

面向APT的网络威胁情报知识图谱构建研究

宋国宝

(江西理工大学信息工程学院,江西 赣州 341000)

摘要: 针对高级持续威胁(APT)情报未能得到有效利用,且缺乏自动化方法生成结构化知识并构建APT攻击知识图谱的问题,提出一种结合实体识别与关系抽取的APT攻击知识自动抽取方法。首先,针对APT攻击特点设计11种实体类别和9种关系类别;其次,分别提出CySecBERT-BiLSTM-ATTN-NCRF实体识别模型和CySecBERT-ATLOP文档级关系抽取模型进行APT情报的实体与关系抽取;最后,构建APT威胁情报知识图谱以清晰呈现实体及其关系。实验结果显示,相较于基线模型,该方法在实体识别和关系抽取方面的F1分数分别提高了4.65和14.52;在Neo4j中展示的知识图谱亦证明了其有效性。

关键词: 知识图谱;实体识别;关系抽取;威胁情报

DOI: 10.11907/rjdk.241277

中图分类号: TP391

文献标识码: A

开放科学(资源服务)标识码(OSID):

文章编号: 1672-7800(2025)005-0179-07



Research on the Construction of Network Threat Intelligence Knowledge Graph for APT

SONG Guobao

(School of Information Engineering, Jiangxi University of Science and Technology, Ganzhou 341000, China)

Abstract: Aiming at the problem of ineffective utilization of Advanced Persistent Threat (APT) intelligence and the lack of automated methods to generate structured knowledge and construct APT attack knowledge graphs, a method for automatic extraction of APT attack knowledge combining entity recognition and relationship extraction is proposed. Firstly, design 11 entity categories and 9 relationship categories based on the characteristics of APT attacks; Secondly, the CySecBERT-BiLSTM-ATTN-NCRF entity recognition model and the CySecBERT-ATLOP document level relationship extraction model are proposed separately for entity and relationship extraction in APT intelligence; Finally, construct an APT threat intelligence knowledge graph to clearly present entities and their relationships. The experimental results showed that compared to the baseline model, this method improved the F1 scores in entity recognition and relationship extraction by 4.65 and 14.52, respectively; The knowledge graph presented in Neo4j also proves its effectiveness.

Key Words: knowledge graph; entity recognition; relationship extraction; threat intelligence

0 引言

近年来,随着互联网技术的迅速发展,社会各界获得了巨大便利,但也面临着严重的安全威胁,如0day攻击、蠕虫感染、勒索软件攻击和拒绝服务攻击等。其中,高级持续威胁攻击(Advanced Persistent Threat, APT)因具有极强的组织性、隐蔽性和威胁性,会造成比传统攻击更严重的损害。安全专家提出利用网络威胁情报进行防御,因为威胁情报能够提供APT组织的动态信息,并协助溯源网络攻

击事件。然而,APT情报的有效运用需要获取、汇总和传递网络中分散、多样、片段化的大量安全威胁信息。这些信息散布在网络空间的各个角落,涵盖安全公司报告、威胁指标库、网络博客等多个平台。如何对海量情报数据进行合理有效的提取与归纳成为一个亟需解决的问题。知识图谱(Knowledge Graph, KG)以图形化的形式有效组织、控制和展示多源异构动态信息,是当前大数据环境下的一个重要研究方向,将其引入APT威胁情报领域具有一定的必要性。

收稿日期:2024-03-19

扫描二维码阅读全文:

作者简介:宋国宝(1998-),男,江西理工大学信息工程学院硕士研究生,研究方向为网络威胁情报。



1 相关工作

1.1 网络威胁情报实体识别

命名实体的准确识别是构建知识图谱节点的基础。命名实体识别(Named Entity Recognition, NER)旨在从非结构化文本中识别出具有特定类型和意义的实体。Simran 等^[1]针对网络安全 NER 提出基于 BiGRU-CNN-CRF (Bidirectional Gated Recurrent Unit-Convolutional Neural Network-Conditional Random Fields)的模型,并在公开的网络安全 NER 数据集上评估了一系列基于深度学习的 NER 方法。Qin 等^[2]提出一种结合特征模板 (Feature Template, FT) 与神经网络的 FT-CNN-BiLSTM-CRF 安全 NER 模型,在大规模网络安全数据集上的 F-score 可以达到 86%,虽然取得了不错效果,但特征提取仍然存在不足。Gao 等^[3]为解决网络威胁情报中难以识别的稀有实体和复杂实体问题,提出结合双向长短期记忆网络 (Bidirectional Long and Short-Term Memory, BiLSTM) 与领域词典的 NER 方法,通过构建领域词典识别稀有实体,利用注意力机制和 BiLSTM 模型丰富文本的局部特征,帮助识别复杂实体。

