基于 Stanza 模型的非结构化病毒知识图谱构建

罗养霞[1,2+], 武晨明[1], 王佳怡[1], 丁涵[1]
1.西安财经大学 信息学院,陕西省 西安市 710100
2.计算机应用与信息安全研究中心，西安 710100
+ 通信作者 E-mail：yxluo8836@163.com

摘　要：本研究设计并完善了一种基于自然语言处理工具包 Stanza（Stanford NLP）的知识图谱构建方法，主要用于从非结构化文本中提取病毒名称、传播途径、危害程度等信息，并结合图数据库 Neo4j 实现多源异构数据的融合与关联。主要处理流程有：通过规则驱动的文本清洗与标准化对数据进行预处理；利用 Stanza 进行命名实体识别和依存句法分析，精准识别病毒相关实体及其复杂关系；将提取的三元组导入图数据库，实现快速检索和可视化分析。最后分析实验结果可得，该方法在准确性和覆盖率方面具有较好的性能，显著提升了威胁情报的分析效率。

关键词：病毒知识图谱；自然语言处理；Stanza；实体关系抽取；Neo4j

文献标志码：A　　**中图分类号**:TP309

Construction of Unstructured Virus Knowledge Graph Based on Stanza Model

Luo Yangxia[1,2], Wu Chenming[1],　Wang Jiayi[1], Ding Han[1]
1.School of Information, Xi'an University of finance and economics Xi′an, China 710100;
2. Computer Application and Information Security Research Center, Xi'an 710100

Abstract：A method based on the Stanford NLP toolkit Stanza is proposed to construct a viral knowledge graph, focusing on effectively extracting key information such as virus names, transmission pathways, and severity levels from unstructured texts. This approach integrates multi-source heterogeneous data in combination with the Neo4j graph database. First, the raw data are preprocessed and standardized through rule-based cleaning; next, Stanza is employed for named entity recognition and dependency parsing to accurately identify virus-related entities and their complex relationships; finally, the extracted triplets are imported into the graph database for fast retrieval and visualized analysis. Experimental results demonstrate a favorable balance between accuracy and coverage, significantly improving the efficiency of malware detection and threat intelligence analysis.

Key words：viral knowledge graph; natural language processing; stanza; entity relation extraction; neo4j

[1]收稿日期: 2025-5-xx。
项目基金：国家自然科学基金（No. 62301304）；陕西省重点研发计划项目资助（No.2024GX-YBXM-545）；西安财经大学 2023 年研究生创新基金项目(No.23YC033); 2024 年国家级大学生创新训练项目(202411560029).
通信作者：**罗养霞**，教授，研究生导，主要研究领域：系统安全、数据分析与智能计算.

1 引言

在信息安全领域，随着网络攻击手段的不断演进和复杂度的持续提升，恶意软件检测已成为研究的热点问题[1]。传统依赖特征库匹配与行为分析的检测技术，受限于病毒变种的激增与攻击策略的快速变化，正面临着检测能力下降的挑战，难以有效应对新型威胁[2]。

近年来，知识图谱作为一种集数据表示与推理能力于一体的技术，在多个领域展现出广泛应用潜力[3]。其在恶意软件分析中的引入，为整合和建模病毒特征、行为及其关联关系提供了全新思路。然而，现有知识图谱构建方法大多依赖于结构化数据，较少关注来自漏洞披露报告、国家漏洞数据库等非结构化文本中的潜在威胁信息，这使得知识图谱难以涵盖最新的攻击动态与防御策略[6]。

为弥补这一不足，本文提出了一种融合结构化与非结构化信息的知识图谱构建方法。该方法以正则表达式进行初步信息筛选，并结合斯坦福大学开发的自然语言处理工具包 Stanza 对非结构化文本中的关键信息进行抽取[8]。Stanza 在命名实体识别、依存句法分析及共指消解等任务中具备优异表现，采用双向长短期记忆网络（BiLSTM）与条件随机场（CRF）模型相结合，能够精准识别病毒相关实体[9]，并借助图神经网络构建实体间的语义关系。

在此基础上，本文通过 Neo4j 图数据库实现了知识图谱的高效构建、可视化呈现及交互式查询，进一步提升了图谱在网络安全分析场景下的实用性。本研究实现了恶意软件相关知识的自动化整合与动态更新，为安全分析人员提供了一种更加高效、智能的辅助工具。

