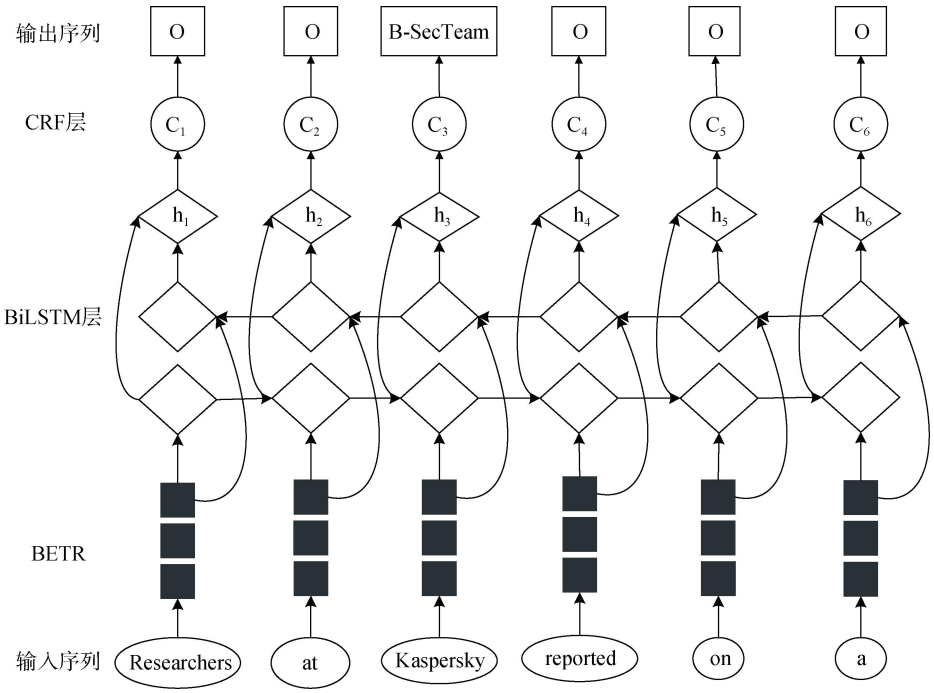

图 11 BERT-BiLSTM-CRF 模型架构图

Figure 11 BERT-BiLSTM-CRF model

Wang 等人[77]提出对非结构化网络威胁情报进行语义提取, 从而实现自动标记。他们使用词性标注器和地名词典实现实体识别, 并利用相关资源构建威胁情报本体。

但由于基于字典的方法需要使用预定义的术语列表, 因此无法识别以前没有见过的威胁情报数据。此外, 该方法也无法区分具有相同名称但含义不同的数据[78]。

4.2.2 基于规则

2015 年, Balduccini 等人[79]提出基于知识表示的机器学习交叉工作的架构, 使用正则表达式和本体结合的方式, 实现日志文件中的实体提取。

2016 年 Liao 等人[80]基于规则从开源威胁情报中识别 IOC, 他们提出一种全自动 IOC 提取的系统 iACE, 它能够从安全博客中自动提取 IOC, 并生成机器可读的结构化格式, 从而发现威胁情报中的内在关联。该系统的思路来源为文章中的 IOC 一般是以可预测的方式进行描述, 即通过稳定的语法关系连接到一组上下文术语。该系统使用安全报告的文章作为数据集。

2018 年, Zhu 等人[81]提出四阶段模型来定义活动, 使用 IOC 描述每个阶段。提出 ChainSmith 系统实现了一套多分类程序来提取 IOC, 在 14155 篇安全文章中训练 IOC 分类器, 将提取的 IOC 实体分类到自己定义的攻击活动的四个阶段, 达到了较高的准确率。

由于威胁情报数据没有固定的模式, 有些数据以字符和数字的任意组合表示, 还有一些实体又多个单词组成。因此, 使用正则表达式这样的规则是不合适的[78]。

4.2.3 基于监督学习

Mulwad 等人[82]对 web 文本中的相关漏洞和攻击信息进行实体提取, 他们提出使用支持向量机(Support Vector Machine, SVM)算法来实现这一目的。原型系统使用来自维基百科的通用知识库提取描述漏洞和攻击的概念, 并将其映射到 Dbpedia, 从而利于检测新漏洞。该方法能够监视社交媒体或聊天室的文本流, 从而识别潜在的新攻击和漏洞。

Lal 等人[83]也基于 SVM 算法识别安全相关术语实体, 从而解决非结构化文本的问题。该模型的数据源于大量开放式博客和安全公司的公告, 使用手动注释的数据进行训练。该系统能够帮助企业和政府加强安全性, 还可以应对零日攻击。

Sabottke 等人[84]设计了基于 Twitter 的漏洞检测器, 使用 SVM 实现, 该系统能够提取出漏洞相关的信息, 并对网络保险进行风险建模。

4.2.4 基于深度学习

Dong 等人[70]提出自动化系统 VIEM, 用来检测标准化的 NVD 数据库和非结构化的 CVE 描述之间的不一致信息。该系统基于深度学习构建, 能够从非结构化文本中提取出易受攻击的软件名称和版本。

Nuno 等人[85]为了处理开源情报平台所产生的信息, 提出使用 BiLSTM-CRF 模型来提取与威胁情报相关的 IOC 实体。在进行实体提取之前加入了一个基于卷积神经网络(CNN)的二分类器, 若文本与威胁情报相关, 则进行 IOC 实体抽取, 若无关则丢弃该文本。Zhao 等人[86]基于卷积神经网络提出了 TIMiner 自动化框架, 该框架能够从社交媒体的 CTI 中提取出 IOC。

开源网络威胁情报能够以结构化的形式存储在网络安全知识图谱中, 知识图谱对安全分析师检测网络威胁大有帮助。图谱中包含了大量的三元组信息, 其中包括两个网络安全实体和一对实体间关系。Pingle 等人[87]提出使用深度学习的方法来提取三元组。

之前工作的重点在于提取孤立的 IOC 数据, 没有考虑到它们之间的关系。为了深入了解威胁的特征, 还应该考虑威胁情报数据之间存在的语义关系。为了解决上述问题, Jo 等人[78]提出新的 CTI 系统 Vulcan, 该系统能够实现从收集非结构化数据到利用提取的 CTI 数据分析威胁的整个过程。基于 BERT 模型实现命名实体识别, 在语言模型基础上堆叠 BiLSTM 网络层和 CRF 层, 从而识别出与勒索软件攻击相关的命名实体。

虽然基于深度学习的方式能够学习单词的上下文信息, 但模型学习时需要大量的标记数据进行训练。因此, 基于提示学习进行威胁情报领域命名实体识别的方式或许是有前景的。

4.2.5 各种技术混合使用

2018 年 Zhou 等人[88]提出使用端到端的序列标记进行 IOC 实体识别, 模型使用人工设计的特征和注意力机制, 结合 BiLSTM 网络构建。

2019 年 Han 等人[89]提出使用生成对抗网络和 BiLSTM-Attention-CRF 模型相结合的方式, 获取大规模的 CTI 标记数据。通过对抗网络找到标记的共同特征, 然后结合领域字典和句子依赖作为附加特征进行命名实体识别。

