恶意软件知识图谱的构建与研究

罗养霞[1] ， 李 浩，武晨明

（西安财经大学 信息学院 陕西西安 710100）

摘 要：近年来，知识图谱在恶意软件领域应用广泛，但是多数学者着重于构建恶意软件 API 知识图谱，利用知识图谱去检测恶意代码，而利用 API 知识图谱解释性较弱，专业性较高。针对上述问题，本文提出通过 NER 模型去抽取恶意软件名称、发现地等文本实体信息，以此构建恶意软件知识图谱，并通过知识图谱发现其多样性、演化路径、威胁方式与分类关联等。本文首先研究了恶意软件知识图谱构建方法，完成数据预处理、模式层构建与数据层构建。其次对恶意软件结构化与半结构化数据进行实体标识与规范化，完成本体构建（实体、关联与附加属性），通过模式层指导数据层的方法，利于 BERT-BiLSTM-CRF 模型进行知识抽取，最后，利用 Neo4j 图数据库对知识图谱进行存储与可视化。同时利用病毒库数据对所建模型进行仿真验证，实验结果表明此模型相比同类模型效果更好，性能指标更优异，对推进网络安全知识简易化和防御体系知识普及化具有重要意义。

关键词：知识图谱；恶意软件；知识抽取

中图分类号：TP309 **文献标识码**： A **文章编号**：

Construction and Research of Malware Knowledge Graph

LUO Yang-xia， LI Hao，WU Chen-ming

(Department of School of Information, Xi'an University of Finance and Economics, Xi'an 710100, Shaanxi)

Abstract： In recent years, knowledge mapping has been concerned by many scholars, and it has been widely used in the field of malware. Most of them focus on constructing knowledge mapping of malware API and using knowledge mapping to detect malicious code. Different from the above, this paper proposes that the knowledge map of malware can analyze the text information such as the name, the place of discovery, and the features of the malware through the NER task, to discover its diversity, evolution path, Threat Mode and classification association, and to complete the construction of Data pre-processing layer, pattern layer and data layer by studying the construction method of malware knowledge map. Specifically, entity identification and normalization are carried out on structured and semi-structured data of malicious software, and ontology construction (entities, associations, and additional attributes) is completed. The data layer is guided by the pattern layer method, and knowledge extraction, fusion, and storage are based on BERT BiLSTM-CRF to construct a knowledge graph. Finally, use the Neo4j graph database to highly visualize and reproduce the knowledge graph. The simulation verification of graph construction using virus library data shows that this model has better performance and superior performance indicators compared to similar models. It is of great significance to promote the construction of intelligent detection and defense systems for network security.

Key words： knowledge graph; malware; knowledge extraction

收稿日期：2024-3-XX

[1]**基金项目：**国家自然科学基金(No. 62372373；No. 61972314)；陕西省重点研发计划项目资助（项目号：2024GX-YBXM-545）；西安财经大学 2023 年研究生创新基金项目(23YC033)； 2024 年国家级大学生创新训练项目（No. S202411560032）.

通信地址：710000 陕西省西安市长安区常宁大街 360 号西安财经大学信息学院

Address: School of Information Technology, Xi 'an University of Finance and Economics, 360 Changning Street, Changan District, Xi 'an, Shaanxi 710000, China

1 引言

随着互联网的普及，恶意软件的数量和复杂性持续攀升。根据 AV-TEST 的最新数据[1]显示， 到 2023 年恶意软件数量已增加至 12.3 亿个，不仅数量上呈指数型上涨，形式多以及变种的能力也越来越强。同时，随着多态病毒引擎与代码混淆等技术的发展，使得识别和揭示恶意软件分类与多样性、病毒演化路径与关联、特征与家族变种等更为困难，对网络安全、信息系统造成严重威胁。

构建恶意软件以及病毒变种之间的关系，使之形成一个“知识库”，可以方便掌握各恶意软件以及种类之间的关联，通过病毒分类、发现时间、变体形式等等去归类，并且在该家族节点下，描述特征属性、应对方法、删除指令、指导网络安全员在发现后第一时间去查询、定位或消灭。

知识图谱[2]是一种多边关系图，它由不同实体之间的连接（边）以及这些实体的节点组成。这种图谱依赖于各类相关数据库，并利用 NLP 等相关技术对数据资源进行实体关联和抽取。其通过 Neo4j 图数据库提供可视化和查询功能。Neo4j 将结构化数据存储在网络上，是高性能的 NoSQL（Not only SQL）图形数据库，是目前知识图谱领域内应用较为广泛的图数据库之一。研究应用知识图谱的构建便于在知识图谱上开展下游任务的延伸，结合人工智能和机器学习，构建相关联的防御体系和提高智能化的病毒检测。