1.2 网络威胁情报实体关系抽取

实体关系抽取旨在通过一些方法自动识别实体对之间的某种语义关系,是构建知识图谱的关键步骤。Wang 等^[4]提出一种名为 NR-RL-PCNN-ATT (Reinforcement Learning-Piecewise Convolutional Neural Network-Attention) 的关系抽取模型,通过提高示例选择器对语句的筛选效果,并降低由标注错误引起的噪音,改善了关系抽取效果。Pingle 等^[5]提出一种基于前馈神经网络模型 (Feed Forward Neural Network, FFNN) 的关系抽取方法 RelExt,能有效获取网络安全实体关系。

1.3 网络威胁情报知识图谱构建及应用

在网络威胁情报的处理与分析中,知识图谱能够提供统一视图、揭示隐藏关系、支持推理与预测,从而帮助相关人员与部门更好地应对网络安全威胁。Milajerdi 等^[6]提出一个名为 POIROT 的系统,利用内核审核日志构建溯源图以检测 APT 攻击。Satvat 等^[7]利用网络威胁情报提取攻击行为相关的实体和关系,并自动生成查询图,然后借助 POIROT 系统实现了威胁搜索的自动化。Han 等^[8]针对 APT 攻击具有缓慢可持续以及使用 Oday 漏洞的特点,构建了一种基于溯源图的 APT 攻击检测系统 Unicorn,其能在没有事先定义攻击特征的情况下识别潜在异常行为。Sarhan 等^[9]提出一种开放式网络威胁情报知识图谱框架,首先采用基于注意力的开放式信息抽取框架 (Open Information Extraction, OIE) 提取无关领域的实体关系三元组;其次进行领域实体标注获取网络安全实体;最后使用基于词嵌入的融合技术规范提取到的数据,并导入图数据库完成知识图谱构建。Sun 等^[10]提出一种基于多特征的 BiL-

STM-CRF 方法用于实体识别,应用基于 BiGRU 的跨句子关系提取模型来抽取实体关系,然后构建 APT 知识图谱。Narayanan 等^[11]开发了一种认知系统,通过整合来自传统传感器、动态在线文本和知识图谱的信息进行网络安全事件的早期检测。Kiesling 等^[12]通过一个基于查询的案例展示了如何将来自网络入侵检测系统的警报与网络安全知识图谱连接起来,以便更深入地理解潜在威胁和正在进行的攻击。

目前,APT 情报知识图谱的构建面临以下挑战:①与一般领域的实体不同,APT 情报领域的实体具有很强的领域特征,如威胁实体包括攻击组织、攻击方式、恶意软件等,很难通过通用领域的实体识别模型直接识别;②APT 情报文本结构复杂,句子较长,实体之间的关系通常是由多个句子推断而来;③缺乏公开的 APT 情报数据集。

为克服上述挑战,本文将自然语言技术应用于 APT 情报领域,提出一种结合实体识别与关系抽取的 APT 知识图谱构建方法,主要贡献包括:①该方法实现了从 APT 非结构化文本报告中提取结构化知识并生成知识图谱的能力,利用图数据库 Neo4j 存储知识图谱,明确展现了 APT 报告中实体及其之间的关系,为安全研究人员提供了分析攻击事件和部署防御的决策支持;②提出一种基于数据增强的 APT 攻击实体识别方法,利用 CySecBERT-BiLSTM-ATTN-NCRF (CyberSecurity Bert-Bidirectional Long and Short-term Memory-Attention-N Conditional Random Fields) 模型对 APT 报告进行实体识别,并通过数据增强方法提高泛化能力;③提出一种适用于 APT 领域的实体关系抽取方法,利用 CySecBERT-ATLOP (CyberSecurity Bert-Adaptive Thresholding and Localized cOntext Pooling) 模型进行实体关系抽取。

2 构建方法

2.1 图谱设计

结构化威胁信息表达 (Structured Threat Information Expressions, STIX) 是用于交换网络威胁情报的语言和序列化格式,是共享威胁情报的行业标准之一^[13]。本文根据 STIX2.1 标准定义了威胁情报的实体和关系类型,具体如图 1 所示。

2.1.1 实体定义

APT 攻击实体识别任务旨在识别有具体意义的攻击实体,是构建 APT 知识图谱的基础。例如,输入文本序列 “Andariel deploys Maui ransomware” 中的 “Andariel” 是一个攻击组织,“Maui” 是一个恶意软件,而 “ransomware” 则是一个恶意软件属性,实体识别即在给定文本序列中识别出 APT 攻击实体。本文针对 APT 攻击特点,同时参考 STIX2.1 标准设计出 19 种命名实体,分为模型识别实体和正则识别实体两类。其中,模型识别实体有 11 种,包括攻