2020 年 Wu 等人[90]提出在 BiLSTM-CRF 模型的基础上, 结合基于本体的领域字典匹配校正, 可以显着减少对手动定义特征的依赖, 并且提高实体识别的效果。武涵[91]针对非结构化 CTI 数据中存在的实体定位难、识别难等问题, 提出基于 Attention-BiLSTM-CRF 架构的情报实体识别模型。

2022 年 Wang 等人[92]针对网络威胁情报的领域特征, 例如实体边界模糊、一词多义等, 提出在 BERT-BiLSTM-CRF 架构的基础上, 添加网络威胁情报词典模板对推理结果进行最大匹配校验, 从而提升 CTI 领域实体识别的准确度。Zhou 等人[93]提出用于非结构化威胁情报文本提取和分析的自动化系统 CTIView, 该系统能够处理大量异构的 CTI 数据, 使用正则表达式结合黑白名单机制, 提取威胁情报中的 IOC 和 TTP 信息。模型在 BERT-BiLSTM-CRF 架构的基础上引入 GRU 层, 从而提升模型的性能。

4.3 基于触发词的命名实体识别研究现状

通常情况下, 训练命名实体识别的神经网络模型时需要大量的人工标注, 而往往这些工作是非常耗时耗力的。那么如何才能最大程度上利用这些标注数据, 成了学者们需要考虑的问题。于是, 出现了实体触发词(Trigger)这一研究方向。所谓触发词, 就是在序列中, 通过某个或某些词可以帮助解释实体识别过程。例如“I am eating apple.”, 很容易看出实体为“apple”, 那对于单词“eating”, 可以通过它推测出后面的单词是一种食物。因此, 把这个概念推广到通用领域, 以及网络安全领域, 通过标注数据中的触发词可以使实体识别更有效率, 更加准确。而对于已经在数据集中标注过实体的工作人员来说, 他们已经熟悉了语料库中的句子, 所以再进行触发词标注时应该很容易。但是, 目前基于触发词的研究只出现在通用实体识别领域[94-95]和基于触发词的威胁情报文本分类研究上[96], 没有用到细粒度的威胁情报 NER 任务上, 下文总结这三篇工作。

Liu 等人[96]提出一个基于触发词向量的网络威胁情报发现系统 TriCTI, 用于发现攻击指标(Indicators of compromise, IOC)和活动阶段之间的关系。他们使用活动触发词(Campaign trigger)解释活动阶段, 该触发词能够指导活动阶段的关键词识别, 增加关键词权重, 从而提高分类模型的性能。该方法能够识别所有活动阶段的网络威胁情报, 从而达到识别和抵制网络威胁的目的。该系统输入网络安全报告, 选择IOC出现的句子, 对该句子所属的活动阶段分类, 从而确定 IOC 的活动阶段。本文所使用的触发词, 是由安全专家凭借专业直觉所标注的对活动阶段具有高解释性的单词或短语, 共 3012 个触发词短语。此外, 还使用数据增强技术扩充数据集, 通过实验证明了该系统的有效性。

2020 年 Lin 等人[94]引入了实体触发词的概念, 使用基于实体触发词的方法 TriggerNER, 使人工标注的数据通过具有成本效益的方式获得监督, 用来提高 NER 模型的效率。将句子中的一组词定义为实体触发词, 通过触发词可以更好的识别实体。本文实验在两个数据集上标记了 14k 个触发词。实验结果表明, 使用 20%的触发词标注语句可以达到使用

70%的常规标注语句相似的性能。

然而他们的触发词标注, 仍然需要大量的人工成本和丰富的专家经验。2021 年 Wu[95]提出使用句法分析器DepTrigger, 用于自动标注句子中的实体触发词。实验结果表明, 该方法与参考文献[94]中的NER 模型性能相当。Lee 等人[97]提出名为AUTOTRIGGER 的框架, 通过自动生成和使用实体触发词, 减少了获取额外信息的成本, 只需要很少的人工参与, 从而提高实体识别的性能。

4.4 网络安全领域其他 NER 相关工作

此外, 图 12 还列出了最近几年网络安全领域命名实体识别的其他相关工作。可以看出在网络安全领域, 随着时间的发展, 大多数的 NER 工作都是通过使用深度学习方法来实现的。

图 12 最近几年网络安全领域 NER 相关工作

Figure 12 NER-related work in the field of cyber security in recent years

从 2016 年的自动化技术 TACE[80], 到 2019 年的自动化系统 VIEM[70], 再到 2022 年的 CTIView[93], 不断有学者研究如何将 NER 技术实现自动化, 据本文了解, 目前该技术的发展还不成熟, 自动化 NER 技术以及自动化联合提取技术可能会成为学者们为之努力的方向。

同时, 在这些工作中还可以看到, 除了神经网络经常被用于 NER, 还有基于图的算法[98]、注意力机制[88-89]、人工设计的特征[88]、特征模板[99]、对抗学习[89]、领域字典[90]、相似性度量[100]等等也被单独或结合用于 NER。未来还有哪些技术可以用于提高 NER 的实验效果还有待研究。

目前, 命名实体识别只在有限的实体类型中取得了较好的成绩, 例如人名、地名、组织机构名的识别。然而这些技术无法很好地迁移到其他特定领域中。

不同领域的数据往往具有其领域的独特特征, 有些领域资源匮乏, 导致缺少标注数据集, 以至于模型训练很难直接开展。通过采用半监督学习、远程监督学习、无监督学习方法可以实现资源的自动

构建和补足, 以及迁移学习等技术都可以作为解决该问题的核心研究方向。

在某些特定领域中, 有时会由相关领域专家创建特定领域知识资源, 以方便进行信息处理。威胁情报领域有其自身的特点, 例如恶意软件描述的域文本通常伴随着一组域元数据。根据威胁情报领域的特点, 可以使用该领域相关知识进行模型改进和训练。

本章首先总结了威胁情报领域命名实体识别的现状, 并概括了该领域所面临的一些问题。其次, 介绍了威胁情报领域经常使用的几种有代表性的 NER 基线模型, 并且分类整理了该领域的一些 NER 研究工作。最后整理了近年来该领域的一些其他工作。

不难发现上述任何使用特征工程或机器学习、深度学习的方法都需要带注释的训练数据集来训练模型。这样就会出现两个挑战。首先, 需要一定数量的带标注的文本才能训练出性能不错的模型; 其次, 文本由人工进行标注, 不仅耗时耗力, 还有可能出现标注错误或标注遗漏等问题。为了最大程度的减少人工标注的工作量, 很多方法被逐渐提出。例如, 自动标注方法[110]、特征工程方法[111-112]、深度学习方法[113-114]和迁移学习方法[70]。

自动标注是指利用模型和手动构建的训练集, 训练出能够对语料库之外文本中应当标注的命名实体进行识别并标注的模型。虽然标注工具 Brat 系统在当下非常流行并且使用方便, 但人工标注总存在着标注不规范等问题。自动标注技术可以有效避免人工标注不规范和疏漏的问题, 为NER的工作带来方便。