2 知识图谱构建方法

本文提出的病毒知识图谱构建方法包括数据预处理、特征提取、实体关系解析、知识图谱构建四个核心环节。

2.1 数据预处理

2.1.1 数据收集

此本研究选取McAfee曾公开的一份病毒信息库作为数据来源。可通过互联网档案馆(Wayback Machine)错误!未找到引用源.访问。从该信息库中提取超过4万条病毒相关记录。所获取的数据包含病毒的基本特征、传播机制、危害评估及其清除指引等多内容，为本研究的知识图谱构建提供数据基础。

2.1.2 数据清洗与标准化

经过清洗的数据以 JSON 和 MySQL 数据库的格式进行存储，便于后续的分析与模型构建。为了将原始数据转化为规范化、结构化数据集，进行清洗与标准化，步骤如下:

1. 字段一致性处理：剔除字段中多余空格、换行符及冗余信息，确保字段在语义层面和格式层面的一致性，消除潜在的误差源。

2. 数据归类与整合：对于多值字段，采取分隔符拆分策略，并将其进行拼接存储，确保数据在表征和存储上的规范性和完整性。

3. 日期标准化：设计并实现日期格式解析器，将所有日期数据转换为统一的 ISO 标准格式(YYYYMMDD)，确保时间信息的一致性。

4. 缺失值填充：对于缺失的字段数据，采用默认值或空字符串替代，确保数据的完整性，防止因缺失数据导致的分析偏差。标准化后记录如表 1 所示。

表 1 标准化数据示例

Table 1 Examples of standardized data

信息项	内容概述
Field Name	Details
Virus Name	Abal.758
Aliases	Abal, Abal758, AL.758
Discovery Date	1/1/95
Origin	Unknown
Length	758 Bytes
Type	File Infector
Risk Assessment	Medium
Minimum DAT	4002
DAT Release Date	12/02/1998
Virus Characteristics	Infected files grow by 758 bytes; virus resides at file start; visible string: ABAL 758.
Symptoms	Spreads via infected files from disks, downloads, networks, etc.
Method of Infection	Requires execution of infected files.
Removal Instructions	Use SCANPM /ADL /CLEAN /ALL; disable System Restore in Windows ME if needed.
Recommended Updates	Install Office and Outlook updates; restore macro protection in Office after cleaning.

2.2 特征提取

为了从非结构化和半结构化数据中提取高信息量、

结构化且富有语义价值的特征，研究结合自然语言处理(NLP)技术与特征工程方法，设计并实现了一套系统化的特征提取流程，以提取病毒的基本属性、语义特征、与其他实体之间的复杂关联等内容，以提升病毒数据的分析深度和知识表示的准确性。

2.2.1 基本属性提取

基本属性提取是对病毒数据进行初步表征，涵盖病毒的基本信息和特征，包括但不限于病毒名称、别名、发现日期、病毒类型、起源地等信息。本研究对原始数据进行严格的预处理，确保提取过程的高效性与准确性。通过规则驱动的特征提取方法，逐步提取出上述属性信息。具体过程如下。

1. 病毒名称提取：结合词典查找和匹配的方式，应对不同表述方式所带来的潜在歧义，确保提取出的病毒名称符合标准化要求。

2. 别名提取：通过模式识别，结合文本中的上下文信息和已知知识，通过精细化的规则设计对病毒别名进行提取。避免由于区域差异或描述方式不同导致的歧义，确保不同来源的别名能够统一为规范形式。基于深度学习和模式识别技术，进一步优化别名提取过程，使其对新出现的、未被广泛认知的病毒别名同样具有较强的识别能力，保证病毒别名在知识图谱中能够全面且准确的表示。

2.2.2 语义特征提取

为了从非结构化文本中抽取和整合深层次含义的实体、属性和关系，为知识推理提供基础，研究采用改进的 Stanza 模型进行文本中的实体识别和依存句法分析，实施方法如图 1 所示。

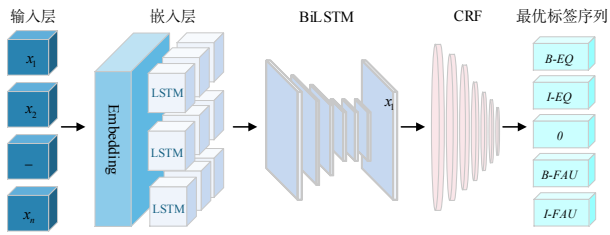

图 1 模型结构图

Fig.1 Model structure diagram

1．双向长短期记忆网络(BiLSTM)