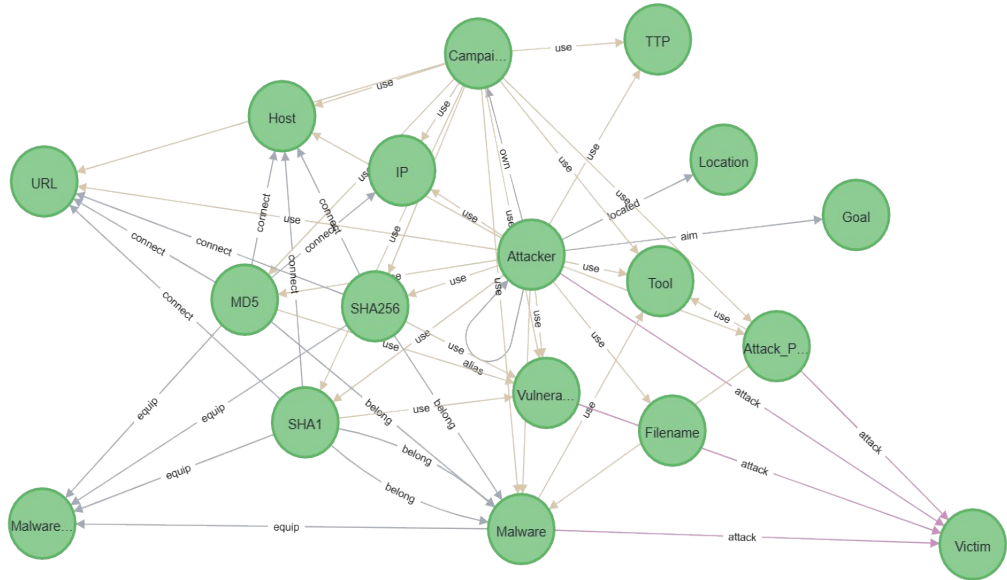


Fig. 1 Threat intelligence entities and relationship types
图1 威胁情报实体和关系类型

击组织、攻击工具、攻击手法、攻击漏洞(非 CVE)、攻击事件、攻击目标、攻击技战术、攻击意图、区域位置、恶意软件和恶意软件属性;正则识别实体有 8 种,具体如表 1 所示。

Table 1 Regular recognition expression
表 1 正则识别表达式

序号	实体类型	正则表达式
1	URL	$\backslash \mathbf{b}([a-z]\{3,\}:[/[\S]\{16,\}\})\backslash \mathbf{b}$
2	Host	$\backslash \mathbf{b}([a-z0-9-]\{2,\}[\.]\{?\}\})\backslash \mathbf{b}$
3	IP	$\backslash \mathbf{b}(\backslash \mathbf{d}\{1,3\}.\backslash \mathbf{d}\{1,3\}.\backslash \mathbf{d}\{1,3\}.\backslash \mathbf{d}\{1,3\})\backslash \mathbf{b}$
4	MD5	$\backslash \mathbf{b}([a-f0-9]\{32\})\backslash \mathbf{b}$
5	SHA1	$\backslash \mathbf{b}([a-f0-9]\{40\})\backslash \mathbf{b}$
6	SHA256	$\backslash \mathbf{b}([a-f0-9]\{64\})\backslash \mathbf{b}$
7	CVE	$\backslash \mathbf{b}(\text{CVE}-[0-9]\{4\}-[0-9]\{4,6\})\backslash \mathbf{b}$
8	Filename	$\backslash \mathbf{b}([a-zA-Z]:((?:[a-zA-Z0-9()])*)\backslash \mathbf{b}$

2.1.1.2 关系定义

APT 关系抽取任务旨在识别攻击实体对之间存在的关系,是构建 APT 知识图谱的关键。例如,输入句子“Andariel deploys Maui ransomware”描述了攻击组织 Andariel 使用恶意软件 Maui 的关系事实,可采用关系三元组(Andariel, use, Maui)表示。由于 APT 攻击的特殊性,针对上述 19 种攻击实体,本文设计 9 种 APT 实体关系,包括攻击(attack)、使用(use)、属于(belong)、位于(located)、拥有(own)、目的(aim)、请求(connect)、属性配备(equip)和别名(alias)。

2.2 实体识别

APT 相关网络威胁情报实体在多个方面都具有高度的特定性,其丰富的实体类型无法非常准确地被基本模型捕获。因此,本文引入 CySecBERT 预训练模型,将 BiLSTM-CRF 作为基准模型进行实体识别^[14]。实体识别模型架构如图 2 所示。

使用预训练模型 CySecBERT 作为编码器,在文档开头添加[CLS]和[SEP]标签。威胁情报中的句子可表示为

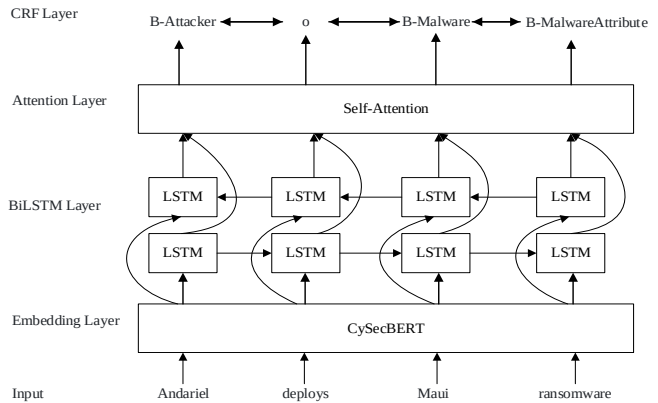


Fig. 2 Structure of entity extraction model
图2 实体识别模型架构

$X = (x_1, x_2, \dots, x_l)$, x_i 表示句子 X 中第 i 个单词。采用 Cy-SecBERT 对句子进行编码并获取上下文表示 H 。表示为:

$$H = \text{CySecBERT}([x_1, x_2, \dots, x_l]) = [h_1, h_2, \dots, h_l] \quad (1)$$