5 网络安全领域的语料库

在网络安全领域中, 常用的结构化数据集包括 CVE 数据集、NVD 数据集、推特数据集, 还有恶意软件数据集等。其中 CVE 数据集包含漏洞描述、漏洞类型、漏洞等级等信息, 由 MITRE 进行维护。NVD 数据集是美国国家标准技术研究所下属的一个漏洞信息库, 包含与漏洞相关的结构化数据。恶意软件数据集包含与恶意软件相关的结构化数据, 例如恶意软件名称、类型等, 可以通过 VirusTotal 等网站获取。推特数据集包含与推特社交媒体相关的结构化数据, 例如推特文本、用户信息等, 常用的数据集有 Twitter API 数据集等。

网络威胁情报领域也有很多半结构化数据集, 例如 VirusTotal、MISP、OpenIOC, 以及 AlienVault OTX。其中 VirusTotal 是一个在线的恶意软件分析平台, 能够对用户提交的可疑文件进行分析。MISP 是一种开源的情报共享平台, 用户能够在上面共享恶意软件样本等情报信息。OpenIOC 用于描述恶意软件的格式, 能够在不同工具间共享数据, 它提供的标准化格式便于将情报数据导入不同的安全工具中。AlienVault OTX 作为开源的情报分享平台, 用户可以在平台上共享和访问各种网络威胁情报。同时, 平台提供的多种数据, 能够帮助安全团队更好地了解威胁情报。

这些结构化或半结构化数据集能够提供有价值的情报数据, 帮助安全分析人员识别潜在的威胁, 更好地对抗网络攻击。

近年来, 由于深度学习技术的成功, 越来越多网络安全领域的学者使用基于深度学习的方法来训练模型, 而网络安全领域缺乏大规模公开可用的有关实体识别的数据集, 但好的数据集对模型来说由为重要。因此, 为方便使用, 本文汇总了该领域一些高质量的标注语料库。常用的语料库标注方式有 Sang 等人提出的“BIO”[115]和 Ratinov 等人提出的“BIOES”[116]两种。

在表 3 中能够发现这些语料库都包含了多个实体类型, 最多的有 22 种, 并且已经被注释。这些数据来源不尽相同, 主要来自一些安全论坛、安全供应商等。此外, 最近这几年的数据集基本集中在 2020 年, 以后再没有高质量的数据集, 这就导致最近几年新兴产生的实体没有包含在数据集中, 检测起来也很困难。况且近些年随着科技的发展和世界局势的多变, 网络攻击变得愈发频繁和严重, 在威胁情报中会出现许多新的实体, 如果我们不能识别检测到它们, 那对于攻击预防会变得更加困难。

6 NER 评估指标

命名实体识别技术通常通过对比预测标签和标注标签之间的损失进行评估, 包括两种评估方法, 分别是精确匹配评估和宽松匹配评估。使用精确匹配评估时需要同时正确识别实体的边界和类别。使用宽松匹配评估时只要实体被分配了正确的类型, 不管其边界是否正确, 则正确的类别被计入; 或者只要实体的边界正确, 不管其类别是否正确, 则正确的边界被计入[121]。由于 NER 任务通常包括识别实体类别以及实体边界两部分内容, 因此大多数情况都是使用精确匹配评估。NER 评估过程包含三个重要的指标:

(1) 精确率(Precision): 指正确预测为正样本的数量占所有被预测为正样本数量的比例, 公式如下:

$$Precision = \frac{TP}{TP+FP} \tag{4}$$

表 3 CTI 领域命名实体识别数据集汇总

Table 3 Summary of Named Entity Recognition Datasets in the CTI field

参考	实体数量	数据来源	数据规模	*TR*:*VA*:*TE*	是否开源
Bridges 等人[110](2013)	15	从 2010 年 1 月到 2013 年 3 月的所有 CVE/NVD 描述	包含 830 739 个单词, 其中 249 828 个实体, 约占 30%	7:1:2	https://github.com/stucco/auto-labeled-corpus
Yi 等人[117](2020)	22	14 000 篇网络安全领域的论坛、软件供应商公告、博客等非结构化文本	7413 个标签	7:0:3	否
Wang 等人[118](2020)	13	开源威胁情报网站收集和分析的 300 多份威胁情报报告	包含 175 220 单词, 其中 36 412 个实体, 约占 21%	7:1.5:1.5	https://github.com/SCreaM-xp/DNRTI-A-Large-scale-Dataset-for-Named-Entity-Recognition-in-Threat-Intelligence
Kim 等人[119](2020)	10	160 篇非结构化 PDF 文档	包含 498,000 个单词, 其 15 720 个实体, 约占 3.2%	8:0:2	http://github.com/nlpai-lab/CTI-reports-dataset
Zhao 等人[86](2020)	6	75 个与威胁相关的数据源, 包括安全博客、安全供应商公告以及黑客论坛中发布的帖子	包含 15 000 个标签化的威胁情报描述	7:2:1	否
Wang 等人[120](2022)	21	安全公司发布的网络安全相关文章和博客中关于恶意软件、漏洞、APT 的相关描述, 以及开源威胁情报报告	包含 260, 124 个单词, 其中 39 565 个实体, 约占 15.2%	7:1.5:1.5	https://github.com/wangxu-ren/APTNER

(注: *TR* 代表训练集, *VA* 代表验证集, *TE* 代表测试集。其中 *VA* 等于 0 则代表没有划分验证集。)

(2) 召回率(Recall): 指正确预测为正样本的数量占实际上所有正样本数量的比例, 公式如下:

$$Recall = \frac{TP}{TP+FN} \tag{5}$$

(3) F1 分数(F1-score): 是精确率和召回率的调和平均值, F1 分数越高, 表明模型性能越好, 通常是 NER 任务最值得参考的指标, 公式如下:

$$F1 = 2 * \frac{Precision*Recall}{Precision+Recall} \tag{6}$$

还包含两个使用相对不多的指标, 但近些年的使用频率也在增加, 它们都考虑了跨多个实体类型的综合结果:

(4) 宏观 F1 值(Macro-averaged F1-score): 首先使用公式(7)和(8)分别独立计算各个实体类型的准确率和召回率, 取其平均值。然后使用公式(9)计算 $F1_{macro}$:

$$Precision_{macro} = \frac{\sum_{i=1}^{n} Precision_i}{n} \tag{7}$$

$$Recall_{macro} = \frac{\sum_{i=1}^{n} Recall_i}{n} \tag{8}$$

$$F1_{macro} = 2 * \frac{Precision_{macro} * Recall_{macro}}{Precision_{macro} + Recall_{macro}} \tag{9}$$

其中, n 代表实体类型总数, i 代表其中的第 i 类。由于 $F1_{macro}$ 没有考虑数据量的问题, 所有准确率和召回率高的实体类别对 $F1_{macro}$ 的影响较大。

(5) 微观 F1 值(Micro-averaged F1-score): 统计所有实体类型的真阳性、假阳性和假阴性。先使用公式(10)和(11)计算所有类别总体的准确率和召回率, 接下来使用公式(12)计算得到 $F1_{micro}$:

$$Precision_{micro} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FP_i} \tag{10}$$

$$Recall_{micro} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FN_i} \tag{11}$$

$$F1_{micro} = 2 * \frac{Precision_{micro} * Recall_{micro}}{Precision_{micro} + Recall_{micro}} \tag{12}$$

