中图分类号：TP391

论文编号：10006SY1808116

北京航空航天大學

专业硕士学位论文

面向非结构化数据的威胁情报知识图谱构建方法

作 者 姓 名　　何 润 龙

学 科 专 业　　计算机技术

指 导 教 师　　郎　波　教 授

严寒冰　高 工

培 养 学 院　　计 算 机 学 院

Unstructured Data Oriented Construction Method of Threat Intelligence knowledge Graph

A Dissertation Submitted for the Degree of Master

Candidate：He Runlong

Supervisor：Prof. Lang Bo

Prof. Yan Hanbing

School of Computer Science & Engineering

Beihang University, Beijing, China

中图分类号：TP391

论文编号：10006SY1808116

硕 士 学 位 论 文

面向非结构化数据的威胁情报知识图谱构建方法

作者姓名	何润龙	申请学位级别	专业硕士
指导教师姓名	郎 波	职 称	教授
	严寒冰	职 称	高级工程师
学科专业	计算机技术	研究方向	网络安全
学习时间自	2018 年 9 月 1 日	起至	2022 年 5 月 30 日
论文提交日期	2022 年 6 月 2 日	论文答辩日期	2022 年 5 月 30 日
学位授予单位	北京航空航天大学	学位授予日期	2022 年 6 月 日

关于学位论文的独创性声明

本人郑重声明：所呈交的论文是本人在指导教师指导下独立进行研究工作所取得的成果，论文中有关资料和数据是实事求是的。尽我所知，除文中已经加以标注和致谢外，本论文不包含其他人已经发表或撰写的研究成果，也不包含本人或他人为获得北京航空航天大学或其它教育机构的学位或学历证书而使用过的材料。与我一同工作的同志对研究所做的任何贡献均已在论文中作出了明确的说明。

若有不实之处，本人愿意承担相关法律责任。

学位论文作者签名：何润龙　　日期：2022 年 6 月 2 日

学位论文使用授权书

本人完全同意北京航空航天大学有权使用本学位论文（包括但不限于其印刷版和电子版），使用方式包括但不限于：保留学位论文，按规定向国家有关部门（机构）送交学位论文，以学术交流为目的赠送和交换学位论文，允许学位论文被查阅、借阅和复印，将学位论文的全部或部分内容编入有关数据库进行检索，采用影印、缩印或其他复制手段保存学位论文。

保密学位论文在解密后的使用授权同上。

学位论文作者签名：何润龙　　日期：2022 年 6 月 2 日

指导教师签名：邹波　　日期：2022 年 6 月 2 日

摘　要

面对日益严峻的网络安全形势，网络威胁情报（Cyber Threat Intelligence, CTI）越来越受到网络安全从业者和各大企事业单位的重视。在各种类型的威胁情报中，高级持续性威胁（Advanced Persistent Threat，APT）报告具有较高的分析、利用价值，为人们所广泛关注。APT 报告常采用 PDF 等非结构化数据格式发布。目前关于威胁情报信息整合的研究，主要集中在半结构化威胁情报数据的抽取与利用。然而，半结构化威胁情报数据通常仅在公司内部共享，数据规模有限且获取难度较大。

本文对威胁情报语义内涵和相关标准进行分析，研究并实现威胁情报本体、非结构化威胁情报命名实体识别和实体关系抽取方法。论文的主要工作包括：

（1）构建了一种威胁情报本体。基于威胁情报的语义内涵和威胁情报标准 STIX 2.1 设计并构造了面向威胁情报领域的本体。该本体囊括了 13 种威胁实体和 7 种实体间关系，为本文构造威胁情报知识图谱提供指导和依据。

（2）提出了一种基于数据增强和 BERT 的威胁情报命名实体识别方法。该方法通过将知识库中的威胁情报领域词汇填入符合威胁情报上下文环境的模板句子实现数据增强，然后将增强数据与 BERT+Bi-LSTM+CRF 模型结合进行威胁情报命名实体识别。在威胁情报命名实体识别数据集上进行实验，本文提出方法的 F1 分数宏平均值可达 89.315%，更换为 DistilBERT 模型后的 F1 分数宏平均值可达 89.177%，均优于未进行数据增强的模型。

（3）提出了一种融合多元实体信息的威胁情报实体关系抽取方法。首先针对威胁情报句子中存在多个实体和关系的数据特点，改进了 Brat 标注工具使其可以用于生成威胁情报实体关系数据集。然后将实体边界信息和实体类型信息通过在实体两侧添加标签的方式与词向量中的实体语义信息进行融合，从而提高 R-BERT 模型性能。在威胁情报实体关系数据集上进行实验，本文提出方法的 F1 分数宏平均值可达 81.061%。

（4）设计并实现了威胁情报知识图谱管理工具。该工具可从非结构化威胁情报文本数据中自动抽取威胁实体和关系，并能够对抽取结果进行存储、检索和图谱可视化。

关键词：非结构化数据，网络威胁情报，本体，知识图谱，命名实体识别，实体关系抽取

Abstract

With the increasingly severe network security situation, Cyber Threat Intelligence (CTI) has been paid more and more attention by network security practitioners and major enterprises and institutions. Among various types of Threat intelligence, Advanced Persistent Threat (APT) reports have high analysis and utilization value and are widely concerned by people. At present, the research on threat intelligence information integration mainly focuses on the extraction and utilization of semi-structured threat intelligence data. However, semi-structured threat intelligence data is usually shared only within the company, and the data is limited in scale and difficult to obtain.

In this thesis, the semantic connotation and related standards of threat intelligence are analyzed, threat intelligence ontology, threat intelligence named entity recognition and entity relation extraction methods are studied and implemented. The main work of this thesis includes:

(1) A threat intelligence ontology is constructed. Based on the semantic connotation of threat intelligence and the threat intelligence standard STIX 2.1, designed and constructed an ontology for the field of threat intelligence, which includes 13 threat entities and 7 relations between them, providing guidance for the construction of the threat intelligence knowledge graph below.

(2) A threat intelligence named entity recognition method based on data enhancement and BERT is proposed. In this method, the threat intelligence domain terms in the knowledge base are filled into the template sentences in accordance with the threat intelligence context to achieve data enhancement, and then the enhanced data is combined with BERT+Bi-LSTM +CRF model for threat intelligence named entity recognition. Experiments on the threat intelligence named entity recognition dataset show that the macro average F1 score of the proposed method is 89.315%, and the macro average F1 score of the DistilBERT model is 89.177%, both of which are better than the model without data enhancement.

(3) A threat intelligence entity relationship extraction method integrating multiple entity information is proposed. Firstly, aiming at the data characteristics of multiple entities and relationships in threat intelligence sentences, the Brat annotation tool is improved so that it can be used to generate the threat intelligence entity relationship data set. Then, the entity boundary

information and entity type information are fused with entity semantic information in word embeddings by adding labels on both sides of the entity, so as to improve the performance of R-BERT model. Experiments on the threat intelligence entity relationship data set show that the macro average F1 score of the proposed method can reach 81.061%.

(4) Designed and implemented the threat intelligence knowledge graph management tool. This tool can automatically extract threat entities and relationships from unstructured threat intelligence text data, and can store, retrieve and visualize the extracted results.

Keywords: Unstructured data, Cyber threat intelligence, Ontology, Knowledge Graph, Named entity recognition, Relation extraction

目　录

图　目

表　目

第一章　绪论

1.1 研究背景及意义

近年来，随着计算机网络与通信技术的快速发展，互联网与人类日常生活的联系愈发密切。中国互联网信息中心 CNNIC 在《第 49 次中国互联网络发展状况统计报告》[1]中指出，截至 2021 年 12 月，我国网民规模为 10.32 亿，较 2020 年 12 月新增网民 4296 万，互联网普及率达到 73%，较 2020 年 12 月提升 2.6 个百分点。此外，互联网持续助力我国中小企业数字化转型，推动数字经济发展，促使数字红利普惠大众。互联网在智能制造、智慧交通和电子政务等领域得到广泛应用，对各行各业的影响力都在逐年增大。

与此同时，互联网网络安全面临着更为严峻的考验。根据国家互联网应急中心 CNCERT 发布的《2020 年我国互联网网络安全态势综述》，2020 年全年捕获的恶意程序样本数量超过 4200 万个，日均传播次数超过 482 万次[2]。按照攻击目标 IP 地址统计，我国境内受恶意程序攻击的 IP 地址约 5541 万个，占我国 IP 地址总数的 14.2%。同年，国家信息安全漏洞共享平台（CNVD）所收录的安全漏洞在数量上达到历史新高：20704 个，同比增长 27.9%。其中，“零日”漏洞数量为 8902 个，占比 43.0%，同比增长 56.0%。以上数据证明我国网络安全形势不容乐观。

此外，分布式拒绝服务（DDoS）攻击作为一种难以防范的常见攻击手段，近年来发生频率有明显上升趋势。2020 年，我国境内遭受大流量分布式拒绝服务攻击 10.4 万余次，日均发生量高达 285 起，同比增加 29.5%[3]。随着工业物联网技术的发展，越来越多的工业设备接入互联网，2020 年共发现涉及重点行业（电力，石油天然气，轨道交通）联网监控管理系统的高危漏洞 142 个[3]，其影响不可忽视。

在当前网络环境下，尽管防火墙、杀毒软件、入侵检测系统等传统安全技术已被广泛应用并取得了一定的效果，但在面对“零日”漏洞攻击和高级持续威胁（APT）攻击时也表现出一些不足。人们开始把目光投向威胁情报以寻求解决网络安全问题的新思路。威胁情报是一种基于证据的知识，包括上下文、机制、标示、含义和能够执行的建议，这些知识与资产所面临已有的或酝酿中的威胁或危害相关，可用于资产相关主体对威胁或危害的响应或处理决策提供信息支持[4]。

近年来，威胁情报驱动的网络安全主动防御模式作为新的发展方向备受学术界和产业界关注，不少安全组织、安全厂商开始研究和应用威胁情报并据此提供相关服务。通

过主动对数据进行采集，提炼与分析，威胁情报可以整合已经发生或正在发生的网络安全事件信息，做到提早发现网络威胁，进而采取措施以达到事先预防或减少损失和危害的作用。

然而，威胁情报的发展仍然面临诸多挑战。荷兰政府高级网络威胁情报分析师 Kris Ossthoek 在《国际情报与反情报杂志》[5]上撰文指出威胁情报的开发利用具有如下难点：首先，威胁情报缺乏方法论。大多数威胁情报分析都是由警报和日志数据而不是预先确定的方法或假设进行输入驱动的。缺乏方法论使得企业难以分析每天产生的大量失陷指标（Indicators of Compromise，IOC）与特定威胁环境的相关性。其次，威胁情报的共享只是口头上的。出于信任和利益等原因，少部分结构化和半结构化威胁情报仅在企业和组织内部共享。相比之下，大多数威胁情报仍以非结构化方式在互联网上共享。最后，威胁情报领域不存在通用的命名约定，给实体识别和归因带来极大困难。出于营销目的，安全厂商和威胁情报提供商热衷于给各个威胁组织起各种名字。例如，同一个俄罗斯威胁组织拥有多个名字：APT28，Fancy Bear，Sofacy，Sednit，STRONTIUM，Pawn Storm。这导致需要识别的实体数量增多，识别难度加大。

因此，使用计算机相关技术对非结构化威胁情报加以开发利用具有现实意义。本论文将从威胁情报本体的设计、非结构化威胁情报命名实体识别和非结构化威胁情报实体关系抽取三个角度入手，研究如何自动化构建威胁情报知识图谱，设计并实现威胁情报知识图谱管理工具，为威胁情报的研判和共享提供有效支持。

1.2 国内外研究现状

1.2.1 威胁情报知识图谱

知识图谱概念的前身是 Tim Berners-Lee 于 1998 年提出的语义网[6]（Semantic Web），基于图的组织形式对客观世界中的实体关系进行描述。2012 年，Google 公司正式提出知识图谱的概念，用来描述事物之间的语义关系，并被应用于其新一代搜索引擎。此后，知识图谱技术在知识库构建领域得到了广泛应用，例如，Yago[7]，FreeBaese[8]，DBpedia[9]等。知识图谱技术被认为是这些知识库应用的“智能中枢”，支撑着知识图谱的应用。此外，许多国内外研究者还将知识图谱与特定领域的数据相结合，拓展知识图谱创新应用的边界。在金融大数据时代，知识图谱与人工智能、大数据分析等技术结合，被广泛应用于信用评估，风险控制等场景，使智能化金融服务拥有广阔的前景。阿里巴巴集团

旗下的淘宝公司将知识图谱技术引入电子商务领域，电商知识图谱涵盖了商品信息，商品评价，买方，卖方等众多类型的节点，为智能搜索、智能客服等应用场景提供支持。知识图谱已经在金融、电商、医疗等领域初见成效，相信在未来会有更多应用。

近年来，随着 APT 攻击普遍化，网络安全威胁情报逐步升温为一项高级安全服务。与大数据紧密结合，提高智能决策效率是威胁情报的发展趋势。知识图谱因包含实体抽取、关系抽取、语义解析、数据融合等多项技术，可以依托大数据直观呈现事物关系并预测发展趋势，已被研究人员用于威胁情报的挖掘工作。Li[10]在其博士论文中将知识图谱技术应用于网络威胁情报领域，提出了一种基于多特征融合的威胁情报实体抽取方法和一种基于语义特征增强的威胁情报实体关系抽取方法，以构建威胁情报知识图谱，并将强化学习和图卷积网络混合用于威胁知识的推理。Qin[11]在其硕士论文中将神经网络与特征模板结合进行网络安全实体识别，并提出了一种 ResPCNN-ATT 远程监督关系抽取方法，用来识别网络安全实体关系，从而构建网络安全知识图谱。

1.2.2 命名实体识别技术

命名实体识别（Named Entity Recognition，NER）旨在从自然语言文本中识别出具有特定意义的实体，如人名、地名、组织名等，包括正确识别实体的边界和类型。命名实体识别是关系抽取、知识图谱、问答系统等众多自然语言处理任务的基础。

早期的命名实体识别方法主要包括：1.基于规则的方法。该方法依赖人工特征，可以基于特定领域词典[12][13]或特定语法规则[14]来设计规则。当词汇详尽时，基于规则的方法可以很好地工作。然而，设计规则和收集词典需要花费大量时间和精力。2.基于无监督学习的方法。该方法使用词汇在大语料上的统计特性来推断命名实体的出现。Nadeau 等人[15]提出了一个用于构建词典和命名实体消歧的无监督系统，Zhang 等人[14]提出了一种从生物医学文本中抽取命名实体的无监督方法。3.基于特征的监督学习方法。在监督学习框架下，命名实体识别被转化为一个序列标注任务（多分类任务）。使用精心设计的特征和机器学习算法，在标注语料上训练模型，从而识别未知文本中的命名实体。已被用于有监督命名实体识别任务的机器学习算法有决策树，隐马尔可夫模型，最大熵模型，支持向量机和条件随机场。特征工程是有监督命名实体识别方法的关键，却十分依赖人们的经验。

近年来，随着深度学习技术的发展，人们开始使用神经网络进行自动化命名实体识别。使用深度学习方法进行命名实体识别的优点包括：可以从数据中学到更复杂的特征；

避免大量人工特征的构建；可以设计为端到端的结构。基于深度学习的命名实体识别方法已经在通用领域取得了一定的成果。例如，Collobert 等人[16]第一次提出基于词嵌入（Word Embedding）的方法，使用卷积神经网络（Convolutional Neural Network，CNN）提取局部特征，然后使用条件随机场作为解码器来预测实体的类型；Huang 等人[17]首次将双向长短期记忆（Bidirectional Long-Short Term Memory，Bi-LSTM）神经网络应用于序列标注任务；Li 等人[18]将命名实体识别任务重新定义为一个机器阅读理解问题，并通过微调 BERT 模型解决该问题。

深度学习技术同样被广泛应用于威胁情报命名实体识别任务中。Qin 等人[19]将特征模板与神经网络模型相结合以解决中英文混合的威胁情报实体识别问题；Wu 等人[20]将领域词典与神经网络模型相结合，通过构建用于纠错的领域词典来改进威胁情报实体识别；Tikhomirov 等人[21]研究了 BERT 模型在俄语威胁情报命名实体识别任务中的表现，并提出了一种新的数据增强方法来辅助实体识别任务。

1.2.3 实体关系抽取技术

实体关系抽取（Relation Extraction，RE）旨在从自然语言文本中自动提取出实体对之间的语义关系，是知识图谱、文本摘要、自动问答、搜索引擎等众多自然语言处理任务的基础。

经典的实体关系抽取方法可分为有监督，无监督，半监督和弱监督 4 类。有监督的实体关系抽取主要包括基于特征的方法和基于核函数的方法[22]。Zhou[23]等人使用 SVM 分类器研究了词汇、句法和语义特征对实体关系抽取的影响。有监督的方法不仅需要大量人工标注的训练数据，还需要一定的专业知识。因此，人们[24]继而提出了基于无监督、半监督[25]和弱监督的实体关系抽取方法来解决人工标注语料的问题：Hasegawa[26]等人首次提出了一种无监督的实体关系抽取方法；Brin[27]利用 Bootstrapping 方法对实体关系进行抽取；Craven 等人[28]在研究从文本中抽取结构化数据建立生物学知识库的过程中，首次提出了弱监督机器学习思想。

然而，经典方法存在误差传播的问题，极大影响了实体关系抽取的效果。近年来，随着深度学习的崛起，研究者们逐渐将深度学习技术应用到实体关系抽取任务中。基于深度学习的实体关系抽取方法，根据数据标注量级的差异，可分为有监督和远程监督两类。基于深度学习的有监督实体关系抽取方法是近几年关系抽取的研究热点，该方法使用端到端的结构，能有效避免经典方法中人工特征工程等步骤，减少并改善特征抽取过

程中的误差积累问题[22]。Zeng 等人[29]于 2014 年首次使用卷积神经网络（CNN）进行实体关系分类，Katiyar 等人[30]于 2017 年首次将注意力机制（Attention）和双向长短期记忆神经网络（Bi-LSTM）用于联合抽取方法，神经网络模型在有监督实体关系抽取领域的拓展皆取得了不错的效果。基于深度学习的远程监督实体关系抽取方法因能够缓解经典方法中标签误差传播的问题而成为研究热点，其主要基础方法包括卷积神经网络，循环神经网络等网络结构[31][32]。近年来，研究者们在基础方法之上又进行了多种改进，如分段卷积神经网络（PCNN）与多示例学习的融合[33]，分段卷积神经网络与注意力机制的融合[34]等。Ji 等人[35]提出在分段卷积神经网络和注意力机制的基础上添加实体描述信息以辅助学习实体的表示。Huang 等人[36]提出的残差网络（ResNet）、Ren 等人[37]提出的 CoType 模型都在实体关系抽取任务中取得了不错的效果。

1.3 研究目标及内容

本文的研究目标是对 APT 威胁情报报告及相关标准进行分析，设计威胁情报知识图谱本体，利用机器学习和深度学习方法，建立面向非结构化数据的威胁情报命名实体识别和实体关系抽取方法，从而实现威胁情报知识图谱的自动化构建，并最终完成一个威胁情报知识图谱管理工具。

本文的研究内容主要包括威胁情报知识图谱本体的构建、基于数据增强与 BERT 的威胁情报实体抽取方法、融合多元实体信息的威胁情报实体关系抽取方法、以及威胁情报知识图谱管理工具的设计与实现。具体研究内容如下：

（1）威胁情报知识图谱本体的构建

分散在互联网上的非结构化威胁情报数据，其格式不统一，结构多样化，这不利于抽取其中的威胁实体和关系。因此，实现对多源异构威胁情报的统一描述是构建威胁情报知识图谱的基础。现有的威胁情报相关标准没有充分考虑各类威胁情报实体的语义信息和实体之间的语义关联，而本体（Ontology）提供了描述事物间语义信息的功能，可以用来指导对威胁情报的统一描述。本文依据威胁情报领域知识和威胁情报国内外相关标准建立威胁情报本体，进而定义威胁情报知识图谱的实体和关系类型。

（2）基于数据增强与 BERT 的威胁情报实体抽取方法

在威胁情报本体设计完成后，知识图谱的模式层已经确立，随后需要对威胁情报中的威胁实体和关系进行抽取以获取构建知识图谱所需的数据。本文主要针对非结构化威胁情报进行实体抽取。与通用领域文本数据（如新闻）不同的是，威胁情报报告通常将

IOC 以附录的形式在文末给出，这导致域名和 IP 等 IOC 的标签数量稀少。此外，由于威胁情报领域不存在通用的命名约定，同一个 APT 组织或恶意软件可能存在多个名称，导致攻击组织和恶意软件实体的识别难度加大。

对此，本文将对 IOC（漏洞、域名、IP）和标注数据多样性有限的实体类型（攻击组织、恶意软件）进行数据增强，并结合 BERT、循环神经网络、条件随机场等方法，实现对非结构化威胁情报的命名实体识别，此方法能更好地识别威胁情报领域未登录词。

（3）融合多元实体信息的威胁情报实体关系抽取方法

抽取完威胁情报实体后，还需要抽取威胁情报实体间关系以完备地构建威胁情报知识图谱。本文主要针对非结构化威胁情报进行关系抽取。通用领域的关系抽取任务多为句子级关系抽取，即给定一个句子和句子中的两个实体，判断这对实体的关系。而威胁情报的实体关系抽取情况相对复杂：一个威胁情报句子中可能存在多个实体和多个实体间关系。现有的工具无法针对这种情况生成威胁情报实体关系数据集，为应用机器学习方法解决威胁情报的实体关系抽取带来不便。此外，目前应用 BERT 进行威胁情报实体关系抽取的方法只考虑了融合实体语义信息和实体边界信息，并没有考虑实体类型信息，存在一定的局限性。

对此，本文将改进 Brat 标注工具以生成可用于监督学习的威胁情报实体关系数据集，并将实体语义信息、实体边界信息和实体类型信息融入 BERT 以实现对为结构化威胁情报的实体关系抽取。

（4）威胁情报知识图谱管理工具的设计与实现

针对非结构化威胁情报数据，本文将设计并实现一个威胁情报知识图谱管理工具，该工具能自动化抽取非结构化威胁情报文本中的威胁实体和关系，并对抽取结果进行数据存储与可视化，提供威胁情报检索功能，验证本文提出方法的有效性。