为获取单词的上下文语义信息,将上述词嵌入向量作为 BiLSTM 的输入,分别获取前向 LSTM 和后向 LSTM 的输出 $\vec{P}_l$ 和 $\overleftarrow{P}_l$, 并对其进行拼接得到 BiLSTM 的输出序列 $\vec{P}_l$ 。表示为:

$$P = \text{BiLSTM}([h_1, h_2, \dots, h_l]) = [p_1, p_2, \dots, p_l] \quad (2)$$

为减少 BiLSTM 模型在网络威胁情报领域中长序列信息丢失的问题,为上层模型输出的特征向量分配权重,突出有助于实体抽取的特征,弱化无关特征,在 BiLSTM 后添加注意力机制。表示为:

$$A = \text{Attention}([p_1, p_2, \dots, p_l]) = [a_1, a_2, \dots, a_l] \quad (3)$$

为获取威胁情报句子的序列标签,本文引入 CRF 层对最终的预测标签进行合法性限制。对于句子 $X = (x_1, x_2, \dots, x_l)$, 对应实体标签序列 $(y_1, y_2, \dots, y_l)$ 的概率分

布公式为:

$$S(X, Y) = \sum_{i=0}^l (A_{y_i, y_{i+1}} + P_{i, y_i}) \quad (4)$$

式中: A 表示 CRF 模型的状态转移矩阵; $A_{y_i, y_{i+1}}$ 表示 y_i 标签到 y_{i+1} 标签的转移分数; P_{i, y_i} 表示句子中第 i 个单词在对应实体类型 y_i 的标签分数。因此, 输入句子 X 的最终标签由矩阵 P 和矩阵 A 共同决定。

然后使用 softmax 函数计算输入句子 X 所有可能的输出标签序列概率。该序列的概率定义为:

$$p(Y|X) = \frac{e^{S(X, Y)}}{\sum_{\tilde{Y} \in Y_X} e^{S(X, \tilde{Y})}} \quad (5)$$

式中: $\tilde{Y}$ 为真实标签序列; Y_X 为标签集合。

最后通过最大似然估计进行模型参数训练, 获取得分最高的标签序列。表示为:

$$\operatorname{argmax} = \operatorname{argmax} \left(S(X, Y) - \log \left(\sum_{\tilde{Y} \in Y_X} e^{S(X, \tilde{Y})} \right) \right) \quad (6)$$

2.3 关系抽取

在 APT 情报中, 实体之间的关系通常跨越多个句子, 因此本文选择文档级关系抽取方法, 将 ATLOP 作为基准模型^[15]。在编码层, 使用预训练模型 CySecBERT 作为编码器。对于给定的一篇文档 $D = [x_i]_{i=1}^l$ 中的每一个提及, 都会在开头和结尾插入一个特殊标签 "*", 并采用 CySecBERT 对该文档进行编码和获取上下文表示 H 。表示为:

$$H = \text{CySecBERT}([x_1, x_2, \dots, x_l]) = [h_1, h_2, \dots, h_l] \quad (7)$$

对于一个具有提及 $\{m_j\}_{j=1}^{N_i}$ 的实体 e_i , 使用 logsumexp pooling 获取实体嵌入 h_{e_i} 。表示为:

$$h_{e_i} = \log \sum_{j=1}^{N_i} \exp(h_{m_j}) \quad (8)$$

在编码层中, CySecBERT 编码器使用多头注意力矩阵 $A \in R^{H \times l \times l}$ 。矩阵中的 A_{ijk} 表示编码器第 i 个注意力头中从标记 j 到标记 k 的注意力得分, 在提及之前带有 "*" 的注意力被视为该提及的注意力得分。同时, 实体级别的注意力得分 $A_i^E \in R^{H \times l}$ 是通过将同一实体所有提及的注意力得分进行平均得到, 表示从第 i 个实体到所有标记的注意力。此外, 特定实体对 (e_s, e_o) 的重要上下文由它们实体级别的注意力相乘确定, 并且通过以下计算得到局部上下文嵌入 $c^{(s, o)}$:

$$A^{(s, o)} = A_s^E \cdot A_o^E \quad (9)$$

$$q^{(s, o)} = \sum_{i=1}^H A_i^{(s, o)} \quad (10)$$

$$a^{(s, o)} = q^{(s, o)} / 1^\top q^{(s, o)} \quad (11)$$

$$c^{(s, o)} = H^\top a^{(s, o)} \quad (12)$$

式中: H 为式(1)中的上下文嵌入。局部上下文嵌入

是在全局池化的实体嵌入基础上构建的, 这样可以为不同实体对之间构建不同的实体表示。具体的实体对可表示为:

$$z_s^{(s, o)} = \tanh(W_s h_{e_s} + W_{c_1} c^{(s, o)}) \quad (13)$$

$$z_o^{(s, o)} = \tanh(W_o h_{e_o} + W_{c_2} c^{(s, o)}) \quad (14)$$

式中: $W_{c_1}, W_{c_2} \in R^{d \times d}$ 为模型参数。

通过双线性函数和 Sigmoid 激活函数计算关系 r 的概率。该过程可表示为:

$$P(r|e_s, e_o) = \sigma \left(\sum_{i=1}^k z_s^{i\top} W_r z_o^i + b_r \right) \quad (15)$$

式中: $b_r \in R, W_r \in R^{d \times d}$ 为模型参数。

2.4 数据增强

数据增强算法旨在通过对原始训练数据进行多样性变换, 以提高深度学习模型的泛化性能。这一过程包括在标注数据中引入不同形式的噪声或进行多样性转换, 如替换、删除、插入等操作。在实体识别任务中, 一些算法采用标签与单词混合的方式进行数据增强, 如 DAGA 通过语言模型进行数据增强, 以拓展训练集^[16]; MELM 方法则基于预训练的微调掩码语言模型将实体标签线性化到句子上下文中, 用于实体替换以进行数据增强^[17]; EDA 采用同义词替换、随机插入、随机交换和随机删除 4 种简单但强大的数据增强技术提升文本分类任务性能^[18]。

在安全领域, 由于网络威胁情报领域的标记样本数量有限, 且获取和标注这些样本需要耗费大量时间和专业知识, 数据增强成为提升深度学习模型性能的关键策略之一。本文使用比较常用的几种简单数据增强方法, 包括替换网络安全实体、替换上下文、随机删除等, 从而选择适合网络安全领域的数据增强方法。以 B-I-O 标注的样本数据为例, 句子 "Andariel deploys Maui ransomware" 描述了 "Andariel" 部署了 "Maui" 的基本事实, 其中实体 "Andariel" 的标签为 "B-Attacker", "Maui" 的标签为 "B-Malware", "ransomware" 的标签为 "B-Malware_Attribute", 而 "deploys" 的标签为 "O"。表 2 为应用不同数据增强方法得到的不同样例。实体替换是指对上述定义的安全实体和正则识别的实体进行同类替换, 例如上述句子是对句中的攻击组织进行替换, 由 "Andariel" 替换为 "APT1"; 上下文替换是指对非上述网络安全实体进行同义替换, 即替换 "O" 标签的实体, 将句中的 "deploys" 替换为 "launches"; 随机删除则是按照本文设置的 15% 概率随机删除句子中的一部分单词, 例如删除句中的 "ransomware"。由于增强的句子具有随机性, 为防止产生若干相同的句子, 对其进行降重处理。

3 实验方法与结果分析

3.1 实验环境与参数设置

系统开发环境为 Ubuntu 22.04 LTS, GPU 为 GeForce RTX 3060, 开发工具为 VSCode, 开发语言为 Python3.8, 基

Table 2 Enhanced sentence examples
表 2 增强句子样例

数据增强方法	句子样例			
随机删除	Andariel	deploys	Maui	
	B-Attacker	O	B-Malware	
替换实体	APT1	deploys	Maui	ransomware
	B-Attacker	O	B-Malware	B-Malware_Attribute
替换上下文	Andariel	launches	Maui	ransomware
	B-Attacker	O	B-Malware	B-Malware_Attribute

于 Pytorch 实现 APT 威胁情报的信息抽取模型。实体识别和关系抽取相关超参数设置如表 3 所示。

Table 3 Hyperparameter settings
表 3 超参数设置

参数	数值
BERT 模型	CySecBert
batch_size	32
num_epochs	30
Max sequence length	600
embedding_size	768
hidden_dim	512
lstm_layers	1
lstm_dropout	0.3
crf_dropout	0.3
lr for encoder	5e-5
lr for classifier	1e-4
Learning rate	0.1

3.2 数据集构建

本文实验数据的主要来源为 CyberMonitor 的开源情报库以及各大安全厂商发布的情报文本,选择 433 篇与 APT 相关的网络威胁情报进行数据标注并构建数据集,根据其内容是否与 APT 攻击相关进行人工筛选。对于实体识别和关系抽取任务,按照 8:1:1 的比例将句子划分为训练集、验证集和测试集。具体信息如表 4 和表 5 所示。