目前在知识图谱的构建和优化中已有不少研究。史慧洋[3]等人提出由情报搜索、信息抽取、本体构建和知识推理构建威胁情报的知识图谱构建框架，用于发现攻击者的威胁情报；刘善玲[4]提出了基于知识图谱的深层次域名检测方法，从域名、IP 地址、地理位置、解析记录等多个维度抽提构建知识图谱所需的实体信息。Dutta[5]等人通过手工注释的 RDF 三元组建立了威胁情报知识图谱，收集了 83 份威胁情报非结构化的数据，构建的知识图谱适用范围较小。当前，已经出现 Cyc、Dbpedia 等依赖专家系统构建的知识图谱以及搜狗知立方等中文知识图谱 [6]。除此之外 Duoyuan Ma[7]、马铎原[8]等人构建了 Android API 知识图谱，Lianqiu Xu[9]、朱朝阳[10]等人构建了恶意软件行为知识图谱。Aritran Piplai[11]等人利用公开网络安全知识图谱进行了恶意软件分析。总结以上研究，目前大多是针对恶意代码、API、恶意软件行为等信息进行恶意软件分析，并未对恶意软件族群及病毒分类、构建图谱等方面进行探索。

针对目前恶意软件命名复杂，恶意软件记录较分散以及受到恶意攻击后无法系统性的解决提示等问题，本文提出一种基于 BERT-BiLSTM-CRF 模型构建恶意软件知识图谱的方法。

其过程概述是：首先从公开数据集获取恶意软件相关的非结构化文本资料，通过 BERT-BiLSTM-CRF 模型处理为结构化标注文本；其次，按照自顶向下-模式层指导数据层的方法构建恶意软件知识图谱和恶意软件处理方法知识图谱；随后，所构建的两个知识图谱被存储于 Neo4j 图数据库中，并进行了可视化的展示。最后将模型与同类方法进行评价分析。

研究的创新点在于：1.本文利用恶意软件知识库进行恶意软件知识图谱的构建，不同其他学者研究，本文构建的实体为恶意软件名称、发源地、发现时间、相关依赖等信息作为实体，并且将病毒特征、删除指令等详细描述病毒信息文本作为实体属性。2.针对恶意软件知识库，本文提出了计算机安全领域 NER 模型：BERT-BILSTM-CRF，通过训练该模型进行恶意软件实体抽取任务；3.本文通过对知识图谱存储更新可以方便安全相关人员进行恶意软件分析以及对其进行归类。

以下详细介绍知识图谱构建方法、模型实现过程，以及评分与对比。

2 恶意软件知识图谱构建方法

本文首先进行数据的预处理，进行本体数据和描述数据的定义。其次进行模式层的构建，定义实体概念以及实体关系。最后进行数据层的构建，形成恶意软件知识图谱。

2.1 知识图谱构建框架

恶意软件知识图谱构建方法包括 3 种：自顶向下、自底向上和两者想结合的方式[9]。本文采取自顶向下的方法构建知识图谱，首先构建模式层，然后再基于模式层从文本数据中抽取实体构建相应的数据层，如图 1 所示。本文采取自顶向下的方法构建知识图谱，首先构建模式层，然后再基于模式层从文本数据中抽取实体构建相应的数据层，本文分别利用本体数据和特性描述数据，构建本体知识图谱和处置知识图谱，攻击发生后，安全人员可通过本体属性查询，锁定攻击的方式和病毒、恶意软件的名称，实现快速定位。然后通过本体知识图谱和知识图谱之间的联系，可以快速的找到处置方案以及了解病毒威胁等级及关联。其中本

体数据即实体数据包含实体的自身信息如：名称、发现世界、发现地点等。处置数据包含实体的附加信息如：解决办法、删除指令等描述恶意软件附加信息。其中本体数据为非结构化数据，处置数据为半结构化数据。

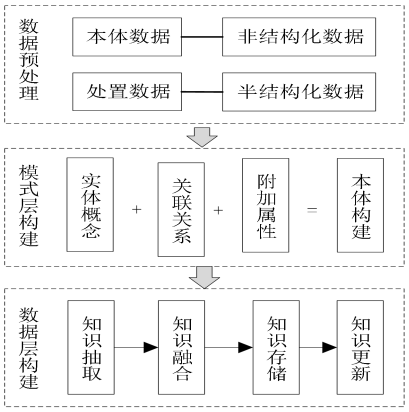

Figure 1 Malware knowledge map construction process
图 1 恶意软件知识图谱构建流程

2.2 数据预处理

恶意软件相关数据可以分为经过安全中心处理后的半结构化数据和公开的未经处理的非结构化数据。针对恶意软件名称、编号、特征等本体信息，因为其数量较多且多为非结构化数据，本文提出病毒领域 NER 模型，对识别实体进行知识加工并整理为三元组。针对发生攻击后采取措施和解决办法等结构化数据，本文采取“机器＋人工”方式进行人工标注，编写规则采用“关键词+文本”形式。数据分类如表 1。

Table 1 Two-source data
表 1 二源数据

数据名称	数据类型	识别内容
本体数据	非结构化数据	名称、别名、时间、地点等
特性描述数据	半结构化数据	解决办法、危险等级等

2.3 模式层构建

模式层构建的目标是定义恶意软件领域内三元组内部直接实体以及属性的关系[12]。通过对数据的实体概念、关联关系和属性的表示抽取，形成其对应的本体规则，用于构建知识图谱的模式层。

2.3.1 本体数据模式层

通过对获取的数据进行分析，得到的实体包括：恶意软件名称、恶意软件别名、特征、首次出现时间、首次出现地点等，根据以上信息建立拓扑关系，抽象出部分实体之间的关联关系，如表 2 所示。