BiLSTM 序列模型包括多个门控机制，分别控制信息的流动和更新。BiLSTM 具有双向学习能力，利用双向计算同时捕获病毒文本序列中的前后文信息，前向(从左到右)和反向(从右到左)处理输入文本，增强对语义的理解和捕捉能力[11]。

例如在处理病毒传播途径时，能够准确理解“电子邮件”作为传播途径与“病毒 X”作为病毒名称之间的关系。

2．条件随机场(CRF)

CRF 是用于序列标注的概率图模型，常用于自然语言处理中命名实体识别(NER)等任务。与传统分类方法不同，CRF 考虑标签之间的条件依赖关系，通过上下文和序列信息来处理文本中的歧义和多样性，能够优化标签的预测精度[12]。其中，CRF 模型的条件概率可以表示为：

$$\mathrm{P}(y\mid x)=\frac{\exp\left(\sum_{t=1}^{n}\sum_{k=1}^{K}\lambda_k f_k\left(y_t,y_{t-1},x_t\right)\right)}{Z(x)} \tag{1}$$

其中，$f_k\left(y_t,y_{t-1},x_t\right)$ 是特征函数，表示当前标签与前一个标签的依赖，以及与当前输入的特征，λ_x 为权重，$Z(x)$ 是归一化因子。

2.2.3 关系抽取与高阶关系建模

在知识图谱构建过程中，关系抽取是连接实体并揭示其相互联系，是从非结构化文本数据中识别并抽取实体间的多层次关系，这些关系为知识图谱提供丰富的语义信息，支持进一步的推理、查询与分析[13]。由于病毒数据本身具有复杂的多维度特征，关系抽取不仅仅局限于简单的实体关系，还需要考虑到高阶复杂关系的建模。本研究结合依存句法分析与规则驱动方法，对病毒与宿主、病毒传播途径、症状等实体之间的关系进行了系统的提取。具体步骤如下。

首先，构建句子的语法树。通过分析动词(如“感染”、“传播”)、名词短语(如“宿主”、“病毒”)以及它们之间的语法依赖关系，提取出潜在的实体关系。例如，句子“病毒感染通过邮件传播”中，动词“感染”与“病毒”之间的主谓关系揭示了“病毒与宿主”的传播关系，而动词“传播”则与“邮箱”形成的介词短语揭示了“病毒传播途径”的信息。这一过程通过规则驱动的方法将

语法结构与语义内容结合，有效提取了病毒数据中复杂的实体间关系。

其次，关系抽取不仅包括基本的实体关系(如病毒与宿主之间的传播关系)，还涵盖病毒传播链条中的高阶复杂关系。高阶关系建模是通过多层次的实体关系抽取，构建出包含多维度信息的传播路径模型。高阶关系的建模能够揭示病毒传播过程中的多个阶段和环节，具体包括病毒传播源、传播途径、感染方式、宿主类型、传播速度以及防护措施等信息[14]。

在本研究中，病毒传播链条的高阶关系主要由多个层次的实体关系组成，例如病毒的传播途径与宿主之间的传染性关系，病毒与症状之间的影响关系等。为了更精确地建模这些复杂的关系，本研究采用了分层次抽取策略：

1. 识别病毒与宿主、病毒与症状之间的直接关系；

2. 通过多维度的实体扩展关系揭示传播源、传播途径、感染路径等深层次的因果关系。

高阶关系的建模不仅可以丰富病毒传播路径的知识表示，还能够为实际应用中的病毒溯源、病毒发展研究以及病毒防护优化提供有力的支持。

2.3 实体关系解析

为了识别文本中不同实体之间的多层次关系，构建实体间的连接，形成一个丰富的知识网络，研究应用 Stanza 的句法依赖分析和语义角色标注技术，深化文本关系抽取的能力。

2.3.1 句法依赖分析

为了识别句子中各词汇之间的依存关系，揭示语法结构的内在联系。应用 Stanza 提供的 BiLSTM 模型，捕捉句子中每个词汇的依赖关系，例如主谓关系、动宾关系、修饰关系等。通过引入上下文信息，捕捉句子中的长距离依赖关系。