Table 4 Entity recognition annotation dataset
表 4 实体识别标注数据集

参数	训练集	验证集	测试集
句子数量	10 458	1 307	1 307
标签个数	19 989	2 504	2 552
Malware	5 589	740	701
Malware_Attribute	1 493	228	192
Victim	5 249	605	652
Location	343	42	40
Attacker	5 302	660	712
Tool	404	43	47
Campaign	382	48	46
Goal	233	22	31
Vulnerability	183	29	25
TTP	673	68	94
Attack_Pattern	138	19	12

3.3 评价指标

使用准确度(Precision)、召回率(Recall)和 F1 值评估模型性能。计算公式为:

$$P = \frac{\text{被正确预测的数目}}{\text{所有识别出的数目}} \tag{16}$$

Table 5 Entity relationship extraction dataset partition
表 5 实体关系抽取数据集划分

参数	训练集	验证集	测试集
文档数量	347	38	48
attack	1 402	213	378
use	2 252	212	234
aim	148	10	13
belong	335	21	16
equip	1 402	108	153
connect	136	17	18
alias	513	32	43
located	233	12	20
own	113	12	18

$$R = \frac{\text{被正确预测的数目}}{\text{所有标记的数目}} \tag{17}$$

$$F1 = \frac{2 \times R \times P}{P + R} \tag{18}$$

在对所有实体及其关系进行效果评估时,通常采用宏平均方法来计算多个实体的 F1 值平均。公式为:

$$Macro - F1 = \frac{1}{n} \sum_{i=1}^n F1_i \tag{19}$$

3.4 实验结果与分析

3.4.1 实体识别性能

将主流模型 BERT-BiLSTM-ATTN-NCRF 作为基线模型,比较通用命名识别领域的 CRF、BiLSTM、BiLSTM-CRF、BERT-BiLSTM-CRF、BERT-BiLSTM-ATTN-CRF 5 种模型与本文模型的性能。结果如表 6 所示。可以看出,本文模型在准确度、召回率和 F1 值上均达到最高水平,分别为 93.68%、79.00% 和 85.47%,表明其在实体识别任务中具有优秀性能,能够更准确地识别实体。

Table 6 Comparison of entity recognition performance among different models

模型	准确度	召回率	F1 值
CRF	79.64	63.12	69.97
BiLSTM	66.57	50.20	54.75
BiLSTM-CRF	75.45	69.92	72.39
BERT-BiLSTM-CRF	90.04	74.02	80.61
BERT-BiLSTM-ATTN-CRF	86.98	72.78	78.80
BERT-BiLSTM-ATTN-NCRF	89.40	74.71	80.82
本文模型	93.68	79.00	85.47

由于通用预训练模型 BERT 无法准确深入理解安全领域文本,本文引入网络安全领域预训练模型 CySecBERT 作为词编码层。表 7 为引入 CySecBert 模型前后的实验结果比较。可以看出,引入预训练模型 CySecBert 后,3 个实体识别模型在准确度、召回率和 F1 值上均有所提升。

尽管本文标注的实体数量已有 2 万多个,但部分实体出现次数仍然很少,如攻击方式仅有 169 个,会影响深度学习模型的泛化能力,因此需引入数据增强方法扩充训练集。在 CySecBERT-BiLSTM-ATTN-NCRF 模型上对常用数据增强方法进行比较,结果如表 8 所示。可以看出,采

Table 7 Comparison of experimental results before and after introducing the CySecBert model**表 7 引入 CySecBert 模型前后实验结果比较 (%)**

模型	准确度	召回率	F1 值
BERT-BiLSTM-CRF	90.04	74.02	80.61
BERT-BiLSTM-ATTN-CRF	86.98	72.78	78.80
BERT-BiLSTM-ATTN-NCRF	89.40	74.71	80.82
CySecBERT-BiLSTM-CRF	91.70	75.62	82.41
CySecBERT-BiLSTM-ATTN-CRF	88.80	74.70	80.70
CySecBERT-BiLSTM-ATTN-NCRF	89.23	77.27	82.59

用替换上下文的数据增强方法,且增强比例为 100% 时,模型的 F1 值最高,为 85.47%,因此本文选择该增强方法。为验证模型是否产生过拟合,本文爬取 20 篇 APT 情报文本对其进行预测,同时进行人工验证计算,得出 F1 值为 85.77%,表明模型并未产生过拟合。

Table 8 Experimental results of different data augmentation methods on CySecBERT-BiLSTM-ATTN-NCRF**表 8 不同数据增强方法在 CySecBERT-BiLSTM-ATTN-NCRF 上的实验结果 (%)**

数据增强方法	数据增强比例	准确度	召回率	F1 值
随机删除	0	89.23	77.27	82.59
	50	91.29	75.13	82.02
	100	93.36	76.76	83.75
	200	92.38	77.67	83.97
替换实体	0	89.23	77.27	82.59
	50	90.42	77.64	83.19
	100	91.68	76.53	82.85
	200	92.35	76.49	83.11
替换上下文	0	89.23	77.27	82.59
	50	90.97	75.96	82.27
	100	93.68	79.00	85.47
	200	90.05	77.59	83.13