其中, n 代表实体类型总数, i 代表其中的第 i 类。由于 $F1_{micro}$ 考虑了各个类别的数量, 所有更适合用于数据分布不均衡的情况, 数量较多的实体类别对 $F1_{micro}$ 影响较大。

在上述所有公式中, *TP*(True positive)代表真阳性, 即预测值与真实值相同的样本数量; *FP*(False positive)代表假阳性, 即预测为实体, 实际上为非实

体或实体边界或标签错误的样本数量; *FN*(False negative)代表假阴性, 即预测为非实体, 实际上为实体的样本数量。

7 非结构化威胁情报 NER 实验

本文采用第五章的四个开源数据集进行实验, 第一个数据集是 AutoLabel 数据集[110]; 第二个数据集是 DNRTI 数据集[118]; 第三个是 CTIReports 数据集[119], 数据集本身是以 8:2 划分为训练集和测试集, 本文在训练集中拆分出验证集, 最终数据集以 7:1:2 划分为训练集、验证集和测试集。最后一个数据集是 APTNER 数据集[120]。

7.1 实验设置

本文用到的一部分实验参数如下表所示。

表 4 实验中所使用的相关参数及其对应数值

Table 4 The relevant parameters used in the experiment and their corresponding values

参数名称	数值
序列长度(SequenceLength)	128
批次大小(BatchSize)	32
隐藏层维度(HiddenDim)	100
迭代次数(Epoch)	20
Dropout	0.5

7.2 实验结果与对比分析

本文首先使用 AutoLabel 数据集和 DNRTI 数据集对比不同预训练模型的效果, 所使用的传统模型为 Word2Vec-BiLSTM-CRF、XLnet-BiLSTM-CRF、BERT-BiLSTM-CRF。此外, 本文还提出使用 BiGRU 代替 BiLSTM 进行实验。实验结果如表 5 所示。其中, 无论是使用 BiLSTM, 还是使用 BiGRU, 效果最好的预训练模型是 BERT, 最差的模型是 Word2Vec。

另外, 观察本文提出的使用 BiGRU 代替 BiLSTM, 对于 AutoLabel 数据集和 DNRTI 数据集在 BERT 模型的基础上, F1 值分别提升了 0.03%和 0.67%。此外, 对于实验效果而言, 神经网络层的选择对实验结果的影响远小于预训练的语言模型。

然后本文通过 XLNet-BiLSTM-CRF、XLNet-BiGRU-CRF、BERT-BiLSTM-CRF 和 BERT-BiGRU-CRF 模型, 对比不同数据集在相同模型下会产生什么结果, 如图 13 所示。我们可以看到, 无论使用哪种模型, 4 个数据集之间的规律是相同的。AutoLabel 数据集的实验效果最好, APTNER 数据集的实验效果最差。此外, 对于 4 个数据集, 实验结果最好的模型都是 BERT-BiGRU-CRF, 这也说明了使用 BiGRU 代替本文提出的 BiLSTM 的有效性。这是因为 BiGRU 能够更好地捕捉威胁情报文本的依赖特征, 从而增强句子的语义表达。同时, BiGRU 在结构和计算上都比 BiLSTM 简单, 能够缩短时间。

本文通过观察每个数据集的数据分布情况, 认为数据集的特点不同, 会导致了在实验结果上的差异。

(1) AutoLabel: 由于实体在数据集中的分布是均匀的, 同时该数据集中的实体数占比约为 30%, 数据量最大, 所以实验结果最好。

此外, 通过实验发现, 对于 AutoLabel, 无论是改变学习速率还是使用不同的预训练语言模型, 实验结果之间的差异都是最小的。这说明该数据集稳定性最强, 数据设计最合理, 最适合模型训练。

虽然 AutoLabel 数据集的实验结果最好, 但该数据集出现在 2013 年, 由于年份过早, 导致它的数据陈旧, 没有包含新兴的网络安全术语。

(2) DNRTI: 观察到标签之间的分类结果跨度不大, 说明它的实体分布比较均匀。此外, 实体标签的数量约占 21%, 因此实验效果位居第二。

表 5 AutoLabel 数据集和 DNRTI 数据集的实验结果

Tabel 5 Experimental results of AutoLabel dataset and DNRTI dataset

模型结构/评估指标	AutoLabel			DNRTI		
	Accuracy	*Recall*	*F1 score*	*Accuracy*	*Recall*	*F1 score*
Word2Vec-BiLSTM-CRF	95.38	93.11	94.15	80.60	73.73	76.88
Word2Vec-BiGRU-CRF	95.15	93.05	94.03	80.06	71.77	75.41
XLnet-BiLSTM-CRF	96.42	96.44	96.42	86.57	86.14	86.27
XLnet-BiGRU-CRF	96.62	96.40	96.51	86.36	87.04	86.62
BERT-BiLSTM-CRF	97.11	97.57	**97.33**	89.19	90.38	**89.73**
BERT-BiGRU-CRF	97.10	97.64	**97.36**	89.72	91.16	**90.40**

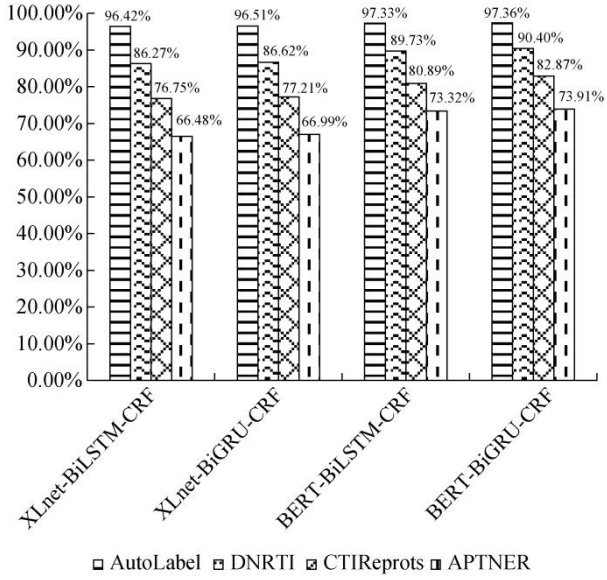

图 13　不同数据集的 F1 得分比较结果

Figure 13　F1 score comparison results of different datasets

但它只包含 13 种实体类型, 根据 STIX2.1 包含的 18 种威胁情报域对象而言, 这不足以满足网络安全领域的需要。

(3) CTIReports: 观察到每个实体类别的分类结果跨度很大, 说明该数据集的数据分布不均匀, 如表 6 所示。此外, 数据集实体所占比例太少, 导致实验结果较差。

表 6　CTIReports 数据集的实体分布情况

Tabel 6　Entity distribution of CTIReports dataset

实体类别	数量	比例(%)
hash	3378	21.80
malware.backdoor	3133	20.22
url.unknown	2614	16.88
malware.infosteal	1507	9.73
url.cncsvr	1049	6.77
ip.unknown	988	6.38
malware.drop	890	5.74
malware.unknown	877	5.66
url.normal	700	4.52
malware.ransom	356	2.30