Table 2 Ontology and relationship
表 2 本体及关系

实体 1	关联关系	实体 2
名称	直连	别名
名称	直连	特征
特征	包含	时间
特征	包含	地点
环境	显示	设备类型

2.3.2 特性描述数据模式层

对恶意软件特性描述数据进行分析，得到的实体有：社会影响、威胁等级、攻击特性、运行环境、传播方式、防御措施等等。再对各实体进行关联关系的分析，抽象出实体间的关联关系，部分安全处置实体和实体间的关联关系，如表 3 所示。

Table 3 Features describe data ontologies and relationships
表 3 特性描述数据本体及关系

实体 1	关联关系	实体 3
运行环境	下连	解决办法
威胁等级	下联	波动指标
攻击特性	下联	防御措施
防御措施	包含	解决办法
传播方式	影响	防御措施

2.4 数据层构建

数据层的构建主要分为知识抽取、知识融合、知识存储与知识更新 4 个步骤。以下详细说明。

2.4.1 知识抽取

知识抽取主要是在知识层的指导下，利用实体识别模型，从非结构化数据抽取出命名实体并整理为结构化数据。本文针对两种不同的结构特征数据采用不同的方法进行知识抽取。

（1）基于 BERT-BiLSTM-CRF 模型的知识抽取

本文所研究的恶意软件本体数据为非结构化数据，构建 BERT-BiLSTM-CRF 实体识别模型，对其进行知识抽取。通过分析数据的结构特征可知，设定的实体类型有：恶意软件/病毒名称、别名、设备类型、时间、地点、环境，如表 4 所示，抽取模型和算法过程见 3 部分。

Table 4 Examples of malware entities
表 4 恶意软件实体示例

实体类别	实体示例
名称	火焰病毒
别名	Flamer、Skywiper
时间	2012 年 5 月
地点	中东
设备	Windows 操作系统
环境	以 Lua 和 C++语言写成

（2）基于多类协同标注的知识抽取

本文所研究恶意软件特性描述数据多为半结构化数据，利用标注平台+人工标注的方式进行知识抽取，识别的实体为：威胁等级、攻击特性、防御措施、传播方式、社会影响，如表 5 所示。

Table 5 Examples of flame virus entities

表 5 火焰病毒实体示例

实体类型	实体示例
威胁等级	* * * *
攻击特性	此类病毒能自主监测并解析自身的网络传输模式，同步实现自动录音功能，记录用户的密码输入行为和键盘按键节奏，并将收集到的数据连同其他关键文件一道，秘密传输至远程控制病毒的服务器端。在完成数据收集使命之后，这类恶意软件还能自我销毁，确保行动痕迹彻底清除。
防御措施	及时更新杀毒软件、不点击下载未了解的邮件链接。
传播方式	通过钓鱼邮件诱骗其点击链接然后进行秘密安装。
社会影响	被认为以军事为目的的攻击，对中东计算机造成大量危害。

2.4.2 知识融合

知识融合指实体链接和知识合并[16-18]，解决不同实体名表示同一实体的问题。本研究在一个统一结构下对源自不同恶意软件类别的异构、多样的知识进行梳理、整合与对应映射。以此达到数据、信息等不同角度的融合，现有研究针对异质性问题，提出了多元化的解决方案，主要可归纳为两个方向：实体整合与实体映射策略。实体整合旨在将源自多个异构数据源的独立实体体系汇聚为一个统一的整体，实现知识结构的标准化与一致性构建。而实体映射则是通过设立一套规则体系，以促进不同实体之间的信息对接和相互转化。包括但不限于选择合适的整合策略、起点的源本体，以及实际执行整合操作。

在本文中，通过计算实体之间文本相似度来进行判断是否应该进行融合。在数据融合的过程中，知识整合的关键层面涉及实体配对和实体合并两大技术手段。实体配对是多源知识深度融合不可或缺的组成部分，其主要作用在于解决由于数据来源多样化引起的实体指向不协调与冲突矛盾问题，旨在通过精准的实体匹配，确保各源头实体信息的一致性和连贯性。本文对抽取的实体采用 Cosine 相似度算法在词典中进行匹配，得到与抽取实体相似度最高的目标实体，避免在知识图谱构建中重复节点造成不同实体对应相同属性问题。计算流程图如图 2。

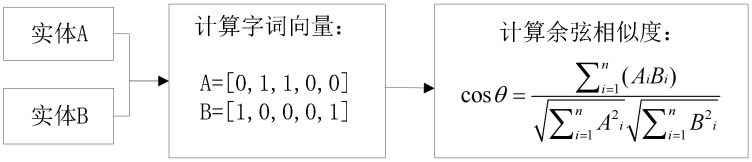

Figure 2 Calculation process of text similarity

图 2 文本相似度计算流程

2.4.3 知识存储

知识存储是利用图数据库等手段将建立的知识图谱可以清晰的展示与保存,本文采用 Neo4j 图数据库技术进行知识图谱的保存。在知识图谱内，信息以高度结构化的形式储存于庞大的知识库存中[19]。这些知识按照类别划分，遵循规范标准，分别储存在知识库存内各自独立的模块单元，有利于深化知识发掘。本研究运用了 Neo4j 图数据库技术以承载这些知识，利用图论为基础的搜索算法，并借助直观的 Cypher 查询语言，使得用户无需手动编写复杂的遍历图结构代码，即可轻松实现对恶意软件相关图形化数据的高效存储和便捷查询服务。如图 3 Neo4j 图数据库构建两个实体范例图。