3.4.2 关系抽取性能

选择文档级关系抽取领域的 GAIN 模型^[19]、SSAN 模型^[20]与 ATLOP 模型进行比较。CySecBERT-ATLOP 模型是在 ATLOP 模型的基础上引入预训练模型进行优化,而本文模型是在引入预训练模型的基础上使用数据增强方法进行优化。各模型关系抽取性能比较结果如表 9 所示。可以看出,在准确度方面,CySecBERT-ATLOP 模型表现最好,达到 88.16%。在召回率和 F1 值方面,本文模型表现最好,分别达到 71.57% 和 77.98%。综合来看,本文模型关系抽取性能最佳。

采用上述替换上下文的数据增强方法,在不同数据增

Table 9 Comparison of relationship extraction performance among various models**表 9 各模型关系抽取性能比较 (%)**

模型名称	准确度	召回率	F1 值
GAIN	75.77	51.40	61.25
SSAN	78.60	52.16	62.71
ATLOP	83.81	51.06	63.46
CySecBERT-ATLOP	88.16	67.70	76.59
本文模型	85.65	71.57	77.98

强比例下对文档级分类模型 CySecBERT-ATLOP 进行实验,结果如表 10 所示。可以看出,当数据增强比例从 0% 到 100% 时,模型的 F1 值逐渐提升,在数据增强比例为 100% 时模型分类效果达到最优,相较于未使用数据增强时 F1 值提升了 1.39%,召回率提升了 3.87%。然而,随着数据增强比例进一步增加到 200% 时,模型产生过拟合现象,模型关系抽取性能不再提高。为验证本文模型增强比例为 100% 时是否产生了过拟合,本文对上述 20 篇 APT 情报文本进行关系预测和人工验证,计算得出其 F1 值为 77.89,表明本文模型并未产生过拟合。

Table 10 Relationship extraction performance of CySecBERT-ATLOP under different enhancement ratios**表 10 不同增强比例下 CySecBERT-ATLOP 的关系抽取性能 (%)**

数据增强比例	准确度	召回率	F1 值
0	88.16	67.70	76.59
25	86.06	69.25	76.74
50	86.57	69.51	77.11
100	85.65	71.57	77.98
200	85.41	70.05	76.97

3.4.3 知识图谱构建

将经过预处理的威胁情报文本段输入到上述训练模型中,并通过实体和关系提取过程生成相应的关系三元组。将三元组插入到 Neo4j 图数据库中,获得的威胁情报知识图谱如图 3 所示。可以看出,相较于其他安全图谱,本文图谱拥有更多的实体和关系,更能直观展现 APT 攻击,为安全人员部署防御提供数据支持。

4 结语

本文提出一种融合实体识别与关系抽取的 APT 攻击知识自动抽取方法,并针对 APT 领域进行了特定优化。该方法结合 APT 攻击的特点设计了 11 种实体类别,构建了融合 CySecBert 与 BiLSTM-NCRF+Attention 的 APT 实体识别模型;同时设计了 9 种关系类别,构建了融合 CySecBert 与 ATLOP 的 APT 实体关系抽取模型,利用替换上下文的数据增强方法对模型进行了优化。实验结果证明,本文模型能有效识别 APT 实体关系三元组,并生成 APT 攻击知识图谱。然而,该图谱在质量和全面性方面仍存在一些不足,未来将尝试应用知识融合技术和知识推理技术进一步完善 APT 知识图谱,并将其应用到攻击事件推理中。

参考文献:

- [1] SIMRAN K, SRIRAM S, VINAYAKUMAR R, et al. Deep learning approach for intelligent named entity recognition of cyber security [C]//Advances in Signal Processing and Intelligent Recognition Systems: 5th International Symposium, 2020: 163-172.
- [2] QIN Y, SHEN G, ZHAO W, et al. A network security entity recognition method based on feature template and CNN-Bilstm-CRF [J]. Frontiers of Information Technology and Electronic Engineering, 2019, 20 (6): 872-884.

