融合知识图谱与图卷积神经网络的非结构文本实体关系抽取

熊文俊[1] 赵 辉[1]
XIONG Wenjun ZHAO Hui

摘 要 非结构文本中的信息通常分散且不连贯，导致实体之间存在关系重叠，影响抽取结果的准确性。为优化非结构文本实体关系抽取效果，文章通过融合知识图谱与图卷积神经网络进行非结构文本实体关系抽取。基于预处理后的非结构文本数据构建一个包含实体和属性的知识图谱，作为后续实体关系抽取的基础。并将此作为 GCN 模型的输入，利用 GCN 对实体和关系进行编码，提取其深层特征。通过与知识图谱的融合，实现了非结构文本中实体关系的抽取。实验结果表明，该方法在处理非结构文本信息量增大的情况下，信息完整性保持得相当稳定，全程维持在约 98% 的高水平，且在 2 ms 以内即可完成 256 个节点的抽取。说明其方法能够完整且准确地抽取非结构文本中的实体关系，应用效果优势显著。

关键词 知识图谱；图卷积神经网络；非结构文本；实体关系抽取

doi：10.3969/j.issn.1672-9528.2025.02.017

0 引言

非结构文本数据如新闻报道、社交媒体内容、学术论文等，构成了互联网信息的主体。这些文本中蕴含着大量实体及其之间的关系，对于知识挖掘、信息搜索等多种应用场景而言至关重要。

实体关系抽取作为自然语言处理的一项基本工作，其核心目标是从文本数据中识别并提取出实体以及这些实体之间所存在的关联。目前，传统的实体关系抽取方法在处理特定领域或特定格式文本时表现出色，但在面对大规模、多样化的非结构化文本时，其泛化能力和准确性往往受到限制。例如，文献 [1] 提出方法通过去中心化、共识算法等原理实现文本实体关系抽取，但存在高能耗、存储空间受限等问题。文献 [2] 提出方法通过选择最有价值样本训练模型实现，但文本实体关系抽取时间成本较高，且准确性较低。

知识图谱作为一种用于存储和组织结构化知识的图形数据库，图卷积神经网络作为一种强大的图数据建模工具，将二者融合在一起，能够充分利用实体和关系之间的结构信息，提升关系抽取的性能 [3]。基于此，本文开展了融合知识图谱与图卷积神经网络的非结构文本实体关系抽取方法研究，以提升关系抽取的准确性和鲁棒性。

1. 河南开放大学资源建设与管理中心 河南郑州 450046

[基金项目] 河南省科技攻关项目“融合知识图谱与图卷积神经网络的在线课程混合推荐模型研究”（242102320174）

1 非结构文本数据预处理

为了保证非结构文本数据可以覆盖广泛的文本类型和主题，其一般来源于社交媒体、电子邮件、文档文件、网页内容等。然而，由此得到的非结构文本数据往往包含大量的噪声，如无关信息、冗余词汇、拼写错误等，导致在处理非结构文本数据时面临计算量大、收敛速度慢等问题。为提供更为准确和清晰的文本输入。对非结构文本数据进行预处理。具体的预处理流程如下：

（1）采用多种技术手段从社交媒体、电子邮件、文档文件、网页内容等数据源中有效收集非结构文本数据，以确保数据来源的多样性和代表性，收集技术方法如表 1 所示。这些方法的选择旨在最大化数据的获取效率和准确性。

表 1 非结构文本数据采集技术方法

技术	适用范围	总收集量占比
网络爬虫	社交媒体和网页内容	60%
API 接口	特定平台的数据获取，如电子邮件 API	25%
数据库查询	已存储的文档文件	15%

（2）在数据收集后，进行初步筛选，删除无关或重复信息。将来自不同数据源的数据进行整合，形成统一的文本数据集。在整合过程中，进行必要的数据清洗，包括删除乱码（如 ISO-8859-1 编码的字符）、特殊字符（如非 ASCII 字符）等 [4]。这一步骤确保了数据的一致性和可读性。删除文本中的无关信息，如无关广告、版权声明、页眉页脚等，进一步减少数据集中的冗余内容。