1.4 论文组织结构

本文围绕面向非结构化数据的威胁情报知识图谱构建方法展开研究并实现验证工具，论文结构如下：

第一章 绪论。本章主要阐述了研究的背景与意义、国内外研究现状、研究目标及主要内容等。

第二章 相关理论与技术。本章首先介绍知识图谱的构建技术，包括本体设计和信息抽取；然后介绍信息抽取的两个子任务，命名实体识别和实体关系抽取；最后介绍本

文所使用的机器学习和深度学习算法，前者包括条件随机场，后者包括长短期记忆神经网络、BERT 和知识蒸馏。

第三章 威胁情报本体构建。本章首先阐述本体的构建思想；然后介绍并分析现有威胁情报相关标准，进而确立威胁实体和关系类型，并在此基础上构建威胁情报本体；最后基于威胁情报本体，确立威胁情报知识图谱的实体和关系类型。

第四章 基于数据增强与 BERT 的威胁情报实体抽取方法。本章首先介绍非结构化威胁情报文本数据特点所带来的问题以及现有 NLP 数据增强方法的不足；然后提出基于知识库和模板填充的数据增强方法，并对知识库与模板设计和模板填充算法进行了详细阐述；接着介绍将数据增强和 BERT 结合的威胁情报命名实体识别方法，并最终通过实验对本文所提方法的有效性进行论证。

第五章 融合多元实体信息的威胁情报实体关系抽取方法。本章首先介绍非结构化威胁情报文本数据实体关系抽取的难点以及现有标注工具的不足；在此基础上，本文对现有标注工具 Brat 进行改进，并用此标注工具生成威胁情报实体关系数据集；接着介绍融合多元实体信息的威胁情报实体关系抽取方法，并对此方法进行了阐述；最后通过实验对该方法的有效性进行验证。

第六章 威胁情报知识图谱管理工具。本章首先介绍工具需求与总体设计思想；然后详细阐述工具各功能模块的实现细节，包括数据采集模块、信息抽取模块、情报存储模块、情报检索模块；最后对工具的功能和性能进行测试。

最后是总结与展望部分。该部分对论文所取得的研究成果进行总结，并对未来的研究方向与工作进行展望。

第二章　相关理论与技术

本章主要介绍知识图谱构建技术、信息抽取、条件随机场和深度学习相关理论与技术。首先介绍了知识图谱构建技术。然后介绍了本文所使用的信息抽取技术，包括命名实体识别和实体关系抽取。接着对本文用到的条件随机场进行了介绍。最后阐述了本文所使用的深度学习技术。

2.1 知识图谱构建技术

知识图谱（Knowledge Graph，KG）是一种基于图数据结构，由许多“实体—关系—实体”三元组组成的语义网络。知识图谱中的实体和关系可以包含属性等信息，不同实体之间通过关系相互连接。高质量的知识图谱通常包含模式层和数据层，构造知识图谱首先要设计本体即图谱的模式层，然后以本体为约束来构建知识图谱的数据层。

本体（Ontology）是对特定领域中某套概念及其相互之间关系的形式化表达，最早来源于哲学领域，现已被广泛应用于信息检索领域[38]、医学工程领域[39][40]和地理领域[41]中。在计算机科学领域，其核心意思是指一种模型，用于描述由一套对象类型（概念或者类）、属性以及关系类型所构成的世界。

在确立本体之后，可依据本体来构建知识图谱的数据层，主要涉及以下几方面技术：知识抽取，知识表示，知识融合以及知识存储。知识抽取旨在从不同数据源中抽取所需的实体和实体间关系，这些信息是构成知识图谱数据层的基础。知识抽取工作主要依赖于命名实体识别和实体关系抽取技术。知识表示旨在将客观知识表示成一种计算机可以接受的用于描述知识的数据结构，例如字典（键-值对）等形式。知识融合技术根据融合对象可分为两类，一类是对抽取得到的实体和关系进行融合，即对冗余的实体进行实体消歧和去重，或对同一实体的不同描述进行共指消解；另一类是对多个知识图谱进行融合，例如融合多个知识库以增强搜索引擎的搜索能力。知识存储旨在将处理后的实体和关系以三元组的形式存储到数据库中。知识图谱通常存储在图数据库中，常见图数据库包括 Neo4j，GraphDB，HugeGraph 等。知识图谱整体构建流程与所涉技术如图 1 所示。

本文主要针对非结构化威胁情报文本数据构建威胁情报知识图谱，主要涉及命名实体识别、实体关系抽取以及知识存储等技术。

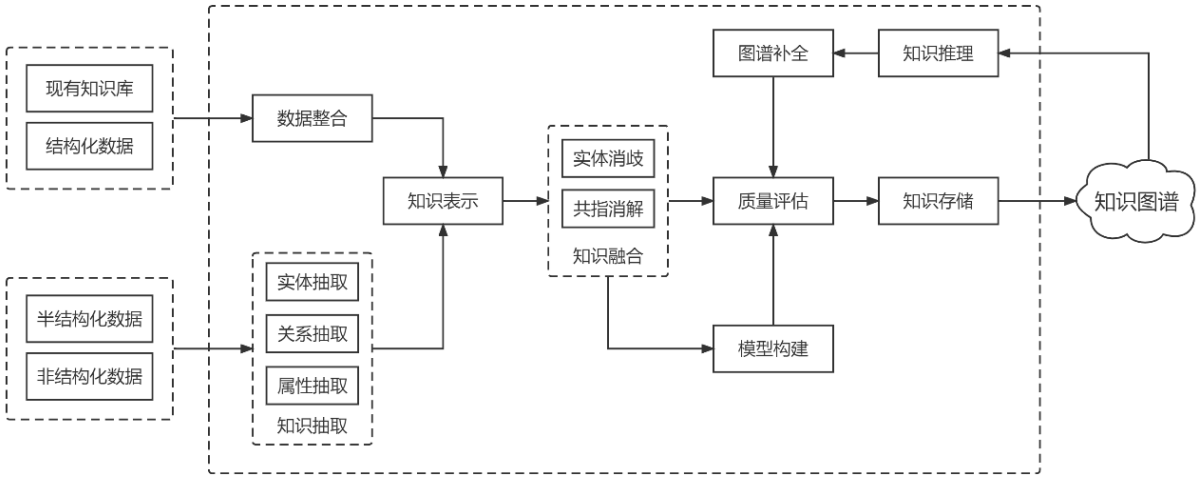

图 1　知识图谱构建流程与技术

2.2 信息抽取

信息抽取（Information Extraction，IE）是从非结构化自然语言文本中提取特定信息的过程，包括命名实体识别、实体关系抽取和事件抽取三个子任务。早期的信息抽取依赖规则、模板和领域专业知识通过人工手段进行建模，浪费时间且可迁移性还差，无法在大数据时代进行推广。近年来，随着深度学习技术的发展，基于神经网络方法的端到端信息抽取已成为新的研究方向。作为知识图谱构建技术中的一环，高效的信息抽取方法推动了知识图谱应用的发展。

2.2.1 命名实体识别

命名实体识别是自然语言处理领域的核心任务之一，它是关系抽取、智能问答、文本摘要等任务的基石。作为自动化构建知识图谱的关键技术，命名实体识别旨在从非结构化自然语言文本中识别出构成知识图谱的实体，例如时间、地点、人物、组织等，对于威胁情报领域则是攻击组织、恶意软件、技术、工具、漏洞等相关类型实体。

传统的命名实体识别方法主要是基于人工规则的方式，这种方式虽然能在小数据集上取得不错的表现，但缺乏灵活性，同时需要领域专家对抽取过程进行设计和检查[42]。基于机器学习的方法利用标注好的语料数据训练一个模型，使用该模型判断新输入语料中的实体边界和实体类型。此类方法将命名实体识别转化为文本序列标注任务处理。基于机器学习的命名实体识别方法按照处理流程可分为以下三个部分：

（1）标注训练语料。现有语料标注模式包括 BIO、BIOES、BMES 等。对于 BIOES 标注模式，B（Begin）表示一个实体的开始，I（Inside）表示一个实体的内部，O（Other）表示非实体词，E（End）表示一个实体的尾部，S（Single）表示由单个词组成的实体。BIO 标注模式中的 BIO 同样表示 Begin，Inside 和 Other。本文使用 BIO 标注模式对非结构化威胁情报文本进行标注。

（2）定义特征。特征是区分事物的关键。在自然语言处理领域，通常用一个高维向量来表示单词的特征，这样的向量被称为词向量或者词嵌入。常见单词特征包括词级别特征、字符级别特征和其他特征等。词级别特征包括语义、词性、是否包含大写等特征。字符级别特征包括单词的拼写特征（组成单词的字母的上下文特征）。其他特征如词典特征，依赖于外部知识库。

（3）训练模型。传统的统计机器学习方法如隐马尔可夫模型和条件随机场等，由于能刻画系统状态的转移过程，被广泛应用于序列标注任务。近年来，深度学习技术和神经网络模型被广泛应用于自然语言处理任务，且在许多任务上的性能表现都超过了传统方法。

2.2.2 实体关系抽取

实体关系抽取旨在根据句子的语义信息推测句子中两个实体之间的关系。通常，句子中的两个实体是给定的，只需判断两个实体之间的关系属于预定义关系类型中的哪一个，可以将该问题建模为多分类问题。例如，给定句子：“Beihang University is located in Beijing.”以及实体“Beihang University”和“Beijing”，模型通过语义得到“located in”的关系，并最终抽取出（Beihang University，located in，Beijing）的知识三元组，实体关系抽取流程示例如图 2 所示。

实体关系抽取按照抽取流程可分为联合抽取（Joint Extraction）和流水线式（Pipeline）抽取。联合抽取是指同时完成命名实体识别和实体关系抽取，流水线式抽取是指先使用命名实体识别模型抽取文本中的实体对，然后使用实体关系抽取模型判断实体对的关系。流水线式抽取方法存在误差传播问题，即命名实体识别的错误无法在后续阶段进行纠正。例如，如果命名实体识别模型将“Beihang University”错误识别为“University”，则实体关系抽取模型只会判断“University”和“Beijing”的关系，最终得到的三元组也会偏离期望得到的结果。联合抽取方法虽然不存在误差传播问题，但有研究[43]指出使用单独的编码器（即流水线式抽取方法）可以学习到更好的特定任务特征。

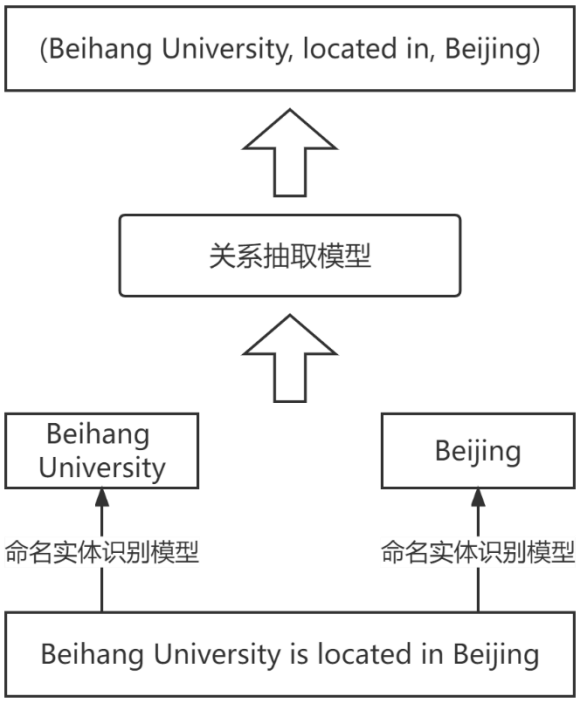

图 2　实体关系抽取流程

实体关系抽取作为一个经典且基础的任务，在过去二十多年里都有持续研究，核方法、图模型、特征工程等曾被广泛应用其中，也都取得过一些阶段性成果。随着深度学习时代的来临，神经网络模型为实体关系抽取任务带来了新的突破。

2.3 条件随机场

条件随机场（Conditional Random Field，CRF）是 Lafferty 等人[44]在 2001 年提出的一种无向图模型，常用于序列标注任务。它描述了在给定一组输入随机变量的情况下，另一组输出随机变量的条件概率，前提是这组输出随机变量的联合概率分布满足马尔可夫性。

马尔可夫性包括成对马尔可夫性、局部马尔可夫性和全局马尔可夫性，三者的定义是等价的。设 u 是无向图 G 中任意节点，V 是与 u 有直连边的所有节点，O 是除 u 和 V 以外的所有节点。用 Y_u 表示节点 u 对应的随机变量，Y_V 和 Y_O 分别表示节点集合 V 和节点集合 O 对应的随机变量组，则局部马尔可夫性是指在给定随机变量组 Y_V 的条件下，随机变量 Y_u 与随机变量组 Y_O 是条件独立的，即：

$$P(Y_u, Y_O|Y_V) = P(Y_u|Y_V)P(Y_O|Y_V) \tag{2.1}$$

局部马尔可夫性示例如图 3 所示。

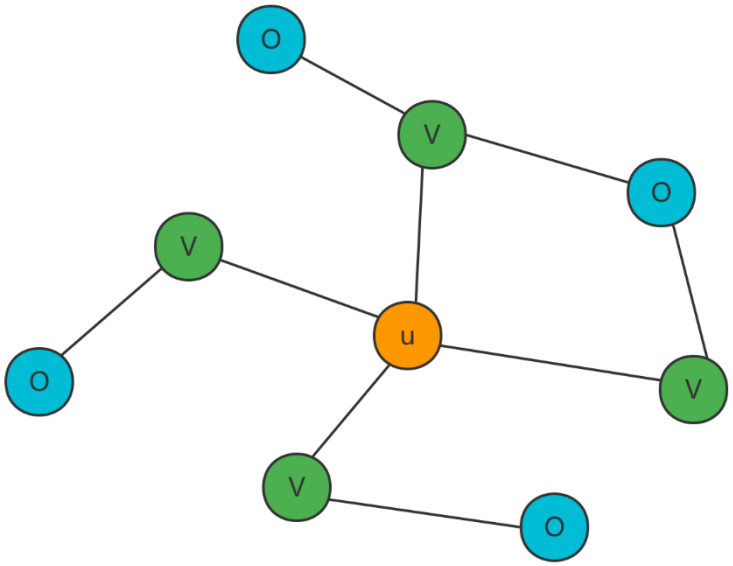

图 3　局部马尔可夫性示例

在$P(Y_O|Y_V) > 0$时，公式(2.2)与(2.1)等价：

$$P(Y_u|Y_V) = P(Y_u|Y_V, Y_O) \tag{2.2}$$

公式(2.2)说明，对于满足马尔可夫性的无向图模型，在计算某个节点的条件概率时，只需考虑其相邻节点，即给定某个节点的相邻节点作为条件即可。

下面给出条件随机场的定义。设 X 与 Y 是随机变量，P(Y|X)是在给定 X 的条件下 Y 的条件概率分布。若随机变量 Y 构成一个可以用无向图 G 表示的马尔可夫随机场（满足马尔可夫性），即：

$$P(Y_u|X, Y_v, v \neq u) = P(Y_u|X, Y_v, v{\sim}u) \tag{2.3}$$

则称条件概率分布P(Y|X)为条件随机场。其中，v≠u 表示除 u 以外的所有节点，v~u 表示与节点 u 有直连边的所有节点 v。

线性链条件随机场（Linear Chain Conditional Random Field）是一种具有特殊图结构的条件随机场，常用于序列标注等问题。本文所使用的条件随机场是 X 和 Y 具有相同图结构的线性链条件随机场，如图 4 所示。

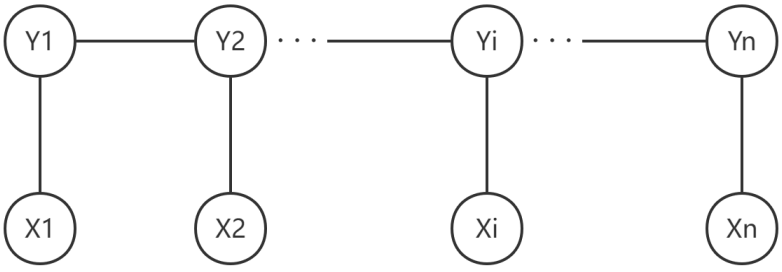

图 4　线性链条件随机场示例

此时，输入随机变量序列 $X = (X_1, \ldots, X_n)$表示需要标注的观测序列，输出随机变量序列 $Y = (Y_1, \ldots, Y_n)$表示标记序列，条件概率模型$P(Y|X)$为目标模型。学习时，通过极大似然估计得到条件概率模型$\hat{P}(Y|X)$；预测时，对于给定的输入序列 x，求出条件概率$\hat{P}(y|x)$最大的输出序列$\hat{y}$。

2.4 深度学习算法

在早期研究中，诸如决策树、支持向量机、条件随机场、隐马尔可夫模型等传统机器学习方法在命名实体识别和实体关系抽取任务上取得了一定成果。然而，这些方法依赖大量领域知识和复杂的特征工程。随着深度学习技术的发展，基于神经网络的方法以其端到端的优点为人们所广泛使用，下面将介绍本文所使用的深度学习算法。

2.4.1 长短期记忆神经网络

长短期记忆神经网络[45]（Long Short-Term Memory Network，LSTM）是循环神经网络的一个变体，可以有效解决循环神经网络的梯度爆炸或消失问题。长短期记忆神经网络的循环单元结构如图 5 所示。

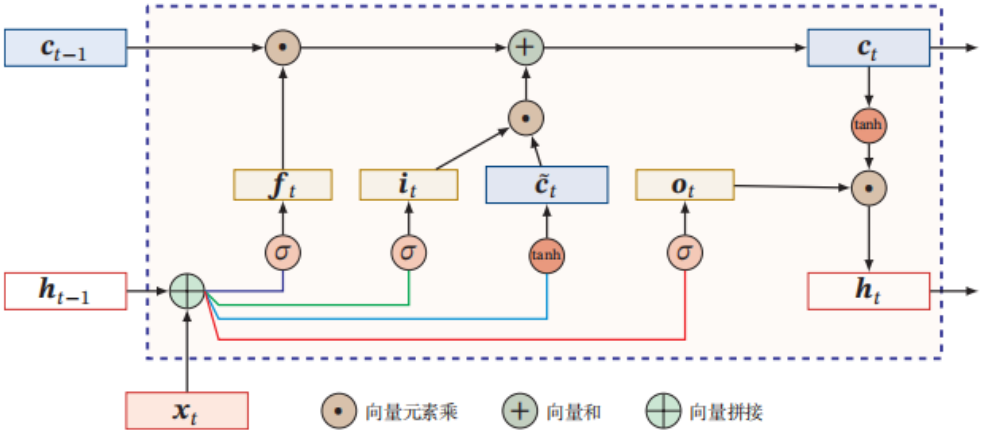

图 5　长短期记忆神经网络的循环单元结构[46]146

长短期记忆神经网络主要在以下两个方面进行了改进。首先，长短期记忆神经网络引入了新的内部状态 $c_t \in R^D$ 来专门传递线性的循环信息，同时将非线性信息输出给隐藏层的外部状态 $h_t \in R^D$。内部状态 c_t 通过公式(2.4)和(2.5)计算：

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{2.4}$$

$$h_t = o_t \odot tanh(c_t) \tag{2.5}$$

其中 $f_t \in [0,1]^D$，$i_t \in [0,1]^D$，$o_t \in [0,1]^D$ 是三个控制信息传递路径的门，$\odot$为向量元素乘积，c_{t-1} 是上一时刻的内部状态，$\tilde{c}_t \in R^D$ 是候选状态，可通过公式(2.6)得到。在任意时

刻 t，内部状态 c_t 记录了到当前时刻为止的历史信息。

$$\tilde{c}_t = tanh(W_c x_t + U_c h_{t-1} + b_c) \tag{2.6}$$

其次，长短期记忆神经网络引入了门控机制（Gate Mechanism）来控制信息传递的路径。公式(2.4)和(2.5)中的三个门分别是遗忘门 f_t，输入门 i_t，输出门 o_t。遗忘门 f_t 控制上一时刻的内部状态 c_{t-1} 有多少信息需要遗忘；输入门 i_t 控制当前时刻的候选状态$\tilde{c}_t$ 有多少信息需要保存；输出门 o_t 控制当前时刻的内部状态 c_t 有多少信息需要传递给外部状态 h_t。三个门的取值在(0, 1)之间，表示以一定比例允许信息通过。三个门的计算方式如公式(2.7)，(2.8)和(2.9)所示。

$$i_t = \delta(W_i x_t + U_i h_{t-1} + b_i) \tag{2.7}$$

$$f_t = \delta(W_f x_t + U_f h_{t-1} + b_f) \tag{2.8}$$

$$o_t = \delta(W_o x_t + U_o h_{t-1} + b_o) \tag{2.9}$$

其中，δ(.)表示 Logistic 函数，其输出区间为(0, 1)，x_t 为当前时刻输入，h_{t-1} 为上一时刻的外部状态。图 5 反映了公式(2.4)至(2.9)的计算过程。

循环神经网络的隐状态 h_t 可以存储的信息是有限的，随着 h_t 存储的信息越来越多，其丢失的信息也越来越多。长短期记忆神经网络通过引入门控机制，其记忆单元 c 可以有选择地保留更多关键信息，但作为循环神经网络的一个变体，长短期记忆神经网络的记忆容量（Memory Capacity）依然是有限的，所以长短期记忆神经网络也没有彻底解决长程依赖问题。

2.4.2 Transformer 与 BERT

Google 公司的机器翻译团队于 2017 年发表了一篇名为《attention is all you need》[47]的文章，并提出了一种只依赖注意力机制（Attention Mechanism）的网络架构，名为 Transformer。时至今日，Transformer 已经成为继多层感知机、卷积神经网络和循环神经网络之后的第四大类深度学习模型，并被广泛应用于各类自然语言处理任务中。

Transformer 是一种编码器-解码器（Encoder-Decoder）模型，其模型架构如图 6 所示。对输入的文本序列进行分词得到单词序列，使用词向量作为单词的表示。Transformer 使用位置编码（Positional Encoding）将文本的顺序信息融入模型。叠加位置向量和单词的词向量得到编码器的输入向量。