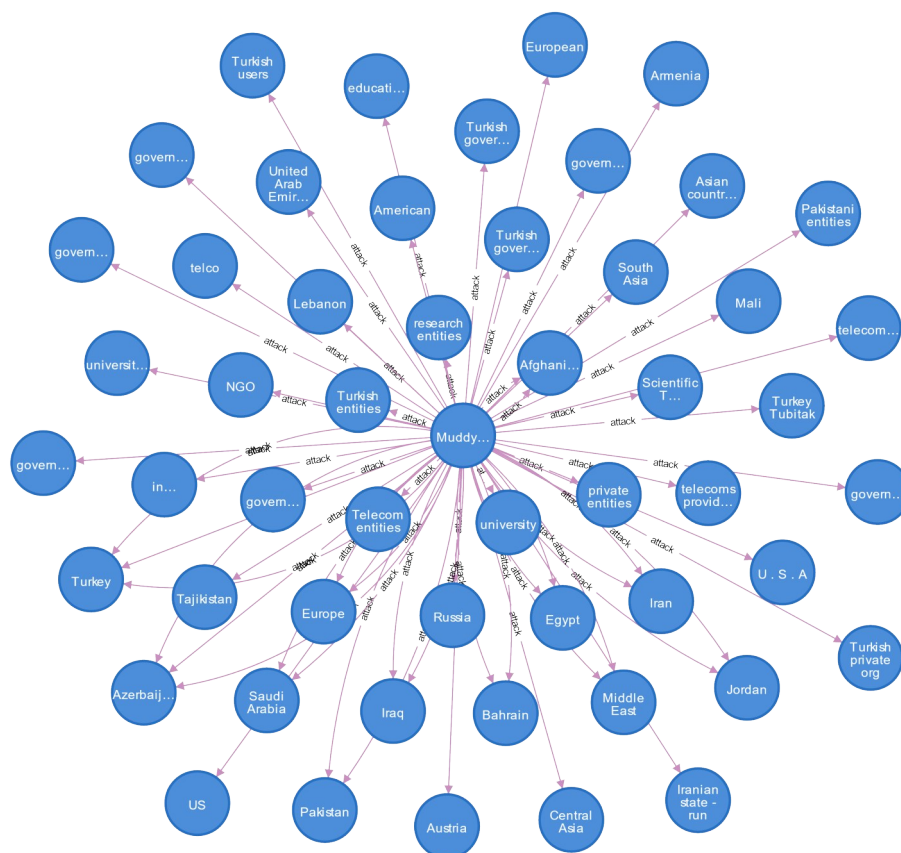


Fig. 3 Threat intelligence knowledge graph

图3 威胁情报知识图谱

- [3] GAO C, ZHANG X, LIU H. Data and knowledge-driven named entity recognition for cyber security[J]. Cybersecurity, 2021, 4(1): 9.
- [4] WANG X, CHEN R, SONG B, et al. A method for extracting unstructured threat intelligence based on dictionary template and reinforcement learning[C]//2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design, 2021: 262-267.
- [5] PINGLE A, PIPLAI A, MITTAL S, et al. Relext: relation extraction using deep learning approaches for cybersecurity knowledge graph improvement [C]//Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2019: 879-886.
- [6] MILAJERDI S M, ESHETE B, GJOMEMO R, et al. Poirot: aligning attack behavior with kernel audit records for cyber threat hunting [C]//Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, 2019: 1795-1812.
- [7] SATVAT K, GJOMEMO R, VENKATAKRISHNAN V N. EXTRACTOR: extracting attack behavior from threat reports [C]//2021 IEEE European Symposium on Security and Privacy (EuroS&P), 2021: 598-615.
- [8] HAN X, PASQUIER T, BATES A, et al. Unicorn: runtime provenance-based detector for advanced persistent threats [DB/OL]. <https://arxiv.org/abs/2001.01525>.
- [9] SARHAN I, SPRUIT M. Open-cykg: an open cyber threat intelligence knowledge graph[J]. Knowledge-based systems, 2021, 233: 107524.
- [10] SUN L, LI Z, XIE L, et al. APTKG: constructing threat intelligence knowledge graph from open-source APT reports based on deep learning [C]//2022 5th International Conference on Data Science and Information Technology (DSIT), 2022: 1-6.
- [11] NARAYANAN S N, GANESAN A, JOSHI K, et al. Early detection of cybersecurity threats using collaborative cognition [C]//2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC), 2018: 354-363.
- [12] KIESLING E, EKEHART A, KURNIAWAN K, et al. The SEPSES knowledge graph: an integrated resource for cybersecurity [C]//The Semantic Web-ISWC 2019: 18th International Semantic Web Conference, 2019: 198-214.
- [13] OpenKG. OpenKG. CN [EB/OL]. [2023-07-28]. <http://openkg.cn/>.
- [14] ZHOU G D, SU J. Named entity recognition using an HMM-based chunk tagger [C]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002: 473-480.
- [15] ZHOU W, HUANG K, MA T, et al. Document-level relation extraction with adaptive thresholding and localized context pooling [C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2021: 14612-14620.
- [16] DING B, LIU L, BING L, et al. DAGA: data augmentation with a generation approach for low-resource tagging tasks [DB/OL]. <https://arxiv.org/abs/2011.01549>.
- [17] ZHOU R, LI X, HE R, et al. MELM: data augmentation with masked entity language modeling for low-resource NER [C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022: 2251-2262.
- [18] DAI X, ADEL H. An analysis of simple data augmentation for named entity recognition [J]. <https://arxiv.org/abs/2010.11683>.
- [19] ZENG S, XU R, CHANG B, et al. Double graph based reasoning for document-level relation extraction [DB/OL]. <https://arxiv.org/abs/2009.13752>.
- [20] XU B, WANG Q, LYU Y, et al. Entity structure within and throughout: modeling mention dependencies for document-level relation extraction [C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2021: 14149-14157.

(责任编辑:尹晨茹)