（3）在此基础上，对文本进行大小写转换，通常转换为小写，以保持文本在后续处理中的一致性。识别并删除文本中的停用词，如“的”“了”“在”等。通过删除部分停用词，减少数据集中噪声，提高后续处理准确性。此外，删除文本中的重复标点、特殊符号等，或将其替换为统一的占位符[5]。这一步骤进一步简化了文本，为后续处理提供了便利。最后，对中文文本进行分词处理。将句子拆分成单词或词组，为后续非结构文本实体关系抽取提供基础。

通过以上流程，能够对非结构文本数据进行全面、系统地预处理。经过预处理后的数据集具有高质量、高一致性和高可用性等特点，为后续非结构文本实体关系抽取提供了有力的数据支持。

2 知识图谱构建

预处理后的非结构文本数据，如果仍然以原始形式存储，可能难以对信息进行高效的检索和利用。特别是当数据以“附件”等形式存在时，检全率和检准率都可能受到影响，难以开展深度的数据挖掘与分析。知识图谱中的节点和边代表了实体之间的关系，这种关系结构使得信息检索更高效。用户可以通过查询节点和边的关系，快速找到所需的信息。此外，知识图谱还支持复杂的查询和分析，使信息利用效率大幅提高。

因此，在非结构文本数据预处理完毕后，可以构建知识图谱。

依据特定学科、行业或领域的知识需求明确知识图谱的主题范围[6]。将这一过程看作是一个集合的确定，表示为：

$$T = \{D_1, D_2, \ldots, D_n\} \tag{1}$$

式中：T 表示主题范围集合；D_n 表示第 n 个学科或领域。

明确主题范围后，基于数据的可靠性、完整性和时效性，识别并确定可用于构建知识图谱的数据源。并在数据准备完毕后，从文本中识别出命名实体，这些实体在知识图谱中作为节点存在，决定了图谱中信息的准确性和丰富性[7]。数据源的集合为：

$$N = \{E_1, E_2, \ldots, E_m\} \tag{2}$$

式中：N 表示节点集合；E_m 表示第 m 个识别出的实体。

除节点外，还需要从文本中抽取出实体的属性及其属性值，使得图谱能够更准确反映属性关系。

在选择知识表示方法时，本文根据应用场景和数据特点，通常会选用 RDF（资源描述框架）或 OWL（网络本体语言）等标准。可以清晰地表达实体和属性，便于知识的共享和互操作。在此基础上，设计如表 2 所示的知识图谱结构。

表 2 知识图谱结构

字段名	数据类型	描述
NodeID	String	节点唯一标识符
NodeType	String	节点类型
Properties	JSON	节点属性及其属性值
Relationships	List<Edge>	与节点相连的边列表
EdgeID	String	边唯一标识符
SourceNode	String	起始节点 ID
TargetNode	String	目标节点 ID
RelationshipType	String	关系类型

将非结构文本相关的知识数据导入上述结构中，完成数据的存储准备。

由此可以构建一个高质量、可扩展的知识图谱，真实反映领域内的知识结构和关系网络，为后续非结构文本实体关系抽取提供有力支持。

3 非结构文本实体关系抽取

在知识图谱中，实体之间的关系可能极其复杂，存在一对多、多对一以及多对多的关系。此外，关系重叠问题也较为常见，例如一个实体可能与其他多个实体存在多种不同的关系。这些复杂的关系模式增加了关系抽取的难度。图卷积神经网络能够处理图结构数据，通过节点之间的信息传递和聚合，捕捉知识图谱中复杂的关系模式，使模型能够更准确地抽取实体之间的关系，特别是存在重叠或复杂依赖的关系。