在病毒数据中，依存句法分析有助于识别如“病毒感染宿主”或“病毒导致症状”的结构，为提取病毒传播途径和影响因素提供了精准的语法基础。

2.3.2 语义角色标注

语义角色标注（SRL）技术的核心目标是识别句子中每个词汇的语义角色，以便理解句子中“谁做了什么，什么受到了影响”的核心信息。具体来说，SRL 可以对句子中的各个成分进行语义分类，包括施事（Agent）、受事（Patient）、工具（Instrument）等角色。在病毒数据的语义分析中，SRL 技术帮助识别出关键的语义关系，例如“病毒感染宿主”、“病毒引起症状”等。上述关系构成了病毒传播链条中的核心部分，能够为进一步的分析和理解提供基础。

Stanza 是一个功能强大的自然语言处理工具包，提供了基于 BERT 的预训练模型来执行语义角色标注。Stanza 的语义角色标注模块利用深层次的上下文信息，能够准确地识别句子中各个实体的语义角色。例如，在描述“病毒感染宿主”的句子中，Stanza 的 SRL 模块能够将“病毒”标注为施事（Agent），将“宿主”标注为受事（Patient），从而明确两者之间的影响关系。这种语义解析使得我们能够更深入地理解文本中的语义结构，捕捉病毒与宿主、症状、传播途径等实体之间的复杂关系。

通过使用 Stanza 提供的 SRL 技术，在更深层次上解析文本中的语义角色标注信息，捕捉病毒与宿主、症状、传播途径等实体之间的关系。

2.3.3 关系抽取与三元组构建

在实体关系解析的基础上，通过关系抽取模型提取文本中的各类关系。这些关系主要涉及病毒与宿主、病毒与症状、病毒与传播途径等多层次的关联。关系抽取的目的是从自然语言文本中识别出结构化的实体关系，为后续的知识图谱构建提供数据支持。

关系抽取方法基于依存句法分析和语义角色标注，采用规则驱动技术，结合动词、名词短语以及句子的语法结构，识别病毒与其他实体之间的关系。

例如，在句子“病毒 X 通过电子邮件传播给宿主 Y”中，系统通过依存句法分析与语义角色标注技术能够提取出“病毒 X”与“宿主 Y”之间的传播关系，并且转化为结构化的三元组。

为了进一步增强关系抽取的准确性，结合多种语法特征和语义信息，如动词的时态、名词短语的修饰

成分等。不仅抽取基本的实体关系，还识别出更加复杂的关系。如病毒传播链中的不同环节(传播源、宿主类型、防护措施等)以及病毒特征之间的相互关系。识别出的实体关系都被转化为三元组形式(subject, relation, object)，并以此为基础构建了病毒知识图谱。例如，“病毒 X 通过邮件传播”会被转化为三元组(Virus X, transmits through, Email)。这种结构化的数据形式便于后续在图数据库中进行存储、推理和查询。

2.4 知识图谱构建

为了从文本中提取出的实体及其相互关系以图结构的形式进行组织与存储，研究通过 Cypher 查询语言实现图数据的插入、查询和存储。知识图谱的构建主要包括三个步骤：数据预处理与节点创建、关系创建与图数据库存储、以及数据验证与质量评估。构建过程如算法 1 所示。

算法 1:构建恶意软件知识图谱

Input：virus_data、output file、Neo4j connection
Output：Neo4j 图数据库、NER 三元组库

```
1. open the output file
2. for each virus in virus_data:
3. virus_triples = [(virus.name, "is_a", "Virus")]
4. for (key, value) in virus.attributes:
5. virus_triples.append((virus.name, key, value))
6. for (key, value) in virus.features:
7. virus_triples.append((virus.name, key, value))
8. doc = stanza.process_text(virus.feature_text)
9. for each sentence in doc.sentences:
10. for each entity in sentence.entities:
11. if entity.type in predefined entity types:
12. virus_triples.append((entity.text, "is_a", entity.type))
13. for each word in sentence.words:
14. if word.text in attribute dictionary:
15. virus_triples.append((virus.name, "has", attribute dictionary[wor-d.text]))
16. if word.deprel in relationship types:
17. subj, obj = extract_dependency(word)
18. virus_triples.append((subj, word.text, obj))
19. write to output_file in format: "{subj}\t{rel}\t{obj}"
20. execute cypher: merge and create nodes and relationships
21. close output_file
22. close Neo4j connection
```