(4) APTNER: 这个数据集的实验效果最差。该数据集包含 21 个实体类别和 85 个实体标签, 存在标签分布不均匀的情况, 一些实体标签的数量太少, 例如 EMAIL 实体在训练集中只有 13 个, 在测试集中只有 5 个, 在验证集中只有 23 个。此外数据量不足, 虽然与其他 3 个数据集相比, 实体类别的数量有所增加, 但词的总数并没有增加。

这也验证了本文第四章提出的威胁情报领域在进行命名实体识别时所面临的问题。

8　研究展望

针对上述内容, 考虑网络安全领域关于命名实体识别的缺陷, 本文提出以下几点研究方向, 供读者参考:

(1) 目前多数模型都是用于通用领域, 并且模型的性能是建立在固定的数据集上, 当改变数据, 或将其用于网络安全领域时, 模型的性能不能保证。因此, 迁移学习可能会成为解决该问题的方法。虽然当下已经有学者开始这方面的工作, 但还远远不够, 关于迁移学习用于 NER 任务的研究还有很长的路。

所谓迁移学习, 就是能够训练出可以跨领域、跨语言的模型, 从而达到节约实体识别资源、提高 NER 检测效率的方法。按训练方式通常可以分为基于数据的迁移学习、基于模型的迁移学习和对抗迁移学习。

(2) 从全文来看, 多数 NER 方法都是基于有监督学习, 这种方法需要大量的标注数据。未来, 可以研究无监督学习或者半监督学习, 从而解决数据标注耗时耗力等问题。

(3) 在数据标注时, 虽然 “BIO”或者“BIOES”的标注方式较为常见, 但仍然存在其他标注方式。各种各样的标注方式使得 NER 模型的移植性变得很差, 并且数据之间结构不一样, 从而无法使用同一模型进行训练。

(4) 数据集实效性低, 尤其是在网络安全领域, 由于威胁情报更新速度快, 会出现新的实体。因此, 通过基于小样本学习或者零样本学习识别出新型实体的研究应该被重视, 例如提示学习。

(5) 在构建大规模数据集时, 为了节约成本, 可以考虑先标注一部分种子, 然后使用半自动化或自动化方式实现标注。

(6) 预训练模型大多数是关于英文的, 其他语言的模型相对较少, 未来可以开发多语言模型, 才能在其他语言的命名实体识别领域有更大的突破。此外, 大多数预训练语言模型都是基于通用跨域的语料库训练的, 基于威胁情报领域语料训练的模型很少。

(7) 使用新的深度学习技术设计更加全面的 NER 模型, 从而提高实体识别能力。

9 总结

本文首先介绍威胁情报的相关概念, 然后分别概括通用领域和威胁情报领域的 NER 研究现状, 最后总结 CTI 领域的语料库, 提出构建该领域数据集所面临的问题, 以及 NER 技术未来的发展方向。

参考文献

[1] Jiang Q J, Gui Q J, Wang L, et al. A Review of the Research Progress of Named Entity Recognition[J]. *Electric Power Information and Communication Technology*, 2022, 20(2): 15-24.
(江千军, 桂前进, 王磊, 等. 命名实体识别技术研究进展综述[J]. *电力信息与通信技术*, 2022, 20(2): 15-24.)

[2] Gao C, Zhang X, Han M T, et al. A Review on Cyber Security Named Entity Recognition[J]. *Frontiers of Information Technology & Electronic Engineering*, 2021, 22(9): 1153-1168.

[3] McMillan R, Pratap K. Market guide for security threat intelligence services[M]. Gartner report (G00259127), 2014.

[4] China Information Security Standardization Technical Committee. GB/T 36643-2018 Information security technology-Cyber security threat information format [S]. Beijing: State Administration for Market Regulation of the P. R. C and Standardization Administration of the P. R. C. 2018 (in Chinese).

[5] Shi Z X, Ma Y R, Zhang Y, et al. Overview of Threat Intelligence Standards[J]. *Journal of Information Security Research*, 2019, 5(7): 560-569.
(石志鑫, 马瑜汝, 张悦, 等. 威胁情报相关标准综述[J]. *信息安全研究*, 2019, 5(7): 560-569.)

[6] Li J, Sun A X, Han J L, et al. A Survey on Deep Learning for Named Entity Recognition[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2022, 34(1): 50-70.

[7] Kruengkrai C, Nguyen T H, Aljunied S M, et al. Improving Low-Resource Named Entity Recognition Using Joint Sentence and Token Labeling[C]. *The 58th Annual Meeting of the Association for Computational Linguistics*, 2020: 5898-5905.

[8] Chen P, Ding H B, Araki J, et al. Explicitly Capturing Relations between Entity Mentions *via* Graph Neural Networks for Domain-Specific Named Entity Recognition[C]. *The 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (*Volume 2*: *Short Papers*), 2021.

[9] Black W J, Rinaldi F, Mowatt D. FACILE: Description of the NE System Used for MUC-7[C]. *Seventh Message Understanding Conference*, 1998.

[10] Collins M, Singer Y. Unsupervised models for named entity classification[C]. *1999 Joint SIGDAT conference on empirical methods in natural language processing and very large corpora*, 1999.

[11] Kim J H, Woodland P C. A Rule-Based Named Entity Recognition System for Speech Input[C]. *6th International Conference on Spoken Language Processing*, 2000.

[12] Quimbaya A P, Múnera A S, Rivera R A G, et al. Named Entity Recognition over Electronic Health Records through a Combined Dictionary-Based Approach[J]. *Procedia Computer Science*, 2016, 100: 55-61.

[13] Fang Z, Cao Y N, Li T, et al. TEBNER: Domain Specific Named Entity Recognition with Type Expanded Boundary-Aware Network[C]. *The 2021 Conference on Empirical Methods in Natural Language Processing*, 2021: 198-207.

[14] Salah R, Mukred M, Qadri binti Zakaria L, et al. Retracted] a New Rule-Based Approach for Classical Arabic in Natural Language Processing[J]. *Journal of Mathematics*, 2022(1).

[15] Nadeau D, Sekine S. A Survey of Named Entity Recognition and Classification[J]. *Lingvisticae Investigationes*, 2007, 30(1): 3-26.

[16] Zhang S D, Elhadad N. Unsupervised Biomedical Named Entity Recognition: Experiments with Clinical and Biological Texts[J]. *Journal of Biomedical Informatics*, 2013, 46(6): 1088-1098.

[17] Brooke J, Hammond A, Baldwin T. Bootstrapped Text-Level Named Entity Recognition for Literature[C]. *The 54th Annual Meeting of the Association for Computational Linguistics* (*Volume 2*: *Short Papers*), 2016: 344-350.

[18] Rabiner L R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition[J]. *Proceedings of the IEEE*, 1989, 77(2): 257-286.

[19] Koeling R. Chunking with Maximum Entropy Models[C]. *The 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning*, 2000.

[20] Lafferty, John D, McCallum, et al. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C]. *The 18th International Conference on Machine Learning*, 2001: 282-289.