因此，在知识图谱构建完毕后，设计图卷积神经网络 GCN，旨在融合知识图谱并高效地抽取非结构文本中的实体关系。本文所设计的 GCN 结构如图 1 所示。

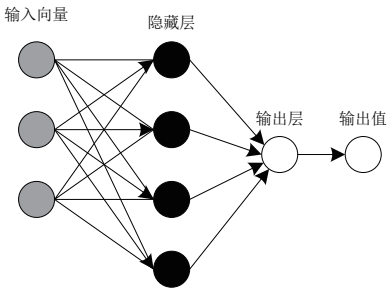

图 1 图卷积神经网络 GCN 结构

图 1 展示了 GCN 的结构设计，该网络包含多层卷积层，每层由一定数量的节点组成，用于处理知识图谱中的实体与关系[8]。

设定 GCN 的层数 L 和每层的节点数 $N_l(l = 1, 2, \cdots, L)$。其中，L 表示网络的总层数；N_l 表示第 l 层的节点数量[9]。这些参数的选择依赖于知识图谱的规模和复杂度。

选择合适的激活函数 $f(\cdot)$ 和损失函数 L_{loss}。激活函数用于引入非线性特性；损失函数则用于衡量模型预测与实际结果之间的差距，常用的有交叉熵损失等。

在 GCN 的输入阶段，将知识图谱中的实体和关系作为输入节点和边，形成图的表示。利用 GCN 对实体和关系进行编码，提取其深层特征。这一过程的计算公式为：

$$\boldsymbol{H}^{(l+1)}=f(\boldsymbol{A}\cdot\boldsymbol{H}^{(l)}\cdot\boldsymbol{W}^{(l)}) \quad (3)$$

式中：$\boldsymbol{H}^{(l)}$ 表示第 l 层的节点特征矩阵；$\boldsymbol{A}$ 表示图的邻接矩阵（或经过某种预处理的矩阵）；$\boldsymbol{W}^{(l)}$ 表示第 l 层的权重矩阵。

将 GCN 的输出与实体关系候选集进行匹配，识别文本中的实体。为进一步提高识别的准确性，利用命名实体识别技术从文本中识别出实体，并对识别出的实体进行标准化处理，以确保与知识图谱中的实体一致 [10]。

在识别出实体的基础上，利用 GCN 提取的深层特征来识别实体之间的关系。这一过程可以通过计算实体特征向量之间的距离来实现。

最后，将抽取出的实体关系整合到知识图谱中，形成完整的实体关系网络。为确保知识图谱的一致性和准确性，对整合后的知识图谱进行校验和修正。

随着新数据的不断产生和旧数据的过时，定期对知识图谱进行更新和维护，并对 GCN 进行持续的训练和优化。这一过程可以通过增量学习或迁移学习等方法来实现，以适应不断变化的数据环境。

综上所述，通过设计合理的 GCN 结构并融合知识图谱，可以实现非结构文本实体关系抽取方法的有效应用。该方法能够为各种智能系统提供准确、丰富的实体关系信息，为后续决策支持、信息检索等任务提供有力支持。

4 实验分析

4.1 实验设置

为验证本文融合知识图谱与图卷积神经网络的非结构文本实体关系抽取方法在公开数据集上的实效性，本实验采用 NYT 和 WebNLG 两个公开数据集。NYT 数据集涵盖了超过 110 万条句子，涉及 24 个关系类别及其对应的索引；WebNLG 数据集则包含 200 多种预定义的关系类型，所有标准语句均由专业注释人员精心编写。适用于评估本文提出的非结构文本实体关系抽取方法的有效性。数据集的具体细节如表 3 所示。

表 3 数据集信息

项目	NYT 数据集	WebNLG 数据集
训练	32 548	5167
验证	3205	420
测试	3674	326

实验数据集准备完成后，在配置有 Intel Core i7 处理器、16 GB 内存、Ubuntu 18.04 操作系统及 NVIDIA GTX 1080 Ti 显卡的计算机上展开实验。实验过程中所构建的知识图谱结构如图 2 所示。