2.4.1 数据预处理与节点创建

在知识图谱的构建过程中，数据预处理的目的是确保后续节点创建和关系抽取正确性。数据预处理完成后进行节点创建。节点代表知识图谱中的实体，每个实体拥有独特的标识符和多个属性。节点的创建步骤如下：

1. 节点标识。为每个实体(如病毒、宿主、传播途径等)赋予一个唯一标识符；

2. 节点属性。节点的属性包括实体的基本信息，如病毒名称、别名、类型、发现日期、症状、传播途径等。属性通过文本中的抽取与标准化处理进行存储

3. 数据标准化。节点的属性会经过标准化处理，如日期格式的统一、异常值的校正等。

节点创建的目标是通过结构化的方式表示实体及其属性，便于后续图数据库的存储和查询。每个节点属性能够提供更细致的实体描述，为后续的图查询与推理提供支持。

2.4.2 关系创建与图数据库存储

关系的创建是将从文本中提取的实体之间的各种关系通过 Neo4j 图数据库中的关系模型，以关系的形式存储。在关系创建中，研究从文本中抽取的实体对之间的关系，通过 Cypher 查询语言完成图数据的插入。具体关系创建步骤如下。

1. 关系建模:在文本中识别出“病毒宿主”、“病毒传播途径”等多种关系类型，并将关系映射到 Neo4j 图数据库中的关系类型，例如：INFECTS(感染)、TRANSMITS_VIA(通过传播)、CAUSES(引起)等。

2. 图数据库存储:数据关系被成功插入后，所有的节点和关系将被持久化存储在 Neo4j 图数据库中。支持快速的图遍历和关系推理；高效处理复杂的多层次关系；在查询时优化数据路径选择，提升查询效率。

知识图谱可视化储存示例如图 2 所示：

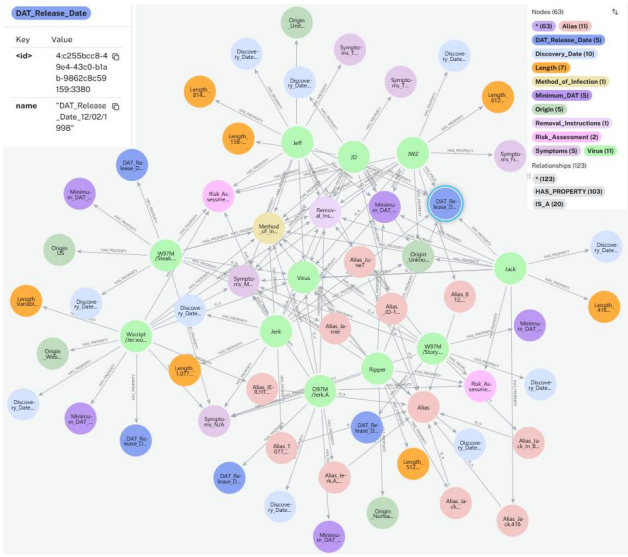

图 2 Neo4j 图数据库示例

Fig.2 Neo4j graph database example

3 实验结果验证

3.1 实验环境设置

本实验采用的测试语料来自手工标注的计算机病毒相关文本数据，数据集采用 BIO(Begin, Inside, Outside)标注格式。BIO 格式是一种常见的序列标注方案，具体方法如下。

1.BXXX 表示某个命名实体的开始，例如 BVIRUS 代表病毒名称的起始部分；

2. IXXX 表示该命名实体的后续部分；

3. O 表示非实体部分。

在数据预处理中，对原始文本进行分词、标注，并构建标准的 BIO 文件作为标注数据(Ground Truth)。

3.2 评价指标

为了全面衡量 NER 模型的识别能力，本文采用了三项评估指标，确率(Precision)、召回率(Recall)和 F1-score[15]。这三项指标共同衡量 NER 模型的性能，Precision 反映了识别结果的准确性，Recall 反映了模型对真实实体的覆盖程度，而 F1-score 作为 Precision 和 Recall 的加权调和平均值，综合评估了模型的整体表现。