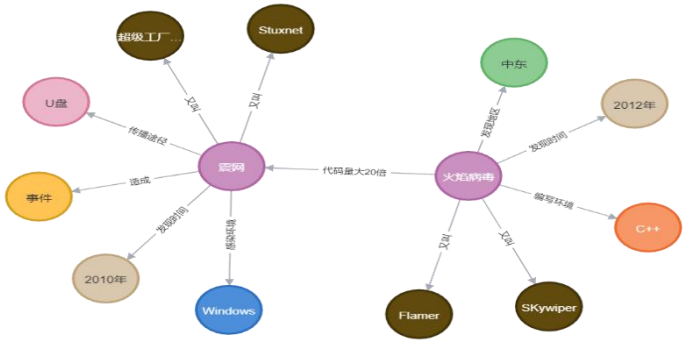

Figure 3 Sample Neo4j diagram of two malware nodes

图 3 两个恶意软件节点 Neo4j 范例图

2.4.4 知识更新

在知识图谱构建完成后，随着恶意软

件和病毒不断的增加，对已构建好的知识图谱实施必要的更新，主要包括模式结构更新与数据内容更新两大部分。模式层的更新是通过增加实体和关系结构进行完善；数据层的更新，是随着实体和关系数据的不断增多，定期增加新出现的新型实体。本文在研究中，已定义了恶意软件模式层，后期只需按照模式层的指导添加新增实体即可。本文构建和输出的恶意软件知识图谱 Neo4j 可视化展示如图 4 所示。

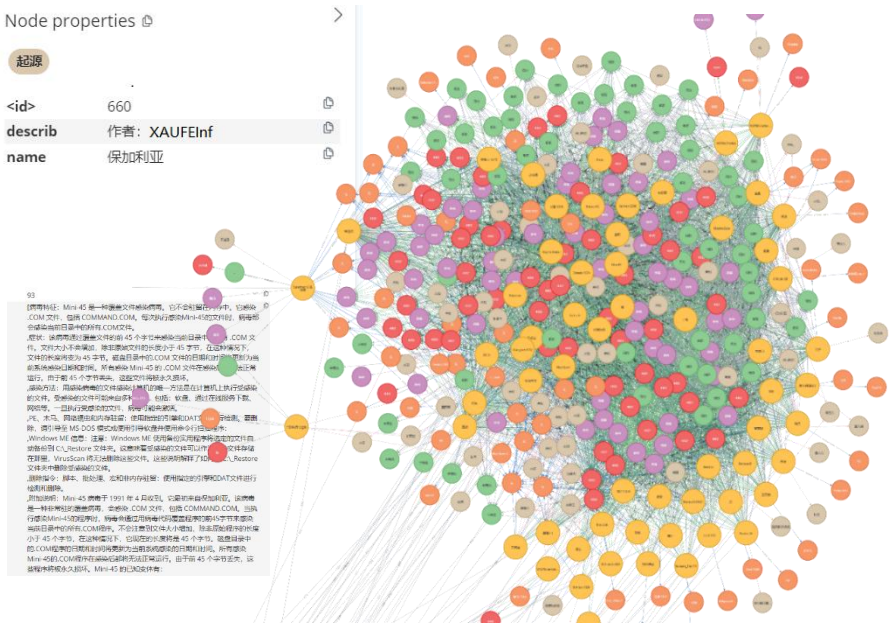

Figure 4 Neo4j visual display of knowledge graph construction
图 4 知识图谱构建 Neo4j 可视化展示图

3 基于 BERT-BiLSTM-CRF 模型的知识抽取

在知识图谱数据获取阶段，大量的文本数据为非结构化文本。故本文提出安全领域的 NER 模型，利用此模型将非结构数据转换为结构化数据。将文本信息转换为相应的词向量嵌入到模型中进行训练，最后输出文本信息中需要的实体名称。本文所构建的模型如图 5。