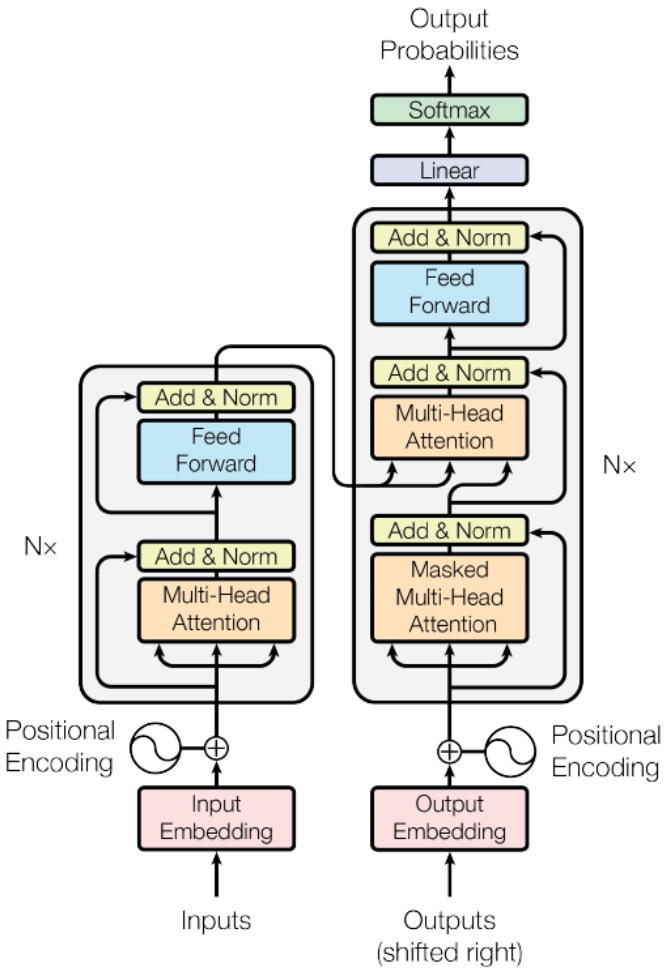

图 6　Transformer 模型结构[47]

编码器模块包含 6 个编码器，解码器模块包含 6 个解码器，即图中 N 的取值为 6。在编码器模块中，每一个编码器的输入是上一个编码器的输出；而在解码器模块中，每一个解码器的输入包括上一个解码器的输出和所有编码器的输出。为了解决梯度消失问题，Transformer 模型在编码器和解码器中都采用了残差神经网络结构，即每一个前馈神经网络的输入包括自注意力机制的输出和原始输入。

每一个编码器由一个自注意力机制（Self-Attention Mechanism）和一个前馈神经网络组成。自注意力机制，又称自注意力模型，旨在捕捉输入数据或特征的内部相关性，从而将有限的计算资源用于处理更重要的信息。自注意力机制在自然语言处理领域中的应用，主要是通过计算单词间的相关性来解决长距离依赖问题。假设输入序列 $X=[x_1, \ldots, x_N] \in R^{Dx*N}$，输出序列 $Y=[y_1, \ldots, y_N] \in R^{Dv*N}$，自注意力机制的计算过程如图 7 所示。

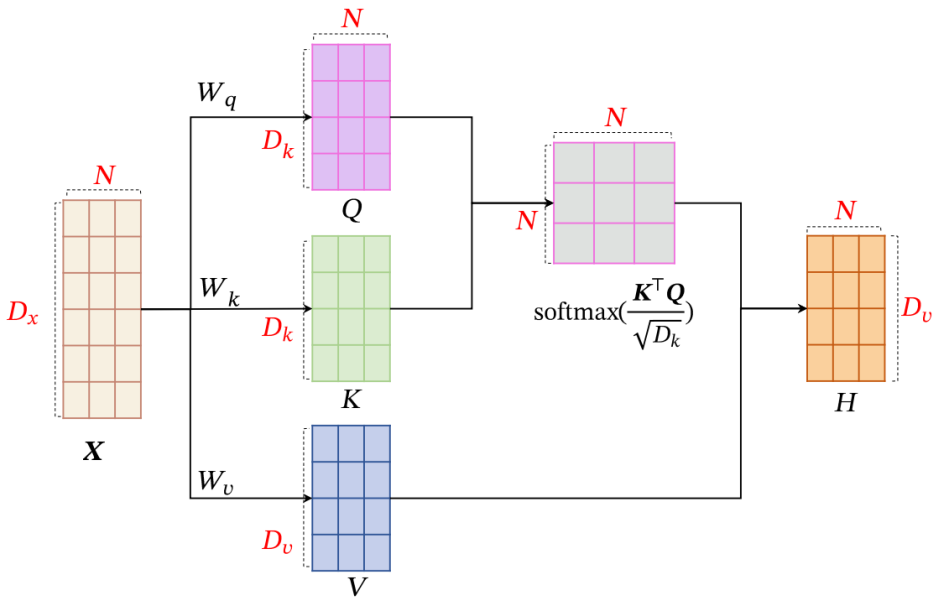

图 7 自注意力机制计算过程[46]204

自注意力机制的计算过程描述如下：

（1）对于每个输入 x_i，将 x_i 线性映射到三个不同空间，得到查询向量 $q_i \in R^{Dk}$、键向量 $k_i \in R^{Dk}$ 和值向量 $v_i \in R^{Dv}$，对于输入序列 X，线性映射过程如公式(2.10)至(2.12)所示。

$$Q = W_q X \in R^{Dk*N} \tag{2.10}$$

$$K = W_k X \in R^{Dk*N} \tag{2.11}$$

$$V = W_v X \in R^{Dk*N} \tag{2.12}$$

其中，$W_q \in R^{Dk*Dx}$，$W_k \in R^{Dk*Dx}$，$W_v \in R^{Dv*Dx}$ 是线性映射的参数矩阵，$Q = [q_1, \ldots, q_N]$，$K = [k_1, \ldots, k_N]$，$V = [v_1, \ldots, v_N]$分别为由查询向量、键向量、值向量构成的矩阵。

（2）对于每一个查询向量 $q_m \in Q$，根据公式(2.13)得到输出向量 h_m。

$$\begin{aligned} h_m &= att((K,V),q_m) \\ &= \sum_{n=1}^{N} \alpha_{mn} v_n \\ &= \sum_{n=1}^{N} SoftMax(s(k_n,q_m)) v_n \end{aligned} \tag{2.13}$$

其中，$m, n \in [1, N]$分别是输出向量序列和输入向量序列的位置，α_{mn}是注意力分布，表示第 m 个输出关注到第 n 个输入的权重。多头注意力机制（Multi-Head Attention Mechanism）使用多个查询向量来并行地从输入信息中捕获多组重要信息，每个注意力关注输入信息的不同部分。近几年的研究集中在提高 Transformer 模型的效率，包括简化自注意力机制的计算[48]和使用稀疏注意力机制[49][50]等。

BERT[51]（Bidirectional Encoder Representation from Transformers）是一个预训练模型，其模型结构为 N 个 Transformer 编码器顺序连接而成，如图 8 所示。BERT 在一个包含 33 亿单词的语料库上通过无监督方式进行预训练。预训练包括两个任务，第一个任务是用掩码[Mask]随机替换 15%的单词，然后让模型预测这些被替换的单词；第二个任务是判断两个句子的关系（是否为上下句关系），每个训练样本是一个上下句，50%的样本，其上句和下句是真实的，另外 50%的样本，其上句和下句是无关的。将两个任务的损失函数加起来作为总损失函数进行优化。

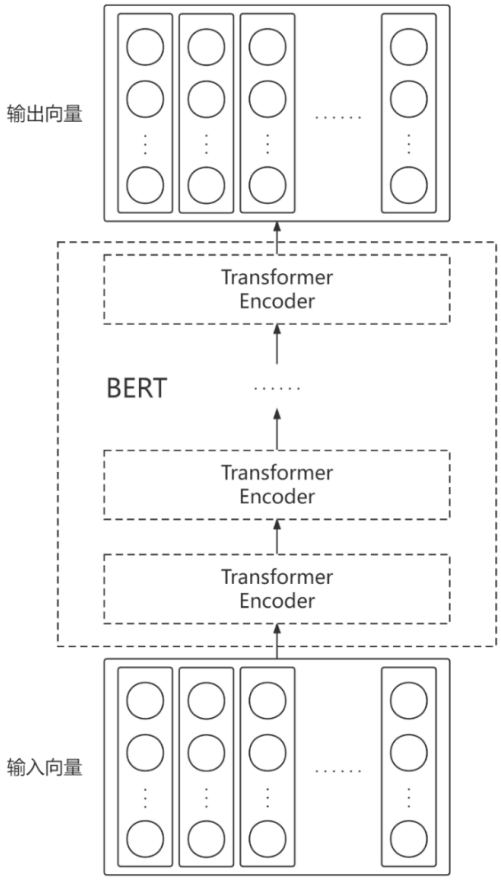

图 8　BERT 模型结构

输入 BERT 的每一个词（token）都会被表示成一个向量，该向量由 3 个部分组成，第一个是 token 向量（token embedding），第二个是表示上下句的向量（segment embedding），第三个是 Transformer 模型需要的位置编码（position embedding），将三个向量相加作为输入。[CLS]表示 Classification，是 BERT 模型用于文本分类任务的标记，它对应的输出向量为包含句子语义信息的句向量。[SEP]表示 Separation，是 BERT 模型用于分隔上下句的标记。BERT 模型的最大输入长度为 512 个 token，其输入表示如图 9 所示。

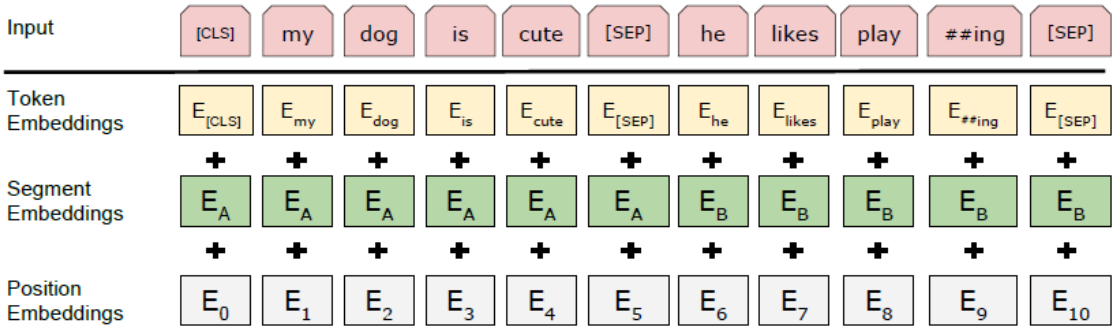

图 9 BERT 模型输入表示[51]

目前有两种方法在下游任务中应用预训练模型的表示：基于特征的方法（feature-based approach）和基于微调的方法（fine-tuning approach）。前者针对具体任务设计特定的模型结构，然后使用数据训练模型参数，训练完成后将参数固定作为表示，例如叠加预训练好的 BERT 模型的后四个隐藏层参数作为输入单词的词向量；后者会在预训练模型的基础上，对模型结构进行小的调整，例如在预训练好的 BERT 模型上面加一个 SoftMax 网络，然后在有监督数据上重新训练（微调），此过程中模型参数会发生改变，训练好的模型可应用于下游任务。

2.4.3 知识蒸馏

在实验室环境中，研究者设计复杂的模型并使用大量计算资源以获得更好的结果，但在实际生产中，由于延迟和部署资源等因素的限制，大模型不方便部署到服务中，模型压缩便应运而生。知识蒸馏（Knowledge Distillation）是一种模型压缩方法，旨在将已经训练好的大模型所包含的知识蒸馏到一个小模型中。

知识蒸馏使用了教师学生模型（Teacher-Student Model），即教师（大模型）是知识的输出者，学生（小模型）是知识的接受者。知识蒸馏的过程可以分为两个阶段，第一阶段训练教师模型，教师模型的输出必须经过 SoftMax 映射，得到相应类别的概率值。第二阶段训练学生模型，学生模型具有相对简单的模型结构和更少的参数，且学生模型的输出同样为经过 SoftMax 映射后的类别概率向量。

知识蒸馏的目的是让学生模型学习教师模型的泛化能力，而非简单地拟合训练集的真值标签（ground truth）。教师模型的泛化能力体现在它输出的 SoftMax 概率值中，因为这些概率值是通过样本与模型参数计算得到的。知识蒸馏就是把大模型输出的类别概率向量作为软目标（Soft Target），让小模型的输出（类别概率向量）与软目标尽量靠近。由于教师模型存在一定的错误率，引入硬目标（Hard Target）即样本的真值标签作为参

考，可以降低错误被传播给学生模型的可能性。蒸馏的目标函数由蒸馏损失（软目标）和学生损失（硬目标）加权得到，如公式 2.14 所示。

$$L = \alpha L_{soft} + \beta L_{hard} \tag{2.14}$$

其中，α 和 β 为权重系数，需要人工设置。L_{soft} 为蒸馏损失，可通过公式(2.15)计算；L_{hard} 为学生损失，可通过公式(2.16)计算。

$$L_{soft} = -\sum_j^N p_j^T \log\left(q_j^T\right) \tag{2.15}$$

$$L_{hard} = -\sum_j^N c_j \log\left(q_j^1\right) \tag{2.16}$$

其中，N 为标签总数，T 为温度，$p_j^T = \frac{\exp(v_j/T)}{\sum_k^N \exp(v_k/T)}$为温度 T 下教师模型的 SoftMax 输出在第 j 类上的值，$q_j^T = \frac{\exp(z_j/T)}{\sum_k^N \exp(z_k/T)}$为温度 T 下学生模型的 SoftMax 输出在第 j 类上的值，$c_j \in \{0,1\}$为第 j 类的真值标签，正标签取 1，负标签取 0。当 T=1 时，p_j^T和q_j^T退化为 SoftMax 函数。引入温度 T 的目的是放大 SoftMax 输出的小概率值所携带的信息，T 值越大，类别概率向量的分布越平滑，负标签携带的信息会被相对放大，模型在训练时将更加关注负标签。

2.5 本章小结

本章主要介绍了知识图谱构建技术、信息抽取、条件随机场、深度学习算法等基础理论知识。首先介绍了知识图谱的构建流程与相关技术。信息抽取是面向非结构化数据构建知识图谱的主要技术，因此本文详细阐述了命名实体识别和实体关系抽取的流程。最后，为实现本文提出的威胁情报命名实体识别方法和威胁情报实体关系抽取方法，对本文方法研究部分用到的深度学习算法进行了介绍。

第三章　威胁情报本体构建

本章主要研究威胁情报领域的本体构建方法并阐述本文构建的威胁情报本体。威胁情报本体包含威胁情报领域各类实体及实体间关系，可以从语义层面实现对威胁实体的组织，指导威胁情报知识图谱的构建。

3.1 本体构建思想

互联网中存在大量非结构化的威胁情报报告，这些报告来源于不同的安全组织和厂商，因而即使报告的内容相似，其格式却不尽相同。这种“数据孤岛”的存在形式使得威胁情报难以被整合。此外，现有的威胁情报表达方式存在着一定不足，无法充分表达威胁情报的语义信息，进而导致某些高价值威胁情报的语义关联信息丢失。因此，构建威胁情报本体在当前情况下非常有必要。

本体是对客观事物的系统性描述，常被用于研究某个领域客观现象的本质。由于本体可以概念化地表达领域知识，描述特定领域的对象类型、概念和关系，因而被广泛应用于多个学科领域，如生物医学领域的基因本体和疾病本体，自然语言处理领域的WordNet 本体等。

现有威胁情报本体在设计时大多参考国外成熟的威胁情报标准，包括指标信息的可信自动化交换（TAXII）、网络可观察表达式（CyboX）和结构化威胁信息表达式（STIX）等。Gao 等人[52]基于 STIX 1.0 和 STIX 2.0 构建了威胁情报领域本体，该本体共包含 6 类威胁对象，分别是攻击目标、攻击者、漏洞、TTP（Tactic、Technique、Process）、观测指标和防御措施，主要对攻击事件进行描述。Wang[53]提出了一个威胁情报本体，该本体包含 6 类威胁对象，分别为攻击组织、攻击成员、攻击目标、攻击意图、攻击手段和攻击控制源，为后续威胁情报实体抽取和关系抽取定义标准。现有威胁情报本体多是针对特定任务而构建的，因此这些本体所涵盖的实体类型较为有限，不具备很好的可移植性和参考价值。此外，随着部分威胁情报标准的更新，现有威胁情报本体在构建时参考的威胁情报标准已经过时。

本文在构建威胁情报本体时，考虑了威胁情报的语义内涵和现有最新的威胁情报相关标准。通过对威胁情报领域的概念与标准进行分析整合，确定了威胁情报本体所包含的 13 类威胁实体和 7 种实体关系类型。本文所构建的威胁情报本体可以在宏观上指导非结构化威胁情报数据的整合以及相关知识图谱的构建工作。

3.2 威胁情报相关标准

本体的目的是对某个领域的知识进行汇总提炼，抽象出该领域的本质概念。因此，在构建威胁情报本体之前，需要首先对威胁情报领域的知识和语义概念进行梳理汇总。本文通过解析威胁情报的语义概念和威胁情报领域现有相关规范标准确定威胁情报本体。

STIX（Structured Threat Information eXpression）是目前威胁情报领域应用最为广泛的标准，由 MITRE 公司与美国国土安全部（DHS）于 2013 年 4 月联合制定发布，发布之初的版本为 STIX 1.0 版本。我国的威胁情报相关标准 GB/T 36643-2018 参考了 STIX 1.0 中的内容。随着不断的更新完善，STIX 于 2017 年 7 月推出了 STIX 2.0 版本，这一代 STIX 融合了 CybOX [54]标准。目前最新的 STIX 是 2020 年 3 月推出的 STIX 2.1[55]版本，该版本涵盖内容更广且威胁实体表达清晰规范。本文以 STIX 2.1 版本作为主要参考的威胁情报标准。

STIX 2.1 总共定义了 18 类威胁实体（STIX 对象），包括 Attack Pattern、Indicator、Campaign、Intrusion Set、Course of Action、Grouping、Identity、Infrastructure、Observed Data、Location、Malware、Malware Analysis、Threat Actor、Tool、Vulnerability、Report、Note、Opinion，下面对这些威胁实体的语义内涵和潜在关系进行分析。第一类和第二类威胁实体是 Attack Pattern 和 Indicator，即攻击模式和攻击指标，他们表示攻击者用来对目标实施侵害的方法、技术和相关特征指标。两者都属于连接攻击者和攻击目标的中心节点，对威胁事件的表达较为重要。第三类和第四类威胁实体是 Campaign 和 Intrusion Set，即攻击活动和入侵集合，入侵集合是一组具有共同属性的敌对行为和资源，被认为是由单个组织精心策划的，而攻击活动属于入侵集合的一部分。两者在威胁事件中难以具体体现，需要对一系列威胁事件进行汇总分析才能发现。第五类威胁实体是 Course of Action，即应对措施，应对措施是指为防止攻击或对正在进行的攻击做出反应而采取的措施，包括对行动的文字描述。应对措施与攻击模式、漏洞、恶意软件等威胁实体存在关联。

第六类威胁实体是 Grouping，即团队协作，团队协作表示在分析和调查过程中产生的数据，如待确认的线索；还可以用来表明某个威胁实体与正在进行的分析过程有关。Grouping 对象与其他 STIX 对象之间并没有明确定义的关系。第七类威胁实体是 Identity，即身份，身份可以表示个人、团伙或组织；也可以表示一类个人、团伙或组织。身份在

描述威胁事件时会有多个角色，它既可以表示攻击者或攻击者所冒充者的身份，也可以表示攻击目标的身份。第八类和第九类威胁实体是 Infrastructure 和 Obeserved Data，即基础设施和可观测数据，前者表示攻击过程中涉及的物理或虚拟资源，如攻击者使用的 C2 服务器，被攻击目标的数据库服务器等，后者表示攻击过程中观察到的网络安全数据，如域名、IP 地址等。两者之间存在对应关系，例如某个 IP 地址对应于某个 C2 服务器。第十类威胁实体是 Location，即地点，地点可以与身份或入侵集合相关联，表示其位置，也可以与威胁主体或恶意软件相关联，表示其目标。第十一和十二类威胁实体是 Malware 和 Malware Analysis，即恶意软件和恶意软件分析，两者存在一一对应的关系，恶意软件分析对应于具体的恶意软件。

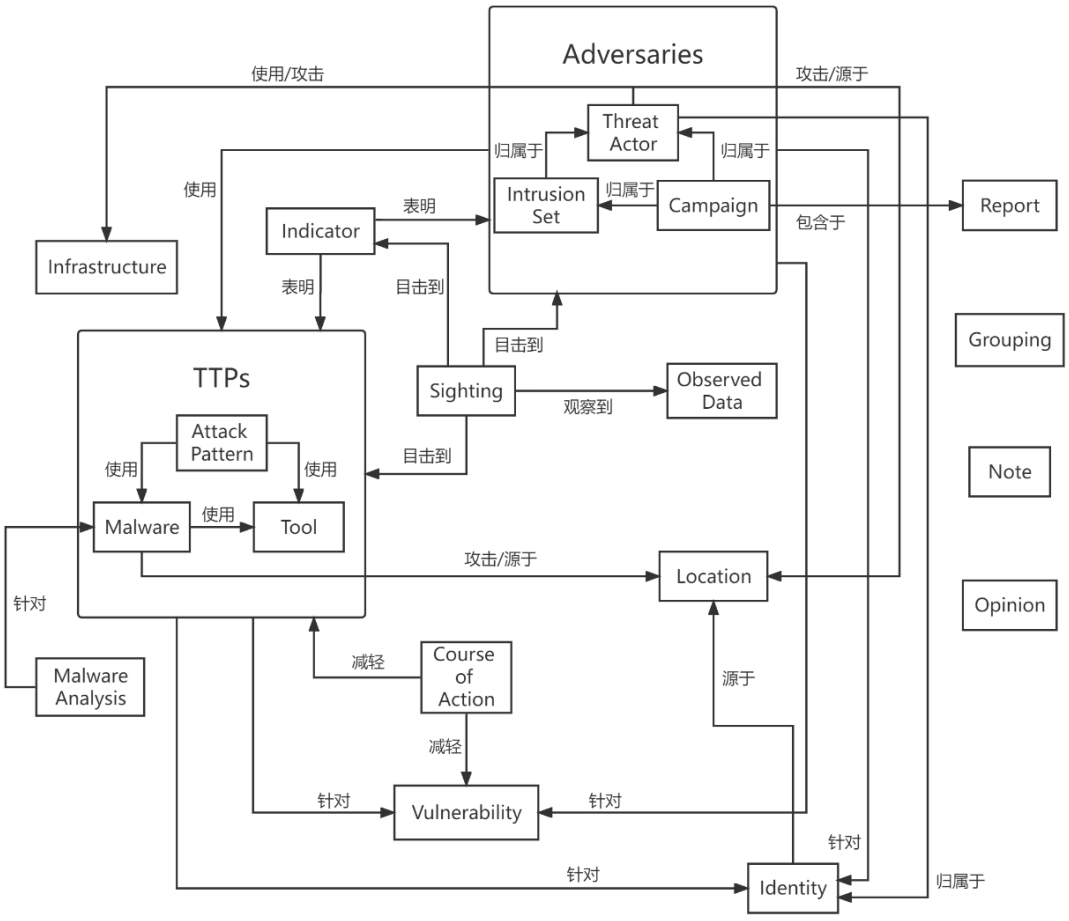

图 10　STIX 2.1

第十三类威胁实体是 Threat Actor，即威胁主体，威胁主体指发起攻击的个人、团伙或组织。威胁主体在实际威胁情报报告中以 APT 组织为主，也是威胁事件的核心对象。第十四类威胁实体是 Tool，即工具，工具指可以被威胁主体用于执行攻击的合法软件，如 Nmap，VNC 等。工具由于存在某些漏洞而被威胁主体利用，与威胁主体、攻击目标、