3.3 实验结果对比

为了评估本文方法的有效性，本文对比了四种 NER 方法的实验结果：

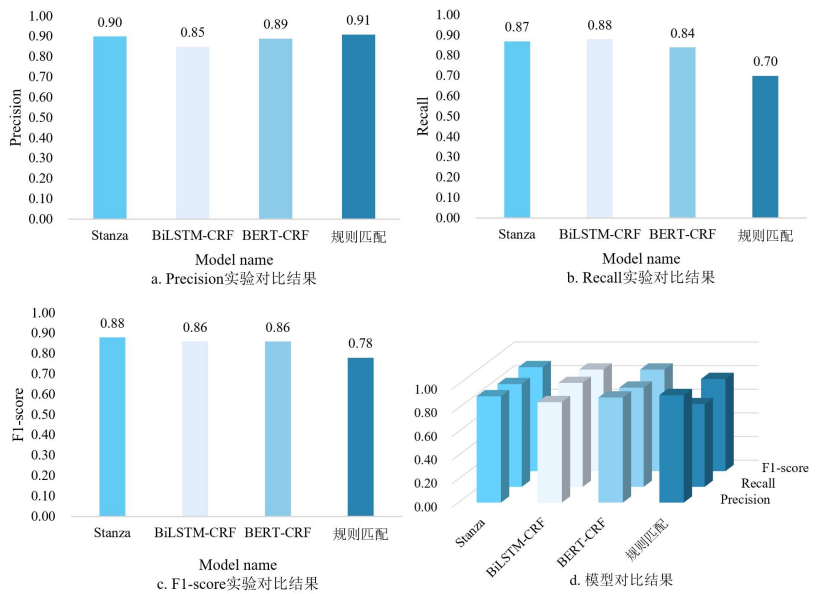

图 3 实验结果

Fig.1 Experimental results

从实验结果分析可知，预训练的 Stanza 模型在整体性能上表现最佳，能够有效适应不同类型的实体识别任务。与性能第二的 BiLSTM-CRF 模型相比，Stanza 模型在 Precision 上提高了 5%，在 F1-score 上提高了 2%。相比之下，BiLSTM-CRF 模型在 Recall 指标上表现较为突出，说明其能够较好地捕捉上下文信息，但

由于训练数据集的限制，其整体性能仍低于 Stanza 模型。BERT-CRF 模型的 Precision 为 0.89，表明该模型能有效识别复杂实体。然而，由于 CRF 层的限制，BERT-CRF 的召回率提升存在一定局限，无法达到 Stanza 和 BiLSTM-CRF 的水平。最后，规则匹配方法依赖人工设计的规则，难以适应多样化的实体类型，因此通常只在特定领域中表现较好。

3.4 对比分析

本文提出了一种基于 Stanza 的命名实体识别方法，并结合关系抽取技术，构建了恶意软件领域的知识图谱。与已有文献相比，本文的方法在实体识别和关系抽取的结合上展现了独特的优势。文献[16]采用 CSKG4APT 方法，侧重于大规模威胁情报数据的整合，利用 BERT、BiLSTM 和 GRU 等多种模型进行实体抽取。然而，Stanza 的应用在恶意软件领域更加精简且高效，避免了多个复杂模型的组合，增强了处理效率。文献[17]基于融合对抗主动学习的网络安全知识三元组抽取，提出了结合 BiLSTM 和 LSTM 的联合实体和关系抽取方法，通过对抗主动学习框架有效减少了标注成本。文献[18]利用基于 BERT 的网络空间安全命名实体识别方法，通过解决子词与标签不匹配问题提升了实体识别效果。但是，上述文献均需要较高的计算开销，且对专家标注的依赖限制其普适性。相比之下，Stanza 模型更加轻量化，适合实时数据处理，并且能够保持较高的识别精度。最后，文献[9]的 APT 攻击实体识别方法结合 BERT 和 BiLSTM-CRF，在 APT 领域表现出色。与本文方法不同，其研究侧重于 APT 攻击的实体对齐，在其他领域模型的精确率则需要进一步验证。

表 2 其他相关实验结果对比

Table 2 Comparison of results with other relevant experiments

模型	Precision	Recall	F1-score
APT 图模型+APT 知识提取[16]	0.8200	0.8200	0.8180
注意力机制+BiLSTM-CRF[17]	0.6570	0.6340	0.6450
MCEL+BERT[18]	0.8743	0.8683	0.8713
BERT+BiLSTM+CRF+APT 实体[9]	0.7800	0.5894	0.6714