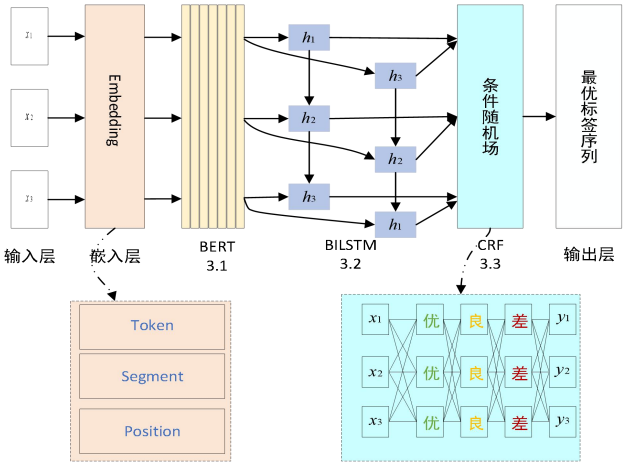

Figure 5 Model frame diagram
图 5 模型框架图

3.1BERT 模型

BERT 模型架构基于双向 Transformer 编码器构建，具备同时参考输入序列前后

的信息能力，这有助于模型高效捕获文本的上下文信息。在预训练阶段，BERT 采用了“遮蔽语言模型”策略，随机隐藏约15%的句子词汇，然后要求模型依据周围环境推测被遮蔽的词语。由此，BERT 模型不仅展现出强大的语义理解与抽取技能，还在实体关系识别上表现出色，尤其擅长解决词汇的多义性问题。BERT 模型生成的词嵌入涵盖三个方面：字符级词嵌入（Token Embeddings）、段落标识嵌入（Segment Embeddings）以及位置嵌入（Position Embeddings）。在这一过程中，首先运用字向量技术，将恶意软件相关的字符序列映射至一个 768 维的词空间，从而实现对恶意软件字符的向量化表达。对于句向量部分，其核心功能在于捕捉并编码句子层面的语义特性，特别是在 Segment Embeddings 层面上，系统采用二元标识法进行处理，即首个句子的所有字符对应的词向量均使用全零向量表示，而第二个句子则以全 1 向量来标识其内部字符的词向量，以此区分不同句子间的语义边界。至于位置向量，则着重于凸显同一句子内不同位置字符的特定语义上下文信息。借助 BERT 的预训练模型架构，能够有效地从丰富的语言环境中抽取出具有深度语义含义的特征，并最终将恶意软件相关文本转化为富含语义信息的序列向量表示，模型图如图 6 所示。

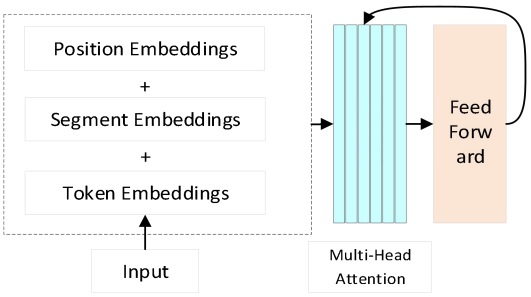

Figure 6 BERT model diagram
图 6 BERT 模型图

3.2 双向长短时间模型（BiLSTM）

为应对循环神经网络训练期间出现的梯度消失和梯度爆炸问题，Hochreiter 等人提出了长短期记忆（LSTM）结构作为解决方案。该结构创新性地采用了门控设计（Gating Mechanism），通过三个关键的调控组件——输入门、遗忘门以及输出门来智能管理信息流的流动与保留。

在 LSTM 中，输入门负责决定在当前时间步应该让多少新的候选状态信息得以保留并整合到单元状态中。另一方面，遗忘门承担着筛选任务，它根据需要选择丢弃上一时间步单元状态中的哪些历史信息。最后，输出门扮演着过滤器的角色，确定当前时间步单元状态中有多少信息应当被输出至网络的下一阶段作为有用信号。

这三个门控组件均采用“软”阈值形式运作，而非硬性的开关，这样能够在连续处理时序数据的过程中更加平滑和精细地调节信息的存储与传播，从而有效解决传统循环神经网络中存在的长期依赖学习难题。即按照一定的概率允许信息通过，计算公式为：

$$i_t = \sigma(W_i X_t + U_i h_{t-1} + b_i) \quad (1)$$

$$f_t = \sigma(W_f X_t + U_f h_{t-1} + b_f) \quad (2)$$

$$o_t = \sigma(W_o X_t + U_o h_{t-1} + b_o) \quad (3)$$

其中 $\sigma(\bullet)$ 为 Logistic 函数，其输出区间（0，1），x_t 为当前的环节的输入，h_{t-1} 为上一环节的外部状态。

LSTM 引入新的内部状态进行线性循环信息传递，其内部状态 c_t 公式为：

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (4)$$

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

其中 $\odot$ 为向量元素乘积，$\tilde{c}_t$ 为候选状态，c_{t-1} 为上一时刻的记忆单元。

记忆单元 c 中的保存信息的周期长于循环神经网络中隐状态存储的历史信息，然而，相较于神经网络中普遍存在的长期记忆能力，其存储容量有限，故此被称为长短时记忆网络。

本文采用双向长短期记忆网络，拼接两个独立的 LSTM，一个用于处理恶意软件数据正序输入，另一个用于处理恶意软件数据逆序输入，分别提取正序和逆序的特征，拼接后形成的词向量作为该次的最终恶意软件图谱特征表达。

3.3 条件随机场（CRF）

条件随机场（Conditional Random Field, CRF）是无向图判别式模型[20]，主要包括特征集合和需要标注的序列。研究通过对标注序列应用特征函数进行概率评估，将集合内所有特征函数针对同一标注序列给出的评分加以整合计算。作为最终的概率值，如公式（6）所示。

$$P(Y|X)=\frac{1}{z}\exp\sum_{t=1}^{T}F(y_{t-1},y_t,x_{1:T})$$
$$=\frac{1}{z}\exp\sum_{t=1}^{T}\left[\sum_{k=1}^{K}\lambda_k f_k(y_{t-1},y_t,x_{1:T})+\sum_{l=1}^{L}\eta_1 g_l(y_t,x_{1:t})\right] \quad (6)$$