[21] Hearst M A, Dumais S T, Osuna E, et al. Support Vector Machines[J]. *IEEE Intelligent Systems and Their Applications*, 1998, 13(4): 18-28.

[22] Bikel D M, Miller S, Schwartz R, et al. Nymble: a high-performance learning name-finder[EB/OL]. 1998: ArXiv Preprint ArXiv: cmp-lg/9803003.

[23] Curran J R, Clark S. Language Independent NER Using a Maximum Entropy Tagger[C]. *The seventh conference on Natural language learning at HLT-NAACL 2003*, 2003: 164-167.

[24] McCallum A, Li W. Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons[C]. *The seventh conference on Natural language learning at HLT-NAACL 2003*, 2003.

[25] Ju Z F, Wang J, Zhu F. Named Entity Recognition from Biomedical Text Using SVM[C]. *2011 5th International Conference on Bioinformatics and Biomedical Engineering*, 2011: 1-4.

[26] Tang B Z, Cao H X, Wu Y H, et al. Recognizing Clinical Entities in Hospital Discharge Summaries Using Structural Support Vector Machines with Word Representation Features[J]. *BMC Medical Informatics and Decision Making*, 2013, 13(Suppl 1): S1.

[27] Gao C, Zhang X, Liu H. Data and Knowledge-Driven Named Entity Recognition for Cyber Security[J]. *Cybersecurity*, 2021, 4(1): 9.

[28] 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.

[29] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[J]. *Advances in neural information processing systems*, 2013, 26.

[30] Peters M, Neumann M, Iyyer M, et al. Deep Contextualized Word Representations[C]. *The 2018 Conference of the North American Chapter ofthe Association for Computational Linguistics*: *Human LanguageTechnologies*, *Volume 1*, 2018.

[31] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[J]. *Computer Science, Linguistics*, 2018.

[32] Devlin J, Chang M W, Lee K, et al. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding[EB/OL]. 2018: 1810.04805. https://arxiv.org/abs/1810.04805v2.

[33] Akbik A, Blythe D, Vollgraf R. Contextual string embeddings for sequence labeling[C]. *The 27th international conference on computational linguistics*, 2018: 1638-1649.

[34] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners[J]. *OpenAI blog*, 2019, 1(8): 9.

[35] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. *Advances in neural information processing systems*, 2020, 33: 1877-1901.

[36] Zheng H H, Hao Y N, Yu H T. Chinese Named Entity Recognition Based on XLnet Embedding[J]. *Journal of Information Engineering University*, 2021, 22(4): 473-477.
(郑洪浩, 郝一诺, 于洪涛. 基于 XLnet 嵌入的中文命名实体识别方法[J]. *信息工程大学学报*, 2021, 22(4): 473-477.)

[37] Yang Z L, Dai Z H, Yang Y M, et al. XLNet: Generalized Autoregressive Pretraining for Language Understanding[EB/OL]. 2019: 1906.08237. https://arxiv.org/abs/1906.08237v2.

[38] Dai Z H, Yang Z L, Yang Y M, et al. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context[EB/OL]. 2019: 1901.02860. https://arxiv.org/abs/1901.02860v3.

[39] Lample G, Ballesteros M, Subramanian S, et al. Neural Architectures for Named Entity Recognition[EB/OL]. 2016: 1603.01360. https://arxiv.org/abs/1603.01360v3.

[40] Straková J, Straka M, Hajič J. Neural Architectures for Nested NER through Linearization[EB/OL]. 2019: 1908.06926. https://arxiv.org/abs/1908.06926v1.

[41] Simran K, Sriram S, Vinayakumar R, et al. Deep Learning Approach for Intelligent Named Entity Recognition of Cyber Security[M]. Communications in Computer and Information Science. Singapore: Springer Singapore, 2020: 163-172.

[42] Wang X R, Xiong Z H, Du X Y, et al. NER in Threat Intelligence Domain with TSFL[M]. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020: 157-169.

[43] Tikhomirov M, Loukachevitch N, Sirotina A, et al. Using BERT and Augmentation in Named Entity Recognition for Cybersecurity Domain[M]. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020: 16-24.

[44] Li F, Wang Z, Hui S C, et al. Modularized Interaction Network for Named Entity Recognition[C]. *The 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021: 200-209.

[45] Ushio A, Camacho-Collados J. T-NER: An All-round Python Library for Transformer-Based Named Entity Recognition[C]. *The 16th Conference of the European Chapter of the Association for Computational Linguistics*: *System Demonstrations*, 2021: 53-62.

[46] Amin S, Neumann G. T2NER: Transformers Based Transfer Learning Framework for Named Entity Recognition[C]. *The 16th Conference of the European Chapter of the Association for Computational Linguistics*: *System Demonstrations*, 2021: 212-220.

[47] Tu J X, Lignos C. TMR: Evaluating NER Recall on Tough Mentions[EB/OL]. 2021: 2103.12312. https://arxiv.org/abs/2103.12312v1.

[48] Zhang R Y, Zhao P Y, Guo W Y, et al. Medical Named Entity Recognition Based on Dilated Convolutional Neural Network[J]. *Cognitive Robotics*, 2022, 2: 13-20.

[49] Liu T Y, Yao J G, Lin C Y. Towards Improving Neural Named Entity Recognition with Gazetteers[C]. *The 57th Annual Meeting of the Association for Computational Linguistics*, 2019: 5301-5307.

[50] Gao X, Li Q C. Named Entity Recognition in Material Field Based on Bert-BILSTM-Attention-CRF[C]. *2021 IEEE Conference on Telecommunications*, *Optics and Computer Science*, 2021: 955-958.

[51] Liou Y T, Chen C C, Huang H H, et al. Dynamic Graph Transformer for Implicit Tag Recognition[C]. *The 16th Conference of the European Chapter of the Association for Computational Linguistics*: *Main Volume*, 2021: 1426-1431.

[52] Ling X, Weld D. Fine-Grained Entity Recognition[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, 26(1): 94-100.

[53] Hedderich M A, Lange L, Adel H, et al. A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios[EB/OL]. 2020: 2010.12309. https://arxiv.org/abs/2010.12309v3.

[54] Ni J, Dinu G, Florian R. Weakly Supervised Cross-Lingual Named Entity Recognition *via* Effective Annotation and Representation Projection[EB/OL]. 2017: 1707.02483. https://arxiv.org/abs/1707.02483v1.

[55] Chen B, Gu X T, Hu Y F, et al. Improving Distantly-Supervised Entity Typing with Compact Latent Space Clustering[EB/OL]. 2019: 1904.06475. https://arxiv.org/abs/1904.06475v1.

[56] Kamnitsas K, Castro D C, Le Folgoc L, et al. Semi-Supervised Learning *via* Compact Latent Space Clustering[EB/OL]. 2018: 1806.02679. https://arxiv.org/abs/1806.02679v2.

[57] Shang J B, Liu L Y, Ren X, et al. Learning Named Entity Tagger Using Domain-Specific Dictionary[EB/OL]. 2018: 1809.03599. https://arxiv.org/abs/1809.03599v1.

[58] Yang Y, Chen W, Li Z, et al. Distantly supervised NER with partial annotation learning and reinforcement learning[C]. *The 27th International Conference on Computational Linguistics*, 2018: 2159-2169.