漏洞存在关联。第十五类威胁主体是 Vulnerability，即漏洞，漏洞是软件或硬件在设计或实现过程中存在的缺陷。在威胁情报领域，漏洞通常以 CVE 编号作为标识符。第十六类威胁实体是 Report，即报告，这里的报告指威胁情报报告，表示与某个威胁事件相关的参考信息。最后两类威胁实体是 Note 和 Opinion，即注释和评估意见，前者包括其他 STIX 对象中不存在的额外信息，如分析人员在攻击活动对象中添加的注释，后者是对威胁实体信息正确性的评估。STIX 2.1 中涉及的威胁实体和关系如图 10 所示。

3.3 威胁情报本体设计

STIX 标准在创建之初的定位是一种用于交换网络威胁情报的语言和序列化格式，其应用场景包括自动化威胁检测和响应、自动化威胁情报交换、协同威胁分析等。各大安全组织和厂商在交换、共享威胁情报数据时常使用 STIX 作为数据表达的标准。但在构建威胁情报本体，尤其是应用于知识图谱领域的威胁情报本体时，部分 STIX 2.1 中的威胁实体可能并不适用。下面就 STIX 2.1 中的 18 类威胁实体是否适用于本文提出的威胁情报本体进行分析。本文构建威胁情报知识图谱所使用的数据以 APT 威胁情报报告为主，因此以下分析会涉及 APT 报告相关内容。

首先是 Indicator、Infrastructure 和 Observed Data。Indicator 表示攻击过程中的失陷指标（IOC）；Observed Data 表示攻击过程中可以观察到的数据，在 APT 报告中以失陷指标为主；Infrastructure 指攻击活动中涉及的物理或虚拟资源，在 APT 报告中通常使用 IP 或域名来描述 APT 组织所使用的资源，如 C2 服务器。三者存在语义上的重叠，都包含失陷指标相关内容，因此将三者细化为 URL、域名和 IP，并归入 Infrastructure 类别。

其次是 Threat Actor、Campaign 和 Intrusion Set。Campaign 是 Intrusion Set 的一部分，APT 报告通常描述某个 Campaign 或某个 Campaign 的部分阶段，因此将两者合并为 Campaign。Threat Actor 作为 Campaign 的主角，与 Campaign 存在参与关系，因此予以保留。

然后是 Attack Pattern、Malware、Malware Analysis 和 Tool。对于 Malware 和 Malware Analysis，威胁情报领域有专门的恶意软件分析报告，但 APT 报告通常只包含恶意软件名称和功能等基本描述，不包括对该恶意软件的分析，因此将 Malware 和 Malware Analysis 合并为 Malware。而 Attack Pattern、Malware 和 Tool 三者属于 TTP 的范畴，在 APT 报告中 Attack Pattern 会被细化为 APT 组织所使用的技术和相应攻击流程的描述，因此使用 Technique 代替 Attack Pattern，并保留 Malware 和 Tool。

接着是 Course of Action、Vulnerability、Identity 和 Location。Course of Action 描述了应对 TTP 和某些安全漏洞的方法和措施；Location 是 APT 报告中的常见信息，也是构成威胁情报知识图谱的重要信息，包括 APT 组织、恶意软件等实体的来源地点和被攻击的地点；Identity 能够以不同身份出现在 APT 报告中，如 APT 组织、被攻击的机构等。因此，对上述 4 类威胁实体予以保留。此外，行业（Industry）作为被 APT 组织攻击的目标，常与地点一起出现，因而将 Industry 添加为威胁实体。

最后是 Grouping、Report、Note 和 Opinion。STIX 2.1 中没有明确定义 Grouping、Note 和 Opinion 与其他 STIX 对象之间的关系，且三者更适用于威胁情报交换或协同威胁分析等场景。对于构建威胁情报知识图谱，这三类威胁实体并不是必须的。Report 在本文中以 APT 报告实例的形式作为数据源，因而不单独将其列为一类威胁实体。因此，以上 4 类威胁实体不予以保留。

本文提出的威胁情报本体包括 13 类威胁实体，分别是 Threat Actor、Campaign、Malware、Technique、Tool、Identity、Location、Industry、Vulnerability、Course of Action、URL、Domain、IP。此外，还有 7 种实体间关系，分别是使用、攻击、源于、相似、相同、拥有、应对。威胁情报本体如图 11 所示。

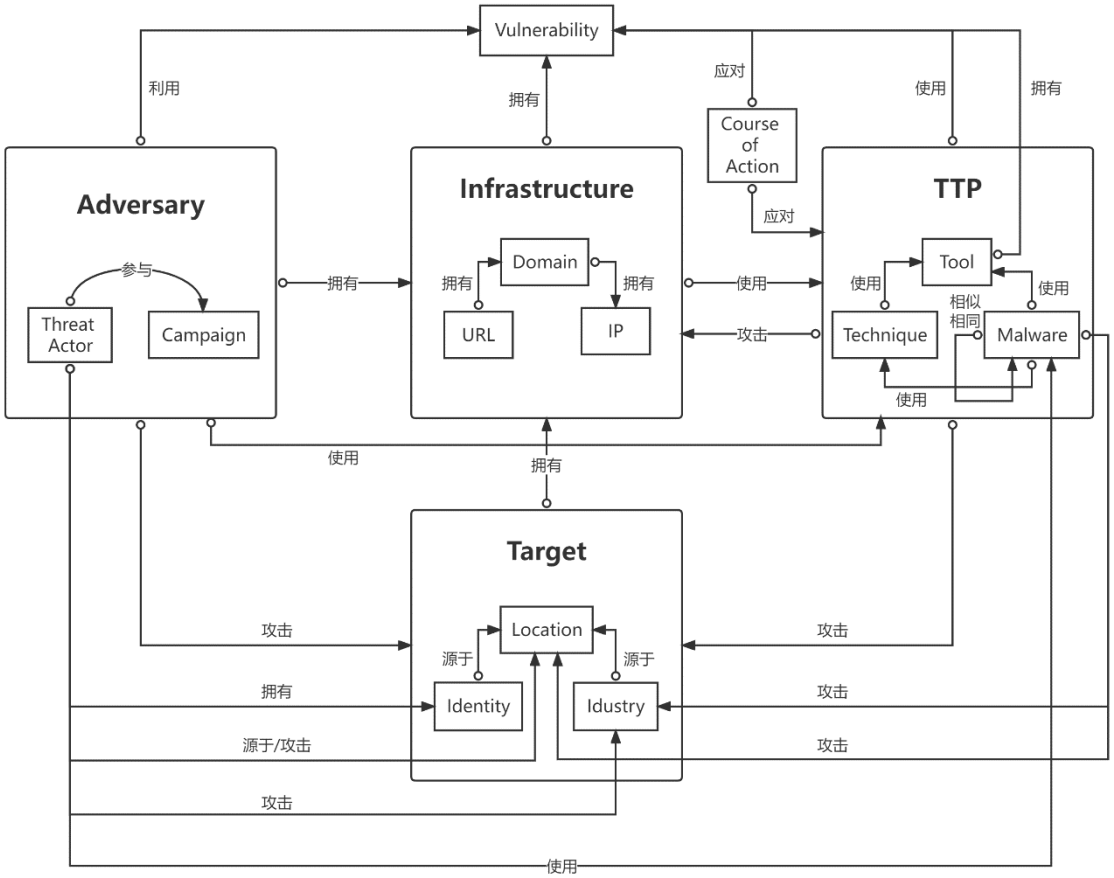

图 11　威胁情报本体

3.4 基于威胁情报本体的知识图谱实体与关系定义

本文构建的威胁情报本体，其目的是指导威胁情报知识图谱的构建。在构建威胁情报知识图谱时，需要明确图谱中包含哪些类型的节点和边，即需要定义威胁情报知识图谱的实体类型和关系类型。本节阐述基于威胁情报本体的威胁情报知识图谱构建方法总体设计。

前文提到本体是对领域中概念的抽象，但在设计威胁情报知识图谱的实体类型和关系类型时，需要结合数据实例进行分析考虑，因为威胁情报知识图谱中的实体和关系来源于威胁情报文本数据，这些实体和关系是本体中概念的具化。部分实体可能同时属于多个概念，或与多个概念有交集。例如，图 12 中的句子表示 APT 组织 TEMP.Periscope 攻击了英国的能源公司，APT 组织 TEMP.Periscope 属于 Threat Actor，同时也属于 Identity。同理，能源公司这个实体既属于 Industry 也属于 Identity。根据本体中的定义，身份可以表示（一类）个人、团伙或组织，当它被实体化为犯罪团伙 TEMP.Periscope 时，与威胁主体概念产生了交集；当它被实体化为能源公司时，与行业概念产生了交集。因此，在参考威胁情报本体设计威胁情报知识图谱中的实体类型和关系类型时，可以对本体中概念存在交集的对象进行合并。由于 APT 报告中 Identity 多以团伙和组织的实例出现，如 APT 组织，行业公司，国家等，本文没有将 Identity 作为实体类别，而保留 Threat actor（Attacker），Location 和 Industry 作为威胁情报知识图谱的实体类别。

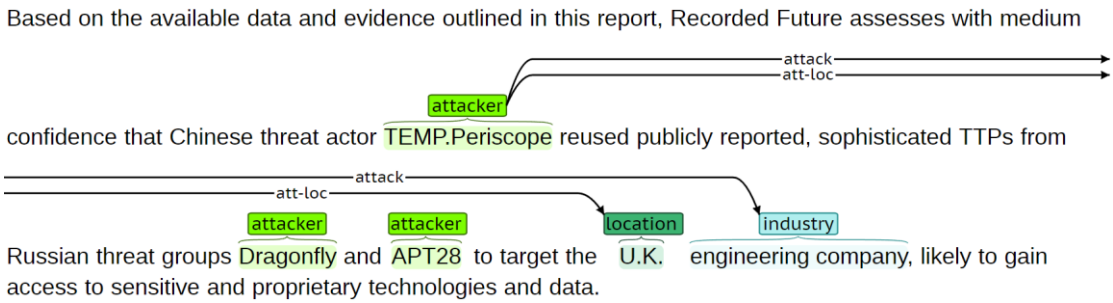

图 12　威胁情报数据示例

对于 Campaign，APT 报告的主要内容确实是对攻击活动的描述，但攻击活动所涵盖的内容较为广泛，它可能包含 Threat actor、Malware、Technique、Tool、Location 等实体。因而在构建威胁情报知识图谱时，Campaign 无法作为图谱中的单个节点，故本文没有将 Campaign 作为威胁情报知识图谱的实体类别。

对于 URL 和 Domain，两者在概念上存在交集。虽然理论上 URL 包含 Domain，但在威胁情报领域，Domain 的使用场景更多，例如，安全机构或厂商会专门收集恶意域名

建立相关知识库，或者根据 APT 组织所使用的恶意域名的命名规则 分析该组织的特征并建立组织画像。因此，本文仅保留 Domain 作为威胁情报知识图谱的实体类别。

对于 Course of Action，APT 报告很少提及攻击的应对措施，更多的是在介绍与 APT 组织攻击手法有关的内容。因此，本文没有将 Course of Action 作为威胁情报知识图谱的实体类别。对于剩下的 Malware、Technique、Tool、Vulnerability 和 IP，以上威胁实体都是 APT 报告的重要组成部分，因为对 APT 组织攻击手法的描述，其主要内容涵盖 APT 组织所使用的恶意软件、技术和工具，APT 组织所利用的漏洞以及 C2 服务器的 IP。因此，本文保留 Malware、Technique、Tool、Vulnerability 和 IP 作为威胁情报知识图谱的实体类别。表 1 梳理了威胁情报知识图谱所包含的实体类型。

表 1　威胁情报知识图谱实体类型

实体类型	中文名称	说明或示例
Attacker	攻击组织	APT 组织
Malware	恶意软件	APT 组织使用的 恶意软件
Technique	技术	TTP 类型之一， APT 组织使用的技术
Tool	工具	APT 组织在攻击活动中 使用的合法软件
Vulnerability	漏洞	APT 组织在攻击活动中 利用的漏洞
IP	IP 地址	例如，C2 服务器的 IP 地址
Domain	域名	例如，钓鱼网站的域名
Location	地点	一般为国家或地区名称 可以是 APT 组织的所在地点，也可以是攻击目标的所在地点
Industry	行业	常作为攻击目标出现

除实体类型外，还需要明确威胁情报知识图谱的关系类型。结合威胁情报数据实例和前文给出的威胁情报本体所包含的对象及其关系，本文将威胁情报知识图谱的实体间关系归纳为以下 5 类：攻击（attack）、使用（use）、源于（from）、相似（similar）、相同（same）。表 2 列举了不同类型实体之间的关系，以三元组的形式给出。

表 2　威胁情报知识图谱关系类型

实体 1	关系	实体 2
Attacker	attack	Location
Attacker	attack	Industry
Malware	attack	Location
Malware	attack	Industry
Attacker	use	Malware
Attacker	use	Technique
Attacker	use	Tool
Attacker	use	Vulnerability
Malware	use	Malware
Malware	use	Technique
Malware	use	Tool
Malware	use	Vulnerability
Technique	use	Vulnerability
Attacker	from	Location
Malware	from	Location
Industry	from	Location
Attacker	similar	Attacker
Malware	similar	Malware
Attacker	same	Attacker
Malware	same	Malware

3.5 本章小结

本章主要介绍了威胁情报本体的构建和威胁情报本体的应用方法。首先介绍了本体的构建思想，包括本体的定义和作用。然后介绍了威胁情报领域应用最为广泛的标准 STIX 2.1，对其中的威胁实体和关系进行了详细阐述。接着，通过分析 APT 报告的特点并对威胁情报标准 STIX 2.1 进行提炼，本文构建了威胁情报本体。最后，本文阐述了如何根据威胁情报本体来定义威胁情报知识图谱的实体和关系类型。

第四章　基于数据增强与 BERT 的威胁情报实体抽取方法

本章介绍论文提出的基于数据增强与 BERT 的威胁情报实体抽取方法。首先对威胁情报的实体特点进行分析，然后针对威胁情报实体数据的缺陷，提出了新的数据增强方法。在方法部分，首先阐述了常见 NLP 数据增强方法的局限性，然后对知识库与模板设计和模板填充算法进行了详细说明，并将提出的数据增强方法与 BERT 结合用于威胁情报命名实体识别模型，最终通过实验对本章提出方法的有效性进行了验证。

4.1 设计思想

本文所使用的威胁情报数据是非结构化的 APT 报告，这类报告主要描述 APT 组织及其攻击活动，包括该组织所使用的技术、工具、恶意代码、目标地区、目标行业等。APT 报告通常将失陷指标（IOC）以附录形式在文末给出，附录 IOC 示例如图 13 所示。

Appendix I – Indicators of Compromise

Note: *The indicators in this section are valid at the time of publication. Any future changes will be directly updated in the corresponding .ioc file.*

File Hashes (malicious documents, trojans, emails, decoys)

Ecipekac loader
be53764063bb1d054d78f2bf08fb90f3 jli.dll P8RAT
cca46fc64425364774e5d5db782ddf54 vmtools.dll SodaMaster
dd672da5d367fd291d936c8cc03b6467 CCFIPC64.DLL FYAnti loader

Encrypted Ecipekac Layer II, IV loader (shellcode)
md5 filename payloads
f60f7a1736840a6149d478b23611d561 vac.dll P8RAT
59747955a8874ff74ce415e56d8beb9c pcasvc.dll P8RAT
4638220ec2c6bc1406b5725c2d35edc3 wiaky002_CNC1755D.dll SodaMaster
d37964a9f7f56aad9433676a6df9bd19 c_apo_ipoib6x.dll SodaMaster
335ce825da93ed3fdd4470634845dfea msftedit.prf.cco FYAnti loader
f4c4644e6d248399a12e2c75cf9e4bdf msdtcuiu.adi.wdb FYAnti loader

Encrypted QuasarRAT
md5 filename payloads
019619318e1e3a77f3071fb297b85cf3 web_lowtrust.config.uninstall QuasarRAT

Domains and IPs

151.236.30[.]223
193.235.207[.]59
45.138.157[.]83
88.198.101[.]58
www.rare-coisns[.]com

图 13　附录 IOC 示例

附录 IOC 主要包括恶意文件名及其哈希值（通常为 MD5）、恶意 IP、恶意域名和恶意 URL。这样做的目的是方便威胁情报用户，如企业、机构等，快速配置其防火墙黑名单，或构建恶意 IP、域名知识库。APT 报告在正文部分介绍攻击活动时，只列举部分具有代表性的 IOC，而许多同源 IOC 则直接在附录中给出，这导致正文部分的 IOC 数量较少。对于命名实体识别任务，其输入通常是一个句子，输出是与该句子对应的标签序列，一般不对附录中的内容进行识别，因为它们不具有上下文环境且格式不统一。所以，APT 报告中的 IOC 实体数量少于其他类型实体的实体数量。表 3 是 306 篇人工标注威胁情报的实体统计信息。

表 3　306 篇人工标注威胁情报的实体统计信息

攻击组织	恶意软件	地点	行业	技术	工具	漏洞	域名	IP
7563	9638	6376	4661	5399	1516	831	281	142

由表 3 可知，IOC（域名和 IP）的实体数量远少于其他类型实体的实体数量，属于稀有类。威胁情报数据存在标签分布不平衡的问题。

威胁情报数据的另一个特点是攻击组织和恶意软件的实体名称过于复杂。导致该问题的原因是威胁情报领域不存在通用命名约定。以 APT 组织为例，出于营销目的，安全厂商和威胁情报提供商热衷于给各个 APT 组织起各种名字。例如，同一个俄罗斯 APT 组织拥有多个名字：APT28，Fancy Bear，Sofacy，Sednit，STRONTIUM，Pawn Storm。根据 MITRE 公司给出的数据，目前全球范围内共有 129 个 APT 组织[56]，但 APT 组织名称却超过 548 个（仅英文）。搜集涵盖这些 APT 组织名称的报告并加以标注，其工作量是巨大的。

本文在标注威胁情报数据时，选取了 15 个 APT 组织相关的报告，共计 306 篇。报告所涵盖的 APT 组织只占 APT 组织总数的 12%，此举是为了降低领域知识带来的标注成本，提高标注效率。这也导致了攻击组织和恶意软件的实体多样性较为有限，不利于模型识别这两类实体的未登录词（Out-of-Vocabulary）。

对于标签分布不平衡问题，删除稀有类是一种解决办法，但 IOC 作为一种重要的威胁情报信息，应当出现在威胁情报知识图谱中，故此办法本文不予采纳；另一种解决办法是增加稀有类的实体数量。对于样本多样性问题，一种有效的做法是通过扩充数据来提高样本多样性，从而让模型更好地学习攻击组织和恶意软件实体的命名特征。因此，

为了解决标签分布不平衡问题和样本多样性问题，本文拟对攻击组织、恶意软件、漏洞、域名和 IP 这 5 类实体进行数据增强，具体方法将在下一节介绍。

4.2 基于知识库和模板填充的数据增强方法

4.2.1 NLP 数据增强方法分析

数据增强（Data Augmentation）是一种从现有训练样本中生成新的训练样本的技术，它缓解了深度学习领域数据不足的问题，在图像领域首先得到广泛应用，近年来延伸到自然语言处理领域，并取得了一定效果。数据增强的主要思路是增加训练数据的多样性，从而提高模型的泛化能力。

现有 NLP 数据增强方法包括：（1）同义词替换，随机将非停用词替换为其同义词。（2）语义嵌入，利用语义向量将单词替换为语义相近的其他单词。（3）回译，把句子翻译成其他语言再翻译回来。（4）随机删除，以某个固定概率随机删除句子中的单词。（5）随机插入，将目标词的同义词随机插入到句子中。

对于威胁情报命名实体识别任务，以上数据增强方法存在不足。首先，部分威胁实体属于专有名词，它们具有特殊含义，不能简单地用同义词或语义相近的单词进行替换。例如，APT 组织 STRONTIUM，该词的中文意思是金属锶，但不能用金属锶的同义词来替换 STRONTIUM，因为在威胁情报领域 STRONTIUM 表示一个 APT 组织而非金属锶。其次，回译通常使用机器翻译，部分威胁情报专有名词在回译过程中会出现错误。例如，APT 组织海莲花，英文名“OceanLotus”，而机器翻译会将海莲花翻译为“Sea lotus”或“Ocean lotus”，导致模型无法正确识别该 APT 组织。最后，随机删除和随机插入会破坏句子内部的上下文信息。在使用预训练语言模型（如 BERT 等）生成词向量时，模型会动态地根据单词的上下文信息来生成该单词的词向量，而随机插入或删除单词会改变句子内部的上下文信息，进而影响词向量的语义准确性。

本文的目的是增加漏洞、域名、IP 这 3 类实体的实体数量，增加攻击组织和恶意软件这两类实体的样本多样性。考虑到现有 NLP 数据增强方法在威胁情报领域存在局限性，本文提出一种面向威胁情报领域的基于知识库和模板填充的数据增强方法。该方法的核心思想是寻找包含待增强类型实体的句子作为模板句子，将知识库中同类型实体填入模板句子生成新的包含特定类型实体的句子，将新生成的句子加入训练集以实现数据增强的目的。

4.2.2 知识库构建与模板设计

知识库（Knowledge Base，KB）的概念来自于两个不同领域，一个来自（人工智能）知识工程领域，另一个来自数据库领域。本文使用的知识库为领域词汇知识库，其概念属于数据库领域。威胁情报领域词汇知识库由威胁情报实体词汇组成，本文共构建了 5 个知识库，涵盖了待增强的实体类型，即攻击组织、恶意软件、漏洞、域名和 IP。

攻击组织和恶意软件知识库可以根据 MITRE 公司提供的 APT 组织名单[56]和恶意软件名单[57]来构建。漏洞知识库可以通过 CVE 漏洞信息库[58]来构建，后者囊括了大部分已经披露的信息安全漏洞或弱点。域名和 IP 知识库可以根据附录 IOC 来构建。攻击组织知识库示例如图 14 所示。

ALLANITE
ANTHROPOIDSPIDER
APT-C-01
APT-C-15
APT-C-34
APT16
APT22
APT27
APT32
APT35
OceanBuffalo
OceanLotus
Anunak
AridViper
ArmaRat
AuroraPanda
Axiom
AyyıldızTim
BITTER
Blackgear
BossSpider
CNC
COBALTJUNO
COBALTKATANA
Callisto
Calypsogroup
Careto
Cobalt
ColdRiver

图 14　攻击组织知识库（部分）

在知识库准备完成后还需要准备模板句子。模板句子来源于 APT 报告，经过筛选得到。筛选模板句子的标准是，如果某个句子包含多个同一目标类型的实体，那么它可以作为模板句子，且归属于相应目标类型。下面举例说明如何构建模板句子。

例句 1：APT28, also known as Fancy Bear, Sofacy, Sednit or Pawn Storm, is a group of attackers operating since 2004.