已知观测序列 X，求 Y 状态序列发生的概率，其中 f_k 为转移特征函数，衡量相邻状态变量之间的影响，g_l 为状态特征函数，衡量状态序列对观测序列的影响。λ_k 和 η_l 为相关参数。Exp 为指数势函数，定义图模型中的概率分布函数，z 为规范化因子，确保此式结果为概率。

在本文中 BiLSTM 的输出向量集作为 CRF 的观测序列，而状态序列即输出标签为{B,I,O}。其中 B 为实体开始，I 为实体结束，O 为非实体，CRF 层学习标签相关信息，进行标签信息约束。例如 I-TIME 不会继续再接 B-TIME。通过运用临近标签信息以获得最优预测序列，从而弥补了 BiLSTM 模型在处理恶意软件词向量间相邻标签依赖关系方面的不足。

4 实验与对比

为了验证本文构建的模型的有效性，进行实验仿真，实验环境如表 6 所示。

Table 6 Operating equipment and environment
表 6 运行设备及环境

设备名称	环境
GPU	V100
CPU	Xeon(R)
环境	Python3.7
框架	TF1.5

本文构建的 BERT-BiLSTM-CRF 命名实体识别模型部分训练超参数如表 7 所示。

Table 7 Model training hyperparameters
表 7 模型训练超参数

超参数名称	值
max_seq_length	203
batch_size	128
learning_rate	0.00001
dropout_rate	0.5
train epochs	100
optimizer	Adam
clip	0.5
lstm_size	64
Num_hidden_layers	12
hidden_size	768

实验中应用的病毒数据来源网站（https://web.archive. org/），数量 53000 多条，一条病毒信息包含 30 句左右描述信息。包括病毒名称、病毒别名、发现日期、删除指令等字段。采用“机器+人工”结合方式对文本数据以字符为最小单位进行 BIO 词性标注，对已标注的数据集实施随机分割，遵循 7:2:1 的比例，将数据集划分为训练、验证和测试三个子集，旨在对模型进行训练和性能评估。

4.1 评价指标

本文选用命名实体识别领域内通用评估方法对本文构建的 BERT-BiLSTM-CRF 模型进行评估，评估指标为精确率 $P_{recision}$、召回率 R_{ecall} 以及 F_{1_score}。

$$P_{recision}=\frac{T_P}{T_P+F_P} \quad (7)$$

$$R_{ecall}=\frac{T_P}{T_P+F_N} \quad (8)$$

$$F_{1\_score}=\frac{2P_{recision}R_{ecall}}{P_{recision}+R_{ecall}} \quad (9)$$

其中，正类样本的正确预测记为 T_P，负类样本的正确预测表示为 T_N，F_P 指的是实际为负类却被误判为正类的样本，F_N 代表实际情况为正类但不幸被预测为负类的样本。

4.2 实验结果分析

本文分别进行 BERT-BiLSTM-CRF 模型、BERT-CRF 模型、BiLSTM-CRF 模型进行对比实验，以此来验证本文所构建模型的效果。模型训练对比结果如图 7 (a)-(d)所示。与其他文献研究实验对比如表 8 所示。

对于相同病毒数据库，本文构建的模型性能都强于同类模型，经过训练后模型的 $P_{recision}$ 达到 94.13%，模型的 R_{ecall} 达到 95.46%，模型的 F_{1_score} 达到 94.79%。其中 BERT-BiLSTM-CRF 模型在各项指标都略高于 BERT-CRF 模型，而另外两个模型结果指标远高于 BiLSTM-CRF 模型，此结果也充分说明数据经过 BERT 模型预处理后充分提取了语料中句子的特征信息，进而提高了实体的识别准确率。BERT-BiLSTM-CRF 模型相比较 BERT-CRF 模型在 F1 值提升 2.46%，Precision 值提升 3.19%，Recall 值提升 1.71%。相比较 BiLSTM-CRF 模型在 F1 值提升 21.83%, Precision 值提升 19.64%，Recall 值提升 24.03。结果表明，在对于病毒类命名实体识别任务中，本文所采用的模型性能最佳。