[59] Liang C, Yu Y, Jiang H M, et al. BOND: BERT-Assisted Open-Domain Named Entity Recognition with Distant Supervision[C]. *The 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020: 1054-1064.

[60] Fries J, Wu S, Ratner A, et al. SwellShark: A Generative Model for Biomedical Named Entity Recognition without Labeled Data [EB/OL]. 2017: 1704.06360. https://arxiv.org/abs/1704.06360v1.

[61] Peng M L, Xing X Y, Zhang Q, et al. Distantly Supervised Named Entity Recognition Using Positive-Unlabeled Learning[EB/OL]. 2019: 1906.01378. https://arxiv.org/abs/1906.01378v2.

[62] Hedderich M A, Lange L, Klakow D. ANEA: Distant Supervision for Low-Resource Named Entity Recognition[EB/OL]. 2021: 2102.13129. https://arxiv.org/abs/2102.13129v2.

[63] Liu P F, Yuan W Z, Fu J L, et al. Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing[EB/OL]. 2021: 2107.13586. https://arxiv.org/abs/2107.13586v1.

[64] Ding N, Chen Y L, Han X, et al. Prompt-Learning for Fine-Grained Entity Typing[EB/OL]. 2021: 2108.10604. https://arxiv.org/abs/2108.10604v1.

[65] Ma R T, Zhou X, Gui T, et al. Template-Free Prompt Tuning for Few-Shot NER[EB/OL]. 2021: 2109.13532. https://arxiv.org/abs/2109.13532v3.

[66] Cui L Y, Wu Y, Liu J, et al. Template-Based Named Entity Recognition Using BART[EB/OL]. 2021: 2106.01760. https://arxiv.org/abs/2106.01760v1.

[67] Chen X, Zhang N, Li L, et al. Lightner: A lightweight generative framework with prompt-guided attention for low-resource NER[EB/OL]. 2021: ArXiv Preprint ArXiv:2109.00720.

[68] Chen J W, Liu Q, Lin H Y, et al. Few-Shot Named Entity Recognition with Self-Describing Networks[EB/OL]. 2022: 2203.12252. https://arxiv.org/abs/2203.12252v1.

[69] Lee D H, Kadakia A, Tan K M, et al. Good Examples Make a Faster Learner: Simple Demonstration-Based Learning for Low-Resource NER[EB/OL]. 2021: 2110.08454. https://arxiv.org/abs/2110.08454v3.

[70] Dong Y, Guo W, Chen Y, et al. Towards the detection of inconsistencies in public security vulnerability reports[C]. *28th USENIX Security Symposium*, 2019: 869-885.

[71] Satyapanich T, Ferraro F, Finin T. CASIE: Extracting Cybersecurity Event Information from Text[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, 34(5): 8749-8757.

[72] R V, Alazab M, Jolfaei A, et al. Ransomware Triage Using Deep Learning: Twitter as a Case Study[C]. *2019 Cybersecurity and Cyberforensics Conference*, 2019: 67-73.

[73] 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.

[74] Gu J X, Wang Z H, Kuen J, et al. Recent Advances in Convolutional Neural Networks[J]. *Pattern Recognition*, 2018, 77: 354-377.

[75] Xie Teng, Yang Jun-an, Liu Hui. Chinese Entity Recognition Based on BERT-BiLSTM-CRF Model[J]. *Computer Systems & Applications*, 2020, 29(7): 48-55.
(谢腾, 杨俊安, 刘辉. 基于 BERT-BiLSTM-CRF 模型的中文实体识别[J]. *计算机系统应用*, 2020, 29(7): 48-55.)

[76] Jones C L, Bridges R A, Huffer K M T, et al. Towards a Relation Extraction Framework for Cyber-Security Concepts[C]. *The 10th Annual Cyber and Information Security Research Conference*, 2015: 1-4.

[77] Wang T Y, Chow K P. Automatic Tagging of Cyber Threat Intelligence Unstructured Data Using Semantics Extraction[C]. *2019 IEEE International Conference on Intelligence and Security Informatics*, 2019: 197-199.

[78] Jo H, Lee Y, Shin S. Vulcan: Automatic Extraction and Analysis of Cyber Threat Intelligence from Unstructured Text[J]. *Computers & Security*, 2022, 120: 102763.

[79] Balduccini M, Kushner S, Speck J. Ontology-Driven Data Semantics Discovery for Cyber-Security[M]. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2015: 1-16.

[80] 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.

[81] Zhu Z Y, Dumitras T. ChainSmith: Automatically Learning the Semantics of Malicious Campaigns by Mining Threat Intelligence Reports[C]. *2018 IEEE European Symposium on Security and Privacy*, 2018: 458-472.

[82] Mulwad V, Li W J, Joshi A, et al. Extracting Information about Security Vulnerabilities from Web Text[C]. *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, 2011: 257-260.

[83] R. Lal. Information Extraction of cyber security related terms and concepts from unstructured text[D]. USA: University of Maryland, 2013.

[84] Sabottke C, Suciu O, Dumitras T. Vulnerability Disclosure in the Age of Social Media: Exploiting Twitter for Predicting Real-World Exploits[J]. *Proceedings of the 24th USENIX Security Symposium*, 2015: 1041-1056.

[85] Dionísio N, Alves F, Ferreira P M, et al. Cyberthreat Detection from Twitter Using Deep Neural Networks[C]. *2019 International Joint Conference on Neural Networks*, 2019: 1-8.

[86] 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.

[87] Pingle A, Piplai A, Mittal S, et al. RelExt: Relation Extraction Using Deep Learning Approaches for Cybersecurity Knowledge Graph Improvement[C]. *The 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2019: 879-886.

[88] Zhou S P, Long Z, Tan L Z, et al. Automatic Identification of Indicators of Compromise Using Neural-Based Sequence Labelling[EB/OL]. 2018: 1810.10156. https://arxiv.org/abs/1810.10156v1.

[89] Zhang H, Guo Y B, Li T. Multifeature Named Entity Recognition in Information Security Based on Adversarial Learning[J]. *Security and Communication Networks*, 2019, 2019: 6417407.

[90] Wu H, Li X Y, Gao Y L. An Effective Approach of Named Entity Recognition for Cyber Threat Intelligence[C]. *2020 IEEE 4th Information Technology*, *Networking*, *Electronic and Automation*

Control Conference, 2020: 1370-1374.

[91] Wu H. Research and implementation of entity recognition model for cyber threat intelligence[D]. Beijing: Beijing University of Posts and Telecommunications, 2020.
(武涵. 网络威胁情报实体识别模型研究与实现[D]. 北京: 北京邮电大学, 2020.)

[92] Wang X R, Liu R S, Yang J, et al. Cyber Threat Intelligence Entity Extraction Based on Deep Learning and Field Knowledge Engineering[C]. *2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design*, 2022: 406-413.

[93] Zhou Y H, Tang Y, Yi M, et al. CTI View: APT Threat Intelligence Analysis System[J]. *Security and Communication Networks*, 2022, 2022: 9875199.