基于例句 1 定义的模板 1：<attacker>, also known as <attacker>, <attacker>, <attacker> or <attacker>, is a group of attackers operating since 2004.

基于模板 1 生成的句子 1：APT32, also known as DarkHotel, IceFog, MERCURY or Turla, is a group of attackers operating since 2004.

因为例句 1 包含 5 个攻击组织类型的实体，所以它可以作为模板句子，且归属于攻击组织类型。在确定例句 1 作为模板句子后，通过将句子中的 5 个攻击组织实体替换为相应类型的特殊标记（本例中是<attacker>）来构建模板句子。模板 1 是将例句 1 中的攻击组织实体替换为<attacker>标记后得到的模板句子，是真正意义上的模板。

在得到句子模板后，需要使用知识库中的领域词汇对模板句子进行填充以生成新的句子。打乱攻击组织知识库中的词汇，从中选取 5 个攻击组织依次填入模板 1 中的标记处，得到的句子 1 即是最终生成的增强数据。一般来说，需要的增强数据量越大，所应收集的模板句子数量也越多。

虽然每个模板句子可以重复使用，但仍应控制每个模板句子在增强数据中出现的次数，这样是为了防止模型过拟合。设每个模板句子在增强数据中出现的次数为 N，则 N 可以表示为：

$$N = \frac{\text{知识库中领域词汇的数量}}{\text{模板句子中所属类型标记的数量}} \tag{4.1}$$

N 值过大会增加模型过拟合的风险。减小 N 值的办法有：使用所属类型标记较多的句子作为模板句子，或者使用知识库的子集对模板句子进行填充。

本文面向 5 类威胁实体定义了共计 126 个模板句子，包括第一批模板和第二批模板，分别用于增强训练集和测试集以验证提出的数据增强方法能够提升威胁情报命名实体识别模型的泛化性能。表 4 列举了 5 类待增强威胁实体所拥有的模板句子数量：

表 4　待增强威胁实体的模板句子数量

	攻击组织	恶意软件	漏洞	域名	IP
第一批	12	13	13	13	12
第二批	13	12	13	12	13

第一批模板句子和第二批模板句子没有交集，即同一类型威胁实体的两批模板句子之间没有重复。例如，攻击组织的第一批 12 个模板句子和攻击组织的第二批 13 个模板

句子完全不同。用于数据增强的威胁情报模板句子（部分）示例如下所示：

模板 1：<attacker>, also known as <attacker>, <attacker>, and <attacker>, is a highly active and prolific APT.

模板 2：<malware>, <malware> and <malware> are among the most frequently observed backdoors used by <attacker>.

模板 3：The RTF file is a weaponized document that attempts to exploit <vulnerability> and <vulnerability> to drop two files to the system.

模板 4：The malware tries to visit a number of gallery pages hosted on <domain> and <domain>, then fetches the image from the page.

模板 5：FireEye identified three VPN connections involving <ip>, <ip> and <ip> IPs between June 30 and July 1, 2018.

其中，模板 1 是攻击组织类型的模板句子，模板 2 是恶意软件类型的模板句子，因为句子中<malware>标签的数量最多。同理，模板 3 是漏洞类型的模板句子，模板 4 是域名类型的模板句子，模板 5 是 IP 类型的模板句子。

4.2.3 模板填充算法

模板句子在填充前需要转化为 BIO 标注模式，示例如图 15 所示。

```
<attacker> B-attacker
, O
also O
known O
as O
<attacker> B-attacker
, O
<attacker> B-attacker
, O
<attacker> B-attacker
or O
<attacker> B-attacker
, O
is O
a O
group O
of O
attackers O
operating O
since O
2004 O
. O
```

图 15　模板句子转化为 BIO 标注模式

为了让每个知识库中的领域词汇尽可能多地出现在不同的模板句子中，本文设计了模板填充算法。模板填充算法如算法 1 所示。

算法 1：模板填充算法

```
Input：template_sentence, KB
Output：filled_template_sentence
filled_template_sentence = []
for token in template_sentence:
    word = token.split(' ') [0]
    tag = token.split(' ') [1]
    if tag = 'O':   //如果是 O 标签，将单词和标签放回列表
        new_line = word + ' ' + tag + '\n'
        filled_template_sentence.append(new_line)
    else:   //如果不是 O 标签，填充模板
        entity_word = get_entity_word(KB)   //从知识库获取领域词汇
        if len(entity_word) = 1:   //实体包含一个单词
            new_line = entity_word[0] + ' ' + tag + '\n'
            filled_template_sentence.append(new_line)
        else:   //实体包含多个单词
            num_of_word = len(entity_word)
            B_tag = tag
            I_tag = 'I-' + tag.split('-')[1]
            for j in range(num_of_word):
                if j = 0:   //  处理 B-标签
                    new_line = entity_word[j] + ' ' + B_tag + '\n'
                    filled_template_sentence.append(new_line)
                else:   //  处理 I-标签
                    new_line = entity_word[j] + ' ' + I_tag + '\n'
                    filled_template_sentence.append(new_line)
return filled_template_sentence
```

算法的输入是一个模板句子的BIO标注结果（template_sentence）和一个知识库（KB，包含一定数量的领域词汇），算法的输出是填充该模板生成的句子（filled_template_sentence）。对于模板句子 BIO 标注结果中的每一行，分别获取它的单词和标签。如果标签是 O，则将单词和标签拼接后存入列表。如果标签不是 O，从知识库中获取一个领域词汇，判断一下领域词汇由几个单词组成，如果领域词汇由一个单词组成，则将领域词汇与相应标签拼接后存入列表，如果领域词汇由多个单词组成，则将

领域词汇的第一个单词与 B-标签拼接后存入列表，将领域词汇第二个及以后的单词与 I-标签拼接后存入列表。返回一个列表，列表中的元素为填充模板生成的句子。

4.3 基于 BERT 的威胁情报命名实体识别模型

本文将提出的数据增强方法与 BERT 结合作为威胁情报命名实体识别模型，模型架构如图 16 所示。模型共包含四个部分，从下至上依次为数据增强层，BERT 词向量层，双向 LSTM 编码层和 CRF 解码层。

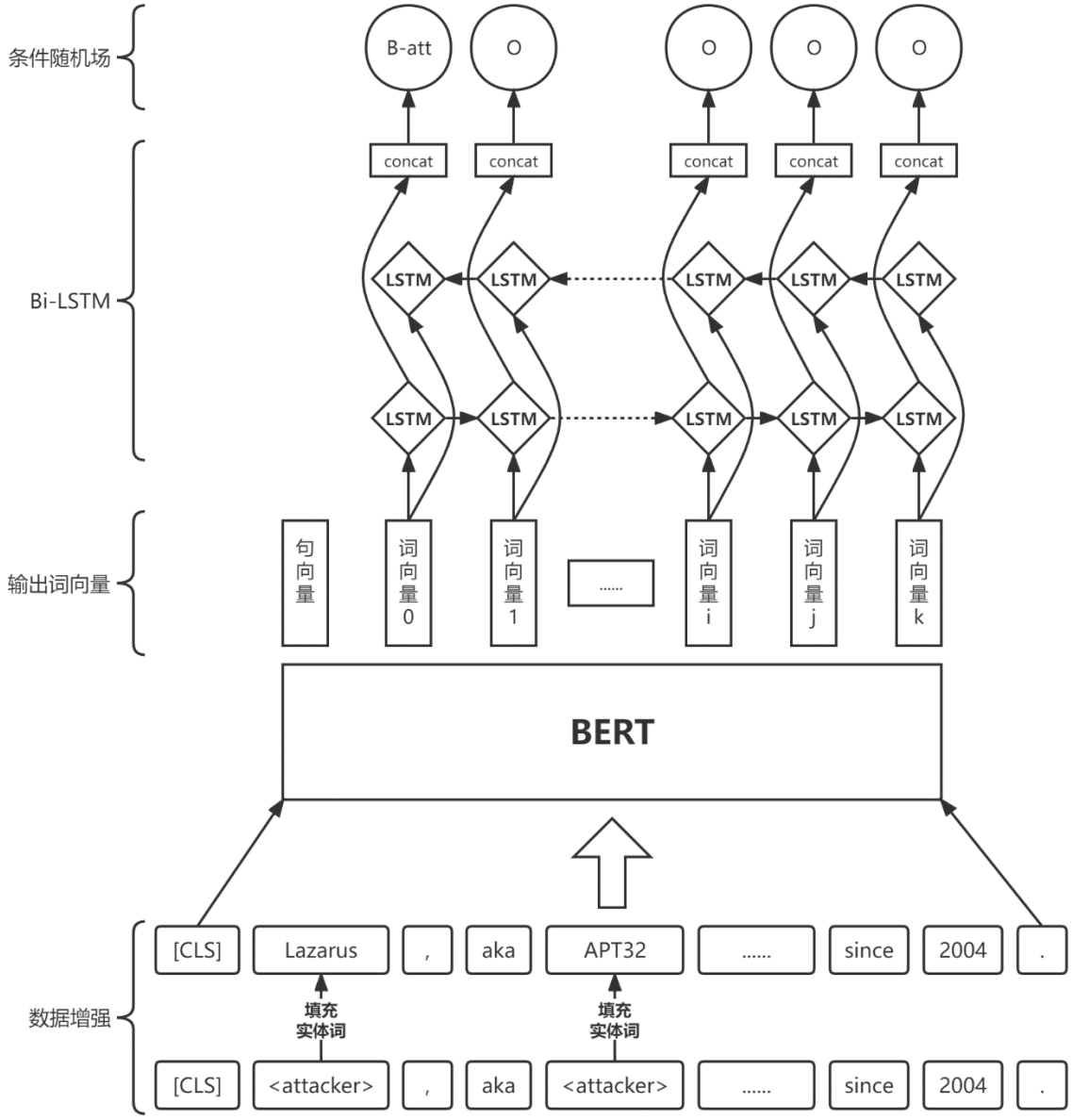

图 16　基于 BERT 的威胁情报命名实体识别模型

在数据增强层，模型输入可以通过参数控制。模型有两种不同的输入模式：无增强输入模式和混合输入模式。无增强输入模式仅使用人工标注的监督数据作为输入；混合

输入模式使用生成的增强数据和人工标注的监督数据作为输入，两者在混合时会被打乱顺序。在 BERT 词向量层，BERT 会根据每个输入单词的上下文动态地为其生成词向量，词向量可以用来表示单词的语义信息。输出的词向量序列将作为下一层的输入。在双向 LSTM 编码层，LSTM 编码器会从两个方向（前向和后向）对输入序列的时序关系进行编码，其输出的隐状态序列将作为下一层的输入。最后，在 CRF 解码层，使用条件随机场对隐状态序列进行解码得到句子对应的标签序列。使用条件随机场作为解码器的原因是它可以学习标签的内在关联性并输出条件概率最大的标签序列。

图中的[CLS]表示 classification，是 BERT 模型用于文本分类任务的标记。[CLS]通过自注意力机制来获取句子级别的信息表示，它对应的向量为包含句子语义信息的句向量。本文中加入[CLS]是 BERT 输入格式需要，与命名实体识别任务无关。

4.4 实验

4.4.1 数据集构建

本文所使用的威胁情报命名实体识别数据集由两部分组成：人工标注的威胁情报 NER 数据集，以及使用数据增强方法生成的威胁情报 NER 数据集。本文将前者称为原始标注数据集，后者称为生成数据集，将两者合并后的数据集称为增强数据集。

原始标注数据集的实体统计信息已在 4.1 节给出，如表 3 所示。将原始标注数据集按 70:15:15 的比例划分为训练集、验证集和测试集，分别称为原始训练集、原始验证集和原始测试集。表 5 列出了原始训练集、原始验证集和原始测试集的实体统计信息。

表 5　原始数据集实体统计信息

	攻击组织	恶意软件	地点	行业	技术	工具	漏洞	域名	IP
原始训练集	5305	6751	4372	3305	3692	1091	606	211	99
原始验证集	1115	1427	975	666	872	195	129	34	23
原始测试集	1143	1460	1029	690	835	230	96	36	20

生成数据集包括一个生成训练集和两个生成测试集。生成训练集和生成测试集 1 使用第一批模板句子生成，生成测试集 2 使用第二批模板句子生成。表 6 列出了生成训练集、生成测试集 1 和生成测试集 2 的实体统计信息。

表 6 生成数据集实体统计信息

	攻击组织	恶意软件	地点	行业	技术	工具	漏洞	域名	IP
生成训练集	1765	2225	141	30	540	122	1979	1388	1950
生成测试集 1	355	424	27	6	100	23	361	262	364
生成测试集 2	308	294	27	0	84	0	334	249	350

本文主要针对攻击组织、恶意软件、漏洞、域名和 IP 等 5 类实体进行数据增强，生成数据集中其他实体类型的少量实体为模板句子自带。生成训练集和生成测试集 1 使用相同的模板句子生成，但填充的领域词汇并没有交集。生成测试集 1 和生成测试集 2 在填充时使用了相同的领域词汇，但两者所使用的模板句子没有交集。

增强数据集包括一个增强训练集和两个增强测试集。增强训练集是原始训练集与生成训练集合并的结果；增强测试集 1 是原始测试集与生成测试集 1 合并的结果；增强测试集 2 是原始测试集与生成测试集 2 合并的结果。增强数据集在合并后都进行了乱序处理。表 7 列出了增强训练集、增强测试集 1 和增强测试集 2 的实体统计信息。

表 7 增强数据集实体统计信息

	攻击组织	恶意软件	地点	行业	技术	工具	漏洞	域名	IP
增强训练集	7070	8976	4513	3335	4232	1213	2585	1599	2049
增强测试集 1	1498	1884	1056	696	935	253	457	298	384
增强测试集 2	1451	1754	1056	690	919	230	430	285	370

4.4.2 评价指标与实验设置

1.评价指标

为评价模型的识别效果，本文使用精度（Precision）、召回率（Recall）和 F1 分数（F1 Score）的宏平均值（Macro Average）来评价多分类任务中模型的整体表现。公式(4-1)、(4-2)和(4-3)描述了精度、召回率和 F1 分数的计算方法，公式(4-4)描述了宏平均值的计算方法。假设类型总数为 N。

$$Pr = \frac{TP}{TP+FP} \tag{4-1}$$

$$Rc = \frac{TP}{TP+FN} \tag{4-2}$$

$$F1 = \frac{2 * Pr * Rc}{Pr + Rc} \tag{4-3}$$

$$macro\ avg = \frac{\sum_{i=1}^{N} eval_i}{N} \tag{4-4}$$

其中，TP（True Positive）表示真正例，TN（True Negative）表示真负例，FP（False Positive）表示假正例，FN（False Negative）表示假负例。N 为实体类型总数，eval 表示精度、召回率和 F1 分数三者中的任一评价指标。

2.实验设置

（1）数据增强方法有效性实验

本实验旨在通过对比训练集增强前后模型在原始测试集上的性能表现来验证本文提出的数据增强方法能够提升威胁情报命名实体识别模型的性能。实验分组 1 使用了原始训练集、原始验证集和原始测试集；实验分组 2 使用了增强训练集、原始验证集和原始测试集。两个实验分组所使用的模型为 BERT+Bi-LSTM+CRF，模型参数设置完全相同，使用的验证集和测试集完全相同，仅训练集不同。

（2）威胁情报命名实体识别模型泛化性能实验

本实验旨在通过对比训练集增强前后模型在增强测试集 2 上的性能表现，来验证本文提出的数据增强方法能够提升威胁情报命名实体识别模型的泛化性能。实验分组 1 使用了原始训练集、原始验证集和增强测试集 2；实验分组 2 使用了增强训练集，原始验证集和增强测试集 2。两个实验分组所使用的模型为 BERT+Bi-LSTM+CRF，模型参数设置完全相同，使用的验证集和测试集完全相同，仅训练集不同。

（3）数据增强方法在不同模型上有效性实验

本实验旨在通过对比训练集增强前后其他模型在原始测试集上的性能表现，来验证本文提出的数据增强方法在不同模型上的有效性。两个实验分组的数据集设置与实验（1）相同，但使用的模型均为 DistilBERT，模型的参数设置完全相同。

4.4.3 实验结果与分析

（1）数据增强方法有效性实验

实验结果如表 8 所示。从表中可以看到，使用增强训练集的威胁情报命名实体识别

模型（实验分组 2）比使用原始训练集的模型（实验分组 1）在原始测试集上有更好的性能表现，其 F1 分数为 85.316%，提高了 0.478%。实验结果表明本文提出的数据增强方法能够提升威胁情报命名实体识别模型的性能，验证了数据增强方法的有效性。

表 8　数据增强方法有效性实验结果

实验分组	Precision	Recall	F1
1	86.954	83.206	84.838
2	**87.944**	**83.322**	**85.316**

（2）威胁情报命名实体识别模型泛化性能实验

实验结果如表 9 所示。从表中可以看到，使用增强训练集的威胁情报命名实体识别模型（实验分组 2）比使用原始训练集的模型（实验分组 1）在增强测试集 2 上有更好的性能表现，其 F1 分数为 79.586%，提高了 3.295%。增强测试集 2 所使用的模板句子和领域词汇与增强训练集所使用的模板句子和领域词汇均没有交集，因此该实验结果表明本文提出的数据增强方法能够提升威胁情报命名实体识别模型的泛化性能。

表 9　威胁情报命名实体识别模型泛化性能实验结果

实验分组	Precision	Recall	F1
1	83.308	73.511	76.291
2	**84.455**	**77.094**	**79.586**

（3）数据增强方法在不同模型上有效性实验

实验结果如表 10 所示。从表中可以看到，即使将威胁情报命名实体识别模型由 BERT+Bi-LSTM+CRF 更换为 DistilBERT，使用增强训练集训练得到的模型（实验分组 2）依然比使用原始训练集训练得到的模型（实验分组 1）具有更好的性能，其 F1 分数为 89.177%，提高了 0.411%，该实验结果验证了本文提出的数据增强方法在不同模型上的有效性。

表 10　数据增强方法在不同模型上有效性实验结果

实验分组	Precision	Recall	F1
1	86.257	**91.426**	88.766
2	**87.134**	91.318	**89.177**

4.5 本章小结

本章主要介绍了基于数据增强与 BERT 的威胁情报命名实体识别方法。首先介绍了威胁情报数据特点所带来的问题，包括标签分布不平衡和样本多样性有限。然后阐述了现有 NLP 数据增强方法在威胁情报领域应用的不足。接着，本文提出了基于知识库和模板填充的数据增强方法，并详细介绍了如何构建知识库与模板，如何填充模板来生成增强数据。在此基础上，本文将数据增强方法与 BERT 结合，作为威胁情报命名实体识别模型，并通过 3 组对比实验验证了本文提出方法的有效性。

第五章　融合多元实体信息的威胁情报关系抽取方法

本章介绍论文提出的融合多元实体信息的威胁情报实体关系抽取方法。首先对威胁情报数据的实体关系特点进行分析，然后针对现有实体关系标注工具的不足，参考现有实体关系数据集格式，对标注工具进行了改进使其可以用于生成威胁情报实体关系数据集。最后本文提出了融合多元实体信息的威胁情报实体关系抽取模型，并在构建的威胁情报实体关系数据集上对本章提出方法的有效性进行了验证。

5.1 设计思想

对于实体关系抽取任务，目前学术界研究较多的是句子级关系抽取，即给定一个句子和句子中的两个实体，判断实体对的语义关系。不少研究工作已经在句子级关系抽取任务中取得了一定效果。然而，威胁情报数据中的实体关系比句子级关系抽取任务要复杂：威胁情报句子中可能包含多个实体和多个关系，任意两个实体之间并不总是存在关系，威胁情报实体关系数据示例如图 17 所示。

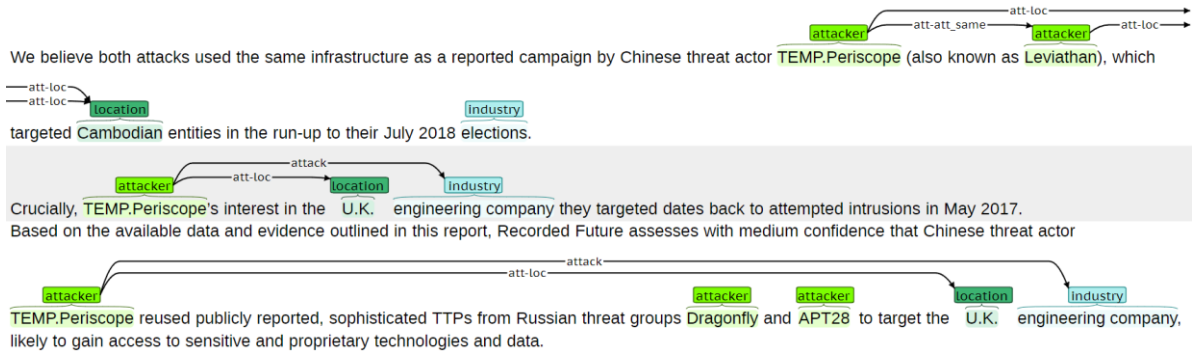

图 17　威胁情报实体关系数据示例

为了构建威胁情报实体关系数据集，需要将包含关系的句子和涉及关系的两个实体提取出来。除此之外，还需要提取一些不包含关系但属于同一个句子的实体对，并将它们的关系设置为其他（other）。这样做的原因是如果一个句子中的两个实体不存在某种预先定义好的关系，模型可以将它们的关系分类为其他，而不是将它们的关系判定为某个预定义的关系。