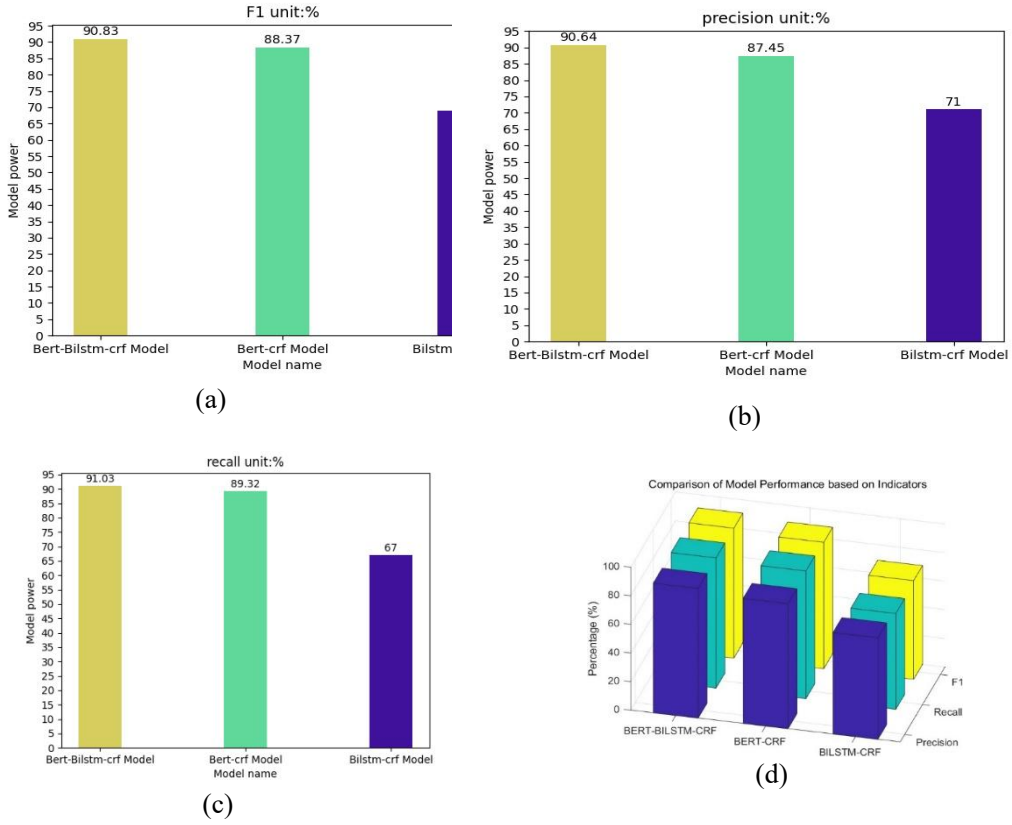

Figure 7 (a) Comparison results of F1 experiment (b) Comparison results of Precision experiment (c) Recall experimental results. (d) Recall model results

图 7 (a) F1 实验对比结果 (b) Precision 实验对比结果 (c) Recall 实验对比结果 (d) 模型结果对比图

Table 8 comparison of results with other relevant experiments

表 8 与其他相关实验结果对比表

文献及年份	关键技术	准确率	召回率	F1 分数
杨秀璋等人[22] 2022	Bert+BiLSTM-CRF+APT 实体	0.78	0.58	0.67
李涛等人[23] 2020	注意力机制+BiLSTM-CRF 模型	0.67	0.63	0.64
Ren[24] 2022	APT 图模型+APT 知识提取	0.81	0.82	0.81
本文模型	实体知识图谱构建+ BERT-BiLSTM-CRF 模型	0.90	0.91	0.90

5 结束语

本文研究并构建了恶意软件知识图谱，本文所作的主要工作流程如下:1.通过数据获取，从开源网站获取数据后进行数据清洗标注工作，本文利用两种标注方法对数据集进行标注。2.本文利用该数据集训练本文所提出的模型。3.利用模型输出实体构建恶意软件知识图谱。4.将所构建三元组利用 Neo4j 图数据库进行保存与可视化。5.进行了模型的仿真与实验对比。

下一阶段将研究与探索如何利用知识图谱延伸下一步的实践应用，使图谱构建结合人工智能和机器学习，提高智能化的病毒检测和防御系统。随着物联网、云计算和大数据等技术的快速发展，结合图谱构建与更新，设计和开发更加高效、可扩展的病毒防御体系。其次，随着区块链技术的逐渐成熟和应用，探索将图谱关联性与病毒防御相结合，构建去中心化的安全网络，通过区块链的分布式特性和图谱关联特征，有效防止混淆病毒、变异传播等，从而提高网络的安全性和可信度。

参考文献：

[1] Malware Statistics & Trends Report | AV-TEST[EB/OL]. [2023-12-26]. https://www.av-test.org/en/statistics/malware/.

[2] SinghalA.IntroducingtheKnowledgegraph:Things,notstrings [EB/OL].[20240119]. https://blog.google/products/search/introducing-knowledge-graph-things-not/ .

[3] SHI huiyang, WEI jinghuan, CAI xingye,et al. Research on Threat Intelligence Extraction and Knowledge Graph Construction Technology [J]. Journal of Xidian University,