[94] Lin B Y, Lee D H, Shen M, et al. Triggerner: Learning with entity triggers as explanations for named entity recognition[EB/OL]. 2020: *ArXiv Preprint ArXiv*:2004.07493.

[95] Wu J X. Low-Resource Named Entity Recognition Based on Multi-Hop Dependency Trigger[EB/OL]. 2021: 2109.07118. https://arxiv.org/abs/2109.07118v3.

[96] Liu J, Yan J J, Jiang J, et al. TriCTI: An Actionable Cyber Threat Intelligence Discovery System *via* Trigger-Enhanced Neural Network[J]. *Cybersecurity*, 2022, 5(1): 8.

[97] Lee D H, Selvam R K, Sarwar S M, et al. AutoTriggER: Label-Efficient and Robust Named Entity Recognition with Auxiliary Trigger Extraction[EB/OL]. 2021: 2109.04726. https://arxiv.org/abs/2109.04726v3.

[98] Gascon H, Grobauer B, Schreck T, et al. Mining Attributed Graphs for Threat Intelligence[C]. *The Seventh ACM on Conference on Data and Application Security and Privacy*, 2017: 15-22.

[99] 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.

[100] Kashihara K, Sandhu H S, Shakarian J. Automated Corpus Annotation for Cybersecurity Named Entity Recognition with Small Keyword Dictionary[M]. Lecture Notes in Networks and Systems. Cham: Springer International Publishing, 2021: 155-174.

[101] 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.

[102] Husari G, Al-Shaer E, Ahmed M, et al. TTPDrill: Automatic and Accurate Extraction of Threat Actions from Unstructured Text of CTI Sources[C]. *The 33rd Annual Computer Security Applications Conference*, 2017: 103-115.

[103] Husari G, Niu X, Chu B, et al. Using Entropy and Mutual Information to Extract Threat Actions from Cyber Threat Intelligence[C]. *2018 IEEE International Conference on Intelligence and Security Informatics*, 2018: 1-6.

[104] Long Z, Tan L Z, Zhou S P, et al. Collecting Indicators of Compromise from Unstructured Text of Cybersecurity Articles Using Neural-Based Sequence Labelling[C]. *2019 International Joint Conference on Neural Networks*, 2019: 1-8.

[105] Lian L Y. Named Entities Recognition of Bi-LSTM+CRF in Cyberspace Security Domain[J]. *Journal of Heilongjiang University of Science and Technology*, 2020, 30(6): 717-722.
(廉龙颖. Bi-LSTM+CRF 的网络空间安全领域命名实体的识别[J]. *黑龙江科技大学学报*, 2020, 30(6): 717-722.)

[106] Liu W G. Network Security Entity Recognition Methods Based on the Deep Neural Network[M]. Advances in Intelligent Systems and Computing. Singapore: Springer Singapore, 2020: 1687-1692.

[107] Chen Y X, Ding J W, Li D S, et al. Joint BERT Model Based Cybersecurity Named Entity Recognition[C]. *2021 The 4th International Conference on Software Engineering and Information Management*, 2021: 236-242.

[108] Xie B, Shen G W, Guo C, et al. The Named Entity Recognition of Chinese Cybersecurity Using an Active Learning Strategy[J]. *Wireless Communications and Mobile Computing*, 2021, 2021(1).

[109] Zhu X Y, Zhang Y, Zhu L, et al. Chinese Named Entity Recognition Method for the Field of Network Security Based on RoBERTa[C]. *2021 International Conference on Networking and Network Applications*, 2021: 420-425.

[110] Bridges R A, Jones C L, Iannacone M D, et al. Automatic Labeling for Entity Extraction in Cyber Security[EB/OL]. 2013: 1308.4941. https://arxiv.org/abs/1308.4941v3.

[111] Collobert R, Weston J, Bottou L, et al. Natural Language Processing (almost) from Scratch[J]. *Journal of Machine Learning Research*, 2011, 12: 2493-2537.

[112] Nguyen T H, Grishman R. Event Detection and Domain Adaptation with 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 2*: *Short Papers*), 2015: 365-371.

[113] Vinayakumar R, Alazab M, Srinivasan S, et al. A Visualized Botnet Detection System Based Deep Learning for the Internet of Things Networks of Smart Cities[J]. *IEEE Transactions on Industry Applications*, 2020, 56(4): 4436-4456.

[114] Vinayakumar R, Alazab M, Soman K P, et al. Robust Intelligent Malware Detection Using Deep Learning[J]. *IEEE Access*, 2019, 7: 46717-46738.

[115] Tjong Kim Sang E F, De Meulder F. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition[C]. *The seventh conference on Natural language learning at HLT-NAACL 2003*, 2003.

[116] Ratinov L, Roth D. Design Challenges and Misconceptions in Named Entity Recognition[C]. *The Thirteenth Conference on Computational Natural Language Learning-CoNLL '09*, 2009: 147-155.

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

[118] Wang X R, Liu X P, Ao S Q, et al. DNRTI: A Large-Scale Dataset for Named Entity Recognition in Threat Intelligence[C]. *2020 IEEE 19th International Conference on Trust*, *Security and Privacy in Computing and Communications*, 2020: 1842-1848.

[119] Kim G, Lee C, Jo J, et al. Automatic Extraction of Named Entities

of Cyber Threats Using a Deep Bi-LSTM-CRF Network[J]. *International Journal of Machine Learning and Cybernetics*, 2020, 11(10): 2341-2355.

[120] Wang X R, He S H, Xiong Z H, et al. APTNER: A Specific Dataset for NER Missions in Cyber Threat Intelligence Field[C]. *2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design*, 2022: 1233-1238.

[121] Grishman R, Sundheim B. Message Understanding Conference-6: A Brief History[C]. *The 16th conference on Computational linguistics*, 1996.

王旭仁 于2004年在中国科学院高能物理研究所获得工学博士学位。现在任首都师范大学副教授。研究领域为网络空间安全、网络威胁情报。Email: wangxuren@cnu.edu.cn

魏欣欣 于2017年在河北师范大学软件工程专业获得学士学位。现在首都师范大学电子信息专业攻读硕士学位。研究领域为网络威胁情报、自然语言处理。Email: 2211002078@cnu.edu.cn

王媛媛 31005部队任工程师, 主要研究空中交通管理数据信息处理、网络空间安全态势分析。Email: wangyuanyuan@sohu.com

姜政伟 于2014年在中国科学院大学获得博士学位。现在任中国科学院信息工程研究所正高级工程师, 中国科学院大学网络空间安全学院岗位教授。研究领域为威胁情报、网络威胁发现与溯源。Email: jiangzhengwei@iie.ac.cn

江钧 于2016年在北京交通大学电子科学与技术专业获得硕士学位。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、威胁情报。Email: jiangjun860@iie.ac.cn

杨沛安 于2018年在中国科学院大学计算机应用技术专业获得博士学位。现任中国科学院信息工程研究所高级工程师。研究领域为网络空间威胁发现与溯源、网络空间威胁情报智能分析。Email: yangpeian@iie.ac.cn

刘润时 现在首都师范大学计算机科学与技术专业攻读本科。研究兴趣包括: 人工智能, 大数据。Email: 1192905008@cnu.edu.cn