现有的标注工具大多只提供基础标注功能，即可以使用标注工具对文本中的实体和实体间关系进行标注，但这些工具并不能帮助用户将包含实体关系的句子从文本中提取出来，进而组织成实体关系数据集。

5.2 威胁情报实体关系标注工具

本文所使用的标注工具为 Brat[59]，Brat 提供了基础标注功能和将标注实体转化为 BIO 标注模式的功能，但并没有提供句子提取功能，即无法将包含实体关系的句子从文本中提取出来。

Brat 标注结果存放在.ann 文件中，标注结果示例如图 18 所示。图中第一列的 T 表示实体，R 表示关系，后面的数字表示序号。例如，“T84 attacker 21795 21809 TEMP.Periscope”表示第 84 个标注的实体，实体类型为 attacker，实体在文本中的索引为[21795, 21809]，实体名称为 TEMP.Periscope；“R4 att-loc Arg1:T44 Arg2:T90”表示第 4 个标注的关系，关系类型为 att-loc（属于 attack 关系），实体 1 是 T44，实体 2 是 T90。为了构建威胁情报实体关系数据集，需要根据 Brat 标注结果将包含实体关系的句子从原文中提取出来。

```
T84     attacker 21795 21809        TEMP.Periscope
T85     attacker 22342 22356        TEMP.Periscope
T86     attacker 23202 23216        TEMP.Periscope
T87     vulnerability 15409 15423   CVE-2017-11882
T88     location 847 855    Cambodia
R1      att-att_same Arg1:T43 Arg2:T35
T89     location 1083 1092  Cambodian
R2      att-loc Arg1:T43 Arg2:T89
R3      att-loc Arg1:T35 Arg2:T89
T90     location 1190 1194  U.K.
R4      att-loc Arg1:T44 Arg2:T90
T91     location 1538 1542  U.K.
T92     industry 1543 1562  engineering company
T93     industry 1195 1214  engineering company
R5      attack Arg1:T44 Arg2:T93
R6      att-loc Arg1:T45 Arg2:T91
R7      attack Arg1:T45 Arg2:T92
T94     technique 1972 1991 command and control
T95     technique 1993 1995 C2
T96     domain 2005 2023    scsnewstoday[.]com
T97     location 2095 2104  Cambodian
T98     industry 2105 2115  government
R8      att-loc Arg1:T48 Arg2:T97
R9      attack Arg1:T48 Arg2:T98
R10     att-dom Arg1:T48 Arg2:T96
T99     industry 2265 2288  critical infrastructure
T100    technique 2393 2403 spearphish
```

图 18　Brat 标注结果示例

5.2.1 现有实体关系数据集样本构成

在明确现有标注工具的不足后，下一步需要确定威胁情报实体关系数据集的格式。本文参考了部分通用领域现有公开实体关系数据集，选取其中具有代表性的两个概要介

绍。

SemEval2010_task8 是句子级关系抽取任务中较为常见的公开数据集，它发布于2010 年，由训练集和测试集组成，其中训练集包含 8000 个样本，测试集包含 2717 个样本。SemEval2010_task8 数据集定义了 19 种关系（9 种区分关系方向的预定义关系+Other 关系），数据集中的每个句子都只包含一对实体和一个关系，其训练集样本示例如下：

"The <e1>burst</e1> has been caused by water hammer <e2>pressure</e2>." Cause-Effect(e2, e1)

标签<e1>和</e1>表明了实体 1 的边界，Cause-Effect(e2, e1)表示实体 2 和实体 1 的关系为 Cause-Effect（关系由实体 2 指向实体 1）。

TACRED 是另一个句子级关系抽取任务中较为常见的公开数据集，它包含 TAC KBP 2009 至 2014 年的数据，由训练集、验证集和测试集组成，其中训练集包含 68124 个样本，验证集包含 22631 个样本，测试集包含 15509 个样本。TACRED 数据集定义了 42 种关系，数据集中的每个句子都只包含一对实体和一个关系，其训练集样本示例如下：

{"token": ["Established", "in", "1875", ",", "Blackburn", "were", "one", "of", "the", "founding", "members", "of", "the", "Football", "League", "."],

"h": {"name": "Blackburn", "pos": [4, 5]},

"t": {"name": "1875", "pos": [2, 3]},

"relation": "org:founded"}

token 为组成句子的单词列表，h 表示头实体，t 表示尾实体，两者均包括实体名称和位置信息，relation 表示关系类型。

考虑到威胁情报原始文本数据为字符串类型，且 Brat 标注结果中的实体位置信息为字符串索引，因此本文设计了如下的威胁情报实体关系数据格式：

{"sentence": "FIN7 is a threat actor group that is financially motivated with targets in the restaurant, services and financial sectors.",

"h_entity": {"name": "FIN7", "type": "attacker", "position": [0, 4]},

"t_entity": {"name": "restaurant", "type": "industry", "position": [79, 89]},

"relation_type": "attack"}

上例中，sentence 表示包含威胁情报实体关系的句子，为字符串类型；h_entity 表示头实体，包括实体名称、实体类型和实体位置信息；t_entity 表示尾实体，其组成与头实体相同；relation_type 表示两实体的关系类型。与 TACRED 数据集不同的是，本文在实

体信息中加入了实体类型，不仅因为命名实体识别是实体关系抽取的前置任务，这样做还可以在威胁情报实体关系抽取模型中引入实体类型信息。

5.2.2 标注工具的改进

在确定威胁情报实体关系数据格式后，需要在 Brat 标注工具的基础上添加句子提取功能，以从原始威胁情报文本中将包含实体关系的句子提取出来，进而生成威胁情报实体关系数据集。句子提取算法如算法 2 所示。

算法 2：句子提取算法

```
Input：S, T, R
Output：sentence_start, sentence_end, Arg1, Arg2
for relation_id in R:
    get Arg1, Arg2∈T
    left_ptr = min(Arg1, Arg2)
    right_ptr = max(Arg1, Arg2)
    while True:    // do left side first
        if left_ptr > 0 and S[left_ptr] = ".":
            if re.match(r'\.[0-9 \t”"()]+[A-Z]', S[left_ptr:]):    // 正则表达式
                get sentence_start
                break
            else:
                left_ptr = left_ptr - 1
        else:
            left_ptr = left_ptr - 1
    while True:    // do right side
        if right_ptr < len(S) and S[right_ptr] = ".":
            if re.match(r'\.[0-9 \t”"()]+[A-Z]', S[right_ptr:]):    // 正则表达式
                get sentence_end
                break
            else:
                right_ptr = right_ptr + 1
        else:
            right_ptr = right_ptr + 1
    return sentence_start, sentence_end, Arg1, Arg2
```

算法的输入包括威胁情报全文文本字符串（S），实体列表（T）和关系列表（R），算法的输出包括目标句子的起始索引（sentence_start），目标句子的结束索引（sentence_end），头实体信息（Arg1）和尾实体信息（Arg2）。对于关系列表中的每一条记录，从实体列表中获取该关系涉及的两个实体。将左指针设为两个实体中靠前实体的开始索引，右指针设为两个实体中靠后实体的结束索引。进入循环直到使用关键字

break 跳出，如果左指针大于零且左指针指向英文句号，判断左指针到字符串末尾是否包含英文句子结尾特征，如果包含则左指针索引为目标句子的起始索引。同理可获得目标句子的结束索引。返回目标句子的起始索引、结束索引，头实体信息和尾实体信息。

5.3 威胁情报实体关系抽取模型

本文基于 R-BERT[60]模型进行改进，提出了威胁情报实体关系抽取模型。R-BERT 模型尝试将实体语义信息和实体边界信息进行融合，本文在此基础上将实体类型信息添加到 BERT 句向量中以帮助模型更好地进行关系分类。模型架构如图 19 所示。

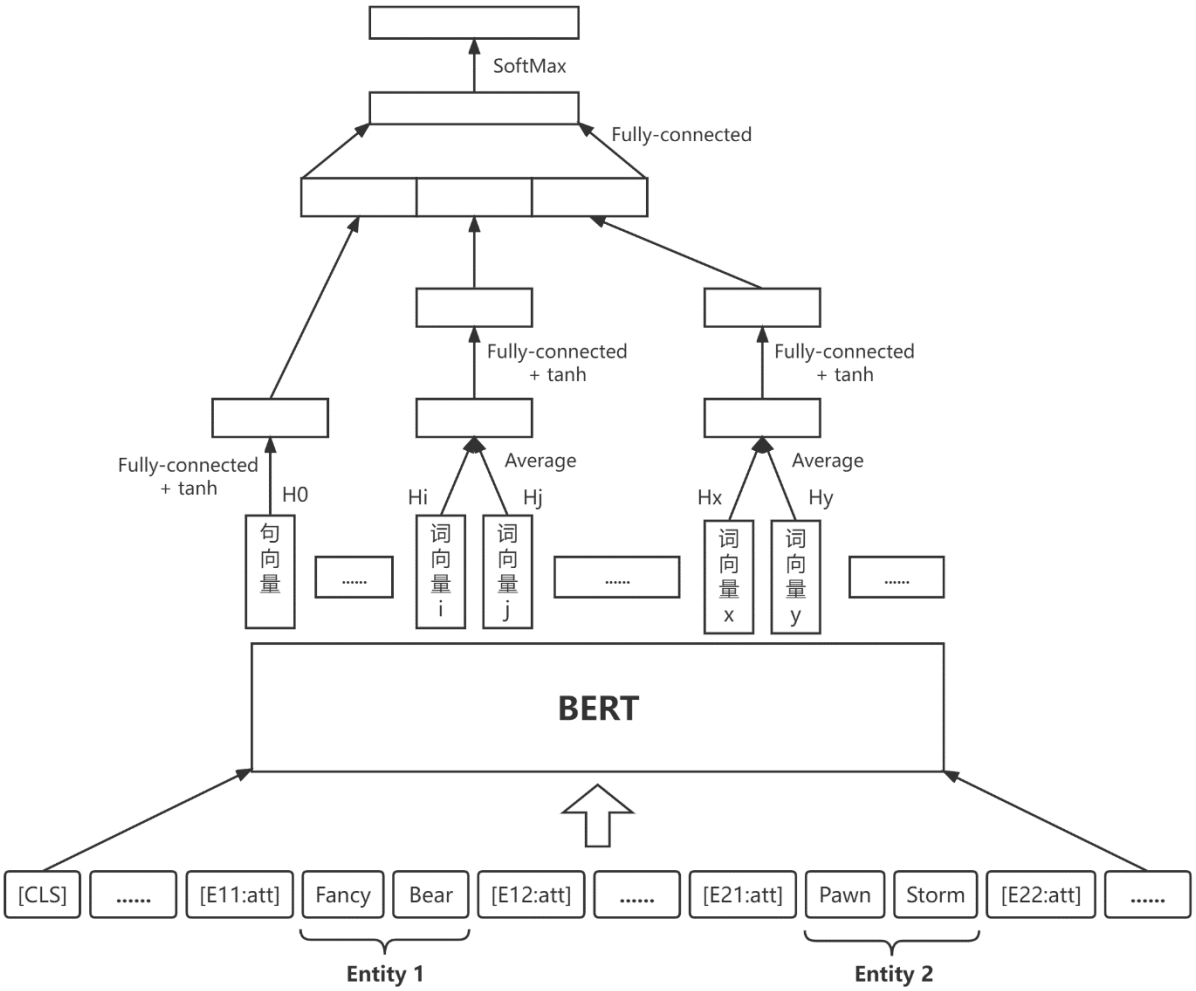

图 19　威胁情报实体关系抽取模型

实体的边界信息和类型信息体现在实体两侧的标签中，[E11:att]表示实体 1 的左侧边界，且该实体属于 attacker 类型；[E12:att]表示实体 1 的右侧边界，且该实体属于 attacker 类型。

给定包含实体 e1 和 e2 的句子 s，BERT 输出的向量用 H 表示。假设向量 H_i 到 H_j 表

示实体 e1 的词向量，向量 H_x 到 H_y 表示实体 e2 的词向量。对组成实体的每个单词的词向量求平均得到该实体的表征向量，然后将两个实体的表征向量送入 tanh 激活函数和一个全连接层，输出分别记为H_1'和H_2'，该过程如公式(5.1)和(5.2)所示。

$$H_1' = W_1\left[tanh\left(\frac{1}{j-i+1}\sum_{t=i}^{j} H_t\right)\right] + b_1 \tag{5.1}$$

$$H_2' = W_2\left[tanh\left(\frac{1}{y-x+1}\sum_{t=x}^{y} H_t\right)\right] + b_2 \tag{5.2}$$

其中，$W_1=W_2\in R^{d*d}$，$b_1=b_2$ 为偏差向量，d 是 BERT 隐状态向量大小。

将[CLS]生成的句向量送入 tanh 激活函数和一个全连接层，如公式(5.3)所示。

$$H_0' = W_0[tanh(H_0)] + b_0 \tag{5.3}$$

其中，$W_0\in R^{d*d}$，b_0 为偏差向量，d 是 BERT 隐状态向量大小。

将H_0', H_1'和H_2'拼接后送入一个全连接层和一个 SoftMax 层，得到输出向量 P，该过程如公式(5.4)和(5.5)所示。

$$H'' = W_3[concat(H_0', H_1', H_2')] + b_3 \tag{5.4}$$

$$p = softmax(H'') \tag{5.5}$$

其中，$W_3\in R^{L*3d}$，L 是关系类别数量，b_3 是偏差向量。输出向量 $P\in R^L$，P 中值最大的项所对应的关系即是模型输出的关系。

5.4 实验

5.4.1 数据集构建

文本所使用的威胁情报实体关系数据集由改进的标注工具生成，总共包含 6 种关系，分别为使用（use）、攻击（attack）、源于（from）、相同（same）、相似（similar）和其他（other）。威胁情报实体关系统计信息如表 11 所示。

表 11　数据集中威胁情报实体关系统计信息

使用	攻击	源于	相同	相似	其他
1274	998	333	121	154	362

本文在标注关系数据之处，为了减少工作量，提高标注效率，只对使用、攻击、源于、相同、相似这五种关系进行了标注，并没有标注其他关系。后来，考虑到威胁情报数据中存在非上述五种关系的情况，本文又标记了一批其他关系的威胁情报数据。由于威胁情报实体关系数据较为有限，本文将数据集以 80:20 的比例划分为训练集和测试集。

5.4.2 评价指标与实验设置

1.评价指标

本实验的评价指标为精度（Precision）、召回率（Recall）和 F1 分数（F1 Score）的宏平均值（Macro Average），指标计算公式见 4.4.2 节。

2.实验设置

本实验属于消融实验，旨在探究实体边界信息和实体类型信息对威胁情报实体关系抽取模型性能的影响。

实验分组 1 的输入句子不添加任何标签，即不添加实体边界信息和实体类型信息。实验分组 2 的输入句子只添加表示实体边界的标签（[E11]和[E12]），即添加实体边界信息但不添加实体类型信息。实验分组 3 的输入句子添加表示实体边界和实体类型的标签（[E11:attacker]和[E12:attacker]），即添加实体边界信息和实体类型信息。

以句子“The U.K. engineering company was targeted by TEMP.Periscope.”为例，句中包含的三元组为（TEMP.Periscope，attack，U.K.）。实验分组 1 的输入为原句，即不添加任何标签。实验分组 2 的输入为“The [E21] U.K. [E22] engineering company was targeted by [E11] TEMP.Periscope [E12] .”，即只添加表示实体边界的标签。实验分组 3 的输入为“The [E21:location] U.K. [E22:location] engineering company was targeted by [E11:attacker] TEMP.Periscope [E12:attacker] .”。

5.4.3 实验结果与分析

实验结果如表 12 所示。

表 12　威胁情报实体关系抽取模型实验结果

实验分组	Precision	Recall	F1
1	72.752	75.558	73.324
2	78.673	81.134	79.510
3	**80.744**	**81.631**	**81.061**

从表中可以看到，只添加实体边界信息的模型和添加实体边界信息与实体类型信息的模型均比不添加任何标签（不添加实体边界信息与实体类型信息）的模型性能表现要好，它们的 F1 分数分别提高了 6.186%和 7.737%，这表明添加额外实体信息有助于提升威胁情报实体关系抽取模型的性能。

对比实验分组 2 和实验分组 3 的结果可以发现，添加实体边界信息与实体类型信息的模型比只添加实体边界信息的模型有更好的性能表现，前者的 F1 分数比后者高 1.551%，这表明实体类型信息的加入确实对威胁情报实体关系分类任务有积极影响。

5.5 本章小结

本章主要介绍了本文提出的融合多元实体信息的威胁情报实体关系抽取方法。首先介绍了威胁情报实体关系数据的特点，包括句子中存在多个实体和多个关系。然后介绍了现有标注工具在威胁情报应用中的不足和现有实体关系数据集的格式，进而阐述了本文针对 Brat 标注工具的改进。接着，本文提出了一种融合实体语义信息、实体边界信息和实体类型信息的威胁情报实体关系抽取模型。最后在实验部分，通过消融实验验证了本文提出方法的有效性。

第六章　威胁情报知识图谱的管理工具

论文基于前述章节提出的方法，设计并实现了一个威胁情报知识图谱管理工具。该工具可以对用户上传的非结构化威胁情报数据进行解析，然后依次对威胁情报文本中的实体和实体间关系进行抽取，并将抽取结果以知识图谱的形式进行展示，同时该工具还提供情报的存储和查询等功能。本章首先介绍了该工具的需求分析和总体设计，然后阐述了工具的实现过程，包括数据解析、实体抽取、关系抽取等多个功能模块，最后给出了该工具的功能测试和性能测试结果。

6.1 工具需求分析

论文需要设计并实现一个威胁情报知识图谱管理工具，该工具应以本论文第三章提出的威胁情报本体为指导，集成本论文第四章、第五章提出的基于数据增强与 BERT 的威胁情报实体抽取方法和融合多元实体信息的威胁情报实体关系抽取方法，并支持对抽取的威胁情报实体和关系进行存储、检索和可视化展示。该工具应当具有如下四点核心功能：

（1）数据解析功能

此功能对用户上传的非结构化威胁情报 PDF 文件进行解析。提取 PDF 文件中的文本信息，并对提取的文本内容进行格式化，为后续的实体抽取和关系抽取工作做准备。

（2）威胁情报实体抽取功能

此功能对非结构化威胁情报文本中的威胁实体进行识别和分类。需要对威胁情报文本进行分句和分词处理，以句（单词序列）为单位送入模型。模型根据学习到的特征和单词的上下文信息对其类型进行判定。

（3）威胁情报关系抽取功能

此功能对非结构化威胁情报文本中威胁实体间的关系进行识别和分类。如果一个句子中抽取出来的威胁实体数量大于等于 2，则将每两个威胁实体和该句子作为关系抽取模型的输入，进而识别出每对威胁实体间的关系。

（4）存储、检索和可视化功能

该工具需要提供对模型抽取结果进行存储、检索和可视化展示的功能。其中，存储和检索功能依赖于数据库的支持，可视化展示功能则需要一个前端交互界面，界面应支持用户上传非结构化威胁情报数据，并根据关键词对抽取的信息进行检索。

6.2 工具总体设计

根据工具的功能需求，本论文设计了如图 20 所示的工具架构。该工具的实现遵循浏览器和服务器架构模式，包括交互层、业务层和持久层。

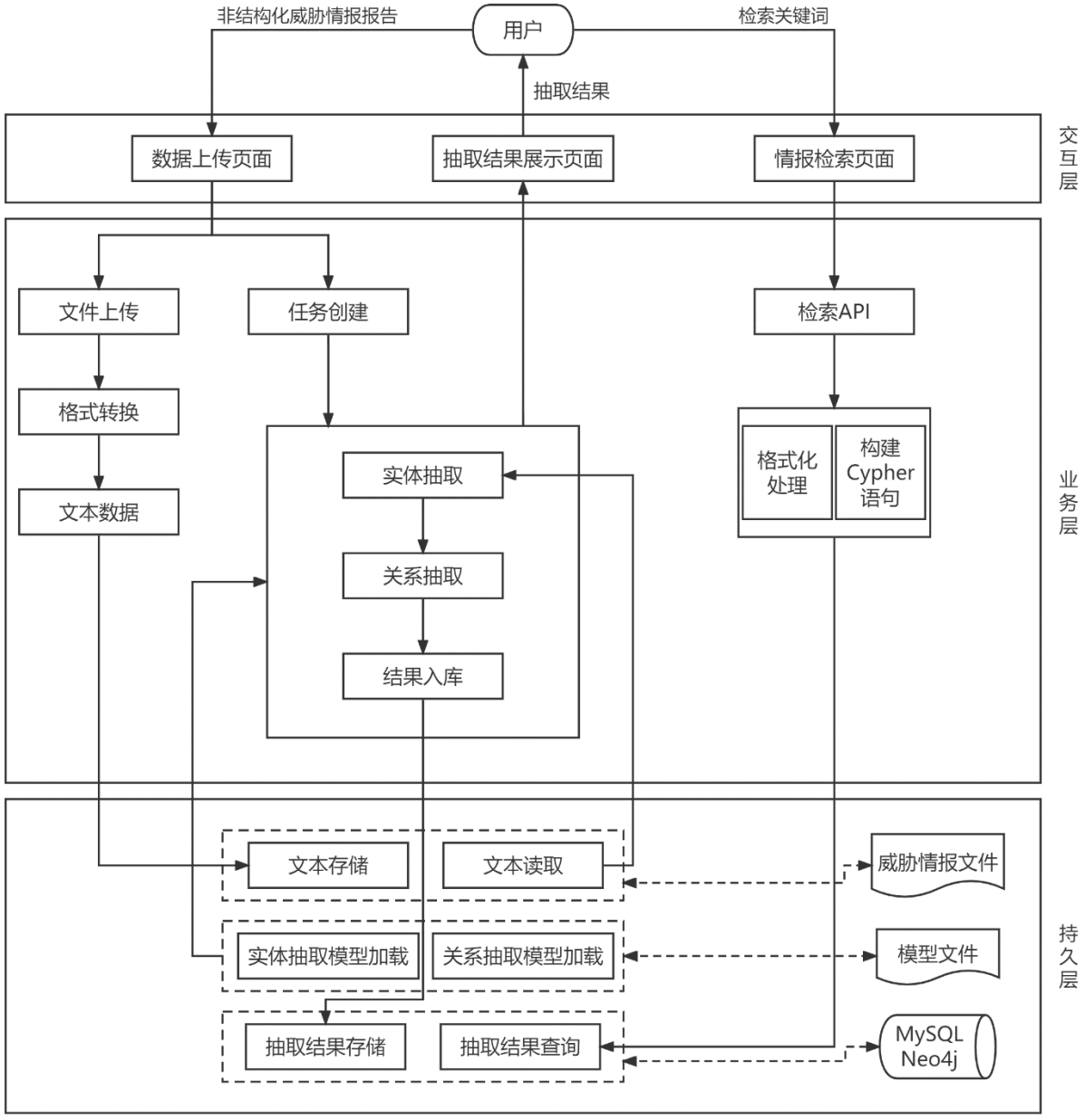

图 20　威胁情报知识图谱管理工具架构

交互层为浏览器端，其主要职责是接收用户输入的内容并展示返回的结果，包括数据上传页面，抽取结果展示页面和情报检索页面。数据上传页面负责接收用户上传的非结构化威胁情报报告文件并启动信息抽取任务；抽取结果展示页面负责将返回的信息抽取结果以文本加高亮的形式呈现给用户；情报检索页面向用户提供关键词检索功能，并以知识图谱的形式展示检索结果。