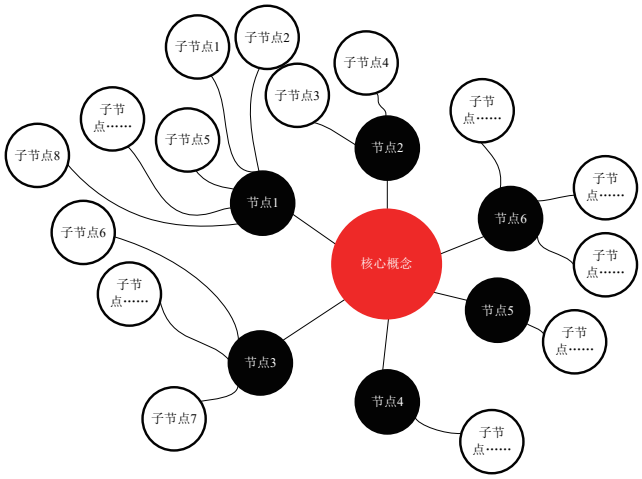

图 2 知识图谱结构

实验参数具体设定如下：

文本长度被限制在 120 个字符以内，隐藏层包含 256 个节点，且采用了两层 RNN 结构。

在上述实验设置的基础上，选取文献 [1] 方法和文献 [2] 方法作为对比方法，全面评估融合知识图谱与图卷积神经网络的非结构文本实体关系抽取方法的效能。

4.2 实验结果

（1）完整性分析

在实验中，利用实验数据集模拟一个遭受外界干扰且非结构文本信息量逐渐增大的环境，以此检验 3 种抽取方法的抗干扰能力。在评估过程中，将非结构文本信息在抽取文本实体关系时的完整性作为衡量各方法抗干扰性能的关键指标，得到的结果如图 3 所示。

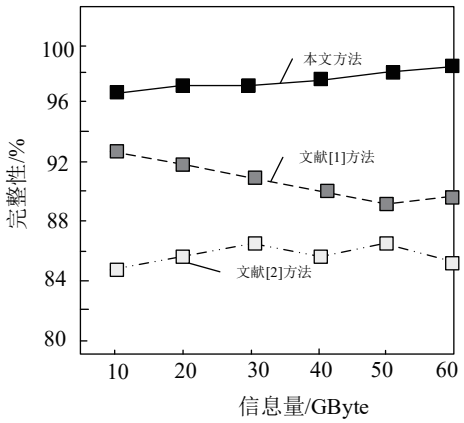

图 3 非结构文本信息完整性对比结果

根据图 3 所示，文献 [1] 和文献 [2] 所提出的方法在应用于非结构文本信息抽取时，随着信息量的增加，其信息完整性均呈现出显著的下降趋势，并伴有较大波动。其中，文献 [2] 的方法在面临外界干扰时，信息完整性出现明显起伏，最终稳定在 84% ～ 86% 的范围内，显示出相对较弱的抗干扰能力。相比之下，本文提出的方法在信息量增大的情况下，信

息完整性保持得相当稳定，未出现大幅波动，且全程维持在约 98% 的高水平。这一结果表明，本文方法具有出色的抗干扰性能，能够较为完整且准确地抽取并存储非结构文本信息，有效抵御外界干扰的影响。

（2）实时性分析

实时性评估可以反映所应用方法在处理大规模非结构文本数据时的计算效率。高效的方法能够在更短的时间内完成实体关系抽取任务，从而节省计算资源和时间成本。3 种抽取方法的实时性结果如图 4 所示。

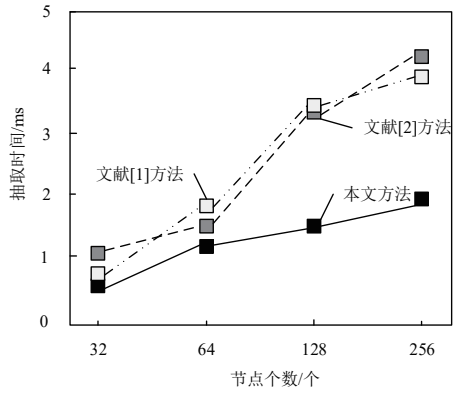

图 4 抽取实时性对比结果

从图 4 中可以看出，随着节点数量的增加，3 种方法所需要的抽取时间均呈现出上升的趋势。其中，本文方法可以在 2 ms 内完成 256 个节点的抽取。因为知识图谱通过结构化表示实体和属性，为模型提供丰富的先验知识和上下文信息。这种结构化的表示方式有助于模型更准确地理解文本中的实体和它们之间的关系，从而提高实体关系抽取的准确性和效率。