2023,50（04）：65-75.
[4] Liu Shanling. Malicious domain name detection based on Knowledge graph in the context of big Data [D]. Jiangsu: Nanjing University of Posts and Telecommunications, 2022.
[5] Sharmishtha Dutta, Nidhi Rastogi, Destin Yee, Chuqiao Gu, et al.Malware Knowledge Graph Generation[J]. Rensselaer Polytechnic Institute.2021.56,66-72.
[6] Zhao Yer Hui, Liu Lin, Wang Hailong and others. Research on Knowledge graph recommendation system [J]. Exploration of Computer Science and Technology, 2023, 17(4):771-791.
[7] Ma, Duoyuan, Y Bai, L Sun,et al. "A knowledge graph-based sensitive feature selection for android malware classification." 2020 27th Asia-Pacific Software Engineering Conference (APSEC). IEEE, 2020.
[8] Matyoyuan. Research on Android Malware family classification based on Knowledge Graph [D]. Tianjin University, 2023.
[9] Xu L, Zhang C, Tang K. A malware analysis method based on behavioral knowledge graph[C]//International Conference on Electronic Information Engineering and Computer Science (EIECS 2022). SPIE, 2023, 12602: 571-579.
[10] ZHU Zhaoyang, Zhou Liang, Zhu Yayun et al. Visual classification algorithm of malicious code based on behavior graph screen. Information Network Security 21.10(2021):54-62.
[11] Piplai, Aritran, et al. "Creating cybersecurity knowledge graphs from malware after action reports." IEEE Access 8 (2020): 211691-211703.
[12] DEVLIN J，CHANG Mingwei，LEE K，et al．BERT：pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: Association for Computational Linguistics，2018：4171-4186．
[13] ZENG D, SUN C, LIN L, et al. LSTM-CRF for drug-named entity recognition [J]. Entropy, 2017, 19(6): 283.
[14] Li Sizhen. Research and implementation of ontology-based industry Knowledge graph construction technology [D]. Beijing: Beijing University of Posts and Telecommunications, 2019.
[15] GUO Rong ， YANG Qun ， LIU Sha ohan ， et al ． Construction and application of power grid fault handing knowledge graph[J]． Power System Technology，2021，45(6)：2092-2100．
[16] Liu Jiao, Li Yang, Duan Hong et al. Overview of knowledge graph construction technology [J]. Computer Research and Development, 2016,53 (3) : 582-600.
[17] Zhang Jixiang, Zhang Xiangsen, Wu Changxu, et al. Overview of knowledge graph construction technology [J]. Computer Engineering, 2022, 48 (3) : 23-37.
[18] Wang Haofen, Qi Guilin, Chen Huajun. Knowledge Graph: Method, Practice and Application [M]. Beijing: Publishing House of Electronics Industry, 2019.
[19] Cao Qian, Zhao Yiming. Technology realization process and related application of Knowledge graph [J]. Information Theory and Practice. 2015.12.026.
[20] Xie Teng, Yang Junan, Liu Hui. Chinese entity recognition based on BERT-BiLSTM-CRF model [J]. Journal of Computer Systems and Applications, 2019,29(7): 48-55.
[21] Ye Xinzhi, Shang Lei, Dong Xuzhu, et al. Research and application of Knowledge Graph for Fault Handling of Distribution Network [J]. Power Grid Technology, 2022,46 (10) :3739-3749.
[22] Yang Xiuzhang, Peng Guojun, Li Zichuan, et al. Research on APT attack entity recognition and alignment based on Bert and BiLSTM-CRF [J]. Journal of Communications, 2022,43(06):58-70.
[23] Li Tao, Guo Erao, Ju Ankang. Triplet extraction of network security knowledge based on fusion adversarial active Learning [J]. Journal of Communications,20,41(10):80-91.
[24] Ren, Yitong, et al. "Cskg4apt: A cybersecurity knowledge graph for advanced persistent threat organization attribution." IEEE Transactions on Knowledge and Data Engineering 35.6 (2022): 5695-5709.

附中文参考文献：

[4] 刘善玲. 大数据背景下基于知识图谱的恶意域名检测[D]: 南京邮电大学, 2022.
[6] 赵晔辉, 柳林, 王海龙等. 知识图谱推荐系统研究综述[J]. 计算机科学与探索. 2023, 17(4):771-791.
[8] 马铎原. 基于知识图谱的 Android 恶意软件家族分类研究[D]. 天津大学, 2023.
[10] 朱朝阳, 周亮, 朱亚运等. 基于行为图谱筛的恶意代码可视化分类算法. 信息网络安全 21.10(2021):54-62.
情报理论与实践. 2015.12.026.
[14] 李思珍. 基于本体的行业知识图谱构建技术的研究与实现[D].北京:北京邮电大学,2019.
[16] 刘峤,李杨,段宏等.知识图谱构建技术综述[J].计算机研究与发展,2016,53(3):582-600
[17] 张吉祥, 张祥森, 武长旭等. 知识图谱构建技术综述[J]. 计算机工程, 2022, 48(3):23-37.
[18] 王昊奋, 漆桂林, 陈华钧. 知识图谱:方法,实践与应用[M].北京:电子工业出版社 2019.
[19] 曹倩, 赵一鸣. 知识图谱的技术实现流程及相关应用[J].
[20] 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.
[21] 叶欣智, 尚磊, 董旭柱等.面向配电网故障处置的知识图谱研究与应用[J].电网技术,2022,46(10):3739-3749.
[22] 杨秀璋,彭国军,李子川,等.基于 Bert 和 BiLSTM-CRF 的 APT 攻击实体识别及对齐研究[J].通信学报,2022,43(06):58-70.
[23] 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.

作者简介：

罗养霞 (1974-)， 女，陕西西安人，博士，研究生导师，研究方向为数据科学与智能计算。Email:yxluo8836@163.com.

LUO Yang-xia(1974-), Female, Xi'an, Shaanxi Province, Ph. D. graduate supervisor, Her research interest includes data science and intelligent computing.

李浩（2001-），男，陕西榆林人，硕士研究生，主要研究领域为知识图谱。

Email:wasd09042001@163.com.

LI Hao (2001-), male, born in Yulin City, Shaanxi province, master candidate. His research interests mainly cover domain knowledge mapping.

武晨明（2004-），男，陕西西安人，学士，主要研究领域为知识图谱构建与应用。

Email:hoshino.isumi.personal@gmail.com.

Wu Chen-ming (2004-), male, Xi'an, Shaanxi, China, B.S. His main research area is construction and application of knowledge mapping.