业务层和持久层为服务器端。业务层的功能包括文件上传、格式转换、任务创建、任务执行和检索。业务层的核心是任务执行，包括实体抽取、关系抽取和结果入库。持久层包括文件系统和数据库，文件系统主要存储非结构化威胁情报报告文件和模型文件，数据库主要存储信息抽取的结果。拟使用的数据库为关系数据库 MySQL 和图数据库 Neo4j，其中 MySQL 用于存储抽取任务信息，Neo4j 用于存储抽取结果信息。

按照逻辑架构，可将工具划分为四个主要功能模块：数据解析模块、命名实体抽取模块、关系抽取模块和可视化展示模块。

数据解析模块主要负责对用户上传的非结构化威胁情报 PDF 文件进行解析，包括提取 PDF 文件中的文本，对提取的文本进行格式化处理，得到可读性较高的非结构化威胁情报文本，以 TXT 形式存储，为后续的信息抽取任务做准备。

命名实体识别模块主要负责识别非结构化威胁情报文本中的威胁实体。将威胁情报文本句子输入威胁情报命名实体识别模型，模型通过学习到的参数和上下文信息判断每个单词的标签，并输出与句子对应的标签序列。

实体关系抽取模块主要负责判断威胁情报句子中的实体对包含何种关系。将包含威胁实体对的句子输入威胁情报实体关系抽取模型，模型根据学习到的参数、上下文信息、实体边界和类型信息判断两个威胁实体之间存在何种关系，输出为三元组，包括两个威胁实体和它们的关系。

可视化展示模块主要负责接收用户输入并呈现分析结果。用户的输入包括非结构化威胁情报报告文件和检索关键词，分析结果以高亮文本和知识图谱的形式呈现，展示内容包括任务进度、实体抽取结果，关系抽取结果和检索结果。

6.3 威胁情报知识图谱管理工具实现

6.3.1 数据解析模块

数据解析模块的功能是将用户上传的威胁情报 PDF 文件中的文本内容提取出来并进行格式化处理以获得可读性较高的威胁情报文本，为后续信息抽取任务做准备。此模块的工作流程图如图 21 所示。

模块的输入是用户上传的威胁情报 PDF 文件，模块的输出是威胁情报文本的 TXT 文件。模块在获取威胁情报 PDF 文件后，使用 pdfminer 库提取其中的文本内容。接着对提取的威胁情报文本进行格式化处理，包括删除多余的空行、去除句子内部的换行符，

去除页码等操作。最后将格式化后的威胁情报文本保存为 TXT 文件。

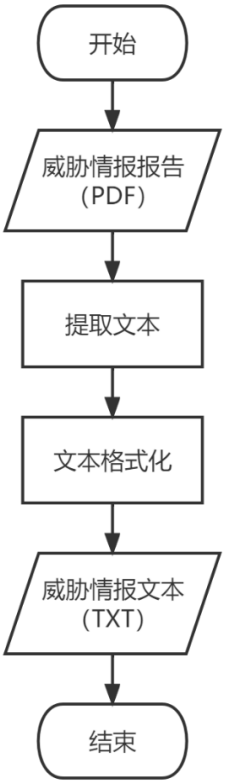

图 21　数据解析模块工作流程图

6.3.2 命名实体识别模块

命名实体识别模块的功能是识别威胁情报文本中的威胁实体，包括威胁实体的边界和类型。此模块的工作流程图如图 22 所示。

模块的输入是威胁情报文本 TXT 文件，模块的输出是包含威胁实体及其类型的 JSON 文件。模块在获取威胁情报文本后，首先对它进行分句处理，将威胁情报报告原文划分为句子。对于不是完整句子的文本，如标题、图片说明等内容，进行丢弃处理。对于以标点符号结尾的句子，有两种不同的处理：如果使用威胁情报命名实体识别模型进行实体抽取，则将句子进一步分解为单词序列，并将单词序列作为实体抽取模型的输入，模型的输出为相应的标签序列，将句子中包含的威胁实体及其标签从序列中提取出来予以保存；如果使用正则表达式进行实体抽取，则将句子文本直接作为正则表达式的输入，其输出为符合正则匹配模式的威胁实体，标签为对应正则表达式的类型。对使用不同抽取方法抽取的威胁实体进行去重处理，然后以 JSON 格式保存，待用户审核后存入数据库。本文使用的正则表达式如表 13 所示。

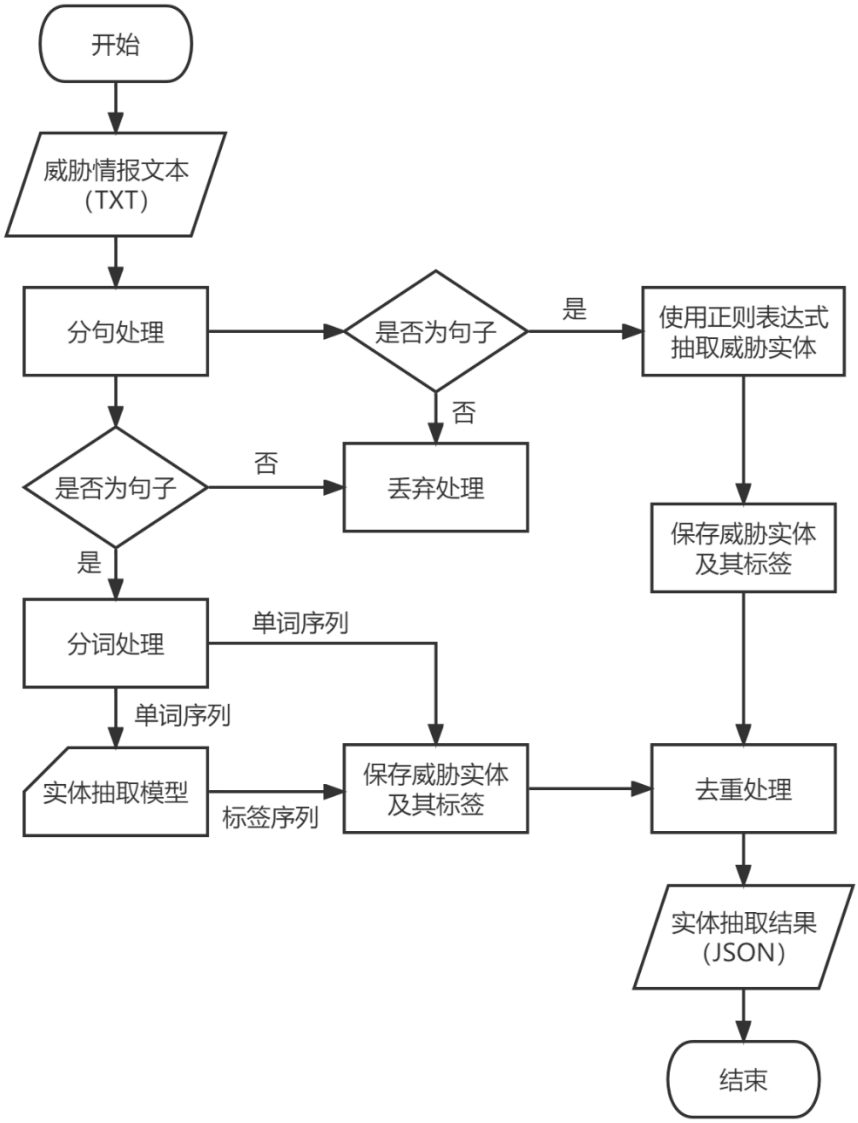

图 22　命名实体识别模块工作流程图

表 13　相关正则表达式

漏洞正则表达式	CVE-[0-9]{4}-[0-9]{4,6}
域名正则表达式	(?<![-a-zA-Z0-9_\\./])([a-zA-Z0-9][-a-zA-Z0-9]{0,61}(?:\.|\[\.\]){1})+(?:com|edu|gov|int|mil|net|org|biz|info|pro|name|museum|network|coop|aero|xxx|xyz|idv|cn|eu|uk|us|fr|de|gs)(?![-a-zA-Z0-9_\\./])
IP 正则表达式	\d{1,3}(?:\.|\[\.\]){1}\d{1,3}(?:\.|\[\.\]){1}\d{1,3}(?:\.|\[\.\]){1}\d{1,3}

6.3.3 实体关系抽取模块

实体关系抽取模块的功能是判断存在于同一句子中的威胁实体对属于何种关系。此模块的工作流程图如图 23 所示。

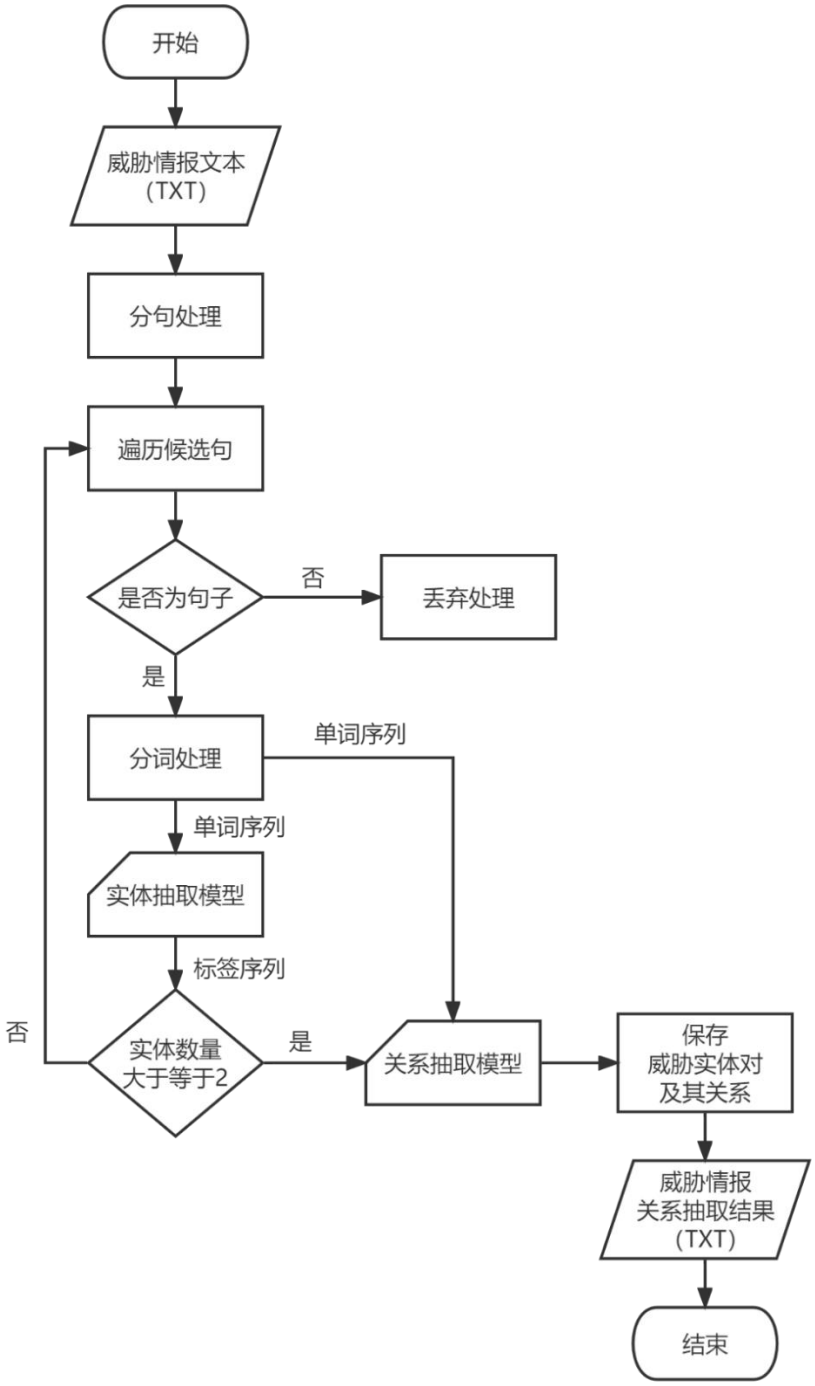

图 23　实体关系抽取模块工作流程图

模块的输入是威胁情报文本 TXT 文件，模块的输出是包含威胁实体对及其关系的 TXT 文件。模块在输入威胁情报文本后，首先对其进行分句处理，将威胁情报原文划分

为句子。对于不是完整句子的文本，如标题、图片说明等内容，予以丢弃处理。对于以标点符号结尾的句子，将它进一步分解为单词序列，并将单词序列送入实体抽取模型，得到相应的标签序列。如果抽取得到的威胁实体数量大于等于 2，则将单词序列和威胁实体对作为关系抽取模型的输入，输出是该威胁实体对的关系。最后将威胁实体对及其关系写入 TXT 文件保存，待用户审核后存入数据库。

6.3.4 可视化展示模块

可视化展示模块的主要功能是与用户交互，包括接收用户指令和呈现后端返回的结果。此模块的用户操作流程如图 24 所示，用户在使用工具时可选择信息抽取和情报检索。

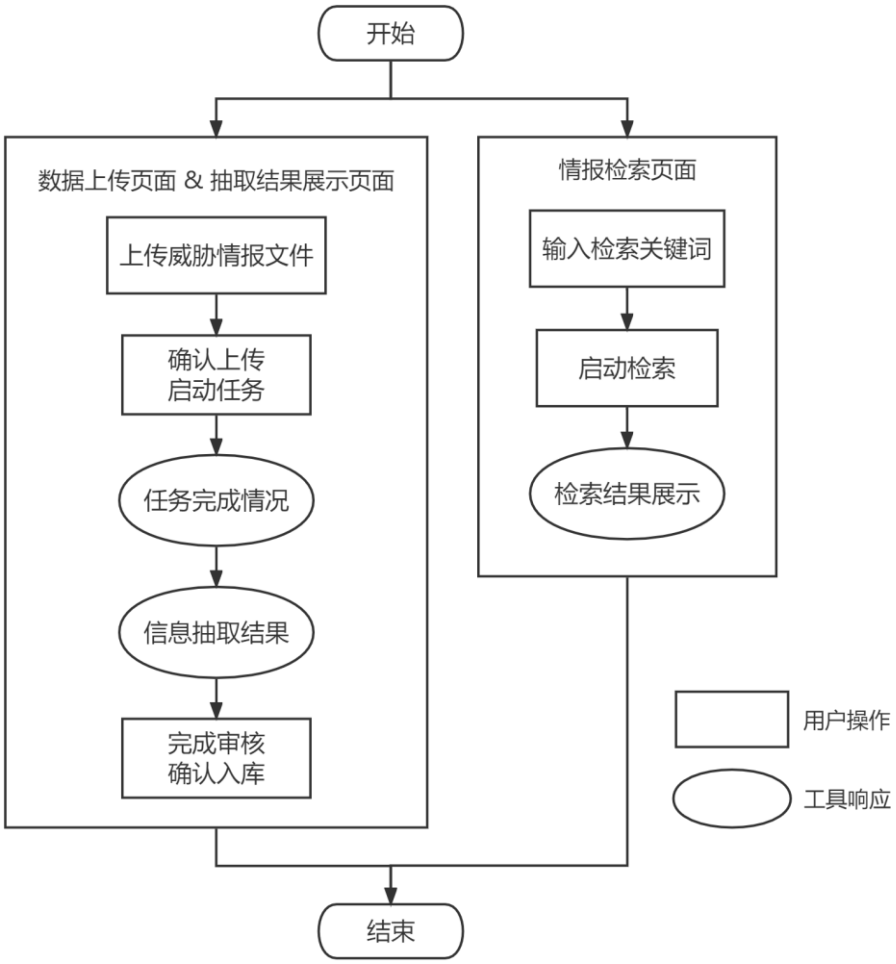

图 24　可视化展示模块用户操作流程图

信息抽取对应数据上传页面和抽取结果展示页面。用户首先上传非结构化威胁情报报告 PDF 文件，然后点击确认上传按钮以启动信息抽取任务。工具在进行信息抽取时，

会阶段性向用户反馈抽取任务的完成情况，用户可选择已完成抽取的文档查看抽取结果，结果以高亮文本和表格的形式在抽取结果展示页面上呈现。其中，高亮文本部分将抽取的威胁实体在威胁情报 PDF 文件中高亮显示，表格部分分类展示了抽取的威胁实体及其关系，工具提供了搜索框支持用户对 PDF 中的威胁实体进行查询。

情报检索对应情报检索页面。用户首先输入检索关键词，点击搜索按钮后，工具会去数据库检索相关数据，并将检索结果返回以知识图谱的形式呈现给用户。用户可以点击图谱中的节点以显示更多相关信息。

6.3.5 数据库设计

1.基于 MySQL 的任务信息存储

关系数据库 MySQL 用于存储抽取任务相关信息。用户上传的每一个 PDF 文件都会和 MySQL 数据库中的项进行匹配，确认无重复后为相关文件创建记录，并启动抽取任务。MySQL 数据库包含一个名为 extraction_task 的数据表，其字段包括报告 id、报告名称、上传者、上传时间、报告发布时间、抽取状态、入库状态和报告路径。抽取任务数据表的详细信息如表 14 所示。

表 14　抽取任务表字段描述信息

字段名	字段释义	字段类型	是否支持 null	键
report_id	报告 id	varchar(255)	否	主键
report_title	报告名称	varchar(255)	否	-
uploader	上传者	varchar(255)	否	-
upload_time	上传时间	varchar(255)	否	-
publish_time	报告发布时间	varchar(255)	是	-
extraction_state	抽取状态	int	否	-
storage_state	入库状态	int	否	-
report_path	报告路径	varchar(255)	是	-

2.基于 Neo4j 的知识图谱存储

图数据库 Neo4j 用于存储威胁情报的信息抽取结果，包括威胁情报实体和威胁情报实体关系，以知识图谱的形式进行保存，其中威胁情报实体作为节点，威胁情报实体关

系作为边。威胁情报知识图谱的节点类型包括攻击组织、恶意软件、技术、工具、地点、行业、漏洞、域名和 IP；边的类型包括攻击、使用、源于、相似、相同。节点类型与边的类型遵循 3.4 节定义的威胁情报知识图谱实体类型和关系类型。

6.4 工具功能与性能测试

6.4.1 功能测试

（1）威胁情报信息抽取

用户在使用工具的信息抽取功能时，会首先进入数据上传页面（如图 25 所示）。工具在执行信息抽取任务前，用户需要点击或拖拽来上传待处理的威胁情报文件。用户点击上传按钮，浏览器会将执行任务请求发送给服务器端。

图 25　数据上传页面

工具所能处理的英文语料示例如图 26 所示。该文件为 PDF 格式，包含非结构化的威胁情报文本。服务器端收到任务请求后，依次进行数据解析、实体抽取、关系抽取、结果入库等操作，抽取结果展示页面会阶段性反馈任务完成的情况。

This blog discusses targeted attacks against the Middle East taking place between February and October 2017 by a group Unit 42 is naming "MuddyWater". This blog links this recent activity with previous isolated public reporting on similar attacks we believe are related. We refer to these attacks as MuddyWater due to the confusion in attributing these attacks. Although the activity was previously linked by others to the FIN7 threat actor group, our research suggests the activity is in fact espionage related and unlikely to be FIN7 related.

The MuddyWater attacks are primarily against Middle Eastern nations. However, we have also observed attacks against surrounding nations and beyond, including targets in India and the USA. MuddyWater attacks are characterized by the use of a slowly evolving PowerShell-based first stage backdoor we call "POWERSTATS". Despite broad scrutiny and reports on MuddyWater attacks, the activity continues with only incremental changes to the tools and techniques.

The Palo Alto Networks Unit 42 research team recently came across a series of malicious files which were almost identical to those targeting the Saudi Arabian government previously discussed by MalwareBytes. Which in turn, closely resembles a previous article by Morphisec. These attacks have also been tracked by several other researchers on Twitter and elsewhere.

The activity has been consistent throughout 2017 and, based on our analysis, targets or is suspected to target, entities in the following countries: Saudi Arabia, Iraq, Israel, India, Pakistan, Turkey, USA.