4 结束语

本文围绕恶意软件知识图谱的构建，提出一种基于 Stanza 的命名实体识别方法，并结合关系抽取技术，构建了恶意软件领域的知识图谱。实验结果表明，本文方法在实体识别和关系抽取方面具有较好的性能，能够有效组织和关联计算机病毒相关信息，为病毒分析和网络安全研究提供支持。通过与其他常用 NER 方法(如 BiLSTM-CRF、BERT-CRF 和规则匹配)的对比，本文验证了 Stanza 在该领域的优越性。Stanza 方法在 F1-score 方面表现突出，且能够平衡 Precision 和 Recall，显示出较强的泛化能力和处理能力。

下一步将优化提升 NER 模型的泛化能力，引入更大规模的预训练模型及结合主动学习机制，进一步提高对新型病毒及未知实体的识别效果。另一方面，改进关系抽取方法，通过深度学习，进一步优化实体间关系的推理，提升知识图谱的准确性和完整性。

参考文献

[1] Amira A, Derhab A, Karbab E B, et al. A survey of malware analysis using community detection algorithms[J]. ACM Computing Surveys, 2023, 56(2): 1-29.

[2] 杨秀璋,彭国军,王晨阳,等.基于动静态语义行为增强的 APT 攻击溯源研究[J/OL].武汉大学学报(理学版),1-14[2025-04-17].https://doi.org/10.14188/j.1671-8836.2024.0126.

[3] 李正洁,沈立炜,李弋,彭鑫.面向文本描述的 CPS 资源能力知识图谱构建[J].软件学报,2023,34(05):2268-2285.

[4] 赖清楠,金建栋,周昌令.基于大语言模型的网络威胁情报知识图谱构建技术研究[J].通信学报,2024,45(S2):33-43.

[5] Li, Hongyi, et al. "Cybersecurity knowledge graphs construction and quality assessment." *Complex & Intelligent Systems* 10.1 (2024): 1201-1217.

[6] Moura G C M, Heidemann J. Vulnerability disclosure considered stressful[J]. ACM SIGCOMM Computer Communication Review, 2023, 53(2): 2-10.

[7] Glyder J, Threatt A K, Franks R, et al. Some analysis of common vulnerabilities and exposures (cve) data from the national vulnerability database

(nvd)[C]//Proceedings of the Conference on Information Systems Applied Research ISSN. 2021, 2167: 1508.

[8] Qi P, Zhang Y, Zhang Y, et al. Stanza: A Python natural language processing toolkit for many human languages[J]. arXiv preprint arXiv:2023.07082, 2023.

[9] 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.

[10] Arora S K, Li Y, Youtie J, et al. Using the wayback machine to mine websites in the social sciences: A methodological resource[J]. Journal of the Association for Information Science and Technology, 2016, 67(8): 1904-1915.

[11] Manoharan P ,Hong W ,Yin J , et al.Optimising Insider Threat Prediction: Exploring BiLSTM Networks and Sequential Features[J].Data Science and Engineering,2024,9(4):1-16.

[12] 王瀛,王泽浩,李红,等.基于深度学习的威胁情报领域命名实体识别[J].东北大学学报(自然科学版),2023,44(01):33-39.

[13] 鄂海红, 张文静, 肖思琪, 等. 深度学习实体关系抽取研究综述[J]. 软件学报, 2019, 30(6): 1793-1818.

[14] Wang, Quan, et al. "Knowledge graph embedding: A survey of approaches and applications." *IEEE transactions on knowledge and data engineering* 29.12 (2017): 2724-2743.

[15] Li, Hongyi, et al. "Cybersecurity knowledge graphs construction and quality assessment." Complex & Intelligent Systems 10.1 (2024): 1201-1217.

[16] Ren Y, Xiao Y, Zhou Y, et al. CSKG4APT: A cybersecurity knowledge graph for advanced persistent threat organization attribution[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 35(6): 5695-5709.

[17] 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.

[18] 韩瑶鹏,王璐,姜波,等.基于预训练模型的网络空间安全命名实体识别方法[J].信息安全学报,2025,10(01):194-204.