（3）消融实验

在 WebNLG 数据集上进行消融实验以探究所提方法的各部分对非结构文本信息抽取的应用性。其结果如表 4 所示。

表 4 消融实验结果

方法使用情况	查准率	召回率	F_1 值
不使用	0.78	0.82	0.83
知识图谱	0.84	0.89	0.88
图卷积神经网络	0.86	0.92	0.91
知识图谱与图卷积神经网络	0.96	0.98	0.96

分析表 4 可知，在非结构文本信息抽取过程中，知识图谱提供了额外的语义和结构信息，有助于模型更准确地理解文本中的实体关系。GCN 能够捕捉图结构中的邻域信息，并通过信息传播和聚合机制提升模型的表示能力。通过结合知识图谱和图卷积神经网络，模型能够更有效地利用文本和图结构中的信息，提高关系抽取的准确性。

5 结语

本文致力于探索融合知识图谱与图卷积神经网络（GCN）的非结构文本实体关系抽取方法，旨在从海量的非结构文本数据中高效、准确地提取出实体及其之间的关系，为知识图谱的构建和更新提供有力支持。实验结果表明，其方法在处理非结构文本信息量增大的情况下，信息完整性相对稳定，且全程维持在较高水平。这不仅验证了该方法的有效性，也展示了其在处理大规模非结构文本数据方面的潜力。展望未来，将继续深化对知识图谱与 GCN 融合技术的研究，探索更多有效的实体关系抽取方法，进一步提升实体关系抽取的准确性和效率。

参考文献：

[1] 杨丽娜 , 刘长胜 , 刘璐璐 . 基于区块链技术的非结构化文本关键信息智能抽取模型 [J]. 信息技术 , 2024,48(2): 154-159.

[2] 丁泓馨 , 邹佩聂 , 赵俊峰 , 等 . 一种基于主动学习的文本实体与关系联合抽取方法 [J]. 计算机科学 , 2023, 50(10): 126-134.

[3] 张仰森 , 刘帅康 , 刘洋 , 等 . 基于深度学习的实体关系联合抽取研究综述 [J]. 电子学报 , 2023, 51(4): 1093-1116.

[4] 王书鸿 , 郑少明 , 刘中硕 , 等 . 面向某地区电网继电保护装置缺陷知识图谱构建的实体关系抽取 [J]. 电网技术 , 2023, 47(5): 1874-1887.

[5] 程顺航 , 李志华 , 魏涛 . 融合自举与语义角色标注的威胁情报实体关系抽取方法 [J]. 计算机应用 ,2023,43(5): 1445-1453.

[6] 吕东东 , 陈俊华 , 毛典辉 , 等 . 农产品标准领域知识图谱实体关系抽取及关联性分析 [J]. 农业工程学报 ,2022,38(9): 315-323.

[7] 何芳州 , 王祉淇 . 基于知识图谱的多数据集成抽取方法仿真 [J]. 计算机仿真 ,2023,40(12):422-427.

[8] 杨美芳 , 杨波 . 融入互注意力的风险领域实体关系抽取研究 [J]. 小型微型计算机系统 , 2023, 44 (5): 991-1001.

[9] 鲁义威 , 杨若鹏 , 殷昌盛 , 等 . 融合预训练模型与注意力机制的军事实体关系抽取方法 [J]. 信息工程大学学报 , 2022, 23 (1): 108-114.

[10] 王思丽 , 刘巍 , 杨恒 , 等 . 基于自然语言处理和机器学习的实体关系抽取方法研究 [J]. 图书馆学研究 , 2021(18): 39-48.

【作者简介】

熊文俊（1989—），女，河南信阳人，硕士，讲师，研究方向：软件工程、计算机应用技术。

赵辉（1989—），男，河南周口人，硕士，讲师，研究方向：计算机网络技术。

（收稿日期：2024-10-22）