图 26　英文语料示例

当抽取任务执行完毕后，抽取结果展示页面就会展示针对威胁情报报告的抽取结果，如图 27 所示。

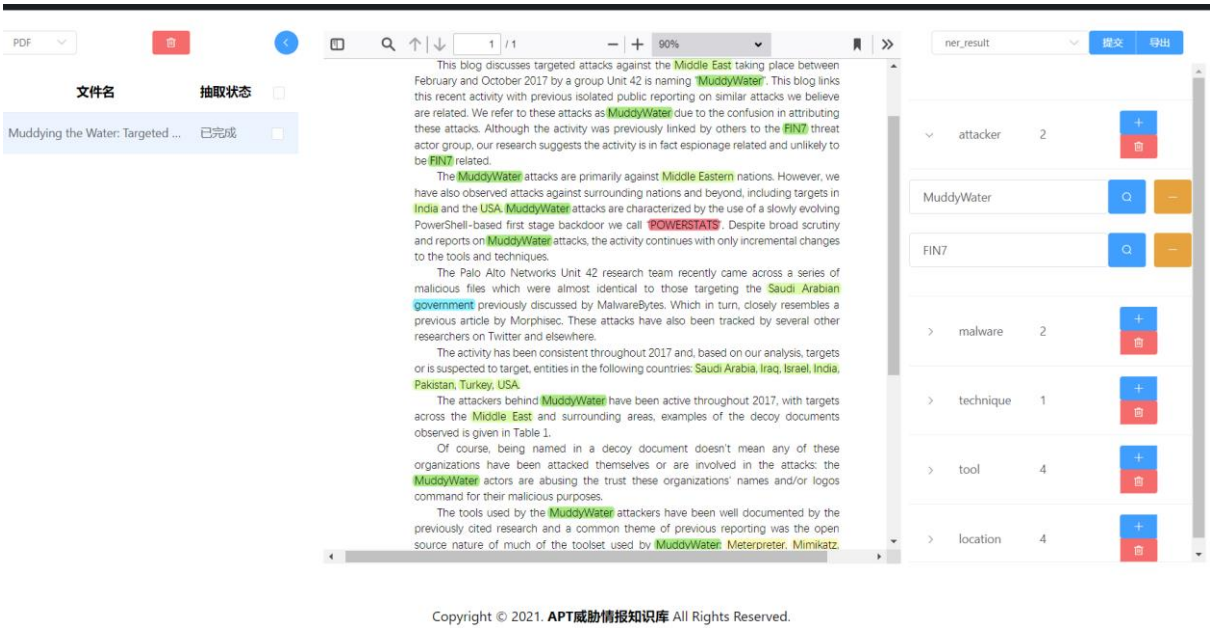

图 27　抽取结果展示页面

页面左侧是上传文件列表及其抽取状态，点击抽取状态为已完成的文件会显示该文件的抽取结果。页面中部是一个 PDF 阅读器，实体抽取的结果被高亮显示在 PDF 中，

用户也可以使用阅读器提供的搜索功能来快速定位威胁实体在 PDF 中的位置。页面右侧将实体抽取和关系抽取结果以表格形式呈现给用户，用户可点击相应类别的实体或关系来查看其包含的数据。如果用户希望纠正抽取结果，可直接在表格中进行编辑操作，表格支持撤销和自动保存功能。用户在完成信息审核后，可点击右上方的提交按钮将抽取结果写入数据库。

（2）情报检索

用户点击导航栏中的情报查询与分析后，会进入情报检索页面（如图 28 所示）。此页面允许用户使用关键词对抽取的威胁情报实体及其关系进行查询。

图 28　情报检索页面

输入关键词后，点击右侧的搜索按钮，会跳转到情报检索结果展示页面（如图 29 所示）。页面左侧显示检索关键词的相关信息，如输入关键词为 APT32（攻击组织），则页面左侧显示 APT32 的相关信息，包括别名、所属地区、攻击地区和攻击行业。页面中部显示与检索关键词有关联的报告、域名、IP 和其他 IOC 等信息。页面右侧是以检索关键词为核心的威胁情报知识图谱，图谱中深蓝色节点表示报告，绿色节点表示攻击组织（含别名），黄色节点表示背景（地点和行业），红色节点表示失陷指标（IOC），浅蓝色节点表示技术。节点之间的关系显示在边上。

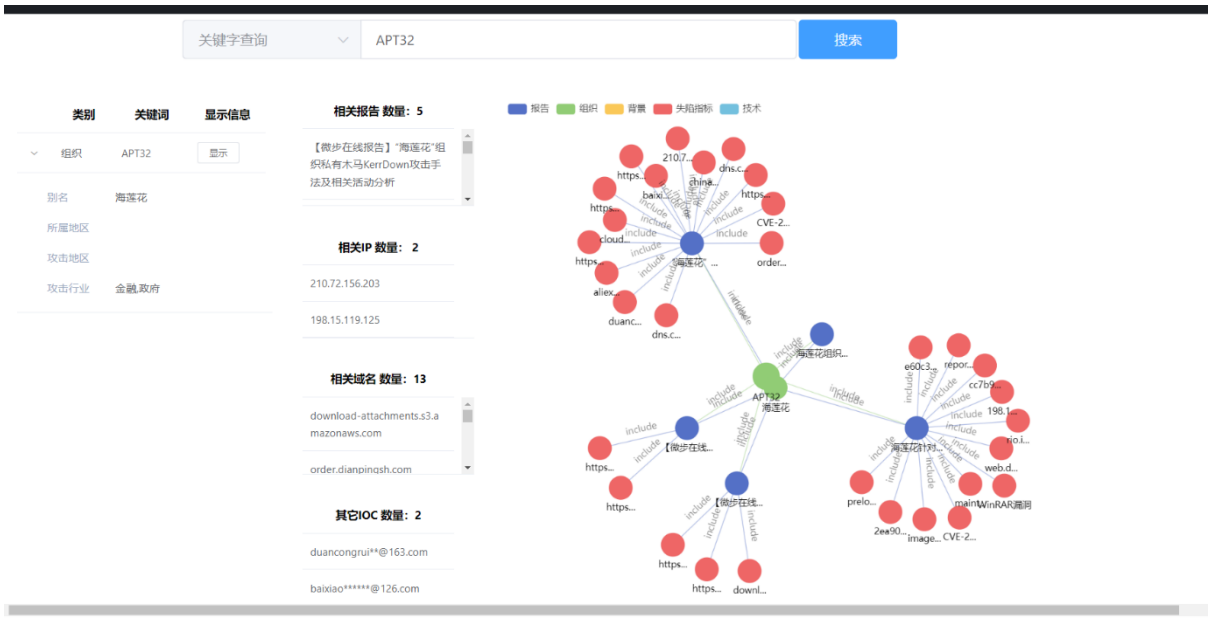

图 29　情报检索结果展示页面

6.4.2 性能测试

工具性能测试包括信息抽取的准确性测试和任务执行的耗时测试。

本文第四、五章的实验结果与分析部分已经给出了实体抽取和关系抽取的准确性指标。其中，威胁情报实体抽取的 F1 分数宏平均值可达 89.315%，威胁情报关系抽取的 F1 分数宏平均值可达 81.061%。

对于工具的耗时测试，本文使用了多个不同威胁情报文件作为测试对象，对实体抽取耗时、关系抽取耗时和总耗时进行了记录，耗时结果如表 15 所示。

表 15　工具耗时测试结果

文件大小 (MB)	文件页数	实体抽取耗时 (秒)	关系抽取耗时 (秒)	总耗时 (秒)
0.45	5	15.7	47.2	63.6
1.81	18	68.4	117.6	186.9
3.68	47	181.3	377.1	559.2

表中的总耗时略高于实体抽取耗时与关系抽取耗时之和，因为总耗时还包括模块交互和函数调用的时间。

6.5 本章小结

本章主要介绍了威胁情报知识图谱管理工具的设计和实现。该工具以第三章的威胁情报本体为指导，整合了第四和第五章提出的威胁情报命名实体识别方法、威胁情报实体关系抽取方法和训练好的模型，对用户上传的非结构化威胁情报文本中的威胁实体和关系进行了抽取，并将抽取结果以知识图谱的形式呈现给用户，同时还为用户提供了情报检索功能。最后，在工具功能与性能测试部分验证了本工具的良好性能表现，也说明了本文提出方法的有效性。

总结与展望

网络安全的重要性日益凸显，由于高级持续性威胁攻击和零日漏洞层出不穷，给传统的安全技术，如防火墙、杀毒软件、入侵检测系统等带来了巨大挑战。威胁情报是对APT组织的分析，包括其能力、方法、动机和目标。威胁情报的意义不仅在于利用基本威胁指标辅助防御，还在于帮助企业或机构对安全局势做出前瞻性判断和决策，并且能够很好提升传统安全技术的防御能力。因此，威胁情报领域吸引了众多企业、机构乃至国家进行布局。传统的威胁情报研究集中在半结构化威胁情报数据的利用，然而半结构化威胁情报数据的规模有限且表达形式多局限于文本描述，难以直观反映威胁实体间的多层次关系，不利于高价值情报的挖掘。对此，本文使用知识图谱描述威胁情报实体和实体间关系，研究面向非结构化威胁情报数据的知识图谱构建方法。

针对威胁情报知识图谱未定义实体类型和关系类型的问题，基于现有威胁情报标准，本文提出了一个威胁情报本体，该本体涵盖了13类威胁情报领域主要概念和7种概念间关系，为下文定义威胁情报知识图谱实体和关系类型提供指导依据；针对威胁情报领域实体标签分布不平衡和实体样本多样性有限的问题，本文提出了一种基于数据增强和BERT的威胁情报命名实体识别方法，该方法通过将威胁情报领域词汇填入具有威胁情报上下文环境的模板句子来生成监督数据，并与BERT结合实现了威胁情报领域的命名实体识别；针对现有标注工具无法用于生成威胁情报实体关系数据集的问题，本文改进了brat标注工具，将包含威胁情报实体对和关系的句子提取出来构成威胁情报实体关系数据集，在此基础上，本文提出了一种融合多元实体信息的威胁情报实体关系抽取方法，该方法将实体语义信息、实体边界信息和实体类型信息融入模型中，实现了威胁情报的实体关系抽取。

本文的主要研究成果如下：

（1）针对威胁情报知识图谱实体和关系范围不明确的问题，构建了一个威胁情报本体。本文基于威胁情报语义分析和相关标准STIX 2.1设计并构建了面向威胁情报领域的本体，该本体涵盖了13类威胁实体和7种实体间关系，为下文定义威胁情报知识图谱实体和关系类型提供指导依据。

（2）针对威胁情报领域实体标签分布不平衡和实体多样性有限的问题，本文提出了一种基于数据增强和BERT的威胁情报命名实体识别方法。该方法首先将威胁情报领域词汇填入符合威胁情报上下文环境的模板句子来生成增强数据，提高了稀有类的实体

数量和部分类的样本多样性，然后将增强数据与 BERT+Bi-LSTM+CRF 模型结合进行威胁情报命名实体识别。实验表明，本文方法在威胁情报命名实体识别数据集上取得了89.315%的 F1 分数宏平均值，更换为 DistilBERT 模型后进行实验，取得了 89.177%的 F1 分数宏平均值，均优于数据增强前的模型表现。

（3）针对现有标注工具无法用于生成威胁情报实体关系数据集的问题，本文改进了 Brat 标注工具，提取包含威胁实体对和关系的句子构成威胁情报实体关系数据集。在此基础上，本文提出了一种融合多元实体信息的威胁情报实体关系抽取方法以实现威胁情报的实体关系抽取。该方法通过在实体单词两侧添加特殊标签的方式，将实体边界信息和实体类型信息融入模型，加上词向量包含的实体语义信息，提升了 R-BERT 模型的性能表现。本文方法在威胁情报实体关系数据集上取得了 81.061%的 F1 分数宏平均值，高于只包含实体语义信息的模型和只融合实体语义信息和实体边界信息的模型。

（4）基于本文提出的威胁情报本体、威胁情报命名实体识别方法和威胁情报实体关系抽取方法，设计并实现了威胁情报知识图谱管理工具。该工具能对上传的非结构化威胁情报文件进行格式解析和清洗，然后自动抽取威胁情报文本中的威胁实体及其关系，将抽取结果存储到数据库中并以表格和知识图谱的形式展示，验证了本文方法的有效性。

本文面向非结构化威胁情报数据的知识图谱构建方法展开研究，在威胁情报本体设计、威胁情报命名实体识别和威胁情报实体关系抽取方面提出了相应解决方案。本文对提出方案进行了大量实验，实验结果表明本文方案达到了良好的抽取效果。在后续的研究工作中，可以主要从以下几个方向开展：

（1）近几年，小样本学习（Few-Shot Learning）逐渐成为研究的热点，它旨在将从其他任务中学习得到的具有先验知识的模型迁移到某个具体任务，从而依靠少量样本便可完成该任务。本文所使用的 BERT 模型可以看作是具有通用领域（如新闻等）先验知识的模型迁移到威胁情报领域任务。在未来研究中，可以将具有威胁情报先验知识的模型应用于威胁情报任务，以减少所需的威胁情报标注数据，从而提高模型的实用性。

（2）本文提出的威胁情报实体关系抽取方法是在实验室构造的威胁情报数据集上进行实验的。在后续研究中，可以收集更多威胁情报领域的实体关系数据，在实体信息的融合方式上做进一步扩展并在更大的威胁情报关系数据集上进行验证，从而提高方法的鲁棒性和泛化能力。

参考文献

[1] 国家互联网信息中心. 第 46 次中国互联网络发展状况统计报告[R]. 北京:国家互联网信息中心, 2020.

[2] 国家互联网应急中心. 2019 年我国互联网网络安全态势综述[R]. 北京:国家互联网应急中心, 2020.

[3] 国家互联网应急中心. 2020 年中国互联网网络安全报告[R]. 北京:国家互联网应急中心, 2021.

[4] Gartner. Definition: Threat Intelligence[EB/OL], https://www.gartner.com/doc/2487216/definition-threat-intelligence, 2013.

[5] Oosthoek K, Doerr C. Cyber threat intelligence: A product without a process?[J]. International Journal of Intelligence and CounterIntelligence, 2021, 34(2): 300-315.

[6] Berners-Lee T. Semantic Web roadmap[EB/OL]. http://www.w3.org/DesignIssues/Semantic.html, 1998.

[7] Mahdisoltani F, Biega J, Suchanek F. Yago3: A knowledge base from multilingual wikipedias[C]//7th biennial conference on innovative data systems research. CIDR Conference, 2014.

[8] Bollacker K, Cook R, Tufts P. Freebase: A shared database of structured general human knowledge[C]//AAAI Conference on Artificial Intelligence. 2007, 7: 1962-1963.

[9] Lehmann J, Isele R, Jakob M, et al. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia[J]. Semantic web, 2015, 6(2): 167-195.

[10] 李涛. 威胁情报知识图谱构建与应用关键技术研究[D]. 战略支援部队信息工程大学, 2020. DOI:10.27188/d.cnki.gzjxu.2020.000141.

[11] 秦娅. 网络安全知识图谱构建关键技术研究[D]. 贵州大学, 2019.

[12] Etzioni O, Cafarella M, Downey D, et al. Unsupervised named-entity extraction from the web: An experimental study[J]. Artificial intelligence, 2005, 165(1): 91-134.

[13] Sekine S, Nobata C. Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy[C]//LREC. 2004: 1977-1980.

[14] Zhang S, Elhadad N. Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts[J]. Journal of biomedical informatics, 2013, 46(6): 1088-1098.

[15] Nadeau D, Turney P D, Matwin S. Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity[C]//Conference of the Canadian society for computational studies of intelligence. Springer, Berlin, Heidelberg, 2006: 266-277.

[16] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch[J]. Journal of machine learning research, 2011, 12(ARTICLE): 2493-2537.

[17] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[J]. CoRR abs/1508.01991, 2015.

[18] Li X, Feng J, Meng Y, et al. A Unified MRC Framework for Named Entity Recognition[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics(ACL-20). 2020: 5849-5859.

[19] 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 & Electronic Engineering, 2019, 20(6): 872-884.

[20] Wu H, Li X, Gao Y. An effective approach of named entity recognition for cyber threat intelligence[C]//2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). IEEE, 2020, 1: 1370-1374.

[21] Tikhomirov M, Loukachevitch N, Sirotina A, et al. Using bert and augmentation in named entity recognition for cybersecurity domain[C]//International Conference on Applications of Natural Language to Information Systems. Springer, Cham, 2020: 16-24.

[22] 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.

[23] Zhou G, Su J, Zhang J, et al. Exploring various knowledge in relation extraction[C]//Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL-05). 2005: 427-434.

[24] 黄勋, 游宏梁, 于洋. 关系抽取技术研究综述[J]. 现代图书情报技术, 2013 (11): 30-39.

[25] 刘建伟, 刘媛, 罗雄麟. 半监督学习方法[J]. 计算机学报, 2015, 38(8): 1592-1617.

[26] Hasegawa T, Sekine S, Grishman R. Discovering relations among named entities from large corpora[C]//Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04). 2004: 415-422.

[27] Brin S. Extracting patterns and relations from the world wide web[C]//International workshop on the world wide web and databases. Springer, Berlin, Heidelberg, 1998: 172-183.

[28] Craven M, Kumlien J. Constructing Biological Knowledge Bases by Extracting Information from Text Sources[C]//Proceedings of International Conference on Intelligent Systems for Molecular Biology, 1999: 77-86.

[29] Zeng D, Liu K, Lai S, et al. Relation classification via convolutional deep neural network[C]//Proceedings of COLING 2014, the 25th international conference on

computational linguistics: technical papers. 2014: 2335-2344.

[30] Katiyar A, Cardie C. Going out on a limb: Joint extraction of entity mentions and relations without dependency trees[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017: 917-928.

[31] 焦李成, 杨淑媛, 刘芳, 等. 神经网络七十年: 回顾与展望[J]. 计算机学报, 2016, 39(8): 1697-1716.

[32] 周飞燕, 金林鹏, 董军. 卷积神经网络研究综述[J]. 计算机学报, 2017, 40(6): 1229-1251.

[33] Zeng D, Liu K, Chen Y, et al. Distant supervision for relation extraction via piecewise convolutional neural networks[C]//Proceedings of the 2015 conference on empirical methods in natural language processing. 2015: 1753-1762.

[34] Lin Y, Shen S, Liu Z, et al. Neural relation extraction with selective attention over instances[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016: 2124-2133.

[35] Ji G, Liu K, He S, et al. Distant supervision for relation extraction with sentence-level attention and entity descriptions[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2017, 31(1).

[36] Huang Y, Wang W. Deep Residual Learning for Weakly-Supervised Relation Extraction[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017.

[37] Ren X, Wu Z, He W, et al. Cotype: Joint extraction of typed entities and relations with knowledge bases[C]//Proceedings of the 26th International Conference on World Wide Web. 2017: 1015-1024.

[38] Rubin D L, Flanders A, Kim W, et al. Ontology-assisted analysis of Web queries to determine the knowledge radiologists seek[J]. Journal of Digital Imaging, 2011, 24(1): 160-164.

[39] Pesquita C, Ferreira J D, Couto F M, et al. The epidemiology ontology: an ontology for the semantic annotation of epidemiological resources[J]. Journal of biomedical semantics, 2014, 5(1): 1-7.

[40] Roncaglia P, Martone M E, Hill D P, et al. The Gene Ontology (GO) cellular component ontology: integration with SAO (Subcellular Anatomy Ontology) and other recent developments[J]. Journal of biomedical semantics, 2013, 4(1): 1-11.

[41] Smith B, Mark D M. Geographical categories: an ontological investigation[J]. International journal of geographical information science, 2001, 15(7): 591-612.

[42] 牛勇. 网络安全知识图谱构建的关键技术研究[D]. 电子科技大学, 2021.

DOI:10.27005/d.cnki.gdzku.2021.001226.

[43] Zhong Z, Chen D. A Frustratingly Easy Approach for Entity and Relation Extraction[C]// Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021: 50-61.

[44] Lafferty J , Mccallum A , Pereira F . Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C]//Proc. 18th International Conf. on Machine Learning. 2001.

[45] Hochreiter S , Schmidhuber J . Long Short-Term Memory[J]. Neural Computation, 1997, 9(8):1735-1780.

[46] 邱锡鹏, 神经网络与深度学习[M], 机械工业出版社, 2020.

[47] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[48] Hao P, Nikolaos P, Dani Y, et al. Random Feature Attention[C]//Proceedings of the 9th International Conference on Learning Representations, 2021.

[49] William F, Barret Z, Noam S. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity[J]. Journal of Machine Learning Research, 2022, (120):1−39.

[50] Jaszczur S, Chowdhery A, Mohiuddin A, et al. Sparse is enough in scaling transformers[J]. Advances in Neural Information Processing Systems, 2021, 34: 9895-9907.

[51] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[52] 高见, 王安. 基于本体的网络威胁情报分析技术研究[J]. 计算机工程与应用, 2020, 56(11):112-117.

[53] 王通. 威胁情报知识图谱构建技术的研究与实现[D]. 中国电子科技集团公司电子科学研究院, 2019. DOI:10.27728/d.cnki.gdzkx.2019.000039.

[54] MITRE. CybOX specification v1.0[EB/OL], http://cybox.mitre.org/language/version1.0, 2012-04-13.

[55] MITRE. Structured threat information expression version 2.1[EB/OL], https://docs.oasi s-open.org/cti/stix/ v2.1/cs 01/stix-v2.1-cs01.pdf, 2020-03-20.

[56] MITRE. ATT&CK: APT groups. [EB/OL], https://attack.mitre.org/groups/, 2022-04-25.

[57] MITRE. ATT&CK: Software. [EB/OL], https://attack.mitre.org/software/, 2022-04-25.

[58] MITRE. CVE List. [EB/OL], https://cve.mitre.org/, 2021-10-12.

[59] Stenetorp P, Pyysalo S, Topić G, et al. BRAT: a web-based tool for NLP-assisted text annotation[C]//Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. 2012: 102-107.

[60] Wu S, He Y. Enriching pre-trained language model with entity information for relation classification[C]//Proceedings of the 28th ACM international conference on information and knowledge management. 2019: 2361-2364.

攻读硕士学位期间取得的学术成果

[1] 付培国，赵忠华，**何润龙**等人，一种基于群智传感器的网络空间群体性事件感知与检测方法[P]. 中华人民共和国. 发明专利. 申请号 201910360276

致谢

感谢我的校内导师郎波教授。无论是科研上的指导，还是生活上的关心，郎老师都给予了我很大帮助。在科研上，老师对我循循善诱，提出的建议常常使我的研究思路豁然开朗。在生活中，总能感受到老师用心的关怀和鼓励，使我的研究生生活倍感温暖。

感谢我的校外导师严寒冰老师。严老师鼓励大家选择自己喜欢的研究方向，并组织同学们进行学术讨论。在学习和生活上，严老师能够给予大家足够的关心和帮助，为同学们提供电脑和工位等学习必需品。

感谢我的父母，个人的成长离不开父母提供的物质和精神支持，父母的关心让我勇敢面对生活的挫折与挑战，希望他们能够拥有健康的身体，长命百岁。

感谢自己，多年来坚持不懈的努力，其中的艰辛与痛苦只有自己知道。尽管未来的路还很长，尽管已经感到些许疲倦，自己依然会选择战斗。只有不断与困难搏斗，才会成就更强大的自己，更好地活下去。

最后感谢北航，它给予了我丰富的专业知识，培养了我踏实认真的学习态度。愿遇到的所有人都有美好的生活，祖国永远繁荣